

Special issue on

Languages of offence in digital perspectives

Guest edited by

Barbara Lewandowska-Tomaszczyk and Marcin Trojczak

TABLE OF CONTENTS

Barbara Lewandowska-Tomaszczyk, Marcin Trojczak, Introduction to the Special Issue: Exploring offensive language in digital perspectives	219
Olga Dontcheva-Navrátilová, Renata Povolná, Offensive language in media discussion forums: A pragmatic analysis	223
Giedrė Valūnaitė-Oleškevičienė, Linas Selmistraitis, Andrius Utko, Dangis Gudelis, Offensive language in user-generated comments in Lithuanian.....	239
Kholoud Boukhris, Online gaming and language aggression in a Tunisian Arabic context.....	255
Mariya Chankova, Evaluations of appropriateness through impoliteness in political discourse reframed for entertainment purpose.....	279
Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová, Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojczak, Renata Povolná, Linas Selmistraitis, Andrius Utko, Dangis Gudelis, LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications.....	301
Chaya Liebeskind, Natalia Vanetik, Marina Litvak, Hebrew offensive language taxonomy and dataset	325
Anna Bączkowska, Implicit offensiveness from linguistic and computational perspectives – a study of irony and sarcasm	353
Kristina Š. Despot, Ana Ostroški Anić, Tony Veale, “Somewhere along your pedigree, a bitch got over the wall!” A proposal of implicitly offensive language typology.....	385
Milana Bolatbek, Shynar Mussiraliyeva, Detection of extremist messages in web resources in the Kazakh language	415

Talya Natanya, Chaya Liebeskind , Clickbait detection in Hebrew	427
Barbara Lewandowska-Tomaszczyk, Chaya Liebeskind, Anna Bączkowska, Jurate Ruzaite, Ardit Dylgjeri, Ledia Kazazi, Erika Lombart , Opinion types and opinion markers in English social media discourse.....	451

Barbara Lewandowska-Tomaszczyk
University of Applied Sciences in Konin

Marcin Trojczak
University of Applied Sciences in Konin

INTRODUCTION TO THE SPECIAL ISSUE: EXPLORING OFFENSIVE LANGUAGE IN DIGITAL PERSPECTIVES

In today's rapidly evolving digital landscape, it has become increasingly important to examine the nuances and effects of offensive language. With the rise of social media platforms, online forums, and digital communication channels, offensive language has gained prominence, often blurring the lines between freedom of expression, harmful speech, and its impact on individuals and society.

This Special Issue aims to delve into offensive language in digital perspectives, providing a comprehensive exploration of its various dimensions. By analysing the ways in which offensive language manifests itself online, we can better understand its origins, consequences, and potential remedies.

The internet, in particular social media, has become an unprecedented breeding ground for offensive language with the intensified use of derogatory language, hate speech, and cyberbullying. This collection of papers recognizes the need to unpack the complexities of offensive language through two basic lenses. Contributors will explore thematic areas such as the sociocultural and pragmatic aspects of offensive language online on the one hand, and the ways in which computational methodology can help identify and classify offensive language types. By examining these perspectives, we hope to provide a well-rounded understanding of the intricacies involved.

One crucial aspect of this volume is its focus on promoting constructive dialogue and proposing potential solutions. As technology continues to evolve, it is essential to strike a balance between identification of positive and negative opinionated texts and mitigating the harmful effects of offensive language.

This Special Issue contains eleven papers, which are a fruit of the workshop on languages of offence organised as part of the conference *Contacts & Contrasts 2023* by the Department of Research in Language and Communication of the Faculty of Humanities and Social Sciences at the University of Applied Sciences in Konin. Furthermore, some of the contributions have been prepared within the framework of European Cooperation in Science and Technology (COST Actions) including Nexus Linguarum (CA18209) and Opinion (CA21129).

The first article “Offensive language in media discussion forums: A pragmatic analysis” authored by Olga Dontcheva-Navrátilová and Renata Povolná both from

Masaryk University in Brno, Czech Republic presents an analysis of selected offensive language categories found in Czech Corpus of Offensive Language. By exploring the nuances of the contextual cues for offensive speech act identification, the authors show that the annotation system devised to detect offensive language need to take into account this type of information. Otherwise, as indicated in the article, the delimitation and interpretation of the instances of offensive language may be largely inadequate.

Four scholars from Lithuania – Giedrė Valūnaitė-Oleškevičienė, Linas Selmistraitis and Dangis Gudelis – from Mykolas Romeris University in Vilnius, and Andrius Utka, from Vytautas Magnus University in Kaunas present a study on “Offensive language in user-generated comments in Lithuanian”. The aim of their research is to investigate the feasibility of identifying offensive language in Lithuanian by utilising the Simplified Offensive Language Taxonomy (SOLT), basing on the dataset of user-generated comments collected from a Lithuanian portal.

In the next paper Kholoud Boukhris (Anglia Ruskin University) focuses on “Online gaming and language aggression in a Tunisian Arabic context”. The paper examines ways how impoliteness, language aggression and conflict are used in two online gaming platforms in a Tunisian Arabic setting. It further contributes to understanding of the sources of conflict and various aspects of impoliteness.

“Evaluations of appropriateness through impoliteness in political discourse reframed for entertainment purposes” is the contribution authored by Mariya Chankova (South-West University N. Rilski, Blagoevgrad, Bulgaria). It analyses a French-language YouTube videos that reframe political discourse through the prism of various categories of evaluation. By doing so, the author shows the roles played by offensive language and humour in this type of online communication.

A team of eleven scholars Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin), Anna Bączkowska (University of Gdańsk), Olga Dontcheva-Navrátilová (Masaryk University, Brno), Chaya Liebeskind (Jerusalem College of Technology), Giedrė Valūnaitė-Oleškevičienė (Mykolas Romeris University, Vilnius), Slavko Žitnik (University of Ljubljana, Slovenia), Marcin Trojczak (University of Applied Sciences in Konin), Renata Povolná (Masaryk University, Brno), Linas Selmistraitis (Mykolas Romeris University, Vilnius), Andrius Utka (Vytautas Magnus University, Kaunas) Dangis Gudelis (Mykolas Romeris University, Vilnius), from four countries (Czech Republic, Israel, Lithuania, and Poland) present their findings in the paper “LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications”. The focus of the paper is the presentation of a Simplified Offensive Language (SOL) Taxonomy, and its annotation application and testing on English, Czech, Lithuanian, and Polish. As the results of the annotation are contained within the accepted statistical values, its outcomes will be presented in terms of the Linguistic Linked Open Data (LLOD).

The paper “Hebrew Offensive Language Taxonomy and Dataset”, authored by Chaya Liebeskind (Jerusalem College of Technology), Natalia Vanetik (Shamoon

College of Engineering, Beer-Sheva) and Marina Litvak (Shamoon College of Engineering, Beer-Sheva) introduces a streamlined taxonomy for categorizing offensive language in Hebrew. In this way, it addresses the perennial problem of offence in a language from a Semitic family. The research findings indicate a need to consider the language data both from linguistic and cultural variationist perspectives.

In her paper “Implicit offensiveness from linguistic and computational perspectives – a study of irony and sarcasm”, Anna Bączkowska (University of Gdańsk) sheds light on the linguistic concept of implicit offensiveness and the gradability of various forms of covert meaning. Irony is differentiated by the author against sarcasm and analysed by means of both linguistic as well as computational tools of non-contextual embeddings.

The next paper “‘Somewhere along your pedigree, a bitch got over the wall!’ – A data-driven approach to a typology of implicitly offensive language” by Kristina Š. Despot (Institute for the Croatian Language), Ana Ostroški Anić (Institute for the Croatian Language), and Tony Veale (University College Dublin), continues the topic of implicit offense by discussing various problems in detecting implicit offense, including definitional problems, narrowing a set of possible implicitness devices, and focusing mainly on limited linguistic forms.

Milana Bolatbek and Shynar Mussiraliyeva from al-Farabi Kazakh National University, Almaty present the paper “Detection of extremist messages in web resources in the Kazakh language”. In their study, the authors develop semantic analysis models to identify extremist messages in the Kazakh language, a Turkic language spoken in Central Asia, and build effective machine learning models based on bigrams and word input methods.

The paper “Clickbait detection in Hebrew” by Talya Natanya and Chaya Liebeskind from Jerusalem College of Technology investigates an approach to clickbait detection in Hebrew, and proposes multiple machine learning models to classify Hebrew headlines as clickbait or non-clickbait: a vector machine (SVM) model and heBERT, a deep learning model.

The last paper in this Special Issue “Opinion types and opinion markers in English social media discourse” by Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin), Chaya Liebeskind (Jerusalem College of Technology), Anna Bączkowska (University of Gdańsk), Jurate Ruzaite (Vytautas Magnus University, Kaunas), Ardita Dylgjeri (“Aleksandër Xhuvani” University of Elbasan), Ledia Kazazi (“Aleksandër Xhuvani” University of Elbasan) and Erika Lombart (Université catholique de Louvain) investigates the concept of opinion as opposed to evidence-based statements. The authors propose a research methodology which combines a linguistic analysis of opinions, with the automatically generated lexical embeddings of positive and negative lexicon acquired from the analysed opinionated texts. The definition of the concept of opinion, as the authors propose, is perceived in terms of an

opinion event, with a number of necessary and characteristic properties of an opinion prototype.

By investigating offensive language in digital perspectives, we hope to shed light on how it intersects with other socio-political issues such as radical content, clickbait or more nuanced – implicit types of offensive language and a distinction between positive and negative opinions in social media. Moreover, we hope that in this way we can foster further research towards building more effective models of negative and offensive language expression.

In conclusion, this Special Issue seeks to unravel the intricacies surrounding this pervasive issue by bringing together pragmatics and computational linguistics scholars with the objective to shed light on the selected facets of offensive language online. It can be hoped that this can provide new materials to propose effective strategies for its prevention and management. Through this exploration, we also hope to contribute to the ongoing conversations and actions aimed at creating a more inclusive digital world for all.

Barbara Lewandowska-Tomaszczyk and Marcin Trojczak
Special Issue Editors

About the Editors

Barbara Lewandowska-Tomaszczyk is Professor Ordinarius Dr Habil. in Linguistics and English Language at the Department of Language and Communication at the University of Applied Sciences in Konin (Poland). Her research focuses on cognitive semantics and pragmatics of language contrasts, corpus linguistics and their applications in translation studies, lexicography and online discourse analysis. She is invited to read papers at international conferences and to lecture and conduct seminars at universities. She publishes extensively, supervises dissertations and also organizes international conferences and workshops.

e-mail: barbara.lewandowska-tomaszczyk@konin.edu.pl
ORCID: 0000-0002-6836-3321

Marcin Trojczak holds PhD in Linguistics and MA in Philosophy. He is Assistant Professor at the University of Applied Sciences in Konin (Poland). He is also actively cooperating with University of Lodz and University of Economics and Human Sciences in Warsaw. His research interests include metaphorical conceptualisations of mental and emotional processes, the impact of translation technologies on translation education, normativity and genericity in language and cognition, as well as offensive language.

e-mail: marcintrk@gmail.com
ORCID: 0000-0003-3062-9670

Address

Department of Language and Communication, University of Applied Sciences in Konin
1, Przyjazni str.
62 510 Konin, Poland

Olga Dontcheva Navratilova
Masaryk University

Renata Povolná
Masaryk University

OFFENSIVE LANGUAGE IN MEDIA DISCUSSION FORUMS: A PRAGMATIC ANALYSIS

Abstract

This study intends to contribute to the delimitation of selected offensive language categories based on an analysis of a corpus of contributions to discussion forums in Czech online national newspapers and news platforms called *Czech Corpus of Offensive Language* (CCOL). It endeavours to study three problematic areas (i) delimitation between the speech acts performed, (ii) lexical realisation of specific properties of the target and (iii) identification and categorisation of implicit offence (e.g. figurative semantic shifts) by exploring contextual cues for the speech act identification, the keywords indicating the properties of the target and the types of semantic shifts in implicit expressions of offence. The findings indicate that annotation systems that do not use context information for the detection of offensive language may face problems with adequate interpretation of the language means under investigation.

Keywords

offensive language, context, speech acts, keywords, implicit offence, figurative semantic shifts, Czech media discussion forums

1 Introduction

The spread of easy access to Internet mobile data technologies across the world in the last two decades has led to a huge increase in online news. While speeding up the process of production, consumption and sharing information, this resulted in a proliferation of posts on informational and social media platforms, which “allow all members to express their opinions and perspectives irrespective of age, level of expertise or any other feature” (Mahdu et al. 2023: 1). The lack of restrictions on the content and language that contributors post, together with the anonymity of the contributors, created fertile grounds for the use of defamatory language and abusive content, “which can induce mental health problems such as depression, anxiety, or suicide, in the worst-case scenario” (Plaza-del-Arco et al. 2022: 1) or lead to the incitement of hateful chain reactions against vulnerable groups. This has called for the detection of offensive language in order to maintain social interaction without threat and abuse of individuals and groups.

Offensive language and hate speech are actively researched, yet there is no satisfactory definition and delimitation of related terms and the unveiling of these phenomena is still to be refined (e.g., Davidson et al. 2017; Mahdu et al. 2023; Menini et al. 2021; Moulson 2016). This study intends to contribute to the delimitation of selected offensive language categories based on an analysis of a corpus of contributions to discussion forums in Czech online national newspapers and news platforms (*Czech Corpus of Offensive Language*). Discussion forums in online newspapers have been selected as the material for this study since they represent a form of asynchronous “post-medial interactions, which follow the production and reception of the medial text” (Chovanec 2018: 510), i.e., the audience contribution does not affect the original media text. The contributors to discussion forums engage with one another based on the media content (Ayass and Gerhardt 2012) without being constrained by the media conventions of impartiality, objectivity and political correctness. This leaves ample space for the use of offensive language directed at individuals or groups. The rare occurrence of non-targeted comments may also be typically interpreted as implicitly targeted at a contextually identified person or group. The use of offensive language in media discussion forums is highly context-dependent as it is affected by sociocultural factors and by the topic the medial text addresses, as well as by the participants’ social background, ideology and values.

Previous annotation campaigns on specific datasets (e.g., Dontcheva-Navratilova and Povolná 2023; Lewandowska-Tomaszczyk et al. 2023; Zampieri et al. 2019) applying different offensive language taxonomies and cross-dataset classification studies (e.g., Fortuna, Soler-Company and Wanner 2021; Waseem et al. 2018) have revealed major problems in three areas (1) delimitation between the speech acts performed, e.g., insult and discredit, (ii) the lexical realisation of specific properties of the target and (iii) the identification and categorisation of implicit offence (e.g., figurative semantic shifts). Our study endeavours to contribute to the study of these three problematic areas by exploring contextual cues for speech act identification, keywords indicating the properties of the target and semantic shifts in implicit expressions of offence. We argue that without collecting contexts from longer stretches of comments in discussion forums (cf., Mahdu et al. 2023), the detection of offensive language is highly problematic.

2 Offensive language

The term offensive language (Kogilavani et al. 2021) is used in this article as an umbrella term to include different concepts used in the literature to refer to the occurrence of abusive language on online news and social media platforms, such as hate speech, aggressive or toxic language. The lack of terminological clarity and unity is one of the major problems that researchers into offensive language face. For instance, Nobata et al. (2016), define “abusive” language as strongly impolite, rude, or hurtful language using

profanity, which shows intense emotions and may devalue the target of abuse, while Wiegand et al. (2019) regard offensive language as the use of hurtful, derogatory or obscene utterances addressed by an individual or group to another person or group. Toxicity is associated with attacks, insults that target the race, ethnicity, nationality, political ideology, religion, gender, and sexual orientation of individuals or groups (Taulé et al. 2021). Probably the most frequent label used to refer to abusive language on social media is hate speech. According to the General Policy Recommendation No. 15 of the European Commission (<http://hudoc.ecri.coe.int/eng?i=REC-15-2016-015-ENG>), hate speech is “the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of race, colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation, and other personal characteristics or status”. The overlapping concepts and the absence of clear delimitation criteria of sub-types of offensive language have resulted in numerous taxonomies applied to various datasets and systems for detecting offensive language and hate speech on online news platforms and social media, some of them acknowledging the need to consider larger spans of context for class assignment of offensive language items. Our study takes a linguistics approach to the delimitation of speech acts, properties of the target and the expression of implicit offence which are context-dependent and may cause problems for systems that do not use context information for the detection of offensive language.

3 Data and method

3.1 Corpus

The *Czech Corpus of Offensive Language* (CCOL) comprises 500 comments totalling 8,150 words. Each comment includes one to three adjacent utterances, extracted from online discussion forums in ten Czech national newspapers and news platforms, namely Seznam Zprávy (<https://www.seznamzpravy.cz>), Seznam Médium (<https://medium.seznam.cz>), Idnes.cz (<https://www.idnes.cz>), Forum24 (<https://www.forum24.cz>), Novinky.cz (<https://www.novinky.cz>), HlídacíPes (<https://hlidacipes.org>), Neviditelný pes (<https://neviditelnypes.lidovky.cz>), Lidovky.cz (<https://www.lidovky.cz>), Super.cz (<https://www.super.cz>) and Sport.cz (<https://www.sport.cz>). The corpus is sampled to represent discussions on a variety of topics, including home and foreign news, home and foreign politics, sport, celebrities, crime, finance, travelling, health and weather. Most of the discussions are associated with home news and politics, such as presidential elections, different political causes and scandals, or foreign news and politics, often referring to the war in Ukraine, which are

indicative of the topics with which Czech readers are likely to feel involved. The composition of the CCOL corpus is summarised in the following table:

Table 1: Composition of the CCOL corpus

Topics	Comments No
Celebrities	31
Crime	26
Foreign news	75
Home news	193
Home politics	114
Finance	15
Sport	26
Travelling	10
Health	8
Weather	2
Total	500

The corpus was annotated with the assistance of the semantic *annotation* tool INCEpTION (<https://github.com/inception-project/inception>) and classified following the Simplified Offensive Language Taxonomy (SOL) proposed by Lewandowska-Tomaszczyk et al. (2021) as part of research work undertaken within COST Action NexusLinguarum WG 4.1.1. The annotation was carried out by two linguists who share a similar social background, age, and profession. Prior to annotating the corpus, the two annotators carried out several training sessions to discuss the offensive language taxonomy, practise annotating samples, compare their results and resolve disagreements. These training sessions, together with the similar social background and knowledge of cultural and political context that the annotators had, proved to be extremely important since the inter-rater agreement achieved during the annotation campaign (February-March 2023) was relatively high in comparison with previous annotation campaigns (e.g., Lewandowska-Tomaszczyk et al. 2023). In the Czech Offensive Language annotation campaign, the speech act type identification for insult, discredit and hate showed substantial agreement (0.74), but the agreement for threat was weak (0.11). Interrater agreement for specific properties of the target was also substantial ranging from 0.61 to 0.70 and for the identification of the types of semantic shift in implicit expressions of offence it was only moderate (0.52) (Dontcheva-Navratilova and Povolná 2023); the Cohen's Kappa measure was calculated according to the scale of the strength of agreement suggested by Landis and Koch (1997) and Sim and Wright (2005).

3.2 Methods and analytical framework

A mixed method approach combining quantitative corpus analysis of frequency list, keywords and collocates and qualitative analysis of speech acts and implicit expressions

of offence via figurative semantic transfer of meaning were used in the analysis of speech acts, lexical realisations of offensive language and implicit offence, since it is believed that quantitative and qualitative methods are complementary to give a comprehensive picture of the language means used for the detection of offensive language.

3.2.1 Speech acts

The division of speech acts into different categories goes back to the work of Austin (1962) and Searle (1969, 1976). In later work, Searle (1976) distinguishes five basic categories of speech acts: representatives, directives, commissives, expressives and declarations. Offensive language conveys the feelings of the speaker and his/her intentions of future actions aiming at the target, i.e., offence is conveyed via the expressive speech acts of insult, discredit and hate, and the commissive speech acts of threat, which are the focus of the qualitative analysis in our study.

Expressives are defined as speech acts that convey the speaker's feelings about themselves or the world (Searle 1976: 12). As Archard (2014: 128) notes, expressive speech acts conveying offensive language are public since they are communicated in public contexts such as meetings or through public institutions such as the press and media and some can even be interpreted as threats depending on when, where, how, by and to whom they are performed. Previous research (e.g., Ronan 2015; Clark 1996; Norrick 1978) distinguishes up to nine categories of expressives, such as apologizing, thanking, congratulating, condoling, deplored (censoring), disapproving, lamenting, welcoming, forgiving and boasting. It is deplored (censoring) defined by Norrick (1978: 287–288) as a speech act by which the addressee is criticised for an event which had a negative impact on the speaker or a third person and disapproving as defined by Guiraud et al. (2011: 1037) as the speech by which an agent expresses disapproval to the addressee about something that is not considered ideal that are relevant to our study of offensive language. As already mentioned, our taxonomy comprises three types of expressives – insult, discredit and hate, which differ mainly in the intensity of the negative evaluation. In contrast to expressives, which indicate emotions and attitudes, the commissive speech act of threat implies an intended action, since uttering a threat can be defined as an offence where a person “knowingly utters, conveys, or causes harm to any person to receive a threat” (Greenspan 1994: 442, as quoted in Walton 2000: 101).

3.2.2 Frequency list and keywords

The quantitative analysis takes as a starting point the frequency list approach as it is efficient in identifying linguistic features such as lexical items characteristic of a language variety (Miller 2020: 77). In this study, a single word has been used as the unit of analysis in frequency list generation. The frequency list was created with the SketchEngine tool, which was also used to identify keywords statistically characteristic of the corpus (Culpeper and Demmen 2015: 90) in comparison with another general

corpus (cf., Rayson and Potts 2020: 121). The use of quantitatively defined keywords enhances the objectivity of the research as it “[delegates] the initial task of identifying items for analysis to a computer algorithm [which] ensures that this stage of the research is completely insulated from researcher bias” (Groom 2010: 60). We use a combination of the frequency list and keyword approaches to identify lexical items which can potentially be regarded as lexical triggers of offensive language.

3.2.3 Implicit expressions of offence via figurative semantic transfer of meaning

The expression of offensiveness can be performed both explicitly and implicitly (Wiegand et al. 2021), i.e., it may be conveyed not only by literally abusive words like “freak”, “monster”, “fool”, “scumbag”, “dumbass”, but also by different subtypes of implicitly abusive lexical items typically associated with the figurative semantic transfer of meaning. In our study, we investigate figurative transfer based on metaphor, simile, exaggeration and irony, as these seem to be the most frequent types of implicit realisations of offence. While identifying the specific categories of implicit offence is problematic (Lewandowska-Tomaszczyk 2023, forthcoming), studying implicit realisations of offensive language is highly important as “implicated offences may be perceived as stronger than those expressed explicitly” (Piskorska 2017: 51).

4 Results

4.1 Speech acts

Differentiating between the four types of speech acts – insult, discredit, hate and threat – has been indicated as problematic by several annotation campaigns. Since the differences between these speech acts are rather thin and heavily context-dependent, we propose two main criteria for differentiating offensive language speech acts: intentionality and intensity. While the commissive speech act of threat tends to be always intentional, the expressive speech acts need not be, depending on the contextual circumstances which must be taken into consideration when interpreting the given acts. The degree of intensity tends to differ across the four speech acts and may be seen as a cline ranging from the most intensive threat, followed by hatred, to the less intensive insult and discredit. The last two speech acts seem to be the most difficult to distinguish both in terms of intensity and intentionality.

The most aggressive type of speech act, the commissive speech act of threat, is relatively rare in our dataset. This type of speech act is always intentional. Merriam Webster online describes it as “an expression of intention to inflict evil, injury, or damage”. Example 1, in which the author of the post threatens the commies to be swung from the gallows 100 times a day, is one of the rare examples of threats in our corpus. The intensity of the emotion is further enhanced by the hateful naming *komouši*

[‘commies’], the colourful metaphor *houpat na šibenicici* [‘swinging from the gallows’] and the exaggeration *100x denně* [‘100 times a day’], the use of an exclamation mark and above all the context in which this comment is used, which helps to express the hatred of communists deeply rooted in modern Czech society:

- (1) **Komouši** vymysleli jen vraždy, zabíjení, rabování! Měli by se pro výstrahu **100x denně houpat na šibenici!** (CZ-OL-185)
[‘The commies only invented murder, killing, looting! They should be **swinging from the gallows 100 times a day** as a warning!’]

The strongest and most hurtful of the studied types of expressive speech acts is hate, characterized as the “intense hostility and aversion usually deriving from fear, anger, or sense of injury” or “a systematic and especially politically exploited expression of hatred” (Merriam-Webster online). This is illustrated in Examples 2 and 3 showing the culture-bound feelings of hatred and active hostility towards the Russians and the Fascists. In Example 2, this is clearly expressed in the collocation *ruská šovinistická prasata*, where the Russians are compared to *šovinistická prasata* [‘chauvinist pigs’], in which the expression *prasata*, even when used on its own, is an instance of hatred or disgust.

- (2) Pokud se nechovají jako **ruská šovinistická prasata**, tak s nimi nemají sebemenší problém (CZ-OL-127)
[‘Unless they’re acting like **Russian chauvinist pigs**, they don’t have the slightest problem with them.’]

An emotional aversion coupled with enmity towards the Russians is indicated in Example 3 by comparison to Hitler. This is combined with the expression of strongly negative feelings towards communists, clearly displayed by the ironical vocative *komunisto!* [‘communist!’]. The implied hateful comment clarified by the larger context of the discussion thread concerns the actions of the Russians in the context of the war in Ukraine.

- (3) A Hitler dělal to, co ted’ Russáci, **komunisto!** (CZ-OL-127)
[‘And Hitler did what the Russians are doing now, **communist!**’]

One of the most typical types of expressive speech acts is that of insult. According to Merriam-Webster online, an insult is “an instance of insolent or contemptuous speech or behaviour” with the intention to deliberately cause humiliation, hurt pride, or shame, which can even cause bodily injury. Example 4 comprises an insult referring to the addressee’s mental disabilities associated with a metaphorical transfer of meaning (*vy máte vypnuty mozek*) which damages the authority of the target and deprives it of good reputation. The intensity of this insult is fostered by the following comment, which

claims that this information has been published on the web and Wikipedia, thus making it sound like an objective fact.

- (4) *Vy máte vypnutý mozek. To bylo oficiálně na webech a i Wikipedii.* (CZ-OL-129)
 ['Your brain's off. It was officially on the web and even Wikipedia.']}

The expressive speech act of insult is sometimes rather difficult to distinguish from that of discredit. In the case of the latter, the definition provided in the Merriam-Webster online dictionary is to cause “loss of credit or reputation” or induce “lack or loss of belief or confidence”. Example 5, taken from the online news on celebrities, refers to the popular TV show titled *Wife Swap*, very popular among TV viewers. It clearly discredits the TV show participants by referring scornfully to their IQ and comparing it to the intelligence of a pumpkin. The couple’s decision to have a baby after a two-month relationship is ridiculed by the ironical expression *chytrost* ['wisdom'], which is intensified by noting that this would pass their ‘wisdom’ on to their descendants.

- (5) *Tohle je jen další tupý pár. Oba mají IQ tykve, a tak se rozhodli udělat si po 2 měsících dítě a předat tu chytrost svým potomkům.* (CZ-OL-118)
 ['This is just another dumb couple. They both have the IQ of a pumpkin, and so they've decided to make a baby after 2 months and pass that wisdom on to their descendants.']}

However, the above example could also be interpreted as an insult because it is clearly intentional and expresses a strong negative attitude. We argue that only the broader context can help discriminate between insults and discredits, although it should be acknowledged that the interpretative decision is likely to be highly subjective.

4.2 Frequency list and keywords

Starting with the frequency list generated by SketchEngine we have identified potential offensive language words occurring more than five times in the corpus (see Table 2; the translations of the words into English are tentative since they are heavily context-dependent). Since our aim is to consider generally applicable principles for offensive language detection and categorisation, we have excluded from the analysis all personal names, names of countries and other content-related items, although some names, such as, Zeman (the Czech ex-president), Babiš (the ex-prime minister) or Russia, were found to be commonly related to offensive language in our corpus. Then a keyword list was generated (reference corpus csTenTen, 2017) and it was verified that the selected twelve offensive words were listed between the first 50 key terms in the corpus.

Table 2: List of the most frequent keywords in the corpus

Keywords in Czech	English translations	Frequency
hnus	disgust	12
banda	gang	9
hlupák	fool, dumbass	8
kolaborant	collaborator	7
lež	lie	6
špína	dirt, filth	6
demence/dement	dementia/dement	6
zrůda	freak, monster	5
komouš	commie, communist	5
desolát	desolate	5
zloděj	thief	5
lhář	liar	5

The selected twelve keywords are *hnus*, *banda*, *hlupák*, *kolaborant*, *lež*, *špína*, *demence*, *zrůda*, *komouš*, *desolát*, *zloděj*, *lhář*, which clearly express negative stance and thus can be regarded as indicative of Czech offensive language.

The keywords can be grouped according to the characteristics of the target that they address. These are the most prominent target characteristics identified within the Czech offensive language annotation campaign (cf., Dontcheva-Navratilova and Povolná 2023) carried on our corpus.

Probably the most typical targeted feature in the *Czech Corpus of Offensive Language* are physical and mental disabilities (see Examples 4 and 5 above), which often acquire political connotations in home news discussion forums. The keywords *hlupák* and *dement* tend to express this kind of offence with the intention to discredit or insult the targeted group or individual. For instance, in Example 6, the offensive item *hlupáků* ['fool's'] is enhanced by its near synonym *pitomec* ['goofy', 'nerd', 'moron'] to refer to foolish people holding high political offices who are more interested in financial gains (*větší koryto* ['bigger trough']) rather than serving the society. In a different non-political context, *hlupák* in Example 7 combines with *lempl* ['slacker', 'idler'] and *hloupost* ['stupidity'] to indicate irresponsible behaviour which is estimated dangerous enough to incite the contributor to threaten the target to be brought to justice.

(6) *Mě se to líbí, tady je **hlupáků** ráj, čím větší pitomec, tím větší koryto.* (CZ-OL-131)
 ['I like it, here is the **fool's** paradise, the bigger the goofy, the bigger the tough.']

(7) *Toho **lempila** postavit před soud jako kohokoli jiného. Každý **hlupák** ví, že nemůže chodit s nataženým uzávěrem a nábojem v komoře, bez důvodu, tak jaképak školení, byla to **hloupost** dotyčného!* (CZ-OL-452)

[‘Put the **slacker/idler** on trial like anyone else. Any **fool** knows that they can’t walk around with the slide cocked and a round in the chamber without any reason, so what kind of training, it was the **stupidity** of the man in question!?’]

The second most prominent group of offensive expressions relate to ideology, which may be the result of the presidential elections, which took place in the period in which our corpus was built and thus represent a massive part of the home news discussion posts. The three offensive words representing this category – *kolaborant*, *komouš*, *desolát* – are mostly related to insults and discredit targeting the alleged communist past of the target. As Example 8 shows, the author of the contribution refers to the communist background of politicians representing several political parties running in the elections (*smečka komoušů* [‘a bunch of commies’]), which is also indicated by the metaphorical item (*s rudou chátrou* [‘with the red rabble/mob’]); the intensity of the insult is fostered by the metaphorical comparison to *krvelačné bestie* [‘bloodthirsty beasts’].

- (8) *Je vidět, že tady není vůle se s rudou chátrou srovnat jednou provždy. Ty krvelačné bestie to tu řídí dodnes. Např. v ODS je pořádná smečka komoušů. A strany NF KDU-ČSL a TOP09 jsou jim v patách.* (CZ-OL-463)
 ['Clearly, there is no will, to get even with the red rabble once and for all. Those **bloodthirsty beasts** still run the place today. For example, there is a hell of a **bunch of commies** in ODS. And the parties of NF KDU-ČSL a TOP0 are right on their heels.']}

Offensive language oriented towards the social characteristics of the target is also often occurring in political home news. The qualification of the discourse of political actors creatively labelled as *Burešbabišovy* (a combination of the names of two prominent politicians) *lží* [‘lies’] discredits these politicians in combination with the use of figurative means, i.e., the metaphor *vyplouvají na světlo* [‘are coming to light’] and the simile *jako doměček z karet* [‘like a house of cards’], underlying the transparency of the lies.

- (9) *Burešbabišovy lží pomalu vyplouvají na světlo a jeho obhajoby se sypou jak domeček z karet.* (CZ-OL-375)
 ['Lies by Bureš and Babiš are slowly **coming to light** and their defence is falling apart **like a house of cards.**']

Another sub-type of socially oriented offensive words comment on the physical and mental abilities of celebrities. The failed cosmetic surgery undergone by the targeted celebrity is evaluated as *hnus* enhanced by the simile *rypák jak kačer* [‘snout like a duck’], and further elaborated on by the exaggeration *nesoudného a trapného strašidla* [‘shameless and embarrassing scarecrow’] a *mozkových buněk neměla dost ani před operacemi* [‘she didn’t have enough brain cells even before the surgeries’], as shown in the following example:

- (10) *No to je teda pěknej hnus – rypák jak kačer, jak se může takhle zohyzdit? Její dcera má mnohem více rozumu a pokud jí to vydrží, tak nepůjde ve šlépéjích nesoudného a trapného strašidla, kterým se její matka stala. Myslím, že mozkových buněk neměla dost ani před operacemi.* (CZ-OL-07-09)
[Well that's pretty **disgusting** – **snout like a duck**, how can she get so ugly? Her daughter has a lot more sense and if she can keep it up, she won't follow in the footsteps of a **shameless and embarrassing scarecrow** that her mother has become. I believe **she didn't have enough brain cells even before the surgeries.**']

The last highly frequent keyword *banda* is easily associated with other keywords, such as *banda zlodějů* ['a bunch of thieves'], *banda hlupáků* ['a bunch of fools'] and other near synonymous words, e.g., *banda špinavců* ['a bunch of scumbags'], *banda srabů* ['a bunch of cowards']. Depending on the broader context in which they are used, these collocations express different degrees of negative meaning, thus intensifying the overall strength of offensiveness of the comment.

4.3 Implicit expressions of offence

The four types of implicit offence considered in this study pertain to different types of semantic transfer the decoding of which is heavily dependent on contextual cues. Metaphors and similes are based on perceived similarity and have the potential to enhance the strength of offence by making it more vivid, colourful and immediate for the reader. In the case of simile, the comparison is explicitly described; however, in the case of metaphor it is only implied as one field of reference is mapped onto another, commonly “to make sense of relatively complex, abstract or unfamiliar experiences in terms of more familiar ones” (Wales 2011: 265). Thus, while simile may be seen as primarily having an intensifying function, metaphor allows for ambiguity and may generate multiple meanings as the interpretation of the conceptual mapping is context-dependent and may differ to some extent across individuals. The use of simile is illustrated in Example 11, where the first part comprises the imperative and vocative written for ironic purposes in Russian *Vyděržajte, pionéry* ['Hang in there, pioneers'], while the second part of the example intensifies the dumbness of the target by the descriptive comparison *ne blbej jak blbeček* ['not dumb as a dumbass'], which is easy to process.

- (11) *Vyděržajte, pionéry, už to máme skoro za sebou. Člověk musí být trošku chytřej a ne blbej jak blbeček.* (CZ-OL-325-326)
[‘Hang in there, pioneers, it’s almost over. You have to be a little smart, and not dumb as a dumbass.’]

The context-dependency and fuzziness of metaphorical meaning is shown in Example 12, in which the teacher is labelled as *vodkou stará struktura přechlastaná* ['an old structure blasted with vodka']. This may be interpreted as referring through the

metaphorical transfer of meaning to someone who is a communist devoted to Russia and the Russians (*stará struktura*) and who drinks a lot, especially vodka traditionally drunk in great quantities in Russia (*vodkou přechlastaná*). However, other interpretations are also possible, especially if the reader lacks the relevant cultural knowledge.

- (12) *Pani učitelka je očividně vodkou stará struktura přechlastaná.* (CZ-OL-132)
 [‘The teacher is obviously an old structure blasted with vodka.’]

Exaggeration, which results from a violation of Grice’s (1967) maxim of quantity and quality, is frequently employed in offensive language to indicate emphasis or to increase the emotions expressed, as clearly illustrated in Example 13, in which Russia is equalled to *lež*, *svrab*, *špína*, *násilí* and *tupost* [‘lies’, ‘scabies’, ‘filth’, ‘violence’ and ‘dullness’], thus expressing intensively the hatred of Russia and the Russians.

- (13) *Rusko je lež, svrab, špína, násilí a tupost.* (CZ-OL-122)
 [‘Russia is lies, scabies, filth, violence and dullness.’]

Irony is clearly the most context-dependent type of semantic transfer as it occurs when what is said appears to contradict the literal sense of the expression in a given context that is presumably intended by the speaker, i.e., it violates Grice’s maxim of quality. The high contextual dependency of irony is indicated by the fact that it may be not perceived and thus it may lose its force altogether. For example, the readers who do not know Golem and the purpose of *šém* [‘reviving ball’], which can awaken Golem to life, may not understand the ironical transfer of meaning, in the post which claims that the politician Válek needs *šém* to be able to operate in a normal and effective way as expected in his political position.

- (14) *Vložte Válkovi zpět jeho šém do čela a třeba začne fungovat* (CZ-OL-101)
 [‘Put the reviving ball back in Valek’s forehead and maybe it’ll work.’]

All the examples illustrated above indicate that all types of figurative language transfer of meaning are often applied in combination, which, depending on the contextual cues, contributes to the intensification of the offensive message intended by the author of the given post.

5 Conclusion

This study has explored the delimitation of the problematic offensive language categories of speech acts, lexical realisations of specific properties of the target and identification and categorisation of implicit offence as displayed in the *Czech Corpus of Offensive*

Language comprising discussion forums in Czech online national newspapers and news platforms.

The qualitative analysis has shown the importance of context in the identification of specific types of speech acts expressing offence (Madhu et al. 2023) and the interpretation of implicit offence based on a figurative transfer of meaning. In addition, the lack of cultural background knowledge may blur the intended speech act types or their intensity and thus lead to a misinterpretation of offensive language categories. Our findings also support Plaza-del-Arco et al.'s (2022) conclusion that offensive language detection can be optimized by taking into consideration a combination of factors such as the target group or person, figurative language, vulgar language and emotions.

The quantitative analysis suggests that although keywords may be used as indicators of offensive language they cannot be automatically understood as implying offence, as the context may contradict such a reading. As Fortuna et al. (2021) stress, a selection procedure based on keywords may only lead to the collection of more explicit offensive language. This is why we argue that annotation systems that do not use context information for the detection of offensive language may face problems with adequate interpretation of the language means under investigation.

Obviously, the results of this study cannot be overgeneralised because it explores only one language and the size of the corpus used is rather limited. Further research taking into consideration different languages and larger corpora of offensive language should verify the reported findings.

References

- Archard, David. 2014. Insults, free speech and offensiveness. *Journal of Applied Philosophy* 31(2). 127–141.
- Austin, John L. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.
- Ayass, Ruth & Cornelia Gerhardt. 2012. *The Appropriation of Media in Everyday Life*. Amsterdam: John Benjamins.
- Chovanec, Jan. 2018. Participating with media: exploring online media activity. In Colleen Cotter & Daniel Perrin (eds.), *The Routledge Handbook of Language and Media*, 505–522. London: Routledge.
- Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.
- Culpeper, Jonathan & Jane Demmen. 2015. Keywords. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 90–105. Cambridge: Cambridge University Press.
- Davidson, Thomas, Dana Warmsley, Michael Macy & Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* 11(1). 512–515. DOI: <https://doi.org/10.1609/icwsm.v1i1.14955>.
- Dontcheva-Navratilova, Olga & Renata Povolná. 2023. Czech Offensive Language: Testing a Simplified Offensive Language Taxonomy. In Sara Carvalho, Anas Fahad Khan, Ana Ostroški

- Anić, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch & Ana Salgado (eds.), *LDK 2023 proceedings*, 627–632. Lisbon: NOVA CLUNL Available at: <https://aclanthology.org/2023.ldk-1.68.pdf> (accessed 30 October 2023).
- Fortuna, Paula, Juan Soler-Company & Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* 58(3). 102524.
- Grice, Paul H. 1967. Logic and conversation. In Paul Grice (ed.), *Studies in the Way of Words*, 41–58. Cambridge MA: Harvard University Press.
- Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3). 257–277.
- Groom, Nicholas. 2010. Closed-class keywords and corpus-driven discourse analysis. In Marina Bondi & Mike Scott (eds.), *Keyness in Texts*, 59–78. Amsterdam: John Benjamins.
- Kogilavani, Shanmuga V., Subramanian Malliga, K. R. Jaiabinaya, M. Malini, M. Manisha Kokila. 2021. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, Vol. 81, Part 2. 630–633. Amsterdam: Elsevier. Available at: <https://doi.org/10.1016/j.matpr.2021.04.102> (accessed 30 August 2023).
- Landis, Richard & Gari Koch. 1997. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Chaya Liebeskind, Geidre Valunaite Oleskevicienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, Olga Dontcheva-Navratilova, Agnieszka Borowiak, Kristina Despot & Jelena Mitrović. 2023. Annotation scheme and evaluation: The case of OFFENSIVE language. *Rasprave* 49(1). 155–175.
- Madhu, Hiren, Shrey Satapara, Sandip Modha, Thomas Mandl & Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications* 2015, 119342.
- Menini, Stefano, Alessio Palmero Aprosio & Sara Tonelli. 2021. Abuse is contextual, what about nlp? The role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.
- Miller, Don. 2020. Analysing frequency lists. In Magali Paquot & Stefan Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 77–98. Cham: Springer.
- Moulson, Geir. 2016. Zuckerberg in Germany: No place for hate speech on facebook. Available at: <http://abcnews.go.com/Technology/wireStory/zukerberg-place-hate-speech-facebook-37217309> (accessed 10 September 2023).
- Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad & Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, 145–153. Geneva: International World Wide Web Conferences Steering Committee.
- Norrick, Neal R. 1978. Expressive illocutionary acts. *Journal of Pragmatics* 2(3). 277–291.
- Piskorska, Agnieszka. 2017. On the strength of explicit and implicit verbal offences: A relevance-theoretic view. In Silvia Bonacchi (ed.), *Verbale Aggression: Multidisziplinäre Zugänge zur verletzenden Macht der Sprache*, 51–72. Berlin: De Gruyter.
- Plaza-del-Arco, Flor Miriam, M. Dolores Molina-González, L. Alfonso Ureña-López & María-Teresa Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems* 258. 109965.

- Rayson, Paul & Amanda Potts. 2020. Analysing keyword lists. In Magali Paquot & Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 199–140. Cham: Springer.
- Ronan, Patricia. 2015. Categorizing expressive speech acts in the pragmatically annotated SPICE Ireland corpus. *ICAME Journal* 39. 25–45.
- Searle, John. 1969. *Speech Acts*. Cambridge: Cambridge University Press.
- Searle, John. 1976. A classification of illocutionary acts. *Language in Society* 5. 1–23.
- Sim, Juūius & Chris Wright. 2005. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85(3). 257–268.
- Taulé, Mariona, Alejandro Ariza, Montserrat Nofre, Enrique Amigó & Paolo Rosso. 2021. Overview of the DETOXIS task at IberLEF-2021: Detection of toxicity in comments in Spanish. *Inicio* 67. 209–221.
- Wales, Katie. 2011. *A Dictionary of Stylistics*. 3rd ed. Oxon: Routledge.
- Walton, Douglas. 2000. The speech act of making a threat. In Douglas Walton, *Scare Tactics. Argumentation Library*, 101–128. Cham: Springer.
- Talat, Zeerak, James Thorne & Joachim Bingel. 2018. Bridging the gaps: Multi-task learning for domain transfer of hate speech detection. In Jennifer Golbeck (ed.), *Online Harassment. Human–Computer Interaction Series*. 1st ed, 29–55. Cham: Springer.
- Wiegand, Michael, Josef Ruppenhofer & Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, 576–587. Available at: <https://aclanthology.org/2021.naaclmain.48> (accessed 10 August 2023).
- Wiegand, Michael, Josef Ruppenhofer & Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), 602–608. Minneapolis: Association for Computational Linguistics.
- Zampieri, Marcos, Shervin Malnasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 75–86. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

About the Authors

Olga Dontcheva-Navratilova is Associate Professor of English Linguistics at the Faculty of Education, Masaryk University, Czech Republic. Her research interests include English for academic and specific purposes and political discourse. She has published the books *Analysing Genre: The Colony Text of UNESCO Resolutions* (2009), *Coherence in Political Speeches* (2011) and co-authored *Persuasion in Specialised Discourses* (2020). She is co-editor of the journal *Discourse and Interaction*.

Address

Department of English Language and Literature, Faculty of Education, Masaryk University
Poríčí 9, Brno 602 00
Czech Republic

e-mail: navratilova@ped.muni.cz
ORCID: 0000-0002-0378-7975

Renata Povolná is Associate Professor of English Linguistics at the Faculty of Education, Masaryk University, Czech Republic. Her research lies in the area of discourse analysis, pragmatics and conversation analysis. She has published the books *Spatial and Temporal Adverbials in English Authentic Face-to-Face Conversation* (2003), *Interactive Discourse Markers in Spoken English* (2010) and co-authored *Persuasion in Specialised Discourses* (2020). She is co-editor of the journal *Discourse and Interaction*.

Address

Department of English Language and Literature, Faculty of Education, Masaryk University
Poríčí 9, Brno 602 00
Czech Republic

e-mail: povolna@ped.muni.cz
ORCID: 0009-0006-4392-2832

Giedrė Valūnaitė-Oleškevičienė
Mykolas Romeris University

Linas Selmistraitis
Mykolas Romeris University

Andrius Utka
Vytautas Magnus University

Dangis Gudelis
Mykolas Romeris University

OFFENSIVE LANGUAGE IN USER-GENERATED COMMENTS IN LITHUANIAN

Abstract

The aim of the current research is to investigate the feasibility of identifying offensive language in Lithuanian by utilising the Simplified Offensive Language Taxonomy (SOLT). The key principle behind this taxonomy is its ability to complement existing offensive language ontologies and tagset systems, with the ultimate goal of integrating it into publicly accessible Linguistic Linked Open Data (LLOD) resources. The dataset used in the current study is a publicly available corpus of user-generated comments collected from a Lithuanian portal (Amilevičius et al. 2016). The study identified that offensive language predominantly focuses on collective derogatory language rather than individuals. The most common category of offensive language is related to physical and mental disabilities, followed by ideological offenses, xenophobic and sexist remarks, and less frequent categories like ageism, classism, homophobia, and religious discrimination. These results highlight the diverse range of offensive language online and underscore the need to combat discrimination and promote respectful discourse, particularly concerning marginalised groups.

Keywords

offensive language, user-generated comments, offensive language taxonomy

1 Introduction

The rise of social media as a platform for discourse and the expression of contentious remarks has spurred extensive research on hate speech or offensive language instances, with the goal of pinpointing instances of insults and hate speech in online interactions. This imperative to identify offensive expressions extends beyond just social media discourse,

encompassing everyday scenarios such as work, education, and various environments. It becomes crucial not only to detect hate speech but also to address a wider range of offensive behaviours, including milder forms of discrediting, insults, and bullying. Research on offensive language can play a pivotal role in identifying inappropriate content and fostering respectful communication by diminishing derogatory comments.

In the realm of addressing offensive language, there have been numerous efforts to create datasets and apply taxonomies specifically tailored for English, as exemplified by the works of Wulczyn et al. (2017), Qian et al. (2019), and Gomez et al. (2020). The primary aim of this study is to investigate the feasibility of identifying offensive language in Lithuanian by utilising the most recent taxonomy, i.e. Simplified Offensive Language Taxonomy (SOLT) proposed by Barbara Lewandowska-Tomaszczyk (2022). SOLT is proposed as the next annotation model for the 2nd Offensive Language Annotation Campaign in WG 4.1.1. (2023) to be tested on Polish and some other languages. The foundation of the taxonomy is its complementarity to general use of offensive language ontologies and tag-set systems, while the ultimate aim is its integration with the public Linguistic Linked Open Data (LLOD) resources (Lewandowska-Tomaszczyk 2022).

Thus, the current research aims to investigate offensive language categorisation to enhance offense detection, and identification of discernible outcomes when applying this categorization in the Lithuanian context. These categories of offensive language are further refined for practical application within the Lithuanian language. The research contributes to classifying different offense types based on their varying degrees of prevalence in the Lithuanian language usage, ranging from the most to the least frequently used.

The sample under examination comprises a total of 200 comment files with 85 instances of offensive speech (this accounts for 20.5 % of all the sentences examined within our sample), encompassing 415 sentences and containing a cumulative count of 2529 words. This sample underwent a rigorous annotation and curation process utilising the INCEpTION tool (<https://github.com/inception-project/inception>), which yielded a well-structured dataset organised into tab-separated fields. The latter serve the crucial purpose of retaining the annotations associated with instances of offensive speech within the dataset, facilitating detailed analysis and exploration of the offensive language patterns.

2 Review of literature

A significant proportion of studies aimed at devising methods to detect various forms of offensive content predominantly focus on the English language. This biased situation partially stems from the availability of annotated datasets primarily in English.

However, this trend is shifting, with an increasing number of languages embarking on the annotation of their own datasets. There have been recent attempts to identify offensive language multilingually such as detection of hate speech against immigrants and women in Twitter (Basile et al. 2019) and Multilingual Offensive Language Identification in social

media (Zampieri et al. 2020). Natural language processing research focused on both: the application of neural network architectures to corpora of different languages (Liebeskind and Liebeskind 2018; Mitrović et al. 2019), and finetuned pre-trained language models applying BERT and RoBERTa (Liu et al. 2019; Swamy et al. 2019) achieving good results with a pre-trained BERT model for identifying offensive language such as hate speech and abusive language in social media. There are also lexicon-based approaches to research, for example HurtLex (Bassignana et al. 2018) which contains words of offensive, aggressive, and hateful character in more than 50 languages.

In the Lithuanian context, there is an active endeavour to tackle the issue of identifying hate speech in online comments (Ruzaitė 2018). Ruzaitė's work offers early insights from an ongoing investigation into hate speech within Lithuanian online comments. Ruzaitė underscores, that online comments frequently manifest behaviours such as flaming and trolling, signifying a notable inclination toward hostility, indecency, and a heightened presence of offensive and aggressive language. These components possess the capacity to cause harm to individuals or entities.

It is imperative to recognise not only outright hate speech or offensive language instances, but also a wider range of offenses, including subtle discrediting remarks, insults, and instances of bullying.

Researchers observe that language is not inherently offensive and the offense might not be taken even in the cases of the use of potentially offensive language, although Culpeper (2011) observes that there are certain discourses more prone to impoliteness strategies, for example, military or police training. However, he also admits that even in the contexts of sanctioned impoliteness the offense might still be taken.

Haugh and Sinkevičiūtė (2019) propose that offence might be taken as a moral emotion or a social action. When offence is taken as moral emotion, feeling offended might be considered as inherently subjective experience as it is related to the moral evaluation of the expression of the participating party. As social action offense might be oriented to delicate or sensitive social actions, and as such offense is socially bound the offending party is considered to be responsible for moral transgression or for the perceived offensive conduct no matter if the offense has been taken by anyone. Society reacts with strong indignation to morally deviant actions even if they do not directly affect the people expressing their reactions (Gunthner 1995).

Disagreement or opposition between different standpoints may trigger some kind of offense provoking deeper conflicts and arguments. The inherent complexity of offense has been analysed both from the behavioural and linguistic perspectives. Behaviour related typologies focus on the offensive forms of conduct. Durant (2010) identifies prototypically offensive types of behaviour such as bad language, insults, hate speech, incitement, unspeakable ideas and blasphemy. Barrow (2005) also observes that such forms of behaviour as racist language are inherently offensive even if no one has taken an offense although the subject of the notion of inherently offensive is still disputable (Moulinou 2014).

The emotional impact of experiencing an offense can range from mild irritation and annoyance to a profound and deeply felt shock. The severity of the offense determines the depth and lasting impact it inflicts, as noted by Hatzis (2021). Offenses can be deliberate, targeting specific individuals, or they can unintentionally affect others without direct reference to them. A case in point is when a derogatory remark is made on social media about a group defined by characteristics like religion, race, ethnicity, or sexuality. Even though the comment is not directed at a particular individual, someone who identifies with that group may feel insulted. Importantly, the intention to insult might not exist, and offense can arise due to carelessness, ignorance, or lack of consideration (Hatzis, 2021: 8). The offensiveness of particular words or phrases is not inherent; rather, it is conferred by the value assigned to them within a specific language community. Some words, expressions, and ideas have become linked to offensiveness in a way that their usage signals it within that community.

In the context of our study, we adopt the working definition of offensive language proposed by O'Driscol, which characterises it as "any word or sequence of words capable of negatively impacting the sense of self and/or overall well-being of those exposed to it." (O'Driscol 2020). In essence, offensive language has the potential to evoke feelings of discomfort, insult, hurt, or even fear, ranging from mild to intense emotions. Stollznow (2020) further elucidates offensive language as expressions directed at us that we find morally repugnant or personally demeaning. This kind of language penetrates to the core of our values or identity, eliciting a profound impact on a fundamental level. Recognising the need to encompass a wide spectrum of offenses, ranging from more subtle instances of undermining to direct insults, remains crucial within exploration (Risch et al. 2020).

Language-based typologies rely on linguistically expressed offense which is theorised as impoliteness strategies by Culpeper (2016) including cases which could be explicitly intended to offend the receiver of the message, for example direct sayings "you are stupid" or implicitly trying to insult or diminish the hearer of the message which places the offence on the explicit-implicit continuum. Explicit cases of offence are easier to identify, categorise and analyse, thus there are more studies on the explicit offence than on the implicit one as the implicit cases are more difficult to identify.

Exploring offensive language through research can play a vital role in identifying inappropriate content and fostering more courteous communication by curbing derogatory expressions. Research in this field is advancing rapidly, yet numerous unresolved inquiries persist, particularly regarding the classification and detection of offensive language, as highlighted by Pitenis et al. (2020). The prevalence of offensive content on social media is a cause for concern for both corporations and governmental bodies. The task of moderators involves discerning which comments should be removed due to their offensive nature, as discussed by Risch et al. (2020).

3 Research methodology

A recent significant endeavour in the realm of categorising offensive language is the Enriched Offensive Language Taxonomy introduced by Lewandowska-Tomaszczyk et al. (2021). This taxonomy was further refined and expanded based on the findings of annotating experiments aimed at enhancing and evolving the proposed Simplified Offensive Language Taxonomy (SOLT) by Lewandowska-Tomaszczyk et al. (2022). Built on earlier research, the proposed SOLT draws upon the categories initially introduced by Zampieri et al. (2019) to establish a comprehensive set of classes for representing offensive language. The resulting SOLT is a finely-grained framework that systematically categorises the concept of offensive language. It is structured into subcategories, organised in a hierarchical arrangement spanning distinct levels, as illustrated in Schema 1. This iterative refinement and expansion signify the ongoing efforts to comprehensively understand and categorise offensive language.

Schema 1. Simplified Offensive Language Taxonomy (Lewandowska-Tomaszczyk 2022)

1. OFFENSIVE [YES or NO]

2. Target 1

Individual // Group // Ind wrt Gr/Gr wrt Ind [by reference to group stereotypes]

3. Target 2

present//absent

4. Vulgar [YES or NO]

5. Choose either (i) or (ii); Then select (iii) or (iv) or both (iii) and (iv)

(i) INSULT [addressed to: individual or group - varied offense types but not by group stereotypes]

(ii) HATE SPEECH [individual or group; offense by reference to group stereotypes]

(iii) DISCREDIT [individual or group//on various grounds – lying-cheating, immorality, unprofessionalism, unfairness]

(iv) THREAT [individual or group, inducing fear]

6. Aspects – [Choose one or more]

[racist] [xenophobic] [homophobic] [sexist] [profane (religion)] [ageism] [physical/mental disabilities] [ableism]] [social class [classism]] [ideologism] [other]

7. Select categories below – [Choose one or more]

RHETORICAL QUESTIONS

METAPHOR

SIMILE

IRONY

EXAGGERATION

OTHER

While SOLT employs a set of meticulously examined linguistic criteria for identifying offensive categories, it is important to note that their conceptual definitions may not encompass all the necessary and sufficient properties required for a completely error-free

identification of such categories in discourse. These categories inherently possess linguistic constraints that make them somewhat indistinct; however, they maintain cohesiveness and form a larger, albeit intricate, conceptual category system by sharing the common feature of offensiveness. Subsequently, as one moves down the hierarchy, these categories also share at least one, if not more, properties linking the higher and lower ranks.

SOLT presents a promising and potentially effective approach to addressing the challenge of detecting offensive language. It is important to clarify that achieving complete inter-annotator agreement is not to be expected. Linguistic categories, including these, tend to have indistinct boundaries and lack precise definitions. However, the degree of this inherent ambiguity is expected to decrease in direct correlation with the success rate of annotation.

The INCEpTION tool, which provides an effective environment for various text-annotation tasks, is used for annotation of the offensive language. This semantic annotation platform offers intelligent annotation assistance and knowledge management. The platform was configured to enable the annotators to annotate texts on a sentence-level selecting one or a few successive sentences containing instances of offensive language. The selected chunks were annotated using all the parameters defined by the schema. The annotation environment was set to allow the overlapping selections if there was a need to annotate multiple meanings. The annotation was performed independently by two different annotators and later curated to obtain the gold standard for further analysis of the expression of the offensive language in Lithuanian. The sample of 200 items was annotated and curated. The Cohen's Kappa values for different taxonomical categories range between 0.6-0.78 with an exceptional result for category of Vulgar reaching the maximum of absolute agreement 1.00.

4 Dataset

The data for the research has been obtained from the Lithuanian corpus *LITIS v.1* of user-generated comments that is freely available from the CLARIN-LT repository¹. The corpus includes ca. 200 thousand comments collected between 2010-2014 from the two major Lithuanian news portals: [delfi.lt](#) and [lrytas.lt](#). As the aim of the study is not the comparison of two different portals, it was decided to analyse only comments from “Delfi.lt”.

It is noteworthy to mention that until December of 2021 “Delfi.lt” portal had allowed their readers to publicly comment news articles with very little restriction. Commentators could remain anonymous and unregistered. However, their IP addresses could be seen, and the content was manually moderated. Due to such rather liberal policy comment sections of news portals for many commentators have become platforms for expressing their views on many societal issues, sometimes even for spreading disinformation and hatred. It was

¹ <https://clarin.vdu.lt/xmlui/handle/20.500.11821/11>

estimated that commentators were generating 13 thousand comments daily on “Delfi.lt” portal and it was claimed that 8 % of comments that contained vulgar, offensive, and commercial information had been removed. Presently the portal has a much stricter policy for commentators: now in order to comment you need to register and reveal some private details of your identity, which prevent commentators from posting more radical or conflicting comments.

The initial corpus consists of tab separated comment files with the following information: a comment, date and time, nick name of the author, URL and a title of the article commented. For the current research only information from two columns was extracted from the corpus and uploaded to the previously mentioned INCEpTION annotation tool, i.e. file and comment:

Failas<tab>Komentaras
0<tab>gelbedamas savo sikna , idave naparnika -taip jau buna....

English gloss:
File<tab>Comment
0<tab>while saving his ass, he betrayed his partner – that's how it is....

The total size of the sample is: 200 comment files, 415 sentences, and 2,529 words. The sample, which was annotated and curated with the INCEpTION tool, resulted in the vertical data structure of tab-separated fields preserving the annotations of the offensive speech. For example, the annotation of the insulting phrase (*gelbedamas savo sikna , idave naparnika*) in the above comment is given below, where the insulting phrase is marked as directed at an individual person by using metaphorical and vulgar language with the aim of discrediting:

1-1	0-10	gelbedamas	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]
1-2	11-15	savo	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]
1-3	16-21	sikna	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]
1-4	22-23	,	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]
1-5	24-29	idave	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]
1-6	30-39	naparnika	Other[1]	Metaphor[1]	
		Insult[1]	Individual[1]	absent[1]	Yes[1]
		*[1] * [1]	Discredit[1]	*[1] * [1]	*[1]

1-7	40-41	-	-	-	-	-	-	-
1-8	41-45	taip	-	-	-	-	-	-
1-9	46-49	jau	-	-	-	-	-	-
1-10	50-54	buna	-	-	-	-	-	-
1-11	54-55	.	-	-	-	-	-	-
1-12	55-56	.	-	-	-	-	-	-
1-13	56-57	.	-	-	-	-	-	-
1-14	57-58	.	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-

All in all, we found 85 phrases of offensive speech according to the SOLT, which is 20.5 % of all analysed sentences in the sample. We have imported the vertical data structure with all relevant metadata and annotations into *MS Excel* and *Python Pandas DataFrame* for further analysis.

5 Research results and discussion

Our rigorous examination of the samples involved a meticulous annotation and curation process, which was facilitated through the use of the INCEpTION tool. This process resulted in the creation of a structured dataset, comprising tab-separated fields. These fields play a pivotal role in preserving the annotations associated with instances of offensive language within the dataset, ultimately facilitating a thorough and detailed analysis of offensive language patterns.

Table 1 in this study offers an overview of the annotation outcomes pertaining to the primary categories of offensive language. Within the table, information is presented in terms of counts and their respective percentages relative to the total number of sentences analysed. Additionally, the table includes one illustrative example per each category of offensive speech in Lithuanian, accompanied by their corresponding English translations (English gloss).

Table 1: The main categories annotated

Categories	Count	%	Lithuanian example	English Gloss
Group	53	62.35	Panašu į visišką koalicijos neįgalumą.	Looks like total coalition disability.
Individual	20	23.53	Tik moteriška, o veidelis kaip Uspaskio, tik čiuč gražesnis.	Only female, and the face is like Uspaskis', only slightly prettier.
Non-targeted	8	9.41	Bobelių pletkynė	A nitwit women gossip club.
Individual with reference to group / Group with reference to individual	4	4.71	Svarbiausia, kad šikėjui (rinkėjui) patinka...	The most important thing is that the shitty electorate likes...
Absent	83	97.65	UAB „Darbas“ ir toliau sėkmungai „skinasi“ kelią prie valdžios lovio.	Ltd. "Labour" continues to successfully "pave" its way to the governmental trough.
Present	2	2.35	Tu pažiurėk į sudėtį, briede.	You look at the lineup, moose.
Not vulgar	72	84.71	Kol agurkai besmegeniai už Uspaskį balsuoja.	While cucumbers mindlessly vote for Uspaskis.
Vulgar	13	15.29	Gelbedamas savo sikna, idave naparnika -taip jau buna.... eik tu seno kuino subine miegot.	Saving his ass, betrayed his partner - that's how it is... go to sleep, you, old horse but.
Hate speech	1	1.18	Drebékite pedai, Neringa turi sąrašą su visais visais pedofilais Lietuvoje	Tremble fagots, Neringa has a list of all fagots in Lithuania.
Insult	84	98.82	Koks tūpas tas Varkala, jei nežino ką veikia Europos parlamentas.	What a fool that Varkala is, if he doesn't know what the European Parliament does.
Discredit	80	94.12	Dabar tik šitą durnelį į dienos šviesą ištraukė!!!	Now they just brought this turd to the light of day!!!
Threat	1	1.18	Drebékite pedai,	Tremble, fags.

Physical/mental disabilities (ableism)	9	10,59	Panasu i visiska koalicijos neigaluma.	It looks like total disability of a coalition.
Ideologism	5	5,88	Darbietės atsiduoda Rusijai ir siela, ir kūnu	Women of labour party devote themselves to Russia in body and soul.
Xenophobic	4	4,71	Jau vien del slaviskos pavardes tokį reikia kuo toliau nuo tarnybos pasiusti	Just because of his Slavic surname, he should be sent as far away from the service as possible.
Sexist	4	4,71	Bobele tai pobaisei atrodo, ar ne?	That poor old woman is scary looking, isn't she?
Ageism	1	1,18	Sis senis jau gime ministru.	That old man was already born a minister
Social class (classism)	1	1,18	Visi bambaliniai išeis Anapilin.	Those useless alcoholics will pass away.
Homophobic	1	1,18	Drebekite pedai,	Tremble, you, fagots.
Profane (religion)	1	1,18	Kazkokiu dvieju pagaliu sukryztiavimo bijoti, tas neegzistuojantis velnias.	To be afraid of a kind of two crossed sticks, that non-existent devil.
Metaphor	48	56,47	Šita ministerija- amžina klapčiukų prieglauda.	This ministry is a perpetual haven for henchmen.
Other	27	31,76	O valdžios institucijose reikalingi kompetentingi specialistai, kurie po populistines partijas nesitrainioja.	And the authorities need competent professionals who don't hang around populist parties.
Exaggeration	2	2,35	Jei jau debilas, tai iki amžiaus galو.	If one is already a moron, it's to the end of one's life.
Simile	1	1,18	Tik moterishka;) O veidelis kaip Uspaskio, tik ciut grazesnis.	Only feminine;) And the face is like that of the man Uspaskis, but prettier.
Irony	2	2,35	Zalgirio vyriausiaiš treneris bus vitalija katunskyte.	The senior coach of a basketball team will be Vitalija Katunskyte.
Rhetorical question	1	1,18	Ar seniai rašyt išmokai?	How long ago did you learn to write?

The analysis of the annotated dataset reveals that in the TARGET category prevails offensive speech targeted to the group (62.3 %) with the offense targeted to the individual (23.5 %) and minor percentages for non-targeted (9.4 %) and individual with the reference to the group or group with the reference to the individual (4.7 %). Due to the nature of the dataset containing comments in the news portals in most cases the target is absent (97.7 %) with rare cases of the target present (2.4 %) when there are personal answers to the comments. The groups targeted in the offensive comments include political or electoral groups or protected groups related to age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation.

Another interesting observation is that the incidents of offensive speech are seldom vulgar (15.3 %) with prevailing non-vulgar offensive instances (84.7 %) which could be since the comments are public. The awareness of the publicity may induce certain restraint to use vulgarisms. Another interesting observation concerning VULGAR category is the inclination to use euphemisms to avoid the direct harsh vulgarisms like in the example in Table 1 *eik tu seno kuino subine miegot* ['go to sleep, you, old horse butt'] where the word *subine* ['butt'] is the euphemism of the Lithuanian word *šikna* ['ass']. The use of euphemisms also demonstrates the strategy to soften the vulgar expressions.

Also, there is very low percentage of hate speech (1.2 %) with prevailing insult category (98.8 %) which again could be caused by the public nature of the comments as the commenting people are aware of the possible legal consequences of hate speech and hate instigation. A similar situation could be observed while analysing the categories of DISCREDIT and THREAT, where threat makes only 1.2 % and the discrediting cases amount to 94.1 %.

It seems that the public nature of the comments restrains the commenting people from targeting individuals or using abundant vulgarisms or taking the means of hate speech or threats. The smaller proportion of vulgar, hate and threat messages may also mean that such comments have been reported and removed from the comment section by a moderator.

When examining various ASPECTS of offensive language, it was observed that the highest frequency of offensive comments, at 10.59 %, pertains to comments related to physical and mental disabilities. This prevalence may be attributed to the common tendency of language users to employ derogatory depictions of individuals with mental or physical illnesses as a means of diminishing the intrinsic human qualities of others and asserting a higher social status for themselves. In the second position, ideology accounts for 5.88 % of offensive speech occurrences, reflecting the polarization of society along political lines and its consequent influence on the prevalence of offensive discourse.

Offensive speech also encompasses xenophobic and sexist remarks, accounting for 4.71 % of the instances. These derogatory comments are directed towards individuals whose national identity differs from the majority in their country, essentially portraying this segment of society as lacking the same rights as the majority. Sexist remarks, on the other hand, are predominantly aimed at women, subjecting them to inferior treatment and reinforcing gender inequality.

The remaining cases of offensive comments are relatively infrequent, each constituting only 1.18% of the total. These instances are linked to ageism, classism, homophobia, and religion. However, it is noteworthy that these categories have fewer occurrences, often appearing as isolated incidents. This limited prevalence can be attributed to the specific focus of the comments, which typically do not revolve around these particular topics.

Regarding stylistically distinctive categories such as metaphor, rhetorical question, simile, irony, and exaggeration, the predominant category employed is metaphor, comprising a substantial 56.47% of instances. This prevalence can be attributed to the widespread utilization of metaphors as a primary picturesque means of conveying ideas, both in the broader context of language and specifically within commentary. Conversely, the utilization of the remaining stylistic categories is notably less frequent, ranging from 2.25 % to 1.18 %. The remaining texts, constituting 31.76 % of all instances, exhibit a lack of marked stylistic features.

6 Conclusions

The primary objective of this study was to explore the potential for identifying offensive language in Lithuanian using the newly proposed SOLT developed by Lewandowska-Tomaszczyk in 2022. The research provides a linguistic framework for identifying offensive language categories through the use of empirical corpus-based techniques. This framework is supported by a review of pertinent literature and validated using computational methods. Findings revealed that offensive speech primarily targeted groups, with relatively low occurrences of vulgar language, hate speech, and threats due to public and moderated comment sections.

The annotated Lithuanian dataset demonstrates that inherently the offensive speech is targeted to the groups (62.3 %) rather than individuals, highlighting the prevalence of collective derogatory language, e.g., political, electoral or protected groups of the society based on age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation. Due to the nature of the dataset comments being public and probably checked by the moderators, the vulgar speech is avoided or softened by euphemisms. Hate speech and threats are very rare reaching only a minor 1.2 %, as such incidents are legally binding in Lithuania, especially hate incitement or open threats.

The relatively low occurrence of vulgar language (15.3 %), hate speech (1.2 %), and threats (1.2 %) within the dataset indicates that the public and moderated nature of comment sections may discourage users from resorting to more extreme and legally questionable forms of offensive communication. This suggests a degree of self-regulation among online commenters due to the potential consequences of their expressions.

Offensive language pertaining to physical and mental disabilities is the most frequent category (10.59 %), reflecting a disturbing trend where derogatory depictions of individuals with disabilities are commonly used to demean others and assert social dominance.

Ideological offenses (5.88 %) highlight the impact of political polarization on offensive discourse, indicating that divisive political views often fuel offensive language. Xenophobic and sexist remarks (4.71 %) target individuals based on their national identity and gender, illustrating that offensive comments disproportionately affect marginalised groups. This underscores the need for measures to combat discrimination and promote respectful online discourse. Less common categories, such as ageism, classism, homophobia, and religious discrimination (1.18 % each), while less frequent, underscore the diverse range of offensive language present online, often appearing as isolated incidents in comments that primarily focus on other topics.

The analysis reveals a significant predominance of metaphor usage among various stylistically marked categories, with metaphor accounting for a substantial 56.47% of instances. This prevalence can be attributed to the widespread preference for metaphorical expressions as a primary tool for conveying ideas. In contrast, the utilization of other stylistic categories such as rhetorical questions, similes, irony, and exaggeration is notably less frequent. Additionally, a considerable portion of the analysed texts, comprising 31.76 % of all instances, lacks marked stylistic features. These findings underscore the significance of metaphor in language and discourse, shedding light on its prominent role in conveying nuanced meaning and facilitating effective communication.

To summarise, offensive language predominantly focused on collective derogatory language rather than individuals. The most common category of offensive language related to physical and mental disabilities, followed by ideological offenses, xenophobic and sexist remarks, and less frequent categories like ageism, classism, homophobia, and religious discrimination. These results highlight the diverse range of offensive language online and underscore the need to combat discrimination and promote respectful discourse, particularly concerning marginalised groups.

Sources

Amilevičius, Darius & Mažvydas Petkevičius. 2016. *LITIS v.1*, CLARIN-LT digital library in the Republic of Lithuania. Available at: <http://hdl.handle.net/20.500.11821/11> (accessed 13 November 2023).

References

- Barrow, Robin. 2005. On the duty of not taking offence. *Journal of Moral Education* 34(3). 265–275.
Basile, Valerio, Cristina, Bosco, Elisabetta, Fersini, Debora, Nozza, Viviana, Patti, Manuel Francisco, Rangel Pardo, Paolo Rosso & Manuela Sanguinettiet. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 54–63. Minneapolis: Association for Computational Linguistics.

- Bassignana, Elisa, Valerio Basile & Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR workshop proceedings*. Vol. 2253. CEUR-WS, 1–6. Torino: Academia University Press.
- Culpeper, Jonathan. 2011. *Impoliteness: Using language to cause offence*. Vol. 28. Cambridge: Cambridge University Press.
- Culpeper, Jonathan. 2016. Impoliteness strategies. In Alessandro Capone & Jacob Mey (eds.), *Interdisciplinary studies in pragmatics, culture and society*, 421–445. Cham: Springer.
- Durant, Alan. 2010. *Meaning in the media: Discourse, controversy and debate*. Cambridge: Cambridge University Press.
- Gomez, Raul, Jaume, Gibert, Lluis Gomez & Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1470–1478. Ithaca: Cornell University.
- Günthner, Susanne. 1995. Exemplary stories: the cooperative construction of moral indignation. *VS* 70–71. 147–175.
- Hatzis, Nicholas. 2021. *Offensive speech, religion, and the limits of the law*. Oxford: Oxford University Press.
- Haugh, Michael & Valeria Sinkevičiūtė. 2019. Offence and conflict talk. In Matthew Evans, Lesley Jeffries & Jim O'Driscoll (eds.), *Routledge handbook of language in conflict*, 196–214. London: Routledge.
- Lewandowska-Tomaszczyk, Barbara, Slavko, Žitnik, Anna, Bączkowska, Chaya, Liebeskind, Jelena Mitrović & Giedrė Valūnaitė Oleškevičienė. 2021. LOD-connected offensive language ontology and tagset enrichment. In Sara Carvalho & Renato Rocha Souza (eds.), *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*, 135–150. CEUR Workshop Proceedings. Warden: Dagstuhl Publishing.
- Lewandowska-Tomaszczyk, Barbara. 2022. A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies* 10. 213–227.
- Liebeskind Chaya & Shmuel Liebeskind. 2018. Identifying abusive comments in Hebrew Facebook. In *2018 IEEE International conference on the science of electrical engineering in Israel (ICSEE)*, 1–5.
- Liu, Ping, Li Wen & Zou Liang. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, 87–91. Minnesota: Association for Computational Linguistics.
- Mitrović, Jelena, Bastian Birkeneder & Michael Granitzer. 2019. nlpUP at semeval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th international workshop on semantic evaluation*, 722–726. Minnesota: Association for Computational Linguistics.
- Moulinou, Iphigenia. 2014. Striving to make the difference: Linguistic devices of moral indignation. *Journal of Language Aggression and Conflict* 2(1). 74–98.
- O'Driscoll, Jim. 2020. *Offensive language: Taboo, offence and social Control*. London: Bloomsbury.
- Zesis, Pitenis, Marcos Zampieri & Tharindu Ramasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the twelfth language resources and evaluation conference*, 5113–5119. Marseille, France: European Language Resources Association.

- Qian, Jing, Anna, Bethke, Yinyin, Liu, Elizabeth Belding & William Yang Wanget. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.
- Risch, Julian, Robin Ruff & Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 137–143. Marseille, France: European Language Resources Association (ELRA).
- Ruzaitė, Jūratė. 2018. In search of hate speech in Lithuanian public discourse: A corpus-assisted analysis of online comments. *Lodz Papers in Pragmatics* 14(1). 93–116.
- Stollznow, Karen. 2020. *On the Offensive: Prejudice in language past and present*. Cambridge: Cambridge University Press.
- Swamy, Steve, Anupam Jamatia Durairaj & Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 940–950. Hong Kong: Association for Computational Linguistics.
- Zampieri, Marcos, Shervin, Malmasi, Preslav, Nakov, Sara, Rosenthal, Noura Farra & Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *arXiv preprint arXiv:1903.08983*.
- Zampieri, Marcos, Preslav, Nakov, Sara, Rosenthal, Pepa, Karadzhov, Georgi, Atanasova, Hamdy, Mubarak, Leon Derczynski, Zeses Pitenis & Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.
- Wulczyn, Ellery, Nithum Thain & Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on World Wide Web*, 1391–1399. Ithaca: Cornell University.

About the Authors

Giedrė Valūnaitė Oleškevičienė is a Vice-Dean for Scientific Research of the Faculty of Public Governance and Business and a professor at the Institute of Humanities, Mykolas Romeris University. Her scientific interests in humanities include discourse analysis, professional English, legal English, linguistics and translation research, while in the domain of social sciences, her scientific interests include social research methodology, modern education, philosophical issues, creativity development in modern education system, and second language teaching and learning. The researcher coordinated international research projects funded by the EU, publishes scientific articles, participates as a presenter in scientific conferences.

Address

Faculty of Human and Social Studies
Mykolas Romeris University
20 Ateities St.
LT-08303 Vilnius, Lithuania

e-mail: gvalunaite@mruni.eu
ORCID: <https://orcid.org/0000-0001-5688-2469>

Linas Selmistraitis has over 24 years of experience in higher education, specifically in developing and implementing quality assurance systems for higher educational institutions. He earned his PhD in the Humanities. Currently, Professor Dr. Linas Selmistraitis holds the position of Vice-Dean for Studies at the Faculty of Human and Social Studies at Mykolas Romeris University and the position of Professor at the Institute of Humanities at Mykolas Romeris University. His interests in research are semantics, morphology, cognitive linguistics, corpus linguistics. He publishes research articles and gives presentations at conferences.

Address

Faculty of Human and Social Studies
Mykolas Romeris University
20 Ateities St.
LT-08303 Vilnius, Lithuania

e-mail: selmistraitis@mruni.eu
ORCID: <https://orcid.org/0000-0002-8612-3276>

Andrius Utka is an associate professor at the department of Lithuanian studies and a senior researcher at the Institute of Digital Resources and Interdisciplinary Research (SITTI), Vytautas Magnus university (Kaunas). He defended the doctoral dissertation Statistical Identification of Text Functions in 2004 (VMU, Kaunas). He was the head of Centre of Computation Linguistics in 2010-2022. He coordinated a number of national and international research projects. His research interests: statistical text analysis, language resources, computer-assisted translation, automatic summarisation, terminology extraction, and the language of disinformation.

Address

Institute of Digital Resources and Interdisciplinary Research
Vytautas Magnus University
23-216 V. Putvinskio St.
LT-44243 Kaunas, Lithuania

e-mail: andrius.utka@vdu.lt
ORCID: <https://orcid.org/0000-0001-5212-4310>

Dangis Gudelis is a professor at Mykolas Romeris University, specializing in public administration and governance. He earned his PhD in Social Sciences, focusing on performance measurement in Lithuanian municipalities. Gudelis has led and contributed to various national and international research projects, particularly in public governance and public policy. His current research interests include applications of big data and AI technologies in the public sector. He is a prolific writer, with numerous publications in scientific journals and presentations at conferences. He teaches courses at both undergraduate and graduate levels. Additionally, he has played a role in policy analysis and consultancy, advising governmental and non-governmental organizations on strategic development and public sector innovation.

Address

Faculty of Public Governance and Business
Mykolas Romeris University
20 Ateities St.
LT-08303 Vilnius, Lithuania

e-mail: dgudel@mruni.eu
ORCID: <https://orcid.org/0000-0002-9004-8029>

Khouloud Boukhris
Anglia Ruskin University

ONLINE GAMING AND LANGUAGE AGGRESSION IN A TUNISIAN ARABIC CONTEXT

Abstract

This paper intends to examine the development of conflictual interactions, how they might be resolved, and the socio-cultural norms involved, by adopting an analytical framework in an online gaming context. The current paper was inspired by Kádár and Haugh's framework as it enables me to investigate both the macro and micro aspects of (im)politeness. The study's aim is to further examine how impoliteness, language aggression and conflict are realised in two online gaming platforms, namely Fortnite and PUBG Mobile. Thus, I will explore discursively how these phenomena are subjective in a Tunisian Arabic setting and discuss how participants reach their subjective perception of conflict in ways that do not always correspond to the supposed intentions of the ostensible offender. The results indicate that conflict is subjective as it is evaluated in different ways by different gamers and could be a result of a conflictual intention. The findings also reveal that conflict may also be created/ escalated as a result of a non-conflictual intention. Thus, this paper contributes to understanding of conflict, how impoliteness can lead to conflict and the various aspects of impoliteness/the perception of impoliteness.

Keywords

impoliteness, conflict, language aggression, politeness

1 Introduction

Academic interest in Computer-Mediated Communication (CMC) began to develop during the 1970s as the popularity and availability of computers increased (Johansen et al. 1979). This interest soon diverged into various sub-disciplines, such as human-computer interaction (HCI). Since then, issues as diverse as spam, aggressive video games, online stalking, computer-related depression and cybercrime have become the subject of headlines, adverts and political campaigns (Hardaker 2012).

Aggression is defined as intentional behaviours to harm others who do not want to be harmed (Parvaresh and Tayebi 2018). Verbal aggression, in particular, refers to using words with the intention of harming others. A previous study (Veiga Simão et al. 2018) showed that adolescents' normative moral beliefs and self-efficacy beliefs mediate the relationship between their personal moral beliefs and their use of verbally aggressive content they observed in cyberbullying to communicate online.

Interest in CMC in the field of linguistics was much slower to develop. By 2010, few studies had been conducted on CMC conflict, with notable studies being Graham (2007) and Locher (2006). Haugh (2010) observed that, at that stage, there was a dearth of research on CMC politeness and impoliteness in its various forms.

One explanation for the shortage of social, linguistic and psychological research on online conflict may be that CMC is still viewed as less formal, less valid or less 'real' than other, more traditional forms of communication, such as face-to-face (FtF) communication or writing (Hardaker 2012). Synchronous CMC (SCMC), in particular, is perceived as a marginal, frivolous plaything for children (Herring and Nix 1997: 7; Merchant 2001: 295) and a time-wasting distraction in the workplace (Kim 2008). Online behaviour is an ongoing development. Its identification through psychological, linguistic and social meanings will progress randomly as there is no set standardisation. Inevitably, misconceptions and misunderstandings will continue to arise and possibly increase. However, there has been some expansion in CMC research in other directions. Linguistics-based CMC research has not been well developed and little attention has been paid to conflict interactions.

A large number of CMC environments suffer negatively marked online behaviour (NMOB), whether they are small message boards or major social networks. The BBC (2010) and Camber and Neville (2011) have discussed how offensive and provocative online posts aimed at upsetting others have become a worldwide problem. This aspect of CMC has also been extensively studied by researchers. However, Baker (2001) observes that it is difficult to trace any articles concerning trolling or verbal aggression, even as an aside, Hardaker (2010) and Shachaf and Hara (2010) confirm this, stating that trolling, aggression and offence are rarely mentioned directly. Research on politeness and impoliteness follows a similar pattern. Parvaresh (2019) have studied the flagrant use of conflictual interactions in e-communication media. They define conflict as interaction between people with strong opposing views or principles that can result in major disagreement. Those areas of the Internet that are not rigorously controlled lend themselves to being used for virulent behaviour such as blatant rudeness and face-threats. Graham (2007) and Kádár et al. (2019) indicate the growing interest in CMC and impoliteness, but it should be noted that there has been no academic study of CMC impoliteness involving Tunisian Arabic.

Bringing together NMOB research and impoliteness research could be beneficial to both parties (Haugh 2010). Conflict and verbal aggression, in particular, have similar tendencies to offline impoliteness, such as differences in what is perceived as impoliteness and conflict, and the degree of hostility that is consequently evaluated (Graham 2007). Given the overlap between conflict and impoliteness, applying impoliteness theory to conflict could also be insightful. The fact that conflict is regularly addressed by users in regard to other behaviours that are generally the focus of impoliteness research, strengthens this case even more (e.g., trolling, hostility, aggression) (Hardaker 2012).

Academics appear to have a limited understanding of not only why conflict occurs, but also what it is, which leads to a number of different evaluations when we examine the flourishing field of (im)politeness¹ (e.g., Locher 2006; Bousfield 2008; Bargiela-Chiappini and Kádár 2010; Culpeper 2011). Thus, this limited understanding of conflict drives the investigation of (im)politeness. The concept of politeness first gained popularity in the 1970s through the works of Lakoff (1989), Brown and Levinson (1978, 1987), Leech (1983) and others. Beyond academia, rudeness has been a focus in British politics, with Tony Blair's Respect Agenda (Culpeper 2011) and the introduction of Anti-Social Behaviour Orders (ASBOs) aimed at addressing harassment, alarm or distress (Wikipedia 2007) that is caused, amongst other things (Hardaker 2012), by abusive and manipulating language (Home Office 2009).

Despite the growing interest in impoliteness and CMC, as previously mentioned, few studies to date have examined impoliteness in CMC (for example, Graham 2007; Hardaker 2012; Parvaresh 2019). In asynchronous CMC (ACMC), verbal aggression has been the main focus of attention. When employing ACMC, the speaker does not presume that the hearer will respond instantly. Usenet, blogs and social media are typical forms of ACMC (e.g., Hatipoglu 2007; Hardaker 2012). When employing SCMC, the speaker usually expects the hearer to be online at the same time and to respond quickly. As a result, ACMC and SCMC can be used for a variety of interactions and norms; however, research into SCMC, such as chatrooms, instant messengers and virtual worlds, is significantly more limited (Fung and Carter 2007, cited in Hardaker 2012).

The structure of this paper is as follows. In section 2, I will provide an account of what conflictual interactions essentially mean, thereby paving the way for a definition of language aggression on the Internet. This will be followed, in Section 4, by an account of how the current study was conducted. The results and discussion of the study will be summarised in Section 5. Finally, conclusions and suggestions for further research will be given in Section 6.

2 Conflictual interactions in CMC

Many studies have considered the link between bloodthirsty video games and violent/aggressive behaviours among children and teenagers. Several studies have concluded that playing video games may temporarily boost offensive and brutal behaviours, hostile perceptions and aggressive feelings (Möller and Krahé 2009; Teng et al. 2011; Diaz et al. 2016). For example, personality characteristics, such as brutality, viciousness, neuroticism, anger and hostility, reinforce the positive relationship between playing violent video games and aggression (Greitemeyer and Sagioglou 2017; Kim et al. 2008; Lull and Bushman 2016). Furthermore, many teenagers play video games with their friends, which, consequently, influences the form and level of their aggressive behaviour, in the same manner as a collective reaction in non-digital life (Lansford et al.

2003). Furthermore, Wright (2019) examines the influence of friendship quality on console-gaming aggression (i.e., verbal aggression, camping, trolling) and how such associations might be mediated by anger and revenge planning. His study indicates that a low-quality friendship is related to increased verbal aggression, trolling and camping. In addition, these relationships are mediated by anger and revenge planning. These results suggest that friendship quality, anger and revenge planning are crucial aspects of the aggressive behaviours displayed by adolescents through first-person shooters.

However, Wright (2019) claims that video games include repetitive play, that they act as socialisation agents and that they strengthen social messages (Crawford 2011; Padilla-Walker et al. 2010). As a result, violent video games may transmit messages about the acceptability or justification of aggressiveness and violence. The literature suggests that exposure to violent digital content via aggressive video game play has links with aggression and violence in the real world and via electronic technologies (e.g., Carnagey et al. 2007; Eastin 2006; Ivory and Kalyanaraman 2007; Kim et al. 2008; Lam et al. 2013).

These connections can be found in experimental, longitudinal and correlational investigations, as well as in diverse groups such as children, adolescents and adults (e.g., Gentile et al. 2011, Möller and Krahé 2009, Willoughby et al. 2012).

Anderson and Bushman (2002) established the general aggression model to explain the links between violent video games, aggression and violence. They define the general aggression model as a social-cognitive paradigm that incorporates situational, personal and biological factors. These variables combine to produce a range of cognitive, emotional, physiological and behavioural effects, including violence and hostility. According to the model, aggression is influenced by situational and individual characteristics. Provocation (e.g., insults, irritation) and cognitive clues are examples of situational circumstances (e.g., the presence of loud music). For instance, Ohno (2022) claims that research on aggression during the playing of video games, especially the new genre of “battle royale” games, has yielded conflicting results. Cho et al. (2017) reported no significant differences in the level of aggression associated with different video game genres. In addition, their study showed no connection between aggression and the use of shooting games (Sferra et al. 2016). On the other hand, Dickmeis and Roe (2019) found a statistically significant positive correlation between physical aggression and fighting/shooting games. They also reported a positive association between violent or competitive games and physical aggression. Kaye and Bryce (2012) focused on negative social gaming dynamics, and found that unpleasant social experiences, for example abuse or swearing by players within a game, increased player aggression. Battle royale games are a new genre of game that combines competitiveness and violence. Drawing on this, my study could make a positive contribution to this conundrum by offering the reader a detailed insight into today’s online gaming environment. Particularly that many people, including teenagers and young adults, are exposed to conflict and language aggression when playing online video games. The active nature of the learning

environment associated with these video games means that their use has become widespread amongst youngsters, and in many cases their parents do not really know what is going on during these games. In other words, the public has little consensus on the rise in online aggression and conflict in video games, and therefore my study can further contribute a modern perspective to original knowledge.

As a result, further empirical research is needed to explore the relationship between the use of violent games and aggression (Ohno 2022). As previously mentioned, little focus has been given to these associations in the literature. Thus, the next section will review language aggression on the internet.

3 Language aggression on the Internet

Language aggression on the internet has long come under scrutiny; however, in terms of language aggression, the spread and development of the internet has led to growing uproar. For instance, the study of anonymity and physical detachment is another pertinent strand of research on Twitter-linked aggression, a complex participation framework brought about by social media. Furthermore, the users of social media platforms lack a physical presence, which has implications for interaction, as discussed in Lorenzo-Dus et al. (2011) and Blitvich (2015); such research has demonstrated that impolite behaviour online might be weakened by complete or partial anonymity and pseudonymity (Hardaker and McGlashan 2016; Vladimirov et al. 2021). Indeed, this anonymity gives individuals more freedom to offend and attack random individuals or groups (Parvaresh and Tayebi 2018). In this framework, Vladimirov et al. (2021) conceptualise the lack of self-awareness and inhibition present in social media through the “de-individuation effect” (Blitvich 2015). De-individuation refers to the phenomenon whereby users feel more empowered to express opinions or expressions aggressively that they would be less willing to do so in face-to-face settings. This idea of de-individuation is crucial to understanding how to analyse online aggression because it explains why these environments can lead to the escalation of hostile behaviour during interactions. Less is known about how verbal aggression, which is defined as a tendency to disagree and argue with others, is affected by the use of digital media. Theoretically speaking, exposure to aggressive language and other verbal assaults on the internet could result in teenagers becoming more verbally aggressive (Anderson and Bushman 2002). Teenagers, for instance, could learn taboo words online and create links in their minds between instances of possible conflict and the use of aggressive language (Linder and Gentile 2009). Only a few studies have examined verbal aggression in connection with internet use, particularly in regard to multiplayer online role-playing games (Appel et al. 2014). While the use of online gaming platforms is increasing, very few studies have examined impoliteness and relational work in these contexts (Graham and Hardaker 2017), and as the field has progressed, researchers have tended to study impoliteness from broadly different perspectives, including postmodern discursive, interactional and social

psychological perspectives. Drawing on this, my study will offer a more in-depth understanding of the way in which gamers function in their online communities, and thus it could make a significant contribution to our knowledge of how conflict impacts, not only the participants' actions, but also their reactions to what they are hearing. Therefore, this paper is a worthy endeavour, since it will enable scholars to explore how (im)politeness can lead to aggression and conflict when it is perceived in different ways by different gamers. In other words, this paper attempts to fill a gap in the intersection between (im)politeness, aggression and conflict, as seen in the methodology section below.

4 Data and methodology

This paper examines naturally occurring data and adopts a qualitative and inductive approach. The study employs video-recordings that include data from gamers/participants who use conflictual and offensive words during their interactions. It also adopts a pragmatic discourse analytical approach as well as a thematic analysis approach to organise and identify regular features in the data, and to provide insights into meaning patterns and themes in the data.

This study relies on naturally-occurring data to record and videotape spontaneous interactions in real-life situations. This is accomplished by paying close attention to everyday discussions and recording significant encounters as soon as they occur on two video game platforms (*Fortnite* and *PUBG Mobile*). Taylor (2001) claims that authentic discourse indicates unofficial talk which would have happened regardless of whether it was being recorded or observed, and which is not influenced by the existence of the recording tool or the observer.

The majority of the ordinary data was gathered through the use of snowball sampling with the assistance of my close network of relatives and friends, who volunteered to take part in this study. Once my relatives and friends provided their consent to participate in the study, they connected me with other gamers who were also interested in taking part. I explained the overall study aim and assured them that their contributions would remain anonymous, an alias would be employed, and their identities would remain confidential at all times. 134 participants were recorded and participated in the study.

When the match finishes, I directly collect my recording devices (the camera and the wireless microphone) from the participants to prevent their speech patterns from changing or behaving in a different manner as a result of the researcher's presence. Each team member was asked to record their interactions when they felt ready, without any restrictions concerning the period or subject of the interactions. In order to reinforce and maintain neutrality, the researcher did not intervene during the matches, performing the role of an observer. Except for participating friends or those individuals who had prior contact with the researcher, the presence of the researcher did not appear to bother them (as they were accustomed to the researcher) or affect them in any way, particularly once the match began.

This was evident as their typical communication style and conduct during the recorded conversations showed consistency. The data was then qualitatively analysed with a view to establishing any recurrent themes and patterns. In pursuit of this aim, I was particularly influenced by Kádár and Haugh's (2013) framework which surpasses the discursive dual comprehension of politeness. In other words, first-order politeness (that of the general user) and second-order politeness (theoretical perspectives) are developed further by the use of this framework (Kádár and Haugh 2013: 86):

1. Participant/meta-participant understandings (first-order).
2. Emic/etic conceptualisations (first-order).
3. Analyst/lay-observer understandings (second-order).
4. Theoretical/folk-theoretic conceptualisations (second-order).

Furthermore, as this model infers from multiple loci of understanding, researchers can study the macro-aspects of politeness, while still focusing on micro-features. This third-wave framework also encompasses the situated understandings of analysts, allowing it to be systematically theorised, as opposed to many other approaches to politeness in which a particular context is used to develop an objective understanding of the subject (Asswae 2018).

5 Results and discussion

I have chosen individual examples to illustrate the aggressive and conflictual interactions that are realised differently between the gamers. Many patterns have emerged from the data, that is why there are different themes which focus on how the triggering of impoliteness and conflict is dynamic in online gaming contexts: 1) Conflict could be created/escalated as a result of a conflictual intention; 2) Conflict could also be created as a result of a non-conflictual intention.

5.1 Conflict creation/escalation as a result of a conflictual intention

This theme involves exchanges where conflicts are caused intentionally and thus are evaluated as conflictual and impolite. As a result, conflict arises between participants. Indeed, in the exchanges below, participants react negatively to conflict, which is obvious in their voice tones and aggressive reactions. Thus, justifiable aggression is common between gamers in the exchanges. This theme aims to classify all the exchanges that include negatively perceived intentions which can lead to the creation/escalation of conflict and/or justifiable aggression between the gamers.

5.1.1 Exchange

In this exchange, the interaction takes place between four randomly selected gamers, referred to here as DrKiller, Sniperk, CandyButcher and MrAboo. Each player was muted and so no player talked in the lobby. When the match starts, DrKiller begins by a conflictual intention, thus raising tensions between the teammates (see Appendix A and B).

٦. DrKiller: تعرف كان تكمبي يا كلب /
{/taʃrif kan tkampi ja: kalb} / Know just camping dog

(.) ٧. Camping is the only thing you know, dog

2. DrKiller: يلعن والديكم ولاد القبيبة: /
{/jalf'sin waldi:kum wla:d qahba} Curse your parents son of bitches

(.) Fuck your parents, sons of bitches

3. CandyButcher: [angry voice] شبيك داخل تسب لابس ما و؟ /
{/ʃbi:k daxil tsib laba:s maw}

✓ Why you enter insulting are you okay

(.) ٩. Why are you insulting us? Are you okay?

4. MrAboo: لعنة الله عليك و الله ما نشوفك نكسر لك وجهك: /
{/laʃnɪt allah ʃli:k vallah ma nʃu:fik nkasirlik wiʒhik} Curse upon you I swear when I see you I will punch your face

(.) ١٠. Curse upon you. Believe me I will punch you in the face when I see you

5. DrKiller: هاني نستني فيكم يا كلاب: /
{/ha:ni niştana fi:kum ja: kla:b} I am waiting for you dogs

(.) I am waiting for you dogs

6. Sniperk: امان علاه ديمها هكا يصبر رانا نفس الteam /
{/ama:n ʃla:h di:ma haka js'i:r rana nafs ti:m} Please why this always happens we are one team

(.) ١٢. Why is it always like that? We are one team

7. CandyButcher: واحد شلاكة حاي بوري في روحه /
{/wahid fla:ka 3a:j jwari fi: ru:hu:} Someone slippers comes to show off

(.) He is just a cheap person who likes to show off

8. MrAboo: تقوه عليه طحان /
{/tfu:h ʃli:h t'fahan} ✓ Spit on him bastard

(.) ١٤. Spit on that bastard

5.1.2 Analysis

Conflict creation:

In turn 1 of this example, the conflict is marked as DrKiller begins in a high, sharp tone with “Camping is the only thing you know, dog”, a conflictual intention.

Conflict avoidance:

On the other hand, the interaction represents a moralising reaction to a moral violation when his teammates continue to avoid conflict by muting their microphones.

Conflict escalation:

The conflict escalates as DrKiller continues to insult them, as seen in turn 2, “Fuck your parents, sons of bitches”. As a result, CandyButcher perceives his conflictual intention as offensive and genuinely impolite. Thus, he replies in an angry, loud voice in turn 3, “Why are you insulting us? Are you okay?”, in an attempt to resolve the conflict. Similarly, in turn 4, “Curse upon you. Believe me I will punch you in the face when I see you”, another conventionalised impoliteness formula (a threat and a curse) is employed by MrAboo, who also perceives DrKiller’s intention to be conflictual. The players believe that by ignoring DrKiller the conflict will be resolved. However, he continues to be intentionally disrespectful towards them by complaining about the way they are playing the game, which escalates the conflict.

Conflict resolution:

In turn 6, “Why is it always like that? We are one team”, Sniperk tries to resolve the conflict even though he perceives DrKiller’s intention to be very conflictual and impolite, as indicated by the content of his words. Also, by saying “We are one team”, Sniperk believes that the conflict will be resolved and confirms that DrKiller has upset the moral order, which has allowed conflict to arise. However, in turns 7 and 8, the interaction is reduced from a violent reaction to a moral trespass and goes beyond his expectations.

Conflict re-escalation:

Despite Snipek’s efforts to avoid conflict and to calm tensions between the players, verbal aggression is used to attack DrKiller - reference is made to his conflictual intention and his poor playing skills - which is again face-threatening for DrKiller. The teammates were randomly selected and therefore did not know one another before the game, and so they perceive DrKiller’s intention as being extremely impolite.

5.2 Conflict creation/escalation as a result of a non-conflictual intention

This theme involves exchanges where conflict could also be created as a result of a non-conflictual intention and/or mocking intention. Consequently, conflict takes place between participants, which, in most cases, causes the conflict to escalate. Indeed, in the exchanges below, participants react negatively to non-conflictual intentions, which is obvious in their voice tones and aggressive reactions. Thus, this theme aims to classify

all the exchanges in which conflict is the result of a non-conflictual intention, but which leads to the creation/escalation of conflict between the gamers.

5.2.1 Exchange

The following interaction is between four randomly selected gamers, referred to here as Killer, Sniper, Assassin and Murder. Killer is mocking his teammates in an attempt to get them to unmute their microphones. Murder decides to keep his microphone muted during the match, leading to the following interaction.

1. Killer: [mocking tone] ياخى باش نقعد نحكي وحدى يازح شيئاً الحالة هاذى؟ {ja:xi ba;ʃ noqfid naħki wahdi ja: zah ſnija ilħala ha:ði}
Are you going to stay talk me alone what the situation this
(.) Am I going to speak to myself? What the hell is this?
2. Sniper: ياخى جايين شنحکيو هنا خلي الواحد بيرکز {ja:xi ɻa:jin ſnaħki:w hna xali ilwaħid yrakiz}
↗ Is it come to speak here let the one focus
(.) ↗ We are not here to speak to each other. Let me focus
3. Assassin: خلينا مرکzin مناش جايين باش نتعرفوا {xali:na mrakzi:n mana;ʃ ɻa:jin ba;ʃ nitħarfu:}
↗ Let us focus we are not coming for acquaintance
(.) ↗ We are not here to get to know each other. Let us focus
4. Killer: [mocking tone] ياخى انا حاجتي بمناظركم ماو باش نعلمكم كيفاه تلعنو {ja:xi ana ha:zti bħħanaðiirkum maw ba;ʃ nħalimkum kifa:h tħallfu:}
Is it me I need your faces just to show you how to play
(.) I don't want to know any of you. I just want to show you the right way to play
5. Sniper: تي هاك اكابر كامبر و تحب تعلمينا كيفاه تلعنوا {ti: ha:k akbir kamper w thib tħallimna ki:fa:h nħalibu:}
You are the biggest camper and you want to teach us how to play
(.) Apparently, you are the greatest camper, and you want to show us how to play?
Murder is still ditching the voice chat.
6. Killer: شكون الكامبر يا طحان؟ {ʃku:n ilkamper ja: f'ħan}
7. Sniper: امك الكامبر ↗ {umuk ilkamper}
- ↗ Your mother the camper
(.) ↗ Your mum is a camper
Assassin mutes his microphone.
8. Killer: لا ما نوريك يا ولد القبيبة ↗ {ila ma nwari:k ja: wild ilqahħba}
- ↗ I will show you oh son of a bitch

- (.) ↗ I am going to show you how things work, son of a bitch
9. Sniper: جرب القاني و اعمل اش تحب {zarib ilqa:ni w a:fmil a;fthib}
Try find me and do whatever you love
(.) If you find me, do whatever you want
10. Killer: تفوه عليك يا مبيون² {tfu:h fli:k ja: mibu:n}
(.) ↗ Spit on you, arsehole
The team loses the game.

5.2.2 Analysis

Conflict creation:

From the beginning of the match, Killer appears to behave sarcastically with the way in which his teammates are playing and muting their microphones as they have failed to act according to his expectations. The mocking intention with which the interaction starts, “Am I going to speak to myself? What the hell is this?”, creates conflict and sheds light on the fact that the other teammates have perceived Killer’s behaviour and mocking tone of voice to be conflictual, “We are not here to speak to each other. Let me focus”. Thus, Killer continues to mock his teammates with “I don’t want to know any of you. I just want to show you the right way to play”. According to his intention, Killer is playing not only to win the match, but also to socialise with different teammates. However, the players go far beyond his expectations and refuse to unmute their microphones, and thus he feels justified in saying that he does not want to get to know them, he just wants to teach them how to play the game. He feels that it is easier to share game tactics and coordinate plans to stay alive. However, from Sniper and Assassin’s point of view, Killer’s intention is conflictual, and the focus should be on the match and not on establishing social relationships, as shown in their loud voice tone (turns 2 and 3), “↗↗ We are not here to speak/get to know to each other. Let me/us focus”.

Conflict avoidance:

Murder chooses to mute his microphone during the first part of the match to avoid conflict. Supposedly, Assassin’s claim that “We are not here to get to know each other” has caused offence and escalated the conflict.

Conflict escalation:

Moving on to turn 4, it can be seen that the conflict between the gamers increases when Killer mocks his teammates by saying, “I just want to show you the right way to play”. Then, as demonstrated in turn 5, Sniper perceives Killer’s intentional behaviour to be offensive as he shows him the correct way to play. Thus, he tries to make Killer feel inferior by describing him as “the greatest camper”.

Justifiable aggression:

Sniper has taken offence as he uses reciprocal impolite exchanges to save his face and justify his belief that Killer mocking aggressive intention is conflictual. It is worth noting

that the impoliteness reciprocity that is evident in this example can often be proven in many impolite situations (Culpeper and Tantucci 2008). The intense negative reactions of Sniper and his angry tone of voice are the result of being abused by the negative vocative assertion “bastard”. The presence of a wide range of insults from both Killer and Sniper, such as “Your mum is a camper”, “son of a bitch”, “arsehole”, etc., has escalated the conflict.

Anonymity encourages Sniper to be more impolite than he might otherwise have been, as he says, “If you find me, do whatever you want”, and his intentional comments are also intended to add fuel to the fire. Graham and Hardaker (2017), (cited in Culpeper et al. 2017), confirm that the “anonymity of the digital environment would lead to chaos and harassment by people who could not be held accountable for their behaviour” (p. 789).

5.2.3 Exchange

The following interaction is between four players. Players 1, 2 and 3 are friends, while Player 4 has been randomly selected. When the match starts, Player 4 tries to mock his teammates and, as a result, tensions increase between the players.

1. Player 1: خلينا ندخلو في الدار هادي: {xali:na nodxlu: fi: aldar haði:}
Let us enter in the house this
(.) Let's enter here
2. Player 2: رد بالك يظهرلي فما شكون: {rod ba:lik yoðhorli: fama ðku:n}
Be careful I think there is someone
(.) Be careful, I think there is someone in here
3. Player 4: رد بالك يظهرلي فما شكون [mocking tone] {rod ba:lik yoðhorli: fama ðku:n}
(.) Be careful, I think there is someone in here
4. Player 1: شبيه هذا يعاود في الكلام {ʃbi:h haða ðʃa:wid fil kla:m}
What's wrong with him he is repeating the words
(.) Why is he imitating us?
5. Player 4: شبيه هذا يعاود في الكلام {ʃbi:h haða ðʃa:wid fil kla:m}
What's wrong with him he is repeating the words
(.) Why is he imitating us?
6. Player 3: شكون يلي يتكلم اصلا والله فيبالى يلعب معانا بوت {ʃku:n jali jitkalim as'lān vallah fiba:li yalfib mʃa:na bot}
Who is speaking originally I swear I thought he plays with us bot
(.) Who is talking by the way? I thought number 4 is a bot³
7. Player 2: كانك تفلاك راو ماسط تفلايك يا 4 ياسر تفوريخ راو: {ka:nik tfadlik raw ma:sit' tfadli:kik ja: 4 ja:sir tfu:rix raw}

- If you are joking it is boring your joking you 4 very childish by the way
(.) If you are joking, be aware that your jokes are so silly. You are so childish
8. Player 3: ظاهر فيه مازال کی تولد: {ð'ahir fi:h maza:l ki touldid}
It seems he is just born
(.) It seems like you are a newborn
9. Player 4: متعروفش مع شکون طایحین تلو نوریکم البوت شباش يعمل يا موينا: {mataðrfu:f mða:jku:n t'ajhi:n ta:w nwarikom ilbot ðba:j jaðmil ja: mwabna}
↗ You don't know with whom you are dealing I will show you what the bot can do
(.) ↗ You don't know who I am. I will show you what the bot is going to do bastards
10. Player 1: يزي من فلة التربية راو ما جييش هكا: {jizi min qilit turbja raw ma: jzi:f haka}
↗ Stop from impoliteness by the way it does not come like that
(.) ↗ Stop being impolite. What you are doing does not make sense
11. Player 3: ماسطها اللعنة يا زاح: {ma:sit'ha luðba ja: zah} ↗ He makes it boring
(.) ↗ He made the match boring
12. Player 4: بلعو فامك: {balðu: fa:mkum} ↗ Swallow your mouths
(.) ↗ Shut the fuck up
13. Player 4: يا ولاد القحبة ما فالحين في شي: {ja: wla:d ilqahba ma: falhi:n fi: saj} ↗ You son of bitches you don't know anything
(.) ↗ You are sons of bitches. You literally don't know anything about the game
14. Player 1: شبيه يحرش في اللغة سي الجحش تي برا يلعن بوك: {ðbi:h jharif fi: lu:ya si: zihf ti: bara jilðan bu:k} ↗ Why he is toughing the language mister idiot just go curse your dad
(.) ↗ Why are you behaving like that? You are a fucking idiot. Fuck your dad
15. Player 2: ملا ميبون طحنا بيه: {mala mibu:n t'uhna bi:h} ↗ Such a bastard with us
(.) ↗ He is such a bastard
16. Player 4: تقوه عليك يا طحانة: {tfu:h ðli:kum ja: t'ahana} ↗ Spit on you, you bastards
(.) Spit on all of you, bastards
17. Player 3 interrupts him: تي برا نايك طحان متاع وندي: {ti: bara najik t'ahan mtað wiðni} ↗ Go and fuck bastard of my ear
(.) ↗ Fuck off, bastard

The team loses the match.

5.2.4 Analysis

Conflict creation:

The initial interaction between Players 1, 2 and 3 is relatively normal, as shown above. Suddenly, Player 4 begins to repeatedly mimic them (using a mocking tone), as shown in turns 4 and 5, resulting in face damage to the other players and the creation of conflict. Player 4's intention is to joke, not to generate conflict (as shown in his mocking tone).

Conflict escalation:

The meta-comments made by Players 1 and 3, and their interpretations of Player 4's behaviour, indicate that they perceive his mocking intention to be impolite, which is clearly evident in turns 10 and 11, "Stop being impolite. What you are doing does not make sense" and "He made the match boring". They perceive Player 4's intentional behaviour as conflictual, and thus impolite, as he has violated a significant norm of politeness, that of showing respect to other people.

Moving on to turn 12, "Shut the fuck up" indicates that the interaction has failed to unfold in the way in which Player 4 would have expected. His use of silencers and his repeatedly aggressive tone of voice escalate tensions between the teammates. He attacks and insults them in a deafening tone of voice with "You are sons of bitches. You literally don't know anything about the game".

Justifiable aggression:

As a result, the other gamers respond in a similarly loud, aggressive and impolite manner by using a personalised negative vocative "You are a fucking idiot", condescension "newborn" and "childish", a personalised negative assertion "He is such a bastard" and a dismissal "Fuck off", which indicates that they have been offended by Player 4 and perceive his behaviour as genuinely impolite.

As Culpeper and Hardaker (2017) claim, "research in aggression has repeatedly shown that verbal insults and taunts are often reciprocated" (p. 213), and thus they lead to an intensifying conflict spiral. The more impolite the reaction, the more aggressive the behaviour of the hearer (the gamer) will be. Consequently, "it is in the interest of both communicators that the other be a worthy opponent—but the goals of the two parties in this case are clearly at odds (one is trying to outdo the other)" (Terkourafi and Kádár 2017: 176).

In the above context, it is quite understandable that offence has been taken when Player 4 creates conflict by using an aggressive manner and verbal insults. Player 4 generates conflict as he provocatively mimics the other players, thus escalating the conflict within the team. Their interactions are generated by conflict and aggression, which are inevitably noticeable and emotionally involved.

5.3 Connections between themes

The themes that have emerged from the data have been linked together to fulfil the purpose of this study, that is, to investigate how conflict, impoliteness and verbal aggression contribute to an interaction, to explore the creative aspect of impoliteness and to propose an analytical framework for impoliteness in online gaming.

The first theme, conflict could be created/escalated as a result of a conflictual intention, is linked to the second theme, conflict could also be created as a result of a non-conflictual intention. Indeed, conflict occurs when the hearer takes offence at the speaker's words, regardless of the latter's intention. For instance, from the exchanges it is evident that the conflicts arising between the gamers involve perception, action and reaction. In other words, even though the gamers perceive conflict in different ways, it always occurs and leads to the escalation of conflict and/or justifiable aggression in many exchanges.

Plus, various types of feelings, such as anger and sadness, and different types of reactions, such as screaming, mockery and teasing, link the themes together as they are present in each one. These attitudes shaped the gamers' behaviour in particular conflicts. In other words, gamers have different norms for conflictual behaviour. Strategies for escalating conflict are also evident in the themes, and include insults, verbal aggression, game ranking power and mockery. The themes also indicate that some gamers' reactions (screams, blows...) are reciprocal/mutual in nature, and so this reciprocal reaction is referred to as justifiable aggression (a form of reactive-expressive aggression) rather than impoliteness, since the person is saving his/her face. In other words, some interactions in the corpus follow a pattern of comparable movements, which are realised by similar rhetorical methods (Kádár et al. 2019).

Furthermore, in the themes, one can observe that conflicts elicit strong moral reactions (the use of negative vocative assertions), and the obvious violations of impoliteness norms (moral order) are interpreted negatively (e.g., face-loss), potentially leading to conflict and subsequent moral judgements. For instance, the failure to reach agreement, a lack of respect and the creation of aggression between gamers can result in loss of face, and therefore some gamers tend to leave conflicts unresolved. However, as the themes reveal, evaluations of mock impoliteness are also perceived in different ways by gamers. Many examples of provocation through the use of mockery or making fun of someone are perceived to be impolite. Kádár et al. (2019) claim that this form of provocation is, at the very least, for teasing purposes. Moreover, the themes reveal that the misinterpretation of mock (im)politeness is different between the gamers, as it ranges from connecting and a minor attack, to out-and-out attacking. Indeed, humour emphasises social convergence, yet also divergence. As revealed in the themes, several exchanges indicate that mock impoliteness can be misinterpreted, and thus excludes the others. In fact, it is possible to socially alienate interactants by making fun of them; however, in these instances gamers perceive mock (im)politeness as being conflictual.

Even amusing banter that includes participants can have a domineering edge while also promoting familiarity. Mock (im)politeness can also be perceived as an operation at multiple relational levels: partly creating familiarity and social convergence, but also carrying aggressive and therefore non-polite undertones (when misinterpreted).

To summarise, the themes reveal that conflict frequently arises between gamers. No matter how conflict intention is perceived (conflictual or non-conflictual), tension between the gamers is always present. Sometimes conflict could be created as a result of mock impoliteness as it is not perceived as such. Sometimes conflict could be created as a result of a conflictual perceived intention. The current paper has examined how conflicts are linguistically achieved, escalated, avoided, and resolved, and has also revealed how impoliteness is perceived and manifested by gamers.

5.4 Discussion

Conflict is a collapse in the ordinary mechanisms of decision making (Nair 2008). In this paper, conflict could be an evaluation of what the hearer perceives from the speaker's intention. The corpus data indicates that there can either be a match or a mismatch between perception and intention. Indeed, perceptions of conflict range from "genuine impoliteness", "mock (im)politeness" to "politeness", while conflict intentions range from "extremely offensive", "offensive", "mocking" to "friendly". Hardaker (2012) upholds this idea by confirming that regardless of how off-the-record the speaker makes her/his statement, it does not mean that her/his purpose can ever be retrievable. Only a few interpretations of her/his intention can become rationally plausible for both the speaker and the hearer. She further claims that the speaker may, for example, attempt to make her/his intentions appear very genuine, but the hearer may (appear to) attach incorrect intentions to the speaker's utterance. Hardaker (2012) also argues that the speaker can also intentionally mislead the hearer about her/his intentions, and the hearer could mislead the speaker about his/her perceptions. In summary, neither the speaker's intentions nor the hearer's interpretations can be certain. Thus, this study's data indicates that intention is subjective and thus (1) a link exists between perception and intention as seen in the participants' metalanguage, which suggests that the more intense the conflict is, the more likely the conflict will escalate, and it might become impossible to resolve; (2) a significant link also exists between perception and intention, indicating that when mockery or sarcasm is used (mock (im)politeness), the more likely it is that the conflict will be resolved or avoided; (3) another significant link exists between perception and intention, which suggests that the more intense the conflict is, conflict will be resolved and (4) a further significant link exists between perception and intention, indicating that when mockery or sarcasm is used (mock (im)politeness), the more likely it is that the conflict will escalate.

Moreover, the data suggests that conflict and impoliteness strategies are more than a feature of words or expressions as they are subjective and are an evaluation of what the

hearer perceives from the interlocutor's intention. In the corpus, there are numerous instances where gamers intend to offend their teammates, and in doing so, they use a loud, furious tone of voice; this intended offence is perceived as genuine impoliteness. Thus, the other teammates become offended, and they reciprocate their offence by escalating the conflict using 'justifiable aggression', as it is termed in the analysis. In line with this study, McDonnell (2011) states that aggressive communication is characterised as being poor in effectiveness, loud in volume and aggressive in tone. Culpeper (1996: 354) also confirms that impoliteness reciprocity may occur for a variety of reasons, including the fact that impoliteness can cause 'emotional arousal'. Parvaresh and Tayebi (2018) confirm this idea by stating that emotional arousal tends to lead to offensive reactions. Various types of feelings, such as anger and sadness, and different types of reactions, such as screaming, mockery and teasing, emerge from exchanges when conflict is created. This result builds on Culpeper's (2011) findings as he claims that every act of impoliteness has emotional implications for at least one participant, usually the target, but sometimes also the impoliteness producer (thus, the focus will be on the target). These attitudes shaped the gamers' behaviour in particular conflicts.

The data analysis also revealed that gamers have different norms for conflictual behaviours and impoliteness strategies. Some gamers tend to avoid conflict, while others are willing to escalate conflict and/or resolve it. As seen in the corpus, strategies for avoiding conflict include muting the microphone, exiting the match and/or surrendering, which in some cases allows the participants to resolve the conflict. However, in other cases, these actions can lead to conflict escalation as the other players consider muting the microphone and exiting the match to be an indication that they are being ignored intentionally, while surrendering can indicate a lack of personality and manhood. This idea aligns with the ideas of Locher and Langlotz (2008), who claim that there is a connection between emotion and cognition which should be examined from two angles: (a) the fundamental relationship between emotion and cognition in emotional processes, and (b) cognition's role in defining expectations about appropriate behaviour in a given setting. The authors also contend that there are three levels on which emotions are expressed in communication: (1) physical symptoms, such as a person's face turning pale or bright red and sweating; (2) nonverbally through gestures and facial expressions, which are not included in my data, and (3) verbally through, for example, differences in style or intonation, the use of interjections or words denoting emotions and so on, as observed in my data (Schwarz-Friesel 2015). Thus, these findings differ from first-order studies as they emphasise that the forms of emotional expression cannot be uniformly and statically associated with politeness evaluations. These associations, on the other hand, are dynamic and linked to the particular discursive contexts in which they take place. While previous research has focused on the hearer's reconstruction of the speaker's malicious intention, the hearer's perception of impoliteness, impoliteness reciprocity and emotions demonstrate that there is either a significant link between perception and intention (the more intense the intention behind the conflict, the more

negative the participant's perception). The results also validate the subjectivity of impoliteness (i.e., how gamers perceive conflict in different ways) and the relationship between conflict and emotions.

6 Conclusion

The results obtained from this study can further develop our understanding of how conflicts arise linguistically and how impoliteness is perceived, manifested and reacted to. Furthermore, the results have also identified the frequently employed impoliteness strategies that are perceived as being impolite. To begin with, this study extends the breadth of many previous studies, which have tended to concentrate on brief excerpts and examine protracted examples of interaction. In other words, this paper attempts to enrich Arabic studies as it contributes new findings on impoliteness in online contexts. Previous studies have tended to adopt classical theories - mainly Speech Act theory (1962) and Brown and Levinson's (1987) theory – while investigating single speech acts. In contrast, my study has examined Tunisian (im)politeness from multiple loci of understanding, thus contributing to a more holistic approach to the analysis of politeness. In addition, previous Arabic studies have tended to use Discourse Completion Tests, role-plays or questionnaires for data collection. In the present study, no created situations were used. Alternatively, (im)politeness has been explored in spontaneous and naturally occurring interactions in a variety of settings. Moreover, conflict in online gaming has become widespread, particularly amongst teenagers and young adults, as the majority of players fall into this age demographic. Therefore, this community is exposed to language aggression and conflict on a routine basis, which makes my study all the more important. Hence, this study has provided an in-depth understanding of the realisation of conflict between gamers, the way in which impoliteness can lead to conflict, the various aspects of impoliteness and its subjective perception, and the boundaries between genuine impoliteness and mock (im)politeness (by developing an analytical framework). However, this study has tackled a research topic that, to date, has received little attention. Furthermore, due to space and time limitations, the data used is not representative of all forms of CMC. While the data used in this study is extremely rich, the corpus does not contain female participants. Indeed, all the participants were male because online gamers, particularly those engaging in Battle Royale games, tend to be male. It is worth mentioning that the most likely reason why females do not play these types of games is the toxic nature of the genre and gaming community. The findings and limitations of this study have opened up new opportunities for the further study and examination of conflictual CMC interactions. Indeed, to develop the current study and pay further attention to the area of conflict and aggression in online settings, one could focus on duplicating the study using a wider range of data.

The first aspect that needs more attention from the field of linguistics involves the boundaries between conflict realisation, conflict management and humour, as well as the emotional reactions of participants when they take offence. Although I emphasised a number of their more significant aspects, psychological sources were used as a research reference. The lack of linguistic research on conflict realisation and emotions left me with few theoretical foundations for this paper. Thus, linguistic research should be open to, and I believe would greatly benefit from, the entire field of conflictual interaction. However, with regard to this study, specific useful developments could include qualitatively and quantitatively profiling the language of the same user across a number of so-called puppet accounts. This could be used to establish crucial linguistic markers that might indicate whether a scenario or an identity is a particular NMOB, and then build up a body of proof for that NMOB.

Appendices

Appendix A: Transcription conventions

Arabic Script	IPA symbol	Symbols used in IVAr
Consonants		
أ - الهمزة	?	2
ب	b	b
ت	t	t
ث	θ	th
ج	ʒ	j
ح	ħ	H
خ	x	x
د	d	d
ذ	ð	dh
ر	r	r
ز	z	z
س	s	s
ش	ʃ	sh
چ	tʃ	ch

ص	s ^r	S
ط	t ^r	T
ض	d ^r	D
ظ	ð ^r , z ^r	DH, Z
ع	ʕ	ʒ
خ	χ	gh
ف	f	f
ق	q	q
ك	k	k
ج	g	g
ل	l	l
م	m	m
ن	n	n
ه	h	h
و	w	w
ي	j	y
ف	v	v

Vowels		
ا	a:	a:
ي	i:, e:	i:, e:
و	u:, o:	u:, o:
فتحة	a	a
كسرة	i	i
ضمة	u	u

Appendix B: Transcription system

Symbol	Use
{ }	Curly brackets denote the Arabic transcription to English.
(.)	A full stop inside brackets denotes the English translation.
↗	Arrow in this direction shows that the pace of the speech has raised up.
↗↗	Arrows in this direction show that the pace of the speech has extremely raised up.
Word	When a word or a sentence is underlined, it denotes that it is used in the analysis.
[]	Square brackets denote the type of the voice tones used.
Word	When a word/ sentence is highlighted, it denotes that it is used in the analysis.

References

- Anderson, Craig & Brad Bushman. 2002. Human aggression. *Annual Review of Psychology* 53(1). 27–51.
- Appel, Markus, Barbara Stiglbauer, Bernad Batinic & Peter Holtz. 2014. Internet use and verbal aggression: The moderating role of parents and peers. *Computers in Human Behavior* 33. 235–241.
- Asswae, May. 2018. *Politeness in Libyan Arabic: A Third-Wave Perspective*. Doctoral Dissertation, University of Huddersfield.
- Baker, Frank. 2001. *The Basics of Item Response Theory*. Available at: <http://ericae.net/irt/baker> (accessed 1 March 2023).
- Bargiela-Chiappini, Francesca & Dániel Kádár (eds.). 2010. *Politeness Across Cultures*. London: Palgrave Macmillan.
- BBC. 2010. Facebook murderer to serve at least 35 years. *BBC News* 8 March. Available at: <http://news.bbc.co.uk/hi/england/weir/8555221.stm> (accessed 1 March 2023).
- Blitvich, Pilar Garcés-Conejos. 2015. Setting the linguistics research agenda for the e-service encounters genre: Natively digital versus digitized perspectives. In María de la O Hernández-López & Lucía Fernández Amaya (eds.), *A Multidisciplinary Approach to Service Encounters*, 13–36. Leiden: Brill.
- Bousfield, Derek & Miriam Locher (eds.). 2008. *Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice* (Vol. 21). Berlin and New York: Walter de Gruyter.
- Brown, Penelope & Stephen Levinson. 1978. Universals in language usage: Politeness phenomena. In Esther Goody (ed.), *Questions and Politeness: Strategies in Social Interaction*, 56–311. Cambridge and New York: Cambridge University Press.
- Brown, Penelope & Stephen Levinson. 1987. *Politeness: Some Universals in Language Usage* (Vol. 4). Cambridge and New York: Cambridge University Press.
- Camber, Rebecca & Simon Neville. 2011. Sick Internet ‘troll’ who posted vile messages and videos taunting the death of teenagers is jailed for 18 WEEKS. *Mail Online*, 14 September 2011. Available at: <https://www.dailymail.co.uk/news/article-2036935/Natasha-MacBryde-death-Facebook-internet-troll-Sean-Duffy-jailed.html> (accessed 1 September 2023).
- Carnagey, Nicholas, Craig Anderson & Brad Bushman. 2007. The effect of video game violence on physiological desensitization to real-life violence. *Journal of Experimental Social Psychology* 43(3). 489–496.
- Cho, Hyun, Sang-Kyu Lee, Jung-Seok Choi, Sam-Wook Choi & Dai-Jin Kim. 2017. An exploratory study on association between Internet game contents and aggression in Korean adolescents. *Computers in Human Behavior* 73. 257–262.
- Crawford, Garry. 2011. *Video Gamers*. Oxford: Routledge.
- Culpeper, Jonathan. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics* 25(3). 349–367.
- Culpeper, Jonathan. 2011. *Impoliteness: Using Language to Cause Offence*. Cambridge: Cambridge University Press.
- Culpeper, Jonathan & Claire Hardaker. 2017. Impoliteness. In Jonathan Culpeper, Michael Haugh & Dániel Kádár (eds.), *The Palgrave Handbook of Linguistic (Im)politeness*, 199–225. London: Palgrave Macmillan.

- Culpeper, Jonathan & Vittorio Tantucci. 2021. The principle of (im)politeness reciprocity. *Journal of Pragmatics* 175. 146–164.
- Culpeper, Jonathan, Michael Haugh & Dániel Kádár (eds.). 2017. *The Palgrave Handbook of Linguistic (Im)politeness*. London: Palgrave Macmillan.
- Diaz, Ruth, Ulric Wong, David Hodges, Carina Chiu & Vina Goghari. 2016. Violent video game players and non-players differ on facial emotion recognition. *Aggressive Behavior* 42(1). 16–28.
- Dickmeis, Anne & Keith Roe. 2019. Genres matter: Video games as predictors of physical aggression among adolescents. *Communications* 44(1). 105–129.
- Eastin, Martin, Bradley Greenberg & Linda Hofschire. 2006. Parenting the Internet. *Journal of Communication* 56(3). 486–504.
- Gentile, Douglas, Hyekyung Choo, Albert Liau, Timothy Sim, Dongdong Li, Daniel Fung & Angeline Khoo. 2011. Pathological video game use among youths: A two-year longitudinal study. *Pediatrics* 127(2). e319–e329.
- Graham, Sage. 2007. Disagreeing to agree: Conflict, (im)politeness and identity in a computer-mediated community. *Journal of Pragmatics* 39(4). 742–759.
- Graham, Sage & Claire Hardaker. 2017. (Im)politeness in digital communication. In Jonathan Culpeper, Michael Haugh & Dániel Kádár (eds.), *The Palgrave Handbook of Linguistic (Im)politeness*, 785–814. London: Palgrave Macmillan.
- Greitemeyer, Tobias & Christina Sagioglou. 2017. The longitudinal relationship between everyday sadism and the amount of violent video game play. *Personality and Individual Differences* 104. 238–242.
- Hardaker, Claire. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research* 6(2). 215–242.
- Hardaker, Claire & Mark McGlashan. 2016. “Real men don’t hate women”: Twitter rape threats and group identity. *Journal of Pragmatics* 91. 80–93.
- Hardaker, Glenn. 2012. An insight into Islamic pedagogy at the University of al-Qarawiyyin. *Multicultural Education & Technology Journal* 6(2). 106–110.
- Hatipoğlu, Çiler. 2007. (Im)politeness, national and professional identities and context: Some evidence from e-mailed ‘Call for Papers’. *Journal of Pragmatics* 39(4). 760–773.
- Haugh, Michael. 2010. Jocular mockery, (dis)affiliation, and face. *Journal of Pragmatics* 42(8). 2106–2119.
- Herring, Susan & Carole Nix. 1997. Is “serious chat” an oxymoron? Pedagogical vs. social uses of Internet Relay Chat. Paper presented at the American Association of Applied Linguistics Annual Conference, Orlando, Florida.
- Home Office. 2009. *Protecting the Public in a Changing Communications Environment* (Vol. 7586). The Stationery Office.
- Ivory, James & Sriram Kalyanaraman. 2007. The effects of technological advancement and violent content in video games on players’ feelings of presence, involvement, physiological arousal, and aggression. *Journal of Communication* 57(3). 532–555.
- Johansen, Robert, Jacques Vallee & Kathleen Spangler. 1979. *Electronic Meetings: Technical Alternatives*. Boston: Addison-Wesley Longman Publishing Co.
- Kádár, Dániel & Michael Haugh. 2013. *Understanding Politeness*. Cambridge: Cambridge University Press.
- Kádár, Dániel, Vahid Parvaresh & Puyu Ning. 2019. Morality, moral order, and language conflict and aggression: A position paper. *Journal of Language Aggression and Conflict* 7(1). 6–31.

- Kaye, Linda & Jo Bryce. 2012. Putting the fun factor into gaming: The influence of social contexts on the experiences of playing videogames. *International Journal of Internet Science* 7(1). 24–38.
- Kim, Eun Joo, Kee Namkoong, Taeyun Ku & Se Joo Kim. 2008. The relationship between online game addiction and aggression, self-control and narcissistic personality traits. *European Psychiatry* 23(3). 212–218.
- Kim, Heejung, David Sherman & Shelley Taylor. 2008. Culture and social support. *American Psychologist* 63(6). 518–526.
- Lakoff, Robin. 1989. The way we were; or; the real actual truth about generative semantics: A memoir. *Journal of Pragmatics* 13(6). 939–988.
- Lam, Lawrence, Zaohua Cheng & Xinmin Liu. 2013. Violent online games exposure and cyberbullying/victimization among adolescents. *Cyberpsychology, Behavior, and Social Networking* 16(3). 159–165.
- Lansford, Jennifer, Michael Criss, Gregory Pettit, Kenneth Dodge & John Bates. 2003. Friendship quality, peer group affiliation, and peer antisocial behavior as moderators of the link between negative parenting and adolescent externalizing behavior. *Journal of Research on Adolescence* 13(2). 161–184.
- Leech, Geoffrey. 1983. *Principles of Pragmatics*. London: Longman.
- Linder, Jennifer & Douglas Gentile. 2009. Is the television rating system valid? Indirect, verbal, and physical aggression in programs viewed by fifth grade girls and associations with behavior. *Journal of Applied Developmental Psychology* 30(3). 286–297.
- Locher, Miriam. 2006. Polite behavior within relational work: The discursive approach to politeness. *Multilingua* 25(3). 249–267.
- Locher, Miriam & Andreas Langlotz. 2008. Relational work: At the intersection of cognition, interaction and emotion. *Bulletin VALS-ASLA* 88. 165–191.
- Lorenzo-Dus, Nuria, Pilar Garcés-Conejos Blitvich & Patricia Bou-Franch. 2011. On-line polylogues and impoliteness: The case of postings sent in response to the Obama Reggaeton YouTube video. *Journal of Pragmatics* 43(10). 2578–2593.
- Lull, Robert & Brad Bushman. 2016. Immersed in violence: Presence mediates the effect of 3D violent video gameplay on angry feelings. *Psychology of Popular Media Culture* 5(2). 133–144.
- McDonnell, Andrew. 2011. *Managing Aggressive Behaviour in Care Settings: Understanding and Applying Low Arousal Approaches*. New Jersey: John Wiley & Sons.
- Merchant, Guy. 2001. Teenagers in cyberspace: An investigation of language use and language change in internet chatrooms. *Journal of Research in Reading* 24(3). 293–306.
- Möller, Ingrid & Barbara Krahé. 2009. Exposure to violent video games and aggression in German adolescents: A longitudinal analysis. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 35(1). 75–89.
- Nair, Nisha. 2008. Towards understanding the role of emotions in conflict: A review and future directions. *International Journal of Conflict Management* 19(4). 359–381.
- Ohno, Shiroh. 2022. The link between battle royale games and aggressive feelings, addiction, and sense of underachievement: Exploring esports-related genres. *International Journal of Mental Health and Addiction* 20(3). 1873–1881.
- Padilla-Walker, Laura, Larry Nelson, Jason Carroll & Alexander Jensen. 2010. More than a just a game: Video game and internet use during emerging adulthood. *Journal of Youth and Adolescence* 39(2). 103–113.

- Parvaresh, Vahid. 2019. Moral impoliteness. *Journal of Language Aggression and Conflict* 7(1). 79–104.
- Parvaresh, Vahid & Tahmineh Tayebi. 2018. Impoliteness, aggression and the moral order. *Journal of Pragmatics* 132. 91–107.
- Schwarz-Friesel, Monika. 2015. Language and emotion. The cognitive linguistic perspective. In Ulrike Lüdke (ed.), *Emotion in Language*, 157–173. Amsterdam: John Benjamins Publishing.
- Sferra, Michala Saint. 2016. *Game type as a moderator of the relationship between pathological video game use, impulsivity, aggression, and general psychopathology* (Doctoral dissertation).
- Shachaf, Pnina & Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36(3). 357–370.
- Taylor, Stephanie. 2001. Locating and conducting discourse analytic research. In Margaret Wetherell, Stephanie Taylor & Simeon Yates (eds.), *Discourse as data: A guide for analysis*, 5–48. London: Sage Publications Ltd.
- Teng, Scott, Gabriel Chong, Amy Cheng & Marko Skoric. 2011. Grand Theft Auto IV comes to Singapore: Effects of repeated exposure to violent video games on aggression. *Cyberpsychology, Behavior, and Social Networking* 14(10). 597–602.
- Terkourafi, Marina & Dániel Kádár. 2017. Convention and ritual (im) politeness. In Jonathan Culpeper, Michael Haugh & Dániel Kádár (eds.), *The Palgrave Handbook of Linguistic (Im)politeness*, 171–195. London: Palgrave Macmillan.
- Veiga Simão, Ana, Paula Ferreira, Sofia Francisco, Paula Paulino & Sidclay de Souza. 2018. Cyberbullying: Shaping the use of verbal aggression through normative moral beliefs and self-efficacy. *New Media & Society* 20(12). 4787–4806.
- Vladimirou, Dimitra, Juliane House & Dániel Kádár. 2021. Aggressive complaining on social media: The case of #MuckyMerton. *Journal of Pragmatics* 177. 51–64.
- Wikipedia. 2007. Anti-Social Behaviour Orders. Available at: http://en.wikipedia.org/wiki/Anti-socialBehaviour_order (accessed 12 April 2021).
- Willoughby, Teena, Paul Adachi & Marie Good. 2012. A longitudinal study of the association between violent video game play and aggression among adolescents. *Developmental Psychology* 48(4). 1044–1057.
- Wright, Michelle. 2019. Friends and console-gaming aggression: The role of friendship quality, anger, and revenge planning. *Games and Culture* 14(6). 604–621.

About the Author

Kholoud Boukhris holds PhD in Linguistics and MA in Applied Linguistics and TESOL. She is a Lecturer at Brunel University London. She is also actively cooperating with Anglia Ruskin University in Cambridge. Her research interests include im/politeness-related research, particularly in investigating issues such as offence and language aggression and the use of pragmatic particles as well as aggressive expressions in how people use these expressions to get their messages across.

Address

Department of Humanities and Social Sciences, Anglia Ruskin University
Cambridge East Road
CB1 1PT, United Kingdom

e-mail: kb1070@pgr.aru.ac.uk
ORCID iD: 0009-0001-7566-2004

Mariya Chankova
South-West University N. Rilski – Blagoevgrad

**EVALUATIONS OF APPROPRIATENESS
THROUGH IMPOLITENESS IN POLITICAL DISCOURSE
REFRAMED FOR ENTERTAINMENT PURPOSES**

Abstract

This contribution takes a look at video-sharing platforms to highlight a popular entertainment format which consists in re-framing political discourse for the purposes of entertaining the audience and, at the same time, providing an evaluation of that discourse. Evaluations of political discourse uncover the role and importance imputed to it by those who are outside of the political system, but who are directly impacted by it, that is, the people. A sample of French-language data, collected from YouTube, is examined for the categories of evaluation which are used by the authors, which are conjectured to represent the ideas ordinary citizens have about political discourse. Reframing political discourse carries these evaluations through offensive language and humor. Teasing out the relevant parameters of evaluation can provide a basis for understanding how regular citizens gauge political discourse. Appropriateness appears to be an important evaluation criterion pertaining to the characteristics of political discourse, occurring in a particular context.

Keywords

evaluative language, offensive language, reframing, political discourse, appropriateness

1 Introduction

Digital media provide a venue for people across the globe to participate in large-scale conversations, especially through content-sharing platforms. This liberalization of communication and information conduits ought to facilitate the interpenetration of different audiences, and their heightened visibility. It has equally become easier to use these platforms to share evaluations of different kinds of discourses, especially on the political and cultural arenas. Access to information on the one hand and the affordances of different digital platforms for sharing and creating content on the other have shaped up a qualitatively different way of engaging with discourses in the digital environment. To complicate matters, the digital environment accommodates the layering of multimodal elements which give rise to multiple interpretative paths and contextualization possibilities (see for example Yus 2011 for a focus on discourse; Kress and Van Leeuwen 2001 for perspectives on multimodality).

Content-sharing platforms provide the opportunity to share vast quantities of data; users upload more than 4 billion hours of videos daily (How Much Data is Created on the Internet Each Day? | Micro Focus Blog). This is largely on the increase, just as daily consumption of digital video has been growing, especially due to hand-held smart devices (statista.com).

Digital platforms were expected to facilitate the participation of the masses in the political communication and debate, and many commentators saw in the digitally-based tools a democracy-enhancer: at the start of the XXI century, digital platforms were described as the carriers of democratic change for their perceived role in social uprisings around the globe (for example, Chebib and Sohail 2011). However, critics such as Morozov (2011) warned that it is not the tools themselves that have an inherent democratizing effect, but it is the intentions of the users that shape up the outcomes. The digital tools themselves are as vulnerable to manipulation as other media tools before them (see, for example, Herman and Chomsky 1988 on news media, Colon 2021 on individually-tailored mass-distributed propaganda), not to mention fragmentation and political polarization (Kubin and von Sikorski 2021). The initial fascination with digital tools appears to have ceded the place to other perspectives, as noted by Shullenberger:

Previously, much of the Western commentariat believed that democratized information would enable people to counter top-down misinformation. Now, it denounces the information free-for-all for unleashing a flood of “fake news” and conspiracy theories. (Shullenberger 2020: 48)

Users and commentators alike were coming to recognize that the media change did not bring about a fundamental shift in the quality of the communication as it ushered in a polycentric world of communication which differs greatly as a communication model by its underlying processes and actors (Blommaert 2020); while it did bring the promise of opportunity for all, the promise of participation, the question whether it has delivered on this promise remains open. Top-down information broadcasting has undergone a serious transformation in the new media context (especially with the Participative Web 2.0), which has affected its financial structures and access architectures. Bottom-up information conduits, which will be referred to as user-generated content throughout this piece, struggle to become a valid alternative, partly due to competition with other information outlets (e.g. Allan 2009; Wall 2018). The mainstream vs alternative information conduits have gradually entangled themselves in a codependent relationship as mainstream media has structures, manpower and financial capabilities which far outstrip those of the regular content-producer. Thus, much of the user-generated content has to fall back on mainstream media for grounding and context; at the same time, mainstream media reluctantly bring legitimacy to alternative media outlets by featuring them in order to stay relevant. This shift also

brings a different standard of evaluating political coverage (Jandura and Friedrich 2014).

As for political discourses, this entanglement may be interesting from the point of view of citizen participation, which is the angle adopted in this contribution. The content-sharing and social networking platforms are charitably seen as a democratizing tool for furthering debate, a platform for regular citizens to engage with politicians without mediators and participate truly in their governance. However, this may well turn out to be no more than a figure of speech from the point of view of politicians: studies into the exploitation of those platforms for political communication have suggested that the dynamics of political communication on alternative platforms appear to provide yet another way to condition the reception of political discourse (Colon 2021) by filling in a self-created informational void (Chankova in print-a) or by different legitimization techniques (Chankova in print-b).

Nonetheless, such platforms have greatly facilitated the sharing of different kinds of discourses and their evaluations, especially on the political and cultural arenas. On the backdrop of studies showing a significant drop in the demand for political news in a high-choice media environment (for an overview, van Aelst et al. 2017), the emergence of hybrid formats which feature political content in any way is worth our interest: the more informed the citizens are, the fuller their participation in the political process can be. Even though these formats are not necessarily aimed at the politicians themselves, they are consumed by different media actors and by a vast citizen public, so that they come to be integrated in the larger debate, even if their impact on policy is difficult to ascertain.¹ It is still instructive to see how exactly content-producers position themselves with regard to political dialogue. In this contribution, I look into a popular entertainment format of user-generated content shared on YouTube that features political discourse reframed for entertainment purposes. Not merely a parody or satire (although there are links with these genres that inform an interesting comparison), reframing itself is based on what essentially is an evaluation of political discourse propagated usually top-down on mainstream or traditional media. This is an alternative way to participate in political debates from which the ordinary citizen is usually excluded. Such evaluative discourses allow content-creators to propagate political views, criticism of policy or political public events, but also it allows content-creators to rally like-minded people or conduct debates with political adversaries using the affordances of digital media.

The question explored here is what criteria of evaluation appear to have been applied by the content-producers in order to choose the samples of political discourse that are then reframed to become the object of satire, derision and/or entertainment?

¹ If one construes this kind of content as an expression of a political stance, it is comparable to citizen protests: the protest itself is not enough to bring about a change in the politician's act or policy; the relationship between these ways of expressing public opinions and policy change is very complex and not at all straightforward.

Evaluations of political discourse uncover the role and importance imputed to it by those who are outside of the political system, but who are directly impacted by it, that is, the people. A sample of French-language data, collected from YouTube and transcribed, is examined for the categories of evaluation which are assumed by the content-creators, and how these categories are conveyed through reframed discourse. I will start by describing the data used in the study and the analytical approach (Section 2). I will then proceed to a thematic exposition of the relevant categories by laying the groundwork to frames and reframing (Section 3.1.), reframing as a conduit for evaluations of political discourse (Section 3.2.), followed by discursive strategies to convey those evaluations - offensive language and humor (Section 4) and a brief Conclusions section.

2 Data and method

The data was collected from two French-language channels, *J'suis pas content TV* (I'm not happy TV) and *Juste milieu* (Middle ground) propagated on the video-sharing platform YouTube. The channels work largely as open broadcasting stations, the former self-describes as "une chaîne humoristique, satirique et informative traitant essentiellement de sujets liés à l'actualité politique" (a humoristic, satiric and informative channel on current political topics²), the latter features the formula "l'actualité + la rigolade - le mensonge = Juste Milieu" (news + fun - lies = Middle Ground) in a dedicated series called *La Farce Tranquille* (Tranquil jokes). After watching a considerable number of videos in order to get acquainted with the format – pre-scripted videoclips of a standard length, which have a stable structure, opening and closing segments, opening and closing credits, an organizing satiric thread running through the episode – I selected randomly four episodes, two from each of the channels, with a total running length of approx. 70 minutes. Those episodes were transcribed and then analyzed qualitatively using the pragmatic approach of discourse analysis. In addition to discourse analysis, I turned to Goffman's frames analysis, refined and developed by Tannen, a decision dictated by the peculiarities of the format, which is strongly inspired by satire and stand-up comedy. But instead of looking into the sources of comedy (which might be an interesting object of study itself), I focus on the serious (or the informative) side of reframing and look into the kind of criteria which underlie the evaluations of political discourse.

² This and further translations from French mine.

3.1 Frames and reframing: revisiting Goffman

Frames, frameworks or schemata of interpretation in Goffman's understanding pertain to ways people organize their experiences of the world (1974). These experiences provide the basis for complex structures of background knowledge, which will participate in the interpretation of discourse, shaping or structuring this interpretation. The participants in the discourse are actors who endorse particular roles in the frame which can be adjusted to various keyings in the actual realization of the frame in discourse. They are thus agents who have intentions, objectives, and a certain sway over the frame, but they are also subject to “standards” of social appraisal (...) based on honesty, efficiency, economy, safety, elegance, tactfulness, good taste, and so forth” (Goffman 1974: 22). Social frames are based on rules that are constitutive or regulative³ for the frame: the social frame for political discourse involves both constitutive and regulative rules, as the activity of engaging in political discourse for a politician involves being able to communicate decisions on policy, implement those decisions, act in the interest of those who elected her, be held accountable for decisions taken, positions declared and policy, and generally work in the public interest. There are expectations that are connected to the politician frame; indeed, Tannen (1993) emphasized that frames mainly bear structures of expectations. These expectations may vary according to the individual interpreter of the frame, but are more often than not culturally or socially conditioned. They are also quite dynamic, as they are constantly “checked against experience and revised” (Tannen and Wallat 1987: 207). Frames also bind the discourse conceptually together: Fillmore (2006: 373) talks about a “system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available.” This interconnected system of semantic concepts creates an interpretative grid readily available for the person familiar with the system.

Speakers may use several different frames as they speak; very often, the predominant frame will leave traces in the discourse which will orient the audience as to how the speaker positions herself. The frame may also shift (for a discussion on strategic uses of frame shifting, see Chankova in print-b), which will carry a different set of expectations about the actors and will in turn influence the interpretation of the discourse stretch.

Reframing consists in transplanting discourse from another frame and setting up a different frame which will constitute the interpretative grid. The process that is observed in the data set involves using the speech events that occur in the political discourse frame and putting them into the entertainment frame. The change affects the

³ The two types of rules are well described by Searle (1969): constitutive rules institute the frame (for example, a game of chess is defined by the rules of the game), whereas regulative rules introduce a way of regulating a pre-existing type of activity (as table rules merely regulate eating).

interpretative grid which is used to interpret the speech event, and it brings about a new set of expectations, restrictions and norms against which the speech event is projected. Changing the interpretative grid may be used for several different strategic purposes: for persuasion or manipulation of the audience, for analytical purposes (when there is need to reflect upon a particular matter), for entertainment, etc. To illustrate, the same event will acquire a different reading according to the frame in which it is put: a bird pecking at eggs in a nest may be a predator ravaging the nest or a caring mother helping hatching chicks get out of the shells. The interpretative grid or frame has far-reaching consequences on how the event is constructed and how participants and observers regard it. Reframing allows one to expose what are considered to be deficiencies of the original discourse. Consider the following:

- (1) je veux dire il est président épidémiologiste ah oui c'est son entourage qui le dit il a «
acquis une maîtrise des travaux épidémiologiques sans précédent » pour un président
je pense ils n'ont pas précisé si ça se trouve dans leur tête c'est pour un être humain
ça y est c'est officiel le mec y croit non non non le mec y croit qu'il est meilleur que
les médecins ce qui explique probablement d'ailleurs pourquoi il a envoyé chier le
conseil scientifique (S07E26, Apr. 9, 2021)

[‘I mean he is the president epidemiologist oh yes it’s his close collaborators who say
it he has “acquired unprecedented expertise in epidemiology” I think they mean for a
president they haven’t clarified that but I’m not ruling out that they believe it’s
unprecedented for a human being so it’s official the guy thinks he is better than the
doctors which is probably why he kicked out the scientific council’]

(1) illustrates the mechanism of reframing: there is a particular speech event which is explicitly identified in the material: it is explicitly identifiable by the cited words of the source material (enclosed in the transcript in inverted commas). Visually, the sourcing is accomplished in different ways: by inserting a sound byte, a news article clipping or a screenshot of the webpage, a video clip or any other identifying means of referencing the media propagating the discourse to be reframed. Both channels consistently use one of these techniques to identify the source they are basing their comments on. This allows them to ground their comments in a publicly available information pool, which in turn will provide the necessary context for their audience to facilitate comprehension. This is done in order to avoid several potential pitfalls of reframing: the loss of context for the primary discourse stretch, which could result in misunderstanding or misinterpretation on the part of the audience; the loss of credibility for the speaker as this is what brings the audience to that content; the accusation of manipulation, which, given the affordances of the entertainment frame, can stem from the unrestricted use of offensive language and exaggerations.

The context: the object of reframing here is the news article published in the mainstream media, reporting the words of an unnamed close collaborator to the French president Emmanuel Macron in March of 2021, amidst the COVID-19 sanitary crisis and debates about lockdowns. It may seem on the surface that this piece of news is the

very object of ridicule - an anonymous source providing unbridled flattery to the French president which is reported in mainstream media who also point out the utter irrelevance of the communication (the communication was taken up by *Le Monde*, March 30, 2021, and *Marianne*, March 31, 2021). The actual object of the speaker's criticism is what he perceives to be random, baseless decisions of policy, ignoring organs of expert advice, and/or going against public interest.

The actor in the new entertainment frame: reframing allows the speaker to assume the special position of jester (or joker), whose very role is to ridicule everyone and everything; the frame carries the expectation of non-serious discourse, which provides the speaker with more leeway to be less careful about the expressive means he uses in his comments. The jester can exaggerate, minimize, opt for a literal interpretation, make wild or unsupported by logic connections, be overdramatic, use vulgar or insulting words, etc. In (1), the actor makes an exaggerated suggestion based on the form of the utterance reported ('they may think Macron has acquired epidemiological expertise unprecedented for a human being'), then jumps to the unsupported by the information conclusion that it is Macron who actually believes this of himself, and further suggests that this belief could have been the reason for a specific policy decision (to get rid of the scientific council). The tradition of political satire in France allows the author to use these techniques to convey serious criticism⁴ or at least raise the question about the decisions taken, and/or have the audience consider that possibility. The subsequent segment confirms this; in it the speaker reminds the audience of previous problematic communications by Macron, ending with a criticism of basing policy decisions on what he describes as megalomaniac beliefs:

- (2) franchement c'est plus drôle du tout là pour Jupiter on n'a rien dit les délires mégalomanes on n'a rien dit les délires sur les Romains et les Gaulois on n'a rien dit mais là il y a la vie des gens qui est en jeu mais je veux dire on sait que l'autre il est bloqué en enfance, mais offrez-lui un fucking kit du petit chimiste et arrêtez de mettre la vie des gens en danger
['frankly that's not funny we didn't say anything about Jupiter we didn't say anything about the megalomaniac lunacies we didn't say anything about the lunacies on the Romans and the Gauls but this is about people's lives I mean we know that guy is stuck in childhood so just give him a fucking chemistry playset and stop putting people's lives in danger']

This type of criticism is carried by the exaggerated language ('lunacies', *that guy* referring to Macron, 'stuck in childhood', 'give him a chemistry playset') and rests on the assumption that it is a shared position: the speaker switches between *I* and *we* to mark the difference between the individual speaker/actor and a sort of synthesized common judgment about current affairs. In staging a direct address to the targeted

⁴ In the French tradition, the pamphlet is the iconic genre to accomplish this, in which polemic and aggressive language is used to attack an idea or an authority figure.

politicians the speaker is able to use crude language for polemic attacks on contentious points. This allows him to use direct, unmitigated means of conveying his evaluations and positions. Reframing into entertainment adds the affordances of the novel frame, which effectively carries the evaluations across.

3.2 Reframing as a conduit for an evaluation of political discourse

Evaluation may be approached from different angles: first, in the narrow sense, it is a very important speech act. Evaluation is found in Austin's verdictives (the class of speech acts that include assessments, appraisals, estimates, in which an evaluation of a state of affairs leads to laying down a verdict establishing a rendered judgment) and in behabitives (which include acts such as criticizing, commanding, deprecating), both of which involve the process of evaluation or judgment (Austin 1962). Second, in the larger sense, it can be viewed as a discursive technique in presenting events which incorporate the speaker's point of view. In the terms of Thompson and Hunston:

the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions [statements] that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values. (Thompson and Hunston 2000: 5)

In order to identify and analyze evaluations, I apply the methodology outlined by Thompson and Hunston (2000), who suggest that the evaluation is contained in "comparison, subjectivity, and social value" (p. 13). Evaluation also, importantly, involves "indication of the existence of goals". The implication is that the speaker conveys his own judgment or echoes the assessment by a source which is arrived at by gauging the state of affairs against a set of values, subjective (and/or collectively shared). While usually evaluative language involves the use of lexical means of conveying an evaluation or an attitude (see, for example, Bednarek 2010 for an analysis of evaluative language in journalism), in reframed discourse much of that evaluation is implicit, or conveyed by other means. Consider:

- (3) Jean-Michel Blanquer il a des problèmes de flagornerie oui de flagornerie c'est comme ça des gens ils ont des problèmes de RSA des problèmes de CAF des problèmes d'insécurité ces gens-là ils ont des problèmes de flagornerie ils sont comme ca sur l'autre monde
 ['Jean-Michel Blanquer has flattery problems yes flattery problems it's like that some people have RSA problems or CAF problems⁵ or insecurity problems but those people they have flattery problems it's like that in the other world']

⁵ RSA (Revenu de solidarité active) and CAF (Caisse d'allocations familiales) are social security aids allocated to help with reintegration of the workplace or for family matters.

- (4) C'est nul c'est raté Olivier tu emploies des mots les gens ne savent pas ce que ça veut dire vu qu'en plus ils ne savent pas qui t'es le combo devient ingérable
[‘It sucks it doesn't work Olivier you are using words that nobody knows what they mean on top of that people don't know who you are so the combination is completely moot’]
- (5) ah ben voilà quelque chose de productif voilà on règle le problème en supprimant le mot comme François Hollande il avait supprimé le racisme en enlevant le mot race de la constitution alors là on ne dit plus interdit non non c'est plus ça c'est réservé à c'est dingue cette manie des socialistes de vouloir tout régler avec un changement des titres
[‘Ah there finally something productive we fix the problem by erasing the word just like Francois Hollande eliminated racism by erasing the word race from the constitution so we don't say forbidden no now we say reserved for it's crazy this mania the socialists have to want to fix everything by a word swap’]

In (3) the speaker jokingly contrasts what he calls real people's problems (the need for social security allocations to cope with difficulties, or insecurity) with the problems of “those people”, referring to the politicians. The politicians are from a different world, they do not share in the difficulties of ordinary people. This contrast is used to set up the speaker's attitude towards the political communication (which is sourced to a commentary piece covering a published interview of J-M Blanquer) as being out of tune with the preoccupations of ordinary citizens. It evokes Blanquer being ridiculed by other officials for being too flattering and sending the wrong picture of the entire government. In the interview, Blanquer aimed to provide a positive assessment of Macron's management of the sanitary crisis. The implicit evaluation is that government officials are more preoccupied with their image in publications and public communications than with the appropriateness and efficacy of their policies. Their image is more important than their actions, and certainly more important than the citizens' problems the speaker mentions.

In (4) there is an evaluative phrase which explicitly rates the political communication reframed here: a declaration by Olivier Faure in which he criticized Macron's sanitary policy by accusing him of assuming the posture of a “thaumaturge king” (France 24, *Covid: Faure (PS) critique Macron « le roi thaumaturge »*). The speaker addresses his words directly to Faure, criticizing his choice of words by the phrases “it sucks” and “it doesn't work”. The criticism appears to target the declaration itself, but it is Olivier Faure personally who is affected more by the criticism, the gist of which is that a traditionally important political party – the socialist party (the French left) – has all but lost its former importance at the political scene, largely due to the party's mistakes and inadequacies. Thus “people do not know who you are”.

This latter idea is at the basis of the evaluation in (5) as well, where evaluation is conveyed by irony, consisting of literally saying the opposite of what one means. In this case, the speaker literally commends a politician, where actually he means to criticize

what he describes to be habitual behavior on the part of people affiliated with the socialist party, with former president Hollande being a prime example. The target of the criticism is the approach to problems by members of the left: the yardstick against which this evaluation is made consists in being consistent in one's values and positions, identifying the core of the problem and proposing a solution consistent with those values. The idea that politicians should be consistent in the display of their political and moral stances is explicitly expressed in a subsequent segment:

- (6) *qu'est-ce qui s'est passé Audrey je dis ça parce que je veux dire tu nous tenais des propos il y a pas longtemps que ça qui n'allait pas dans ce sens-là (...) ce n'est pas la même philosophie qui est derrière il me semble*
 ['What happened Audrey I am saying this because not that long ago you used to say things which did not go in the same direction (...) it's not the same philosophy which is behind all this I think']

Humor or entertainment provides the speaker with a way to accomplish serious work: laughing at something, ridiculing, satirizing allows the audience to free themselves from any constraint that political discourse usually lays upon them. Political discourse can be intimidating, incomprehensible, full of technical terms or treating subject matters in a complex manner. Laughter demystifies it, allowing the audience to distance themselves from those constraints and effects. With distance comes understanding: emotional distance allows them to decode or interpret the implications and the objectives of political discourse, and to be able to discern problems. These problems, or deficiencies of political discourse are the object of the subjective evaluation of the speaker. The work that he accomplishes is to offer his evaluation, to propose an interpretation of the speech event.

On the basis of these illustrations, it could be suggested that each assessment or evaluation is based upon values which the speaker projects to his audience, and proposes them for consideration. For (3), (4) and (5), the values that seem to underscore the evaluation are the following: politicians are supposed to work in the interest of the people, be held accountable for policy or declarations of stances they make. Reframing consists in a very intricate enterprise, in which many layers of communication overlap: there is an original speaker (a political figure) who makes a public statement, or gives a press conference, or is interviewed, or makes a television appearance, in other words, a primary political communication which is directed at the public. Its purpose may be to inform (announce, clarify, justify, explain a policy or a decision), it may be political (it seeks legitimisation, it is aimed at an adversary to undermine them, it is public exposure in order to score political points, etc.), it may be purely communicative (a PR campaign, in order to appear more appealing, to win over people's sympathies), and so on. There might be (and usually, there is) a media coverage of the original communication: a news piece reporting on it, sometimes already providing a particular kind of evaluation to it, echoing a median type of attitude towards this communication.

And on top, there is reframing itself, which takes up on these previous communications and recycles (evokes) them in an entertaining way. In (1) the speaker sourced the material to a media piece which already negatively evaluated the description of Macron as “a president-epidemiologist”. But in (4) the media piece which is referred to does not negatively evaluate Faure’s declaration. Ultimately, the speaker’s evaluation is not based on the media’s assessment of these speech events, but often incorporates the implicit criticism of the media coverage itself. Criticism essentializes negative evaluations which are based on comparisons with subjective standards of value. By looking into the kind of speech events that are selected for reframing, these standards can be identified. The position of the jester allows the speakers to expose deficiencies in the entire communication chain involved in propagating political discourse. The speakers choose a particular speech event precisely because they find it deficient in some way. This kind of relationship with traditional media sources alludes to a special role those content-creators wish to adopt, that of a corrective to traditional media sources. (I will return to this point in the final section)

Consider another selection of examples, harvested from an episode of *La Farce Tranquille* (JM-FT, N97):

- (7) La page de la réforme des retraites n'a pas encore été tournée même si les membres du gouvernement se disent bon allez hein on a mis un dernier coup de genou à la démocratie maintenant c'est roue libre mais bon que ça gronde encore un petit peu et d'autant plus le grand sultan Emmanuel Macron fait des siennes
[‘The pensions page wasn’t turned yet even if the government members tell themselves ok we’ve administered one final kick at democracy so now we’re free-wheeling but ok let it rumble for a while longer especially now that the great sultan Emmanuel Macron is up to his usual tricks’]
- (8) On a des suites de l'affaire Marlène Schiappa alors c'est vrai que bon cette histoire de une sur Playboy (...) finalement c'est pas le problème qu'elle fasse la une de Playboy si c'est la ligne édito maintenant de ce magazine jadis dédié aux bimbos qui désormais se recycle dans les abat-jours c'est pas mon problème je ne porte pas de jugement là-dessus mais ce qui est intéressant c'est que ça a complètement caché une affaire scandaleuse
[‘There’s the followup of the case of Marlene Schiappa so it’s true that this thing with the photo-op for Playboy’s cover (...) it’s not the problem that she made the cover of Playboy if this is the magazine’s policy now a magazine that used to be dedicated to bimbos which now apparently repurposed itself in covering lampshades it’s not my problem I am not judging this but what’s interesting is that this scoop has completely concealed a scandal’]
- (9) Le journal Marianne a révélé la semaine dernière que à la suite de l'assassinat de Samuel Paty un fonds a été créé le fonds Marianne qui n'a rien à voir avec le journal un fonds qui avait récolté de l'argent (...) et bon on a appris l'argent on ne sait pas trop où il est allé si ce n'est dans des associations chelou

[‘The newspaper Marianne has revealed last week that after the assassination of Samuel Paty a fund had been created the Marianne fund which has nothing to do with the newspaper a fund which had collected money (...) and well we found out that it was unclear where the money had gone but to shady associations’]

The evaluative discourse is constructed in a more explicit way, through the use of evaluative language. The speaker bases his comments on a publication or a previous speech event for context. In (7), the joke involving the imagined position of the government is the vehicle for the speaker’s evaluation. The evaluation concerns the government policy and public discussion on the pensions reform which had sparked a strong protest movement in France. For context, not only was the reform itself opposed as it raised the retirement age by two years, but also the mechanism which was used by the government to pass the reform was widely criticized for being against democratic principles in spirit as it relied on the invocation of a special constitutional provision - article 49-3 which allows the government to engage its responsibility as a vouch in order to pass a bill into a law circumventing the Assembly vote.⁶ The speaker’s evaluation of these actions by the government is clearly negative: he believes them to be against the democratic principles which are supposed to organize decision-making and the creation of policy (the law-makers have given a kick at democracy by passing that law). The same evaluation - that the actions of the government regarding the pensions reform go against democratic principles - is also discernible in the way the speaker refers to Macron, calling him ‘the grand sultan’, an absolute monarch who does as he pleases without checks and balances. The phrase ‘let it rumble a little while longer’ refers to the protesting movement, for which the speaker expresses his support.

In (8) the speaker explicitly identifies the reason he picked up a particular political event (and speech event): these events are picked out because they are problematic in some way, a reason which is implicit throughout the data set. These political events can be problematic if they infringe upon democratic principles: when they attract attention and have no commensurate outcomes in terms of how they help promote the interest of the people; what is more, the speaker takes issue with how these events can be utilized to divert attention from important issues which are a matter of public interest, namely, the squandering of public funds in a non-transparent manner.⁷

In (9) the speaker provides a brief recall of the case, referring to a media source for the facts; his assessment of the case is seen in the description “shady associations” –

⁶ For lack of space, I will not discuss the historical justification for this provision, but I will note that a motion of no confidence adopted by absolute majority can be opposed to using the provision 49-3, a motion which failed to receive the support of the majority of Assembly members in March 2023 when the reform bill was passed.

⁷ The Marianne fund organized by Marlène Schiappa in 2021 is currently under investigation for corruption and embezzlement after a series of journalistic investigations revealing irregularities in the mechanism of allocation of funds, the activities which appear to have been funded, etc. (France24)

untrustworthy associations who have received the bulk of the funding, which is the opposite of transparency.

Both of the channels position themselves primarily as entertainment, but also, importantly, as information channels, not producing the news themselves, but as providing checks and balances to political news. On the backdrop of a receding interest for political news (for example, Aalberg et al. 2013), the interest for an alternative checks and balances, providing an overview of political news, entertainment and an interpretative blueprint from a fellow citizen who does not claim to possess any special kind of expertise but their own critical evaluation skills, seems to grow. These channels have a clear advantage over their traditional media counterparts: they proudly display their subjectivity, their opinions in order to convey the importance of being critical and alert while consuming content, regardless of the source. These channels are also free from top-down ideological pressure: while traditional media outlets in France depend on state subsidies and on private capital which may lead to complicated ideological dependencies (or suspicions of such dependencies), user-generated content is bottom-up-driven. This means that it is the interest of the audience which is going to determine the viability of such content on the platform. Even if it can be motivated by opportunity on the part of content-creators, as authenticity appears to be quite prized on these platforms (for example, Leppänen et al. 2015), a long-term success on the platform will depend on the content-maker displaying consistency within the projected image. Authenticity, in this respect, would amount to not committing the faults that the content-creators ridicule in other outlets.

The analysis of the data set suggests that the evaluation is based on the general criterion of appropriateness of the discourse stretch to the political discourse frame. The subjective brand of political discourse frame for the two channels involve an idea of how (the form and manner in which political discourse is conducted) and what (the content) is being presented to the audience. It involves a certain image of what a democratic society (especially, the French democratic society) should look like. Transparency, clarity, internal consistency, accountability; integrity – these are some of the facettes of this image which can be inferred from the speakers' comments. In Table (1), a brief selection of more examples from the data set is presented.

The very choice of the reframed discourse indicates that the speech event is in some way non-conforming to the standard they hold for political discourse. In the illustrations provided above, some of the deficiencies which are implicit in the reframed discourse are the preoccupation with matters of political image instead of political action, political discourse is not transparent, relevant, or truthful, it does not address current issues in an adequate manner, is not public-oriented and does not follow public interest. The deficiency evaluation may concern politicians and public figures who do not display honesty, consistency, solid principles, etc. All of these implicit standards are supposed to provide a gauge of appropriateness of the political speech event. These categories are called by Bednarek “evaluative parameters” (2010)

and they concern different semantic dimensions along which speakers can propose an evaluation through language.

Table 1: Evaluative criteria

Criterion	Reframed speech event	Discursive elements and strategies that carry the evaluation
transparency	The affair of the Marianne fund	Un immense scandale (...) complètement caché
appropriateness	The Playboy photoshoot of Marlène Schiappa	Jouer à la mannequin
transparency accountability	Public communication to explain/ justify the Marianne fund scandal	Ce qui est pas bon signe quand tu as ce genre d'affaires t'envoies tout le monde sauf Olivier Véran
appropriateness	Public communication to explain/ justify the Marianne fund scandal	Le mieux c'est d'envoyer un sac plastique à la rigueur tu peux vomir dedans
clarity accountability appropriateness	Public communication to explain/ justify the Marianne fund scandal	Le problème avec Olivier Véran c'est qu'il n'en met jamais une dedans il ment tout le temps
appropriateness relevance	Public communication to explain/ justify the Marianne fund scandal	Quand tu vois les photos de Marlène Schiappa dans Playboy il a plein de choses qui améliorent la vie des Français mais pas ça
appropriateness decorum	Macron's visit to China in april 2023	Emmanuel Macron hein c'est Tintin en Chine il continue son petit tour du monde
decorum transparency truth	Macron's tour of rural France late March 2023	La tournée en France n'a pas très bien marché ouais faute de public apparemment
transparency appropriateness	Macron's visit to China in april 2023	Quel meilleur endroit pour faire de la propagande gouvernementale que la Chine
relevance	Public communication on the pension reform	C'est vrai qu'à moins d'être un pyromane convaincu tu défends pas un incendie que t'as toi-même créé pour la qualité des flammes
integrity	A book promotion by Manuel Vals	Franchement je serais favorable pour que pour la première fois de ta vie tu obéisses à tes propres principes

The primary motivation in presenting those evaluations to the audience is embedded in the purpose of the content-creating enterprise. The primary goal (and a self-professed one) of these channels is entertainment: providing engaging content which caters to the interests of a particular kind of audience while providing them a pleasurable experience. Both of the channels opt for a recognizable formula: each episode follows more or less the same structure, with cues and special segments which

announce and prepare the audience for what is to come. Familiarity, paired with news value, and jokes and comedy make up the entertainment value of the format. The segments have their stable audience, who can tune in to see the new episode at their convenience, using hand-held devices. The evaluations add to the serious engagement with political discourse and the evaluation itself adds informative and communal value to the content: audiences receive a special selection of political speech events which are critically presented by somebody who is their peer (the information does not propagate in a top-down manner). The communal value lies in the ideological ties that are thus potentially built: even though the evaluation proposed by the speakers is subjective, it relies on values that can be expected to be shared by the community who come from the same socio-political tradition. For people outside that tradition, it provides an interesting insight into the cultural and political makeover of French society, as viewed by a regular citizen. Such activity can be inscribed in the type of grassroots citizen journalism (Allan 2009), reflecting a grassroots political activity: as regular citizens are usually the passive recipients of political discourse (their active participation in many cases is limited to vote-casting in elections), such content creation and sharing helps to reflect upon, comment, criticize, and, generally, gauge the kind of political discourse which is propagated daily through various media outlets. It helps to organize interpretative grids which go beyond what traditional media usually offer; they help decode actions, decipher speeches, expose falsehoods, in other words, help build up a critically-minded citizen community. This ties in with what researchers have identified as the role of alternative news media: being “a self-perceived corrective to ‘traditional’ news media” (Holt, Figenschou and Frischlich 2019). Even though the purpose of the channels is not to be news sources, both claim informativity as a characteristic of their enterprise, and providing a different take on what is already considered a piece of news can qualify as informative. Of course, popularity is another objective of the channels, as the number of views of the video helps with its ranking and propagation, increasing the possibility for monetization for the creator.

The function of the evaluation is more difficult to define. There are, however, some indications which can lead to a tentative conjecture. As mentioned above, the serious engagement with the reframed political discourse aims to expose its deficiencies, the identification of which are implicitly presented to be in the best interest of an alert citizen who wishes to understand what is going on on the political scene. Positioning themselves as correctives to traditional media, the only affiliation which the speakers accept for themselves is the affiliation with truth: partisan positions are not expressed in the analyzed data set (nor have I noted any in the other numerous episodes I watched in preparation of the study); no political party is specially targeted for its political or ideological positions; political figures appear to be targeted for what they do and say, not for what their political affiliation is. Denouncing partisanship on the part of traditional media (usually for their support of the ruling center-right), the speakers are committed to not display partisanship themselves. The evaluation does not appear to

have ideological functions in terms of agitprop in favor of a particular political actor or party, in sharp contrast with algorithm-driven individually tailored political content which aims to capitalize on dissent and polarization (for an in-depth discussion, see Colon 2021). It is committed to exposing problems related to the ruling government, which is considered to be what a regular citizen should do. This is in particular the kind of grassroots citizen political activity which is carried by the social networks: instead of a political activism, promoting the development of a critical attitude towards political discourse of any kind.

4 Emotion and the intensity of the evaluation

The evaluation is carried through different discursive means: lexical means, that is through evaluative language of the type seen in illustrations (7) – (9); irony and sarcasm, helped by the oral delivery by the speakers who modulate their voice to provide cues for the audience. Thus, entertaining elements are intertwined with evaluative elements, and this discursive technique allows the very intensity of the evaluation to be judged. The format relies on pre-scripted contributions, which means that the speaker prepared and crafted the message before filming, chose the words and came up with the jokes in order to enhance the entertaining elements, carefully selecting what he says. The message being prepared beforehand, the emotion which is displayed is performed, not felt at the time of filming, this performance also being part of the speaker's role as an entertainer. The emotion that is performed has the function of validating the intensity of the judgment the speaker is trying to convey. The stronger the displayed emotion, the more intense the conveyed attitude. Emotion and attitude cues help with setting the entertainer frame, help establish the actor as a performer who aims to entertain first. But when it comes to the evaluation, these cues carry the intensity of the evaluation. Consider:

- (10)[brings right hand to forehead then to mouth] Oh le niveau de branlette que vous êtes obligés d'atteindre pour justifier votre bousin oh ça ne veut ↑rien dire [makes punctuating gesture with left hand] mais ↑rien du tout [repeats punctuating gesture, brings hand to mouth] c'est c'est c'est quoi l'impôt de trop Gabriel l'impôt de trop pour l'hyper-classe c'est quoi c'est quand c'est fois deux fois trois fois dix fois cent [accompanying gesture of gradation] fois un ↑trilliard TU PEUX L'AUGMENTER AUTANT QUE TU ↑VEUX ILS LE PAIENT PAS DU CON [brings hands in an opening gesture] grâce aux lois de l'évasion fiscale que vous avez votées (JSPC N380)⁸ ['Oh the level of bullshit you have to get to to justify your crap oh but this does not mean anything anything at all what is the one tax too many Gabriel what is the one tax too many for the hyperclass when it's twice as much three times ten times a hundred

⁸ I use symbols from the Jefferson Transcription System to indicate relevant pitch change in this segment.

times as much you can increase it as you want they are not paying it you moron all thanks to the laws of tax evasion that you voted in’]

This illustrates how emotion is performed through an assortment of cues: the voice pitch changes, the rising pitch serving to emphasize the words that the speaker wants to draw attention to. The gestures also participate to build up the performance of emotion, along with the face: the eyes roll, or are closed as if in disbelief, or are fixed straight at the camera intensely as if making eye contact. The words in capital letters stand for a segment which is shouted out, instead of spoken, which also serves to convey an intense emotion. All of these cues assist the performance and help the audience understand the intensity of the judgment the speaker has for the reframed discourse.

Performances of emotion are also carried out by using offensive language: insults and invectives, swear and taboo words are a deliberate breach of politeness expectations in interpersonal communication; they are openly and deliberately transgressive. They are by themselves a sign of intense emotional states, or they can help stage a speaker in a highly emotional psychological state. As the data set consists of pre-scripted material, the choice of the invectives is deliberate and they perform a specific function, namely, to convey the intensity of the speaker’s emotion. In (10) there are several terms which are considered to be pejorative and/or offensive: ‘branlette’, ‘bousin’, ‘du con’, which along with the gradation (twice, three times, ten times, etc.), are markers of increased intensity of the displayed emotion. Table (2) presents an illustrative sample from the data set.

The evaluations that are carried by the offensive terms can be those of the speaker or those of another actor involved in the speech event. Some of the emotionally colored terms can also be evaluative terms, such as ‘naze’ or ‘connard’. Offensive terms become a strategy to carry out the evaluations in the context of reframing, a strategy because they are a deliberate choice on the part of the speaker who stages face-threatening acts unambiguously and directly (according to Culpepper’s taxonomy, 2016). The speaker stages the acts because he is not seriously (in Austin’s sense, 1962) addressing the president or various government members he targets in his satire. This is an etiolated use of language, in which the entertainment frame gives license to the speaker (the actor) to violate usual requirements for what is appropriate to say to interlocutors. The transgressions that he allows himself to make are thus done under the creative license of the entertainment frame and are not taken seriously. They become pure symbols of the intensity of the message - this time the serious message - which satire usually contains, namely, that there is something inappropriate, the kind of non-conformity to some type of standard held by the speaker of the speech events that are reframed.

Table 2: A sample of offensive language from the data set

Offensive language terms from data set	Function	Evaluation expressed
Il a envoyé chier le conseil scientifique	Vulgar phrase for ‘dismiss, ignore’	Disdain or contempt, in this case, on the part of Macron towards the scientific council
C'est un peu des connards	Vulgar word for an idiot	Dismissal
Offrez-lui un fucking kit du petit chimiste	Interjection, intensifier	desperation
Assumez jusqu'au bout que vous n'en avez rien à foutre	Colloquial/vulgar phrase meaning ‘not give a fuck’	Disdain or contempt is imputed to imaginary addressee (the government)
C'est un gros naze	Pejorative word meaning ‘moron’	Disdain or contempt
Ils l'ont choppé ils lui ont éclaté la tronche	Colloquial phrases for ‘get’ and ‘beat up’, used here figuratively	Ridicule, contempt
Pause caca	Appellative, a cultural borrowing from a character nickname (the American pie, 1999)	Disdain or contempt
Il se fait défoncer la gueule	colloquial/ vulgar phrase meaning ‘get the shit beaten out of oneself’ used here figuratively	Disapproval imputed to Olivier Faure towards Macron
Qu'est-ce que vous êtes drôles putain	Interjection, intensifier	desperation

5 Conclusions

The interest in exploring this format on user-generated content-sharing platforms lies in the insight it can provide about the way ordinary citizens, who are usually excluded from political discourse, engage with that discourse in a more active way. First, by drawing attention to deficiencies they see in it; second, by proposing a critical assessment of its content and objectives; third, by building up a community of people who share the ideas they have about the quality of political discourse. This is a grassroots way of ordinary citizens to enact their participation in political discourse, completely consistent with the affordances of digital platforms (Allan 2009).

Positioning themselves as an alternative to traditional media, they explicitly assume the particularities of user-generated content and use them to their advantage, especially in the context of high-choice media environment characterized by, as Van Aelst et al. 2017 put it, a declining supply of political information, paired with a decline in the quality of

political news.⁹ The advantages consist in weaponizing the subjective take on events, as the way information on social networks and platforms is judged differs greatly from the way information is judged on traditional media: social platforms put forth the subjective, personal experience, the individual story (Jandura and Friedrich 2014). The driving motive of this particular format appears to be a shared understanding of democratic processes (at the heart of political discourse) which ought to be inclusive of all members of society, driven by enlightened understanding, and be shaped by active and effective participation (Dahl 1989). Peace, transparency, liberty - these are the core values which appear to shape democratic societies, so that any perceived deviation or betrayal of those principles is judged as undermining the democratic build-up of society. For the ordinary citizen whose perceptions of political processes are shaped by the political communications, there appears to be a fundamental discrepancy between the two sides of political discourse, the politicians and the rest of the citizens, which is perceived as a divide. There is another divergence which concerns the type and extent of political information which is available to both sides of the divide. And while there is an objective reason for the latter divergence, one way of bridging the gap lies in learning to understand it better. This appears to be one of the functions of reframing, in demystifying the political actors and their discourses, exposing their shortcomings, distancing oneself in order to assess them critically.

Political communication is generally propagated through a plethora of different media platforms, traditional (television, radio and newspaper broadcasting) and alternative (different associations of digitally-based platforms, which may provide alternative propagating paths or may enhance traditional outlets); but in a high-choice media environment, especially one which is mainly carried through the internet, where news flows are fragmented, many researchers point out how easy it has become to opt out of political news in favor of other preferred type of content (for example, Prior 2007, Colon 2021). That a hybrid type of content as the one under investigation here is in demand can be an indirect confirmation of that trend; it could be construed as a substitute for news, as what these channels do is provide a synthesis of current affairs, together with contextualization, entertainment and comment/evaluation. In order to select the content, the content-creators have to browse massive amounts of political news, which their audience is spared from. It is highly doubtful that such content can replace the gathering of political knowledge by other means. But what it can do is spark an interest for such knowledge and provide critical perspective, which is generally a positive development for democracy (for recent perspectives, see Butler forthcoming).

This interwoven web of communications, their coverage and assessment provide a fair picture of what political discourse looks like in a high-choice media environment: an intricate mixture of genuine communication which has the purpose to inform the public of policy decisions or what not, and image concerns (damage control in case one of the

⁹ Van Aelst et al. also list increased fragmentation and polarization, decreased diversity of political news, which holds serious concern about the state of democracy in society.

actors involved makes a faux pas, or creates a controversy, or provides an alternative explanation or justification). Such hybrid formats allow people (content-creators and their audiences alike) to enter the debate which unfolds in real time through digital media coverage. What place it is going to carve for itself on the political scene remains to be seen.

References

- Allan, Stuart. 2009. Histories of citizen journalism. In Stuart Allan & Einar Thorsen (eds.), *Citizen journalism: Global Perspectives (Vol. I)*, 17–31. New York: Peter Lang.
- Austin, John L. 1962. *How to do things with words*. Cambridge, Mass.: Harvard U.P.
- Bednarek, Monika. 2006. *Evaluation in media discourse: Analysis of a newspaper corpus*. London: Continuum.
- Blommaert, Jan. 2020. Political discourse in post-digital societies. *Trabalhos em Linguística Aplicada* 59. 390–403.
- Butler, Robert. (ed.) forthcoming in 2024. *Political Discourse Analysis - Legitimisation Strategies in Crisis and Conflict*. Edinburgh: Edinburgh University Press.
- Chankova, Mariya. in print-a. Political discourse strategies in online communication: a case study. To appear in Isabela Nedelcu, Irina Paraschiv & Andra Vasilescu (eds.), *Proceedings of the 21st Conference of the Department of Linguistics, Bucharest, 2021*. Bucharest: Bucharest University Press.
- Chankova, Mariya. in print-b. Political discourse and the new media: new architectures of communication. To appear in Robert Butler (ed.), *Political Discourse Analysis - Legitimisation Strategies in Crisis and Conflict*. Edinburgh: Edinburgh University Press.
- Chebib, Nadine K. & Rabia M. Sohail. 2011. The reasons social media contributed to the 2011 Egyptian revolution. *International Journal of Business Research and Management (IJBRM)* 2(3). 139–162.
- Colon, David. 2021. *Propagande. La manipulation de masse dans le monde contemporain*. Paris: Editions Flammarion.
- Culpeper, Jonathan. 2016. Impoliteness strategies. In Alessandro Capone & Jacob May (eds.), *Interdisciplinary studies in pragmatics, culture and society*, 421–445. Heidelberg: Springer.
- Dahl, Robert A. 1989. *Democracy and its critics*. New Haven: Yale University Press.
- Fillmore, Charles J. 2006. Frame semantics. In Dirk Geeraerts (ed.), *Cognitive Linguistics: Basic Readings*, 373–400. Berlin: Mouton De Gruyter.
- Goffman, Erving. 1974. *Frame analysis*. New York: Harper and Row.
- Herman, Edward S. & Noam Chomsky. 2010. *Manufacturing consent: The political economy of the mass media*. London: Vintage (Random House).
- Holt, Kristoffer, Tine U. Figenschou & Lena Frischlich. 2019. Key dimensions of alternative news media. *Digital Journalism* 7(7). 860–869.
- Hunston, Susan. 2010. *Corpus approaches to evaluation: Phraseology and evaluative language*. New York: Routledge.
- Jandura, Olaf & Katja Friedrich. 2014. The quality of political media coverage. In Carsten Reinemann (ed.), *Political communication*, 351–373. Berlin: Mouton de Gruyter.
- Kress, Gunther & Theo Van Leeuwen. 2001. *Multimodal discourse. The modes and media in contemporary communication*. London: Arnold.

- Kubin, Emily & Christian von Sikorski. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* 45(3). 188–206.
- Leppänen, Sirpa, Janus S. Møller, Thomas R. Nørreby, Andreas Stæhr & Samu Kytölä. 2015. Authenticity, normativity and social media. *Discourse, Context and Media* 8(June). 1–5. DOI: <https://doi.org/10.1016/j.dcm.2015.05.008>.
- Morozov, Evgeny. 2011. *The net delusion: How not to liberate the world*. London: Allen Lane.
- Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.
- Prior, Markus. 2007. *Post-broadcast democracy. How media choice increases inequality in political involvement and polarizes elections*. New York, NY: Cambridge University Press.
- Shullenberger, Geoff. 2020. The new net delusion. *The New Atlantis* 62. 46–52.
- Tannen, Deborah. 1993. What's in a frame? Surface evidence for underlying expectations. In Deborah Tannen (ed.), *Framing in discourse*, 14–56. Oxford/New York: Oxford University Press.
- Tannen, Deborah & Cynthia Wallat. 1987. Interactive frames and knowledge schemas in interaction: examples from a medical examination/interview. *Social Psychology Quarterly* 50(2), 205–216.
- Thompson, Geoffrey & Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston & Geoffrey Thompson (eds.), *Evaluation in text: Authorial stance and the construction of discourse*, 1–27. Oxford: Oxford University Press.
- Van Aelst, Peter, Jesper Strömbäck, Toril Aalberg, Frank Esser, Claes de Vreese, Jörg Matthes, David Hopmann, Susana Salgado, Nicolas Hubé, Agnieszka Stepińska, Stylianos Papathanassopoulos, Rosa Berganza, Guido Legnante, Carsten Reinemann, Tamir Shefer & James Stanyer. 2017. Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association* 41(1). 3–27.
- Wall, Melissa. 2018. *Citizen journalism: Practices, propaganda, pedagogy*. London: Routledge.
- Yus, Francesco. 2011. *Cyberpragmatics: Internet-mediated communication in context*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

About the Author

Mariya Chankova holds a PhD in English Linguistics awarded from RWTH – Aachen. She is Senior Assistant Professor in English and French at the South-West University N. Rilski in Blagoevgrad (Bulgaria). Her research interests include speech act theory, discourse analysis, political discourse, frame analysis, plagiarism and academic integrity, communication in the multimedia environment.

Address

Alliance française, Department of Germanic and Romance Studies
66, Ivan Mihaylov str.
2700 Blagoevgrad, Bulgaria

e-mail : mariyachankova@gmail.com
ORCID: 0000-0003-3017-125X

Barbara Lewandowska-Tomaszyk, University of Applied Sciences in Konin

Anna Bączkowska, University of Gdańsk

Olga Dontcheva-Navrátilová, Masaryk University, Brno

Chaya Liebeskind, Jerusalem College of Technology

Giedre Valunaite Oleskeviciene, Mykolas Romeris University

Slavko Žitnik, University of Ljubljana

Marcin Trojszczak, University of Applied Sciences in Konin

Renata Povolná, Masaryk University, Brno

Linas Selmistraitis, Mykolas Romeris University

Andrius Utka, Vytautas Magnus University

Dangis Gudelis, Mykolas Romeris University

LLOD SCHEMA FOR SIMPLIFIED OFFENSIVE LANGUAGE TAXONOMY IN MULTILINGUAL DETECTION AND APPLICATIONS

Abstract

The goal of the paper is to present a Simplified Offensive Language (SOL) Taxonomy, its application and testing in the Second Annotation Campaign conducted between March-May 2023 on four languages: English, Czech, Lithuanian, and Polish to be verified and located in LLOD. Making reference to the previous Offensive Language taxonomic models proposed mostly by the same COST Action Nexus Linguarum WG 4.1.1 team, the number and variety of the categories underwent the definitional revision, and the present typology was tested in the annotation on the publicly available offensive language datasets of each of the four languages. The results of the annotation are presented and as they are contained within the accepted statistical values on the inter-annotator agreement in the SOL categories and their aspects, we propose this taxonomy as a core ontology which represents the encoding of the supported offensive languages and justify its use on new data in terms of a more universal Linguistic Linked Open Data (LLOD) schema.

Keywords

offensive language, offensive language taxonomy, annotation, LLOD, linguistic linked open data, hate speech

1 Introduction

This paper is a presentation of the background and application of a Simplified Taxonomy of Offensive Language (SOL) (Lewandowska-Tomaszczyk 2022) in the identification of offensive language in four languages: Czech, Lithuanian, Polish, and English. The ultimate objective is to verify the taxonomy for the development of a schema for LLOD.

2 Previous models proposed by the Nexus Linguarum WG 4.1.1. team

2.1 SALLD-1 (Lewandowska-Tomaszczyk et al. 2021)

A definitional revision and enrichment of offensive language typology were the main objectives of Lewandowska-Tomaszczyk et al.'s (2021) publication. We reviewed over 60 existing corpora and compared applied tagging schemas of the existing offensive (also called abusive, toxic, etc.) language tag set systems and exemplified their classes in a proposed schema. Similar schema generated with the Sketch Engine data and non-contextual word embeddings – i.e., Word2Vec, Glove, were consulted to get better insight to their semantic difference in English. In the 2021 paper we developed a taxonomy covering a finite set of categories and aspects of offensive language representation along with linguistically sound explanations. We proposed a core ontology which represented the encoding of the defined offensive language schema. A survey of computational models of detecting offensive language was also presented basing on the HatEval Task 5 of SemEval-2019 (Basile et al. 2019) and on the OffensEval Tasks of SemEval-2019 and SemEval-2020 (Liu et al. 2019; Zampieri et al. 2019a and 2019b). We did not consider using the O-Dang! or similar ontologies (<https://aclanthology.org/2022.salld-1.2/>), as these proposals are rather broad and represent metadata of a corpus.

The ontology of offensive language we propose in this research provides defined classes for each concept. It was originally inspired by a three-level hierarchy of offensive language put forward by Zampieri et al. (2019a, 2019b). Contrary to Zampieri et al. (2019a, 2019b), however, in our research, offensive language is further refined and divided into two basic levels of analysis (Level I and II), and four sublevels (A, B, C, D) within Level I. Level I distinguishes lexical items that are offensive from those that are not (Level A: offensive vs. non-offensive). Secondly (Level B: targeted vs. non-targeted), the question whether the selected items are targeted at some addressee should be answered. If there is no identifiable addressee then the use of offensive language is an example of self-expression, which has an exclamatory function, e.g., the use of swear words to express anger, frustration, pain etc (*abusive swearing* in Andersson and Trudgill 2007: 197). Targeted offensive items are further

divided into either *implicit* or *explicit* cases of offensive language (Level C: implicit vs. explicit language). While implicitness may be encoded by, for example, sarcasm and irony, whereby offence is veiled, explicitness entails more straightforward forms of verbal attack. Classes of explicit targeted categories of *offence* are further subcategorized into types characterized by varying kinds of internal or external targets as well as partly distinct characterization of the lexicon.

2.2 Integrated explicit and implicit offensive language taxonomy (Lewandowska-Tomaszczyk et al. 2023; Bączkowska et al. 2022)

The first attempt at proposing an integrated explicit and implicit offensive language taxonomy was given in Lewandowska-Tomaszczyk et al (2023) and supported by analyses of implicit offensive language categorization as discussed in Bączkowska (2022) and Bączkowska et al. (2022). The implicit offensive language model that we proposed (Bączkowska et al. 2022a, 2022b) is rooted in Grice's (1989) four categories of implicitness, i.e., metaphor, irony, hyperbole (overstatement) and meiosis (understatement), which was enriched by the category of indirectness, understood in the Searlian sense of Indirect Speech Acts (Searle 1975). Additionally, the term *sarcasm* was also added as a subtype of irony, rhetorical question and simile. Overall, 8 main implicit categories were distinguished in our model. Whilst essentially, two models of offence (explicit and implicit) have been proposed in our project, we also identify an area which is in between the two models, and which integrates them. Their status is less obvious in terms of their typology as they seem to share features typical of both explicitness and implicitness, and this primarily includes dead metaphors e.g., *hand of a clock*, *to fall/be in love*, which, having a high frequency of occurrence and, as a result, conventionalized, are no longer seen by the receivers as opaque. In fact, they are perfectly understandable, i.e., explicit in meaning, though at the same time, technically speaking, they are instances of implicit meaning.

In the paper by Lewandowska-Tomaszczyk et al. (2023), the concept of offensive language as a superordinate category was proposed with a number of hierarchically arranged 17 subcategories, taxonomically structured into 4 levels and verified with the use of neural-based (lexical) embeddings, which automatically encode generic semantic relatedness as well as hypernym, synonymy, and other types of relationships. The graphs included in the paper visualize the relationship between the embeddings of the concepts.

Together with a taxonomy of implicit offensive language and its subcategorization levels which received little scholarly attention before, the categories were divided into offensive category *levels* (types of offence, targets, etc.) and *aspects* (offensive language property clusters) as well as the main categories of *explicitness* and *implicitness*.

2.3 A short survey of results of the First Annotation Campaign (Lewandowska-Tomaszczyk et al. accepted for *Rasprave*)

The integrated explicit-implicit offensive language category schema was verified on a large English data, consisting of 25 publicly available English datasets of offensive language with the INCEpTION tool (<https://github.com/inception-project/inception>) – a semantic *annotation* platform offering assistance in the annotation. The annotation categories were defined according to the annotation guidelines. The annotation setup was the same in both campaigns - annotators needed to select one or more consecutive sentences (could also be the whole passage) that were identified as offensive. For the selected sentences then the annotator needed to select appropriate annotation categories.

The list of the English datasets used in this First Annotation Campaign is available as an Appendix in Lewandowska-Tomaszczyk et al. (2023). The results (Lewandowska-Tomaszczyk et al. accepted) partly support the proposed ontology of explicit offence and positive implicitness types to provide more variance among widely recognized types of figurative language (e.g., metaphorical, metonymic, ironic, etc.). However, further results and a series of the annotators' comments in a questionnaire showed that for some of the categories there was low or medium inter-annotator agreement. It was also more challenging for annotators to distinguish between category items than between aspect items such as *offensive*, *insulting* and *abusive*, being the most difficult in this respect. The need for taxonomic simplification measures in this respect was thus recognized for further annotation practice and offensive language identification.

3 A Simplified Taxonomy of Offensive Language (SOL)

3.1 Introduction

The need to simplify the taxonomy for the purposes of computation effective offensive language annotation and recognition was acknowledged and a proposal of a Simplified Offensive Language taxonomy (SOL) was soon put forward (Lewandowska-Tomaszczyk 2022). The linguistic and computational limitations of any attempt to provide water-tight categorization schemas in any language was discussed in the paper and what was proposed there was a carefully supervised simplification of the Extended Model both as far as the number and types of the key categories are concerned. To verify the taxonomy we resorted to the word embedding correlations discussed in 3.2. below.

3.2 Word embeddings for the English SOL taxonomy keywords

We used the approach given in the *integrated explicit and implicit offensive language taxonomy* publication (Lewandowska-Tomaszczyk et al. 2023) to examine the correlations between offensive language categories. To represent the categories as vectors in a lower dimension space, we utilized the Word2Vec word embedding method and calculated the Cosine distance between them. The word embedding learning was done from scratch with a corpus of offensive language. We also experimented with contextual embeddings such as ELMo, BERT, KeyBERT, USE and ConceptNet Numberbatch embeddings. Only the results of the latter revealed some structures, similar to Word2Vec ones, while results of the others were blurred and overlapping in the resulting 2D visualizations.

We calculated our categories and aspects' pairwise cosine similarity first. Cosine similarity heatmap is shown in Figure 1. A heatmap is a color-coded visual representation of data, with red indicating high similarity and blue indicating low similarity. The figure is mostly blue. Since words without similar contexts have a low cosine value, most of these categories and aspects may be readily isolated and form independent offensive categories.

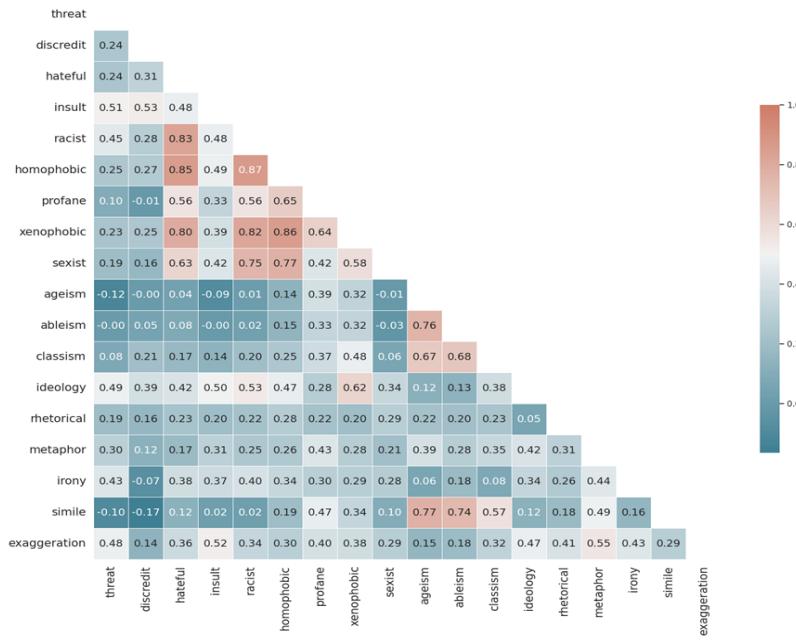


Figure 1: Cosine similarity heatmap

While the categories can be separated readily, certain aspects cannot. Homophobic and racist aspects are the most similar (0.87), and xenophobic is very close to both (0.86 and 0.82, respectively). Sexist is also close to homophobic and racist (0.77 and 0.75, respectively). As evidenced by their proximity to the hateful category, the homophobic, racist, and xenophobic (0.85, 0.83 and 0.80, respectively) aspects are likely the most prevalent expressions of hatred within the corpus.

There is a high degree of similarity between ageism and ableism (0.76) and both ageism and ableism are close to classism (0.67 and 0.68, respectively). Ageism and ableism are also predominate aspects of the implicit category simile (0.77 and 0.74, respectively).

The lemma forms of the categories were then analyzed. We extracted the top 30 most similar terms for each category, excluding words whose substrings are the category, its lemma, or its stem. The t-SNE (t-distributed Stochastic Neighbor Embedding) method was then applied to the embeddings of the categories and their 30 most similar terms.

Figure 2 shows the fifty-to-two-dimensional t-SNE transformation of our embedding vectors. The t-SNE Figure reveals that the discredit and threat categories form relatively well-defined clusters, whereas the hateful and insult categories overlap various aspects of offensive language. The ideology aspect forms a relatively distinct cluster. However, the remaining aspects are dispersed and overlap.

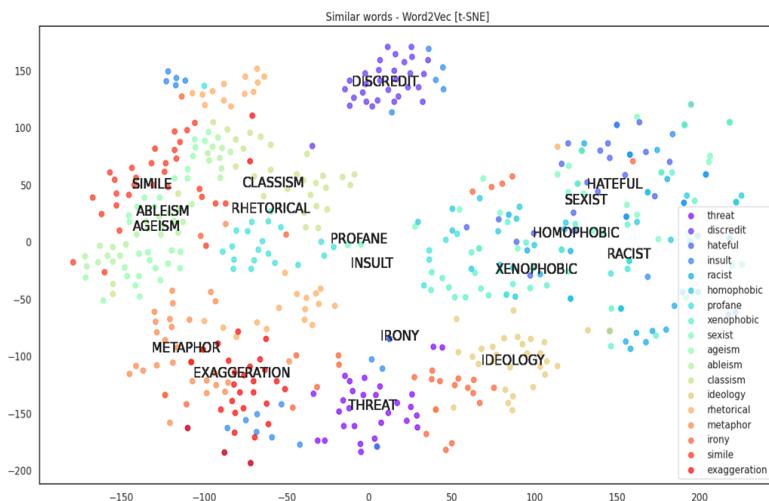


Figure 2: Fifty-to-two-dimensional t-SNE transformation of embedding vectors

3.3 SOL Taxonomy

The simplified taxonomy was presented in terms of the step-by-step hierarchical procedure. This taxonomy was to prepare the ground for the Second Annotation Campaign in a multilingual context (Czech, Polish, Lithuanian, and English).

SIMPLIFIED OFFENSIVE LANGUAGE TAXONOMY

1. OFFENSIVE [YES or NO]

2. Target 1

Individual // Group // Ind wrt Gr/Gr wrt Ind [by reference to group stereotypes]

3. Target 2

present/absent

4. Vulgar [YES or NO]

5. Choose either (i) or (ii); Then select (iii) or (iv) or both (iii) and (iv)

(i) INSULT [addressed to: individual or group – varied offence types but not by group stereotypes]

(ii) HATE SPEECH [individual or group; offence by reference to group stereotypes]

(iii) DISCREDIT [individual or group//on various grounds – lying-cheating, immorality, unprofessionalism, unfairness]

(iv) THREAT [individual or group, inducing fear]

6. Aspects – [Choose one or more]

[racist] [xenophobic] [homophobic] [sexist] [profane (religion)] [ageism]
[physical/mental disabilities [ableism]] [social class [classism]] [ideologism] [other]

7. Select categories below – [Choose one or more]

RHETORICAL QUESTIONS

METAPHOR

SIMILE

IRONY

EXAGGERATION

OTHER

Figure 3: Step-by-step hierarchical procedure
of the simplified offensive language (SOL) taxonomy

The question concerning the overall offensiveness status of the selected sample is crucial to establish. The categorical answer *yes* or *no*, similarly to the other *yes-no* categorization proposed in the categories below, is not a reflection of actual conceptual-linguistic reality (Lakoff 1987) but, rather, indicates annotation requirements to adjust to a computer program to distinguish between dichotomic judgments (e.g., *offensive vulgar* or *not*).

Levels 2 and 3 refer to a selection of individuals or groups, that is, the *targets* of an offensive act – an individual (*target a*), a group (*target b*), or else a *target c*, addressed at a group through a particular individual or else an individual meant to be a group representative. The main criterial property of the latter (*target c*) is the use of gender, race,

etc., *stereotypes* in the offensive language sample and paves the way to the category of *hate speech* as one of the offence types in the hierarchy. *Target 2* is a tag which represents presence or absence of the offensive language target at the *locus* of the interactional encounter.

The selection between *vulgar* and *non-vulgar language* (i.e., words, phrases) is taxonomically connected to the first-level selection between *offensive* or *non-offensive* type, i.e., in further judgments of the vulgarity of a particular sample provided in a larger linguistic context, as in the *Yep usual bullshit* response (Lewandowska-Tomaszczyk 2017). The lower distinctions are definitional with respect to the character of the used offence. i.e., the category of *insult* to determine an individual or group offence, but *not* by reference to any *group stereotypes* (e.g., *it's the state of your own mental health you should be VERY concerned about presents INSULT* and the Aspect of abism2), as juxtaposed to the concept of *hate speech*, whose discriminating property is precisely the reference to a group or individual *via* discriminatory group stereotypes. The *discredit* tag signifies an offensive act addressed at an individual or a group on grounds of accusation of lie, immorality, unprofessionalism, and unfairness, while *threat* is a statement intended to frighten or intimidate a person or a group into believing in prospective harm they will experience (Brenner 2002).

The level of *Aspects* as defined in this model is considered a type of property *target* of a discriminating act, and can be addressed at the offendee's race, gender, age, religion, ethnicity, ability, any other kind of physical property or behavioural conduct (dubbed as *other* in the Aspects compartment), or else any of their combinations.

The last-level distinction refers to a differentiation between linguistically explicit versus implicit types of utterances (category types 7.) and proposed here is a selection of one or more of the linguistically implicit categories (cf. Bączkowska et al. 2022).

4 Second Annotation Campaign and its results

In this section we present results of the Second Annotation Campaign performed with the use of the SOL taxonomy on the four languages independently. In the tables below we quote *positive results of each annotation*, while the few problematic cases are discussed in Section 6. below.

4.1 English – a comparison with the First Annotation Campaign results

The first attempts to identify a satisfying Offensive Language annotation system were carried out by our team on the example of English. In this section we will present a result of the application of the SOL taxonomy on English to see to what extent the Extended and Simplified systems compare. The present annotation of English on the SOL taxonomy was conducted on a smaller sample of 50 samples annotated by two annotators. The results are as follows:

Table 1: SOL taxonomy inter-rater annotation results for English

Annotation type	Agreement
Target 1 – Individual/group	0.82
Target 2 – present/absent	0.84
Vulgar	1.00
Offensive type – hate speech/insult	0.21
Offensive type discredit	0.57
Offensive type threat	1.00
Aspect 05	0.48
Aspect 05a	0.42
Aspect 05b	0.00 i.e., no tags given by annotators
Category 06	0.64
Category 06a	0.78
Category 06b	1.00

The SOL taxonomy annotation for English gives rather solid results. When contrasted with the results achieved in the First Annotation Campaign (Lewandowska-Tomaszczyk et al. accepted for publication), the present outcomes show a higher positive consistency in the annotators' selections of the categories and Aspects, and higher values of their annotation agreement. One category only (Offensive type - hate speech/insult) met with some problems, which may be remedied by means of training sessions more intensive than possible in this first test period. Others are mostly above the standard, which qualifies them to be proposed as a LLOD standard.

The proposed LLOD schema will update the initial schema (Lewandowska-Tomaszczyk et al. 2021). It defines a hierarchical structure of SOL offensive types and categories along with defined lists of targets and aspects. In this work we extended the taxonomy over multiple languages, so each concept will contain *rdfs:label* and *skos:definition* in multiple languages. As the schema represents collection of datasets, we will connect the classes with existing schemas such as DCAT-AP, so that the data is more inter-operable. Apart from the schema, instances will represent exemplars of each concept for all languages that achieved highest annotation agreements and are selected by the curators.

4.2 Czech

The Czech Corpus of Offensive Language comprises 400 comments extracted from online discussions in ten Czech national newspapers and news platforms and is located in the Sketch Engine software tool. The corpus was annotated by two annotators who are linguists and share a similar social background, age, and profession. Prior to annotating the corpus, the two annotators carried several training sessions focused on discussing potential problems in applying the simplified offensive language taxonomy.

Table 2: SOL taxonomy inter-rater annotation results for Czech

Annotation type	Agreement
Target 1 – Individual/group	0.89
Target 2 – present/absent	0.93
Vulgar	0.85
Offensive type – hate speech/insult	0.74
Offensive type discredit	0.74
Offensive type threat	0.11
Aspect 05	0.70
Aspect 05a	0.52
Category 06	0.61
Category 06a	0.53

The Cohen’s Kappa results for inter-rater agreement summarised in Table 2 show that the annotator agreement is high. More specifically, it is almost perfect for the categories Target 1 (0.89), Target 2 (0.93) and Vulgar (0.85), and substantial for the Offensive type categories (0.74 for both insult and discredit); the slight agreement for the Threat category may be explained by its very low occurrence in the annotations. As to Aspects of offensive language and categories of figurativeness, there is substantial agreement at the level of the main category (0.70 and 0.61 respectively), but there is moderate agreement in the sub-classes (0.52 and 0.53). This result is possibly affected by the absence of more specific instructions, as the annotators have ranked the sub-classes differently, for instance, annotator 1 has classified Metaphor as sub-category 6a and Irony as category 6b, while annotator 1 considers Metaphor as category 6b and Irony as category 6b. This is a rather technical issue that needs to be attended to in order to improve inter-rater agreement.

Overall, in the case of the Czech Corpus of Offensive Language, the use of the simplified taxonomy (Lewandowska-Tomaszczyk 2022) has yielded a considerably higher degree of inter-rater agreement in comparison with annotation campaigns using a more elaborate taxonomy of offensive language (e.g., Lewandowska-Tomaszczyk 2022). Other factors leading to a higher level of agreement might be a more intensive training campaign and careful preparation of the annotators.

4.3 Lithuanian

The pilot annotation applying the SOL taxonomy in Lithuanian was carried out on a sample of 200 items taken from the Lithuanian dataset LITIS which is freely available at CLARIN-LT repository <http://hdl.handle.net/20.500.11821/11>. Corpus LITIS contains user-generated comments collected from two Lithuanian news portals: www.delfi.lt and www.lytatas.lt. Each comment is in a separate file (TXT). Each file contains: a comment, date and time, the nick name of the author, URL and a title of the

article commented. Comments from www.delfi.lt amounts to 17,909 items, dated 2014 and comments from www.lrytas.lt amount to 182,000 items, dated 2010-2014.

The Cohen's Kappa results for inter-annotator agreement (Table 3) provide comparable results to the other languages of the annotation experiment.

Table 3: SOL taxonomy inter-rater annotation results for Lithuanian

Annotation type	Agreement
Target 1 – Individual/group	0.4
Vulgar	0.78
Offensive type threat	1.00
Aspect 05	0.45
Aspect 05a	0.65
Aspect 05b	1.00
Category 06	1.00
Category 06a	1.00

The Cohen's Kappa values range between 0.4-0.78 showing solid agreement on the identification of vulgar category with 1.00 reaching the maximum. There also could be observed Cohen's Kappa value 0.00 for offensive type threat and supplemental aspects and categories which were not tagged by the annotators in the annotated sample. The closer look at the annotated data revealed that lower annotator agreement for the categories of Target 1 - Individual/group and Aspect 5 could be explained by the differences in the annotator chosen chunks. It could also indicate a need for annotator pre-training sessions discussing the annotated samples, the offensive language taxonomy, comparing the results, and resolving disagreements.

4.4 Polish

The Polish annotation was performed by two raters on 100 samples derived from two Polish offensive language corpora (Ptaszyński and Masui 2018, Ptaszynski et al. 2019, Troszyński and Wawer 2017). The two datasets provide ca. 10,000 words, which make them suitable for our annotation. The majority of the inter-annotator agreement categories, their types and aspects are satisfying within the range between 0.20-0.30 and 1.00, reaching the perfect agreement.

Table 4: SOL taxonomy inter-rater annotation results for Polish

Annotation type	Agreement
Target 1 – Individual/group	1.00
Vulgar	0.62
Offensive type – hate speech/insult	0.30
Offensive type discredit	0.20
Offensive type threat	1.00
Aspect 05	0.61
Aspect 05a	0.33
Category 06	0.58
Category 06a	0.71
Category 06b	1.00

5 Problematic cases in the annotation

Some of Cohen's Kappa results for inter-annotator agreement show two basic problematic areas. The closer analysis revealed that the values of the inter-annotator agreement could be influenced by the differences in the length of the linguistic material selected for the annotation, from one word to full sentences or larger units. This is by far the basic reason for the discrepancies in the annotation practice. It indicates that there is a need for more careful and more intensive annotator training sessions, which will certainly be considered for the further regular annotator practice in order to resolve disagreements and come to the common grounds prior to the annotation implementation. Another, independent reason, is the uneven sampling length of particular corpus excerpts.

Interestingly, the differences are particularly visible between the standard for Target 2 – present/absent, proposed in SOL taxonomy and the practice in Lithuanian (L), 0.05, are also partly visible in Polish. Lithuanian also showed lower counts on the two other categories Offensive type_hate-speech/insult, and Offensive type discredit, both acquiring values lower than random.

On the one hand, the Target 2 problems can be accounted for by the rather uncertain status of these context-unavailable properties as presented in the language samples, and, on the other, on general problems encountered, as mentioned before, in the automatic division of the language in the data sets into particular samples for the annotation. Besides, the division problems of the main offensive language category into sub-categories: Hate speech, Insult, and – independently – Discredit, as seen in the lower agreement values for these categories in all languages, was not sufficiently considered in the preparatory stages of the annotation process. A more intensive training session might be proposed in the future tasks to remedy that.

Both the Polish annotation results and the other two languages (except for English) showed a lower agreement value (Polish – 0.19) for the second Aspect type – Offensive language 5b Aspect, which is typical of the selection among a larger number of Aspects (10 in SOL) and a rather low number of a variety of exemplary annotated samples.

On the other hand, the general categorization problem very well depicts the question of the absence of the strict boundaries in linguistic categorization, as seen as early as the fifties of the twentieth century in first publications from a philosophical and logical orientation (Wittgenstein 1953; Zadeh 1964), followed up by a surge of such research studies in cognitive linguistics (Lakoff 1987).

Furthermore, as observed by Lewandowska-Tomaszczyk (2011/2012), unless asked for detailed semantic analyses, language speakers generally use rather approximative meanings in their natural interactions than those involving minute sense discrimination. The computer applications as is known, also require a more definite judgement on such issues, although congruent in a more general sense, unlike in a number of previous ontology schemes, with minute sense differentiation achieved in professional linguistic analyses.

6 Gold standard offensive language examples in 4 languages

The present section shows “gold standard” offensive language examples in English, Czech, Lithanian and Polish as identified in a SOL taxonomy and used in the annotation referred to in the previous sections in the tables below.

ENGLISH

TARGET 1	Language: English
Individual	You are such a fucking moron.
Group	Snotty 17 year olds projecting their daddy and mammy issues on the world.
Vulgar	This is the first time I'm actually replying to your shit.
OFFENSIVE TYPE	
hate speech	Black people tend to be quite uncivilised.
Insult	Those wimps are the reason why we're losing more and more rights by the day.
Discredit	It would be irresponsible for Tory MPs to opt for #BorisJohnsonShouldNotBePM.
Threat	Consider yourself reported to the admin.
Aspect 01 racist/xenophobic	This is typical nigger territorial behavior.
Aspect 02 homophobic	He may be good at anal. You never know!
Aspect 03 physical/mental	No wonder I'm being uncivil, when you're stupid.

314 Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová,
 Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojszczak,
 Renata Povolná, Linas Selminstraitis, Andrius Utka and Dangis Gudelis
 LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications

disabilities/behavioural properties	
Aspect 04 Sexist	All women are too emotional and idiotic to form rational opinions so they just copy the opinions of the most dominant male in their lives.
Aspect 05 Social class [classism]	Kabir Singh's character was shown as irresponsible rich spoilt brat.
Aspect 06 Ideologism	Trump bullying Ukraine has as much subtlety as Harvey Weinstein friendly conversations with the women he raped.
CATEGORY (IMPLICIT)	
rhetorical questions	How does it feel to be an unbearable self-centred douche-nozzle?
Metaphor	You must be lower than excrement at the bottom of a municipal sewage system.
Simile	They are almost as bad as CNN.
Irony	I am sad when people don't make fun of him for being Arab.
Exaggeration	Congress has taken Indian polity to a new low.

CZECH

TARGET 1	Language: Czech
Individual	Narozdíl od vás Josefa mi paměť ještě slouží.) [Unlike you, Joseph, my memory still serves me.]
Group	Pro vládnoucí nenažrance, kteří cpou peníze do černých děr a vlastních kapes, je to málo, tak chtějí ozebračit důchodce. [It is not enough for the ruling greedy elite, who stuff money into black holes and their own pockets, so they want to impoverish the pensioners.]
Vulgar	Těžko se diskutuje s proruským trotem. [It is hard to argue with a pro-Russian idiot.]
OFFENSIVE TYPE	
Hate speech	Jednou Rusák, vždycky Rusák. To není národnost, to je diagnóza [Once a Russian, always a Russian. It is not a nationality, it is a diagnosis.]
Insult	Je to banda zlodějů [It's a bunch of thieves]
Discredit	Má asi vyoperovaný mozek. [He must have had brain surgery]
Threat	V hloubi duše doufám, že existuje peklo, kde se tenhle hnus bude navěky smažit. [Deep in my heart, I hope there's a hell where this shit will roast forever.]
Aspect 01 racist/xenophobic	Trapní jsou servilní zádelezci Asiatů ... Ještě že je opilej Eman v tahu. [It is the servile backstabbers of the Asians that are embarrassing ... Good thing that the drunk Eman (Zeman, the Czech ex-president) is gone.]

Aspect 02 homophobic	Morální úpadek společnosti právě nastal. LGBT je zhoubá, slepá vývojová větev. [The moral decline of society has just occurred. LGBT is a blight, a blind branch of evolution.]
Aspect 03 physical/mental disabilities/behavioural properties	Někteří lidé jsou fakt jednodušší. [Some people are really rather simple.]
Aspect 04 sexist	No to je hnus zelenej, místo mozku silikonovou kostku. [Well, that's so disgusting, a silicone cube for a brain.]
Aspect 05 social class [classism]	To je ta stejná lůza, která volila Babiše. [This is the very same mob that voted for Babiš.]
Aspect 06 ideologism	A toto je naprosto stejný případ, kolaborantský hlupáku [And this is exactly the same case, you collaborating fool]
CATEGORY (IMPLICIT)	
Rhetorical questions	Problémy jsou všude ... nejste klaun spíš vy? [There are problems everywhere... aren't you the clown actually?]
Metaphor	Co kdyby jsi šel bojovat na Ukrajinu, mudrci? [Why don't you go fight in the Ukraine, wise guy?]
Simile	Ta by byla dobrá jako záchranný člun na Titaniku, ale asi bych se raději topil, než se jí chytíl. [She would be as good as a lifeboat on the Titanic, but I'd probably rather drown than cling on her.]
Irony	Zato v podhradí je to jeden chytrák vedle druhého ... experti bez jakékoli zodpovědnosti [But in the sub-castle (i.e. government) it's one smart guy after another ... experts without any responsibility]
Exaggeration	Co vy jste to za pablba slepého? [What kind of a blind idiot are you?]

LITHUANIAN

TARGET 1	Language: Lithuanian
Individual	Užsidaryk savo srėbtuvę, nes supuvę dantys matosi (Close your mouth not to show your rotten teeth)
Group	Degradų kompanija tame seime sėdi.... gaila tos LIETUVOS. (The company of degraded persons is sitting in that parliament.... it's a pity for LITHUANIA.)
Vulgar	Šūdas is tos bandos, levakų komanda (Shit from that herd, a team of losers)
OFFENSIVE TYPE	
Hate speech	Bijokite pedai, turime sąrašą su visais visais pedofilais Lietuvoje (Fear pedophiles, we have a list with all pedophiles in Lithuania)

316 Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová,
 Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojszczak,
 Renata Povolná, Linas Selmistraitis, Andrius Utka and Dangis Gudelis
 LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications

Insult	eik tu seno kuino subine miegot (go to sleep, you old horse ass)
Discredit	šita ministerija - amžina klapčiukų prieiglauda. (this ministry is an eternal shelter for flappers.)
Threat	Bijokite pedai (Be afraid pedophiles)
Aspect 01 racist/xenophobic	Jau vien del slaviškos pavardės tokį reikia kuo toliau nuo tarnybos pasiųsti. (Just because of his Slavic surname, he should be sent as far away from the service as possible.)
Aspect 02 homophobic	Tegu vyras su vyru gyvena ménulyje, čia jiems ne vieta. (Let a man with a man live on the moon, there is no place for them here.)
Aspect 03 physical/mental disabilities/behavioural properties	Už grotų tuos daunus ir kuo greičiau! (Put those with Down syndrome behind the bars as soon as possible!)
Aspect 04 sexist	Moterytė tai tik kuklus debesėlis prie to atsakingojo. (The woman is just a modest cloud next to the person in charge.)
Aspect 05 social class [classism]	Manau, kada visi bambaliniai išmirsi (I think when will all the vagrants die out)
Aspect 06 ideology	Tu aišku už raudonsniukius balsuosi. (You obviously vote for the redfaced communists)
CATEGORY (IMPLICIT)	
Rhetorical questions	Ar seniai skaityt išmokai? (Has it been a long time since you learned to read?)
Metaphor	Čia vieno asilo galvos tereikia Motiejūno, jis grauna komandą savo idiotiškais sprendimais ir pasiteisimimais. (One donkey's head is all you need here, Motiejūnas, he destroys the team with his idiotic decisions and excuses)
Simile	Moterėlės veidelis kaip Uspaskio, gal kiek gražesnis. (The woman's face is like Uspaski's, maybe a little prettier)
Irony	pedofilai violetinės spalvos bijo kaip velnių kryžiaus. (pedophiles are afraid of the purple color like the devil of the cross)
Exaggeration	Jei jau debilas,tai iki amžiaus galio. (If he is already a moron, then until the end of the days.)

POLISH

TARGET 1	Language: Polish
Individual	Odezwali się fan Rydzyka. [The fan of Rydzyk has just spoken out.]
Group	Odradzam wszystkim to oszuści, chamy i brudas. [I advise all of you against them they are crooks, yahoos, and slobs.]

Vulgar	Niedojobane katechetki które swoją erotyczną frustrację wyładowują strasząc dzieci ogniem piekielnym. [Retarded catechists who unload their erotic frustration by scaring children with hell fire.]
OFFENSIVE TYPE	
Hate speech	Żydzi to jednak najbardziej podła rasa. [Jews are actually the meanest of all human races.]
Insult	Oszolomy chcą koniecznie oczyścić Rosjan ze współdziałania w tej katastrofie. [Nuts absolutely want to cleanse Russian of their complicity in that catastrophe.]
Discredit	Podobała mu się moja bardzo dobra koleżanka chwalił się jej ile to na tych Żydach zarobił. [He was into my very good friend boasting to her about how much money he made on these Jews.]
Threat	Żydka lepiej nie zabieraj głosu. [Better not speak up, you little Jew, a sheeny.]
Aspect 01 racist/xenophobic	Jest żydkiem tylko dla korzystnych układów. [He is a little Jew, a sheeny only for profit]
Aspect 02 homophobic	A Piroga nie lubię nie dlatego że jest pedzem tylko dlatego że jest kiepskim tancerzem. [And I dislike Pirog not because he is a faggot but because he is a poor dancer.]
Aspect 03 physical/mental disabilities/behavioural properties	Ludzie nie posiadający elementarnej wiedzy bądź niedouczeni czy po prostu tepi znajdują się wszędzie. [People without basic knowledge or ignorants or just simply stupid can be found everywhere.]
Aspect 04 sexist	Najgorętsze są Hiszpanki i Brazylijki. [Spanish and Brazilian women are the hottest.]
Aspect 05 social class [classism]	Pierdalone brudasy z biedaszybów. [Fucking bootleg mining slobs.]
Aspect 06 ideologism	UE to ostoja Żydów i Muzułmanów prowadząca do zbydlęcenia byłych wyznawców chrześcijaństwa. [EU is a Jewish and Muslim stronghold that leads to the bastardisation of former Christians.]
CATEGORY (IMPLICIT)	
Rhetorical questions	No i co panie wielki trenerze rozpracowałeś już Rosjan? [So what Sir Great Coach have you already worked out the Russians?]
Metaphor	Szanuję pracę innych nie jestem typem pasożyta. [I appreciate other hard working people I'm not a parasite.]
Simile	Ja też bardzo nie wiem jak zareagować gdy ktoś lata jak Żyd po pustym sklepie. [I also don't know how to react when someone is running like a Jew in an empty shop.]

Irony	Ten Rosjanin się nadaje do pchania karuzeli jak zabraknie prądu a nie do boksu. [This Russian is good for pushing the carrousel when there is no electricity and not for boxing.]
Exaggeration	Gojowie to padlina która ma służyć Żydom. [Gentiles are carcasses who are supposed to serve Jews.]

7 Conclusions

Taking into consideration the three tests described in section 2. above from the first proposal, via the Extended Integrated system, to the Simplified Taxonomy of Offensive Language (SOL) , targeted towards reaching adequate measures to be considered a standard for the notoriously problematic offensive language categorization, we might propose that the most recent SOL taxonomy model of offensive language, implemented and verified on four languages: English, Czech, Lithuanian and Polish, can be postulated to function as a LLOD standard for Offensive Language taxonomy for computational applications. For the future development we might also try to achieve good performance on a number of target languages in parallel (translated) sets by training on a source language with the use of a multilingual transformer model.

What we proposed here is an ontology schema (Lewandowska-Tomaszczyk et al. 2021) that will be presented in terms of the Linguistic Linked Open Data (LLOD) system with instances from multiple languages to share and commonly (re-)use language resources.

Acknowledgments

The present study has been conducted within the Use Case WG 4.1.1. Incivility in Media and Social Media, COST Action CA 18209 *European network for Web-centred linguistic data science Nexus Linguarum*.

References

- Amilevičius, Darius & Mažvydas Petkevičius. 2016. *LITIS v.1*, CLARIN-LT digital library in the Republic of Lithuania. Available at: <http://hdl.handle.net/20.500.11821/11> (accessed 12 March 2022).
- Andersson, Lars-Gunnar & Peter Trudgill. 1990. *Bad Language*. London: Penguin Books Ltd.
- Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Vivian Patti, Francisco Manuel Rangel Pardo, Paolo Rosso & Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki & Saif M.

- Mohammad (eds.) *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 54–63. Stroudsburg, PA: Association for Computational Linguistics.
- Bączkowska, Anna. 2022. Explicit and implicit offensiveness in dialogical film discourse in Brigit Jones films. *International Review of Pragmatics* 14(2). 198–225.
- Bączkowska, Anna, Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedre Valunaite Oleskeviciene & Marcin Trojczak. 2022. *Implicit offensive language taxonomy and its application to automatic extraction and ontology*. Presentation at *LLOD Approaches to language data research and management*, Vilnius, 21–22 September 2022, Lithuania.
- Brenner, Jennifer L. 2002. True threats: A more appropriate standard for analyzing First Amendment protection and free speech when violence is perpetrated over the Internet. *North Dakota Law Review* 78(4). 753–784.
- Grice, H. Paul. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Lakoff, George. 1987. Cognitive models and prototype theory. In Ulric Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, 63–100. Cambridge: Cambridge University Press.
- Lewandowska-Tomaszczyk, Barbara. 2012. Approximative spaces and the tolerance threshold in communication. *International Journal of Cognitive Linguistics* 2(2). 1–19.
- Landis J. Richard & Garry G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174.
- Lewandowska-Tomaszczyk, Barbara. 2017. Conflict radicalization and emotions in English and Polish online discourses on immigration and refugees. In Stephen M. Croucher, Barbara Lewandowska-Tomaszczyk & Paul A. Wilson (eds.), *Conflict, mediated message and group dynamics: intersections of communication*, 1–24. New York: Rowman & Littlefield.
- Lewandowska-Tomaszczyk, Barbara. 2022. A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies* 10(3). 213–227.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović & Giedre Valunaite Oleskeviciene. 2021. Lod-connected offensive language ontology and tagset enrichment. In Sara Carvalho & Renato Rocha Souza (eds.), *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*, 135–150. CEUR Workshop Proceedings.
- Lewandowska-Tomaszczyk, Barbara, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene & Slavko Žitnik. 2023. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 23(1). 7–48.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Chaya Liebeskind, Giedre Valunaite Oleskevicienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, Olga Dontcheva-Navratilova, Agnieszka Borowiak, Kristina Despot & Jelena Mitrović. (accepted) Annotation scheme and evaluation: The case of OFFENSIVE language. *Rasprave*.
- Liu, Ping, Wen Li & Liang Zou. 2019. nlpUP at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, & Saif M. Mohammad (eds), *Proceedings of the 13th international workshop on semantic evaluation*, 87–91. Stroudsburg, PA: Association for Computational Linguistics.
- Ptaszyński, Michał & Fumito Masui. 2018. *Automatic Cyberbullying Detection: Emerging Research and Opportunities*. Hershey, PA: IGI Global Publishing.

- Ptaszyński, Michał, Agata Pieciurkiewicz & Paweł Dyba. 2019. *Results of the Poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter*. Warsaw: Institute of Computer Sciences. Polish Academy of Sciences.
- Searle, John. 1975. Indirect Speech Acts. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, 59–82. New York: Academic Press.
- Troszyński, Marek & Aleksander Wawer. 2017. Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych. *Przegląd Socjologii Jakościowej* XIII(2). 62–80.
- Wittgenstein, Ludwig. 1953. *Philosophical investigations*. New York: Macmillan.
- Zadeh, Lofti. 1964. Fuzzy sets. *Information and Control* 8(3). 338–353.
- Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, & Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 1415–1420. Stroudsburg, PA: Association for Computational Linguistics.
- Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, & Ritesh Kumar. 2019b. *Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)*. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, & Saif M. Mohammad (eds.), *Proceedings of the 13th international workshop on semantic evaluation*, 75–86. Stroudsburg, PA: Association for Computational Linguistics.

Datasets and tools

- 25 English offensive language and hate speech data sets (for the itemized list cf. Lewandowska-Tomaszczyk et al. 2023 Appendix 1.)
- Sketch Engine corpus Czech Offensive Language. Available at:
https://ske.fi.muni.cz/#dashboard?corpname=user%2Fss_259%2Fczech_offensive_language
(accessed 5 April 2022)
- Amilevičius, Darius & Mažvydas Petkevičius, M., 2016, *LITIS v.1*, CLARIN_LT digital library in the Republic of Lithuania. Available at: <http://hdl.handle.net/20.500.11821/11>. (accessed 12 March 2022).
- Troszczyński, Marek & Aleksander Wawer. 2017. Available at: <http://zil.ipipan.waw.pl/HateSpeech>
(accessed 1 March 2022)
- Ptaszyński, Michał & Fumito Masui. 2018. Available at: ptaszynski/cyberbullying-Polish (accessed 10 April 2022)
- Ptaszyński, Michał et al. 2019. Available at: ptaszynski/cyberbullying-Polish (accessed 10 April 2022)
- Annotation INCEpTION platform. Available at: <https://inception-project.github.io/> (accessed 20 February 2022)
- Sketch Engine webcorpus of English. Available at: <https://www.sketchengine.eu/ententen-eng-lish-corpus> (accessed February 2022)
- Marco A. Stranisci, Simona Frenda, Mirko Lai, Oscar Araque, Alessandra T. Cignarella, Valerio Basile, Viviana Patti & Cristina Bosco. 2022. O-Dang! The ontology of dangerous speech messages. In Ilan Kernerman, Sara Carvalho, Carlos A. Iglesias & Rachele Sprugnoli (eds.) *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, 2–8. Paris: European Language Resources Association.

About the Authors

Barbara Lewandowska-Tomaszczyk is Professor Ordinarius Dr Habil. in Linguistics and English Language at the Department of Language and Communication at the University of Applied Sciences in Konin (Poland). Her research focuses on cognitive semantics and pragmatics of language contrasts, corpus linguistics and their applications in translation studies, lexicography and online discourse analysis. She is invited to read papers at international conferences and to lecture and conduct seminars at universities. She publishes extensively, supervises dissertations and also organizes international conferences and workshops.

Address

Department of Language and Communication
University of Applied Sciences in Konin
1, Przyjazni str.
62 510 Konin, Poland

e-mail: barbara.lewandowska-tomaszczyk@konin.edu.pl
ORCID: 0000-0002-6836-3321

Anna Baczkowska, Dr Habil. Prof. UG, holds MA in English Philology, which she received from Adam Mickiewicz University in Poznan, as well as PhD in linguistics and D.Litt. in English Linguistics, which she received from the University of Lodz. Her research interests revolve around translation studies (film subtitles), cognitive semantics, corpus and computational linguistics, and discourse studies (media discourse). She has guest lectures in Italy, Spain, Portugal, UK, Norway, Kazakhstan and Slovakia, and she has also conducted her research during her scientific stays in Ireland, Iceland, Norway, Austria and Luxembourg.

Address

Institute of English and American Studies, University of Gdańsk
Wita Stwosza 51
80-308 Gdańsk, Poland

e-mail: anna.baczkowska@ug.edu.pl
ORCID: 0000-0002-0147-2718

Olga Dontcheva- Navrátilová is Associate Professor of English Linguistics at the Faculty of Education, Masaryk University, Czech Republic. Her research interests include English for academic and specific purposes and political discourse. She has published the books *Analysing Genre: The Colony Text of UNESCO Resolutions* (2009), *Coherence in Political Speeches* (2011) and co-authored *Persuasion in Specialised Discourses* (2020). She is co-editor of the journal *Discourse and Interaction*.

Address

Department of English Language and Literature, Faculty of Education
Masaryk University
Poríčí 9, Brno 602 00
Czech Republic

e-mail: navratilova@ped.muni.cz
ORCID: 0000-0002-0378-7975

- 322 Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová,
Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojszczak,
Renata Povolná, Linas Selmistraitis, Andrius Utka and Dangis Gudelis
LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications

Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and data mining. Especially, her scientific interests include Semantic Similarity, Language Technology for Cultural Heritage, Morphologically rich languages (MRL), Multi-word Expressions (MWEs), Information Retrieval (IR), and Text Classification (TC). Much of her recent work has been focusing on analysing offensive language. She has published a variety of studies and a few of her articles are under review or in preparation. She is a member of several international research actions funded by the EU.

Address

Jerusalem College of Technology, Department of Computer Science
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel

e-mail: liebchaya@gmail.com
ORCID: <http://orcid.org/0000-0003-0476-3796>

Giedrė Valūnaitė Oleškevičienė is Vice-Dean for Scientific Research of the Faculty of Public Governance and Business and a professor at the Institute of Humanities, Mykolas Romeris University. Her scientific interests in humanities include discourse analysis, professional English, legal English, linguistics and translation research, while in the domain of social sciences, her scientific interests include social research methodology, modern education, philosophical issues, creativity development in modern education system, and second language teaching and learning. The researcher coordinated international research projects funded by the EU, publishes scientific articles, participates as a presenter in scientific conferences.

Address

Faculty of Public Governance and Business, Mykolas Romeris University
20 Ateities St.,
LT-08303 Vilnius, Lithuania

e-mail: gvalunaite@mruni.eu
ORCID: <https://orcid.org/0000-0001-5688-2469>

Slavko Žitnik is Assistant Professor and Vice-dean for Education at the University of Ljubljana, Faculty for Computer and Information Science. His research focuses on natural language processing, information extraction, databases, semantic technologies, and information systems. He is actively collaborating with Université Paris 1 Sorbonne, Harvard University, University of South Florida, and University of Belgrade. He is engaged in multiple research and professional projects. As a chairman of *Slovenian Language Technologies Society* he is organizing lectures related to language technologies and provides grants to students to visit summer schools. He is also Chairman of the Slovene Society *INFORMATIKA*, and organizes national conferences on informatics and is editor of a scientific journal.

Address

Faculty for Computer and Information Science, University of Ljubljana
Večna pot 113
SI-1000 Ljubljana, Slovenia

e-mail: slavko.zitnik@fri.uni-lj.si
ORCID: 0000-0003-3452-1106

Marcin Trojczak holds PhD in Linguistics and MA in Philosophy. He is Assistant Professor at the University of Applied Sciences in Konin (Poland). He is also actively cooperating with University of Łódź and University of Economics and Human Sciences in Warsaw. His research interests include metaphorical conceptualisations of mental and emotional processes, the impact of translation technologies on translation education, normativity and genericity in language and cognition, as well as offensive language.

Address

Department of Language and Communication, University of Applied Sciences in Konin
1, Przyjazni str.
62 510 Konin, Poland

e-mail: marcintrk@gmail.com
ORCID: 0000-0003-3062-9670

Renata Povolná is Associate Professor of English Linguistics at the Faculty of Education, Masaryk University, Czech Republic. Her research lies in the area of discourse analysis, pragmatics and conversation analysis. She has published the books *Spatial and Temporal Adverbials in English Authentic Face-to-Face Conversation* (2003), *Interactive Discourse Markers in Spoken English* (2010) and co-authored *Persuasion in Specialised Discourses* (2020). She is co-editor of the journal *Discourse and Interaction*.

Address

Department of English Language and Literature, Faculty of Education, Masaryk University
Porčíč 9, Brno 602 00
Czech Republic

e-mail: povolna@ped.muni.cz
ORCID: 0009-0006-4392-2832

Linas Selmistraitis has over 24 years of experience in higher education specifically in developing and implementing quality assurance systems for higher educational institutions. He earned his PhD in Humanities. Currently Professor Dr Linas Selmistraitis holds the position of Vice-Dean for Studies at Faculty of Human and Social Studies at Mykolas Romeris University and the position of Professor at Institute of Humanities at Mykolas Romeris University. His interest in research are semantics, morphology, cognitive linguistics. He publishes research articles and gives presentations at conferences.

Address

Faculty of Human and Social Studies
Mykolas Romeris University
20 Ateities St.
LT-08303 Vilnius, Lithuania

e-mail: selmistraitis@mruni.eu
ORCID: 0000-0002-8612-3276

- 324 Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Olga Dontcheva-Navrátilová,
Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojszczak,
Renata Povolná, Linas Selmistraitis, Andrius Utka and Dangis Gudelis
LLOD schema for Simplified Offensive Language Taxonomy in multilingual detection and applications

Andrius Utka is Associate Professor at the Department of Lithuanian Studies and a senior researcher at the Institute of Digital Resources and Interdisciplinary Research (SITTI), Vytautas Magnus university (Kaunas). He defended the doctoral dissertation *Statistical Identification of Text Functions* in 2004 (VMU, Kaunas). He was the head of Centre of Computation Linguistics in 2010-2022. He coordinated a number of national and international research projects. His research interests: statistical text analysis, language resources, computer-assisted translation, automatic summarisation, terminology extraction, and the language of disinformation.

Address

Institute of Digital Resources and Interdisciplinary Research
Vytautas Magnus University
23-216 V. Putvinskio St.
LT-44243 Kaunas, Lithuania

e-mail: andrius.utka@vdu.lt
ORCID: 0000-0001-5212-4310

Dangis Gudelis is a professor at Mykolas Romeris University, specializing in public administration and governance. He earned his PhD in Social Sciences, focusing on performance measurement in Lithuanian municipalities. Gudelis has led and contributed to various national and international research projects, particularly in public governance and public policy. His current research interests include applications of big data and AI technologies in the public sector. He is a prolific writer, with numerous publications in scientific journals and presentations at conferences. He teaches courses at both undergraduate and graduate levels. Additionally, he has played a role in policy analysis and consultancy, advising governmental and non-governmental organizations on strategic development and public sector innovation.

Address

Faculty of Public Governance and Business
Mykolas Romeris University
20 Ateities St.
LT-08303 Vilnius, Lithuania

e-mail: dgudel@mruni.eu
ORCID: 0000-0002-9004-8029

Chaya Liebeskind
Jerusalem College of Technology

Natalia Vanetik
Shamoon College of Engineering

Marina Litvak
Shamoon College of Engineering

HEBREW OFFENSIVE LANGUAGE TAXONOMY AND DATASET

Abstract

This paper introduces a streamlined taxonomy for categorizing offensive language in Hebrew, addressing a gap in the literature that has, until now, largely focused on Indo-European languages. Our taxonomy divides offensive language into seven levels (six explicit and one implicit level). We based our work on the simplified offensive language (SOL) taxonomy introduced in (Lewandowska-Tomaszczyk et al. 2021a) hoping that our adjustment of SOL to the Hebrew language will be capable of reflecting the unique linguistic and cultural nuances of Hebrew. The study involves both linguistic and cultural analysis beyond Natural Language Processing (NLP). We employed manual linguistic analysis to understand the nuances of offensive language in Hebrew.

An accompanying dataset, gathered on Twitter and manually curated by human annotators, is described in detail. This dataset was constructed to both validate the taxonomy and serve as a foundation for future research on offensive language detection and analysis in Hebrew. Preliminary analysis of the dataset reveals intriguing patterns and distributions, underscoring the complexity and specificity of offensive expressions in the Hebrew language.

The aim of our work is to capture the complexity and specificity of offensive expressions in Hebrew beyond what automated NLP methods alone can provide. Our findings highlight the significance of considering linguistic and cultural variations when researching and correcting abusive language online. We believe that our streamlined taxonomy and associated dataset will be crucial in improving research in Hebrew language sociocultural studies, natural language processing, and offensive language detection. Our study also makes a substantial contribution to the study of low-resource languages and can be used as a model for future research on other languages.

Keywords

offensive language, low-resource languages, taxonomy, Hebrew offensive language dataset

1 Introduction

A definition of offensive language often depends on cultural, social, and personal perspectives. What may be offensive in one culture or context may not be in another. However, in a general sense, offensive language can be categorized as any form of communication that may upset, offend, or cause discomfort to individuals or groups. In pragmatic literature, offense is defined as face-attacks, often accompanied by offense perception by the target of offense (Haugh and Sinkeviciute 2019; LewandowskaTomaszczyk 2023).

Offensive language can be categorized into different types. Using words or phrases that are insulting to certain people or groups, such as slurs, insults, or hate speech is known as explicit offensive language. Implicit offensive language, on the other hand, is more subtle, using insinuations, euphemisms, or coded language to convey prejudice or bias without overtly stating it (Caselli et al. 2020). Explicit offensive language is easily recognizable, while implicit offensive language requires deeper cultural and contextual understanding to be identified. Within these two categories, there are multiple varieties of offensive language (Kogilavani et al. 2021; Lewandowska-Tomaszczyk 2023), such as vulgarity (which includes swear words or slurs), hate speech (language that discriminates against, threatens, or incites violence), sexual harassment (unsolicited sexual remarks, threats, or graphic descriptions), threats (language meant to intimidate or harm someone), disparaging remarks (language that belittles or mocks individuals or groups), cultural, religious, or historical insensitivity, trolling (deliberate attempts to upset or provoke), and more.

One of the most common sources of offensive language is social media. In the EU, surveys and reports focusing on young people show an increase in hate speech and related crimes based on religious convictions, ethnicity, sexual orientation, or gender. According to Pan-European anti-racism network (2022) and Wine (2016), 80% of respondents have experienced online hate speech and 40% have felt attacked or threatened.

A survey by the Israel Internet Association published in December 2022 (Bright 2022) and conducted by the “Mindpool” research institute (Technologies 2023) reveals that quite a few Israelis were exposed to online harm on the internet, including hatred, harassment, exposing of intimate details, and violent discourse. The results of the survey also demonstrated how damaging discourse has an impact on individuals using the internet and leads many to forgo conversation and online opinion expression completely. 22% of Israelis who said they had directly encountered profanity and abuse on social media during the previous year, while 18% said they had personally encountered surveillance or harassment during the same period (WALLA! TECH 2022).

There are laws prohibiting false information, hate speech, and other damaging communication in many countries. Companies and platforms that improperly regulate material risk legal repercussions as well as reputational harm if they are known for hosting or failing to stop the spread of harmful content. Users are more inclined to offer their thoughts and engage in deep conversations when inappropriate language is reduced since

it promotes more positive and productive interactions. Identifying and controlling such language can help save people and communities from unwarranted injury and suffering. While detecting inflammatory language is vital, it is also necessary to find a balance between removing damaging information and protecting the right to free speech. False positives, which classify non-offensive language incorrectly as offensive, can result in censorship, while false negatives, which fail to identify truly offensive content, can allow damaging content to pass.

Popular search engines like Google, Yahoo, and Bing, as well as online videosharing sites like YouTube, have safe search options that filter out dangerous information including insulting language, racial and sexual harassment, bullying, and statements of hatred. Users are required to follow the rules set forth by many online businesses. They use regular expressions and blacklists to detect offensive language and then remove the offending messages. However, the work of maintaining a list of derogatory phrases for content filters has become inadequate due to the rise in user-created internet material. As a result, there is an increasing trend towards using Natural Language Processing (NLP) approaches to identify offensive language.

The clear majority of the offensive language detection studies (explored below in Section 2) deal with English partially because most available annotated datasets contain English data (Zampieri et al. 2019b). Despite the great international effort, many low-resource languages such as Hebrew get much less attention than others.

Offensive language detection varies in difficulty across languages due to linguistic, cultural, and contextual differences. Idioms, slang, and double meanings that might be offensive in one language may not have direct translations in another, making automated detection difficult. Additionally, the fluidity of language means that words can adopt new, potentially offensive meanings over time.

Motivated by the lack of relevant resources and research for Hebrew, we decided to fill this gap by adopting a new simplified taxonomy for the offensive language in Hebrew and creating, annotating, and analysing a new annotated Hebrew dataset that covers various forms of explicit offensive language.

This paper is organized as follows. Section 2 covers the related work. Section 3 describes the simplified taxonomy of offensive language for Hebrew. Section 4 explains the collection and construction of the offensive language dataset in Hebrew, and Section 5 describes how the annotation of this dataset was carried out. Finally, Section 6 explains conclusions drawn from our work.

2 Related work

2.1 Offensive language taxonomy

In Zampieri et al. (2019a, b) the authors proposed a three-level hierarchy for categorizing offensive language, especially in the context of social media content. This hierarchy is designed to provide a structured way to differentiate between different types and severities of offensive language. This is a three-level taxonomy that separates the text into offensive and not offensive (level A), then to targeted insults or threats (level B), and then differentiates between target types (level C) – groups, individuals, or others.

The Nexus Linguarum¹ Working Group conducted research of offensive language typology (Lewandowska-Tomaszczyk et al. 2021a). The group evaluated existing datasets and compared the recently applied tagging schemas of offensive language. The authors also presented a hierarchical structure (ontology) that categorizes offensive language into multiple tiers and sub-levels. In contrast to the three-level hierarchy of offensive language introduced in Zampieri et al. (2019a, 2019b), the proposed ontology is composed of two basic levels (I and II), and four sub-levels (A, B, C, D) within Level I. Level I distinguishes lexical items that are offensive from those that are not (Level A: offensive vs. non-offensive). Level B distinguishes between the offensive items that are targeted at some addressee and those without identifiable addressee (targeted vs. non-targeted). Targeted offensive items are further divided into either implicit or explicit cases of offensive language at Level C (implicit vs. explicit language). While implicitness may be expressed by sarcasm and irony, explicitness is usually a straightforward form of verbal attack. Classes of explicitly targeted categories of offense are further classified by different kinds of internal or external targets. The authors also provided a conceptual framework embodying the representation of the offensive language structure and investigated computational models for detecting offensive language, using the proposed ontology.

In Lewandowska-Tomaszczyk et al. (2022), the combined schema for explicit and implicit offensive language was tested using English datasets via the INCEpTION annotation platform. While some results aligned with the proposed framework, challenges arose in terms of consensus among annotators in distinguishing specific categories. The need for streamlining methods was acknowledged for subsequent annotation activities and improved identification of offensive language.

The work in Lewandowska-Tomaszczyk (2023) introduced a unified approach that addresses both straightforward and implied forms of offensive language. This approach was built upon linguistic concepts such as Grice's implicitness categories (Grice 1990). She identified eight principal implicit categories of offensive language and instances where offensive language holds attributes of both apparent and implied meanings. Their proposition included a hierarchical structure that incorporates offensive language

¹ <https://nexuslinguarum.eu/>

categories, facets, and levels, merging directness and impliedness. The conclusion made by the researchers is that there is a need to simplify the offensive language taxonomy.

2.2 Hebrew offensive language detection

Despite multiple works on automated offensive language detection, including early unsupervised lexicon-based approaches (Tulkens et al. 2016), traditional supervised approaches (Davidson et al. 2017), and recent approaches based on deep neural networks (Zampieri et al. 2019b) and transformer models (Liu, Li and Zou 2019; Ranasinghe, Zampieri and Hettiarachchi 2019), the clear majority of the offensive detection studies deal with English. However, since social media has become a popular communication tool worldwide, many researchers started to cope with the “multilingual challenge” by developing multilingual methodologies and annotated corpora in multiple languages. For example, such languages as Arabic (Mohaouchane, Mourhir and Nikolov 2019), Dutch (Tulkens et al. 2016), French (Chiril et al. 2019), Turkish (Çöltekin 2020), Danish (Sigurbergsson and Derczynski 2019), Greek (Pitenis, Zampieri and Ranasinghe 2020), Italian (Poletto et al. 2017), Portuguese (Fortuna et al. 2019), Slovene (Fišer, Erjavec and Ljubešić 2017), and Dravidian (Yasaswini et al. 2021) were explored for the task of offensive content identification. Also, multilingual methods and datasets for offensive language detection were proposed. Hate Speech and Offensive Content Identification (HASOC) 2019 (Mandl et al. 2019) and 2020 (Mandl et al. 2020) were dedicated to evaluating technology for finding offensive language and hate speech in multiple low-resource languages.

Despite the great international effort, many low-resource languages got much less attention than others. For example, only a few works proposed solutions for Hebrew: Liebeskind and Liebeskind (2018) developed a Hebrew corpus of user comments annotated for abusive language and investigated highly sparse n-grams representations as well as denser character n-grams representations for comment abuse classification; then Litvak et al. (2021) introduced the dataset containing annotated Facebook comments written in Hebrew and a publicly available system for offensive text detection using multiple machine learning algorithms; and finally Litvak et al. (2022) presented a union of these two datasets and performed monolingual, cross-lingual, and multilingual experiments with multiple supervised models and text representations for a task of offensive language detection on this dataset. The cross-lingual experiments were motivated by a big portion of low-resource languages in general and a lack of resources for Hebrew in particular. The authors took advantage of rich resources in Arabic, which is a similar language to Hebrew (both belong to the Semitic family of languages). To represent the meaning of a text in different languages correctly, the authors used multilingual sentence embeddings and multilingual pre-trained models, containing both languages.

Recently, Hamad et al. (2023) introduced a new offensive language corpus in Hebrew containing 15,881 texts retrieved from Twitter. The texts were labelled by Arabic-Hebrew

bilingual speakers into one or more of the five available classes, namely abuse, hate, violence, pornography, or non-offensive. Fine-tuning of pre-trained Hebrew LLMs showed that the proposed dataset is useful for the detection of offensive language in Hebrew.

3 Hebrew simplified taxonomy of offensive language

An offensive language taxonomy is important because it provides structured categorization for diverse offensive content and by doing so it aids automated systems in content moderation and response. Such a taxonomy offers a basic structure that not only organizes the complex terrain of online interactions but also serves as a tool to improve the safety and utility of digital platforms since it improves knowledge of online behavior. To define a taxonomy of offensive language for Hebrew, we reformulate and extend the taxonomy proposed by Lewandowska-Tomaszczyk et al. (2023) that in its turn is an extension of the one proposed by Zampieri et al. (2019a, b).

This taxonomy has seven basic levels (or steps) of analysis – levels 1 to 7. The first six levels refer to explicit categories, and the seventh level refers to implicit categories. Explicit offensive language is unambiguous and straightforward, and it is clear to most people that the content is meant to offend, insult, or harm (for example, direct slurs, clear threats, openly racist or sexist remarks, etc.) In implicit offensive language, the offensive content is more subtle or even hidden, and it may require context or cultural knowledge to understand the offensive intent (such as dog whistles or insinuations.) Below we describe the levels in detail.

Level 1 separates texts into *non-offensive* (texts that can be neutral, positive, or can be about any topic that does not contain any offensive content) and *offensive* (texts that contain offensive language or a topic that can be considered offensive in certain contexts.)

Level 2 indicates the target of an offensive text. The target can be an *individual*, a *group*, an *individual w.r.t. the group* (where stereotypes about a group are used to insult someone) or other (where none of the previous cases apply).

Level 3 indicates whether or not the target of an offensive text is explicitly addressed in the text (in this case we say that the target is *present*) and it is clear to whom the offense is directed, or *absent* (the target is not a part of the conversation).

Level 4 indicates whether the offensive language is *vulgar* (i.e., uses vulgar words, expressions, and curses) or *non-vulgar* (no vulgar words or curses are used).

Level 5 specifies the offense strength and it contains four categories, from less severe to more severe, as follows.

An *insult* is a remark or action intended to offend or hurt someone either by demeaning them, expressing contempt, or mocking them. Insults do not use group stereotypes.

Hate speech is a form of offensive language that discriminates against, denigrates, or incites violence or prejudicial actions against a person or a group based on attributes such

as race, religion, ethnic origin, sexual orientation, disability, or gender (i.e., it utilizes group stereotypes.)

Discrediting someone means to harm their good reputation or to cause others to doubt their credibility, abilities, or trustworthiness through spreading false information, highlighting real or perceived flaws, or casting doubt on their intentions.

A *threat* is an expression of an intention to inflict pain, injury, damage, or another bad consequence on someone else.

Level 6 describes different aspects of explicit offensive language, as specified below. A single text may exhibit one or more aspects.

Racism – expressions that belittle, mock, stereotype, or show prejudice against individuals or groups based on their race or ethnicity.

Xenofobia – expressions that show a strong dislike, fear, or prejudice against people from different countries, cultures, or ethnic backgrounds.

Homophobia – showing prejudice, disdain, fear, or hostility towards individuals based on their sexual orientation, especially towards those who identify as gay, lesbian, or bisexual.

Sexism – utterances that belittle, mock, stereotype, or show prejudice against individuals based on their gender or sex.

Profanity – expressions that show irreverence or disrespect towards religious beliefs, figures, objects, or practices.

Ageism – utterances that belittle, mock, stereotype, or show prejudice against individuals based on their age.

Ableism – utterances that belittle, mock, stereotype, or show prejudice against individuals based on their disabilities or perceived abilities.

Classism – expressions that belittle, mock, stereotype, or show prejudice against individuals based on their social, economic, or educational class.

Ideologism – prejudice, discrimination, or bias based on ideology or political views.

Other – any aspect that does not fall into any of the above categories.

Level 7 describes different aspects of implicit offensive language, as specified below. A single text may exhibit one or more aspects.

Indirectness – the use of expressions or statements that imply or suggest something negative or derogatory without saying it outright.

Rhetorical questions – questions that are posed without expecting an answer but are meant to make a point, often in a sarcastic, condescending, or belittling manner.

Simile – a figure of speech that compares two different things using “like” or “as” to highlight a particular quality or attribute; similes can be used to belittle, mock, or demean a person or a group by drawing derogatory or unflattering comparisons.

Metaphor – wherein one thing is seen in terms of another one, according to the definition of conceptual metaphors offered by Cognitive Linguistics (Lakoff and Johnson, 1980)

It may provide (or obscure) clarity or identify hidden similarities between two different ideas.; it can be used to belittle, mock, demean, or stereotype a person, group, or concept by drawing negative or unflattering comparisons.

Irony – saying something in which the intended meaning is opposite to the literal or usual meaning.

Understatement – a figure of speech in which a speaker intentionally downplays or minimizes the significance of something, making it seem less important or severe than it is.

Overstatement (hyperbole) – exaggeration that is used to mock, belittle, or misrepresent someone or something in a derisive manner.

Figure 1 depicts levels 1–6 of this taxonomy that refer to explicit offensive language. It should be noted that it can be harder to identify inappropriate content that is not stated explicitly. Implicitly offensive language makes use of innuendos, euphemisms, or coded terminology that may not be immediately understood as offensive unless the underlying context or meaning is understood. For instance, a phrase may be neutral in one cultural or social setting but insulting in another. It can be challenging to comprehend such subtleties. For that reason, in this paper, we focus on detecting and classifying explicit offensive language in Hebrew. Therefore, we focus solely on levels 1–6 of the above taxonomy for Hebrew. We do not address the analysis of implicit offensive language (level 7) due to its high ambiguity and cultural dependency, heavily affecting the annotation process. This taxonomy, translated to Hebrew, is shown in Figure 2. Translation of implicit offensive terms belonging to level 7 is given in Table 1.

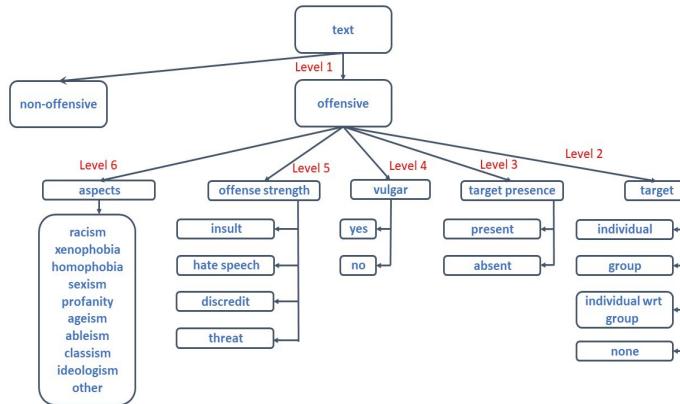


Figure 1: Simplified offensive language taxonomy for English, levels 1–6

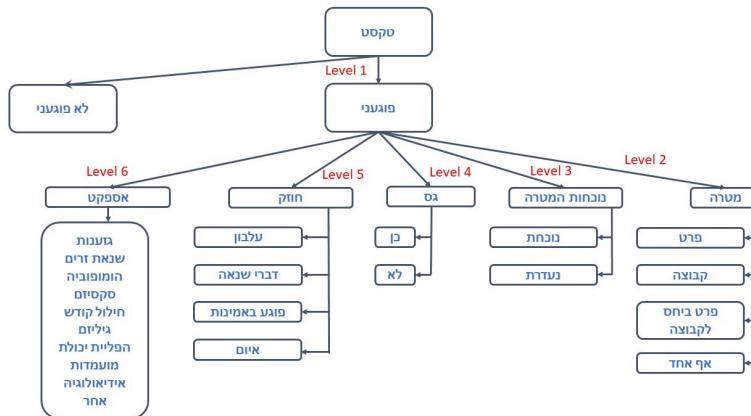


Figure 2: Simplified offensive language taxonomy for Hebrew

Table 1: Hebrew translation of implicit offensive language aspects (level 7)

Implicit aspect	Hebrew translation
indirectness	[akifut]
rhetorical questions	[sheelot retoriyot]
simile	[dimiyon]
metaphor	מטפורה או שם מושאל [metufarach o shem mush'al]
irony	[ironiya]
understatement	[lashon hama'ateh]
overstatement	[te'ur mugzam]

Translating offensive language taxonomy is a delicate task that requires careful consideration for several reasons because different cultures have different norms and taboos. A term that is considered innocuous in one culture might be highly offensive in another. Failing to take this into account can lead to misunderstandings and unintended offenses. Direct translations can sometimes miss the nuance or weight of the original term. For example, there are two main options for translating the word “vulgar” into Hebrew. In Hebrew, the words גז [‘gas’] and גז רוח [‘gas ruach’] seem similar and are related, but they refer to different things and have different meanings. The word גז [‘gas’] mostly refers to something’s size, texture, or quality but can also refer to a choice of words. For instance, when we say an expression is גז [‘gas’], we mean that the expression is vulgar, immodest, blunt, or impudent. The same word can describe a fabric as rough and not soft to the touch, and in another context, we can describe a person as גז [‘gas’] meaning they lack sensitivity

or finesse. The phrase *גָּז רֹעַח* [‘gas ruach’] mainly refers to a person’s personality or behaviour. An expression that is *גָּז רֹעַח* [‘gas ruach’] means that the utterance is arrogant, not gentle, bold-faced, or insolent. The phrase refers to the way one interacts with their social environment, and it is less severe than the word *גָּז* [‘gas’]. Therefore we have chosen the first option for translating this term.

4 Hebrew dataset of offensive language

The purpose of this section is to describe the construction of a dataset that the current study used to annotate documents based on the proposed taxonomy, and that we intend to use in the future for classifying offensive content using machine learning. Our dataset attempts to include all types of offensive language, not just the most prevalent ones. For this purpose, we have developed a dataset construction methodology based on statistical approaches with minimal labelling that can be adapted to any language.

In this section, we outline the stages of our methodology for the creation of the dataset. Using data mining techniques, we then analyse the constructed offensive language dataset.

4.1 Dataset construction methodology

The process for generating our dataset involves the following steps: Initially, a small list of offensive words is extracted, which is subsequently augmented by the application of statistical analysis to examine patterns of co-occurrence. Next, the extended list is categorized according to the predefined categories proposed by the taxonomy. Subsequently, a set of documents is extracted for each category, resulting in a dataset that encompasses diverse forms of offensive language. We provide a comprehensive explanation of each step in subsections below.

4.1.1 Semi-automatic collection of a list of offensive words

The procedure outlined by Liebeskind and Liebeskind (2018) for creating a list of offensive terms was implemented in our study. The list generation procedure began with a list of a few dozen (67) offensive terms. We supplemented this list with a statistical measure of word co-occurrence. Assuming that words that appear frequently together are topically related (Schütze and Pedersen 1997), we retrieved the k (100) most similar word for each offensive term in the first list using the Dice coefficient (Smadja, McKeown and Hatzivassiloglou 1996) and a large unannotated corpus (over than 5 million Facebook comments). The Dice coefficient normalizes the frequency of co-occurrence, or the intersection of the two terms’ document sets, by dividing it by the sum of the individual terms’ frequencies and multiplying it by two, yielding a measure between 0 and 1, with 1 indicating complete co-occurrence:

$$\begin{aligned} \text{Dice}(x, y) &= \frac{|C_x \cap C_y|}{|C_x| + |C_y|} \\ &= \frac{2 * \text{count}(c \in C \mid x \in c, y \in c)}{\text{count}(c \in C \mid x \in c) + \text{count}(c \in C \mid y \in c)} \end{aligned}$$

where C_x is the comment set where term x appears and C_y is the comment set where term y appears. Then, from these candidate lists, 683 offensive terms were manually chosen.

4.1.2. Classification of the word list into taxonomy categories

Following the compilation of the vocabulary containing offensive terms, our investigation proceeded to the categorization of these words into several categories. The categories we have established are the aspects of our taxonomy at level 6. Several instances of such a classification are provided in Table 2.

Please note, that our taxonomy does not have the xenophobia category, because we could not find any example of using xenophobic expressions in our collection. We assume that this phenomenon could be explained by the multicultural specificity of the Israeli society (Tova Hartman 2022). We assign all texts with dislike of or prejudice against people from other countries to racism.

Table 2: Instances of offensive word categorization

Category	Offensive words
Racism	ג'זן (gizan) racist, כושי (kushi) black, יהודונאצי (yehudonatzi) Judeo-Nazi
Homophobia	הומו (homo) Homosexual, קוקסינל (coccinel) drag queen, טראנסית (transit) transgender (female)
Sexism	אנאש (anas) rapist, זונה (zona) prostitute, פרוצזה (protza) whore
Profanity	ארור (arur) cursed, ספר (kupher) atheist, שטן (satan) devil
Ageism	תרח (trach) old man, מתרשלך (mtushelach) antique, זקן (zaken) old
Ableism	מטומטם (metumtam) stupid, מפער (mephager) intellectual disability, דפיק (daphuk) not right in the head
Classism	אנאלאפabetית (analphbetim) illiterate person, אינפאנטיליסטי (infantilim) infantile, תפרני (taphranim) down-and-out
Ideologism	שמאלי (smolan) left-hand side, פאשיסטי (fashist) fascist, ביביסטי (bibist) Bibi supporters
Other	הזרוי (hazuy) imaginary, גנב (ganav) thief, בזוי (bazuy) worthy of contempt

We have adopted a permissive classification approach to allow for the inclusion of offensive expressions in several categories. It is deemed satisfactory if such expressions possess even a minimal associative connection with a given category. For instance, the

word “!דָפְעַק” (daphuk) meaning “not right in the head”, which was classified as an ableism, was also placed in the ‘other’ category because it can also be used to describe something terrible or faulty.

4.1.3. Retrieving documents containing the offensive words

To construct our Hebrew offensive dataset, we utilized the classified word list and employed the Twitter API² to identify offensive tweets containing those words (corresponding to each category of offensive language.) Subsequently, the tweets were collected and the textual content of each tweet was extracted.

Table 3 presents the number of documents retrieved for each category, as well as the number of offensive words employed in the search for offensive tweets within each category.

4.2 Data analysis

Word embedding is a computational technique applied in the field of natural language processing (NLP) to encode words as numerical vectors, enabling computers to process and analyse textual data. Word embedding techniques are employed to assign high-dimensional vectors to words or phrases within a specific vocabulary. These vectors are learned from a corpus and aim to capture both semantic and syntactic relationships among the linguistic units.

Table 3: The number of retrieved documents and the number of offensive words used for retrieval in each category

Category	No. of offensive words	No. of documents
Racism	55	113040
Homophobia	37	43745
Sexism	155	1573764
Profanity	23	1432455
Ageism	8	1386664
Ableism	82	1365564
Classism	71	582791
Ideologism	110	219428
Other	317	1101039

² <https://developer.twitter.com/en>

Following Lewandowska-Tomaszczyk et al. (2021b), we employed two word embedding methods: Word2Vec (Mikolov et al. 2013) and FastText (Bojanowski et al. 2017), to analyse the constructed Hebrew dataset of offensive language. Word2Vec and FastText share the conceptual objective, which is to learn word vector representations by training a neural network with a single hidden layer to predict a target word based on its context (neighboring words). In contrast to Word2Vec, which uses words as the basis for word prediction, FastText operates at a more granular level by using character n-grams, which are sets of co-occurring characters within a specified window. In FastText, the representation of words is based on the summation of the vectors of their character n-grams. While Word2Vec is unable to generate embeddings for words that are not present in its training data, FastText can generate embeddings for out-of-vocabulary (OOV) words. In addition, it has been observed that FastText exhibits superior performance in generating embeddings for morphologically rich languages, such as Hebrew, in comparison to word2vec.

For the analysis, we used the level 6 taxonomy aspects in their lemma forms. In each word embedding method, 30 neighbouring words were retrieved for each aspect. The dimension (vector length) of the word embeddings was 100. A filtering process was then conducted, where neighbouring words that contained the aspect or its lemma, or a stem as a substring, were omitted.

For data visualization, we used two common dimensionality reduction techniques in machine learning and data analysis: Principal Component Analysis (PCA) (Shlens 2014) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Belkina et al. 2019). PCA focuses on reducing dimensionality by finding orthogonal axes (principal components) that maximize data variance. This method is well-suited for linear relationships and simplifies data to 2D or 3D, making it easy to visualize. On the other hand, t-SNE is designed to preserve pairwise similarities, which makes it more effective at capturing nonlinear structures and clusters in high-dimensional data. It excels at revealing intricate patterns that might not be apparent in the original data. However, t-SNE can be sensitive to hyperparameter choices, leading to varying results across different runs.

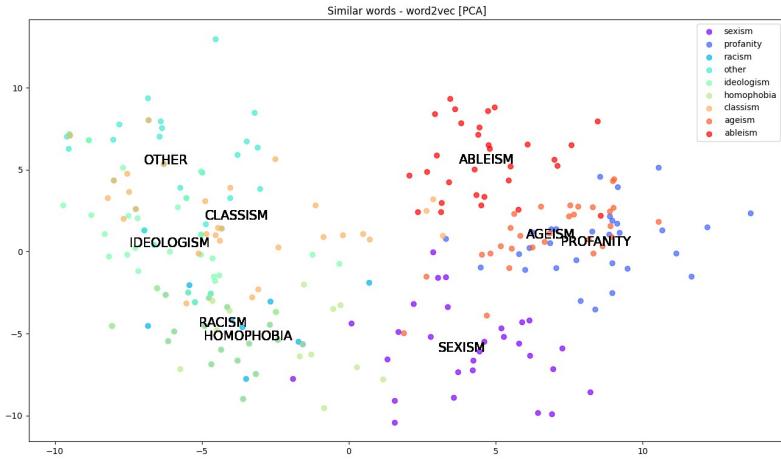


Figure 3: Word2Vec top 30 neighbouring vectors visualization using PCA

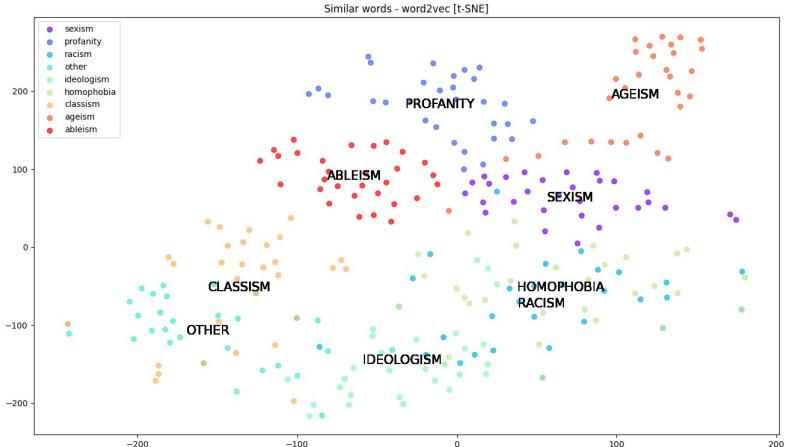


Figure 4: Word2Vec top 30 neighbouring vectors visualization using t-SNE

It can be seen in Figures 3–6 that there is a clear distinction between the neighboring words that appear in the various offensive content aspects, i.e., they are easily distinguished. In both Word2Vec and FastText, the t-SNE visualization distinguishes aspects marginally better than PCA, likely due to its capacity to capture non-linear structures.

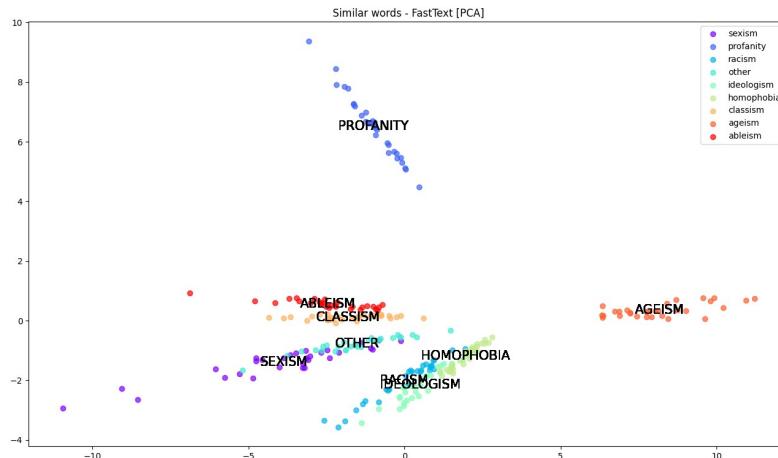


Figure 5: FastText top 30 neighbouring vectors visualization using PCA

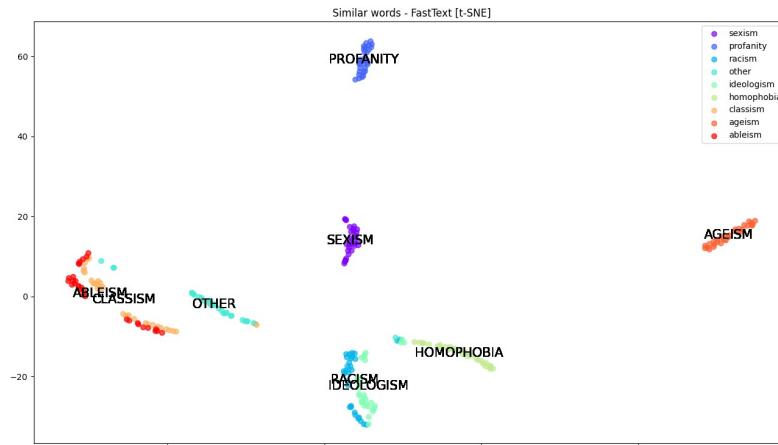


Figure 6: FastText top 30 neighbouring vectors visualization using t-SNE

In the illustration of the FastText method (Figures 5 and 6), there is a significant difference between the various aspects; however, there are categories that almost converge due to their close relationship in reality, such as racial and ideological (making racism an ideology) or ableism and classism (referring to a person in a different social position as disabled).

The Word2Vec illustration (Figures 3 and 4) depicts a separation between the aspects, albeit to a lesser extent. This observation can be attributed to the fact that Hebrew is a morphologically rich language. The FastText model, which operates at the character level,

is able to effectively capture and represent the intricate morphological complexity of Hebrew.

5 Hebrew annotation campaign

5.1 Annotation setup

We used the INCEpTION platform (Klie et al. 2018) to produce annotations. This platform is designed to facilitate interactive and intelligent annotation for various NLP tasks and to aid human annotators by integrating machine learning assistance and leveraging knowledge bases.

We have defined the entire multilevel taxonomy for Hebrew in the INCEpTION tool, including implicit offensive language categories. However, the annotators were instructed to annotate the explicit levels only (levels 1–6 described in Section 3). A snapshot of our corpus in the tool is given in Figure 7.

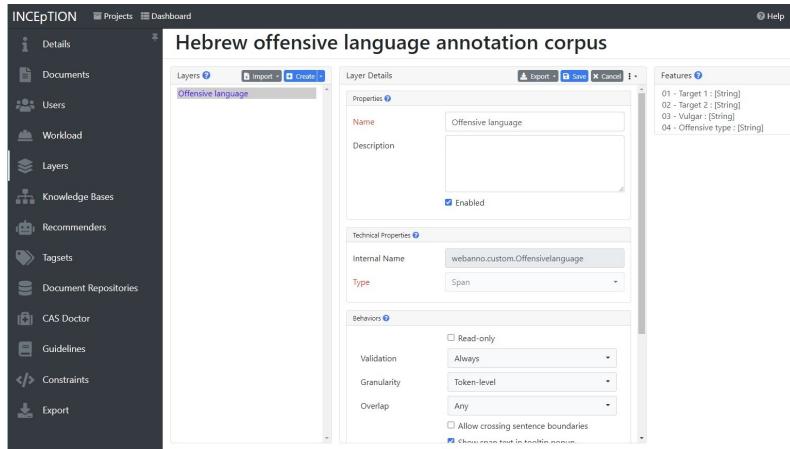


Figure 7: The INCEpTION tool

The data is divided into 9 files (one file per category) with 50 texts in every file, 450 texts in total. Our annotators were unaware of this division.

5.2 Annotation process

Two annotators who are native Hebrew speakers were asked to annotate the comments independently and not to consult one another. Our annotators received detailed guidelines (full text of guidelines appears in the Appendix) that instructed them to annotate the

comments in a top-down manner according to our taxonomy. The guidelines are written in English but contain examples in Hebrew; the annotators have received the original and translated taxonomies for reference.

The annotators were first asked to decide whether or not a text is offensive and then to proceed to levels 2–5 (named targets 01–04 in the system). Because we focus on explicit offensive language in this paper, the annotators were required not to address implicit offensive categories (level 7 of the taxonomy in Section 3). The annotation process is depicted in Figure 8.

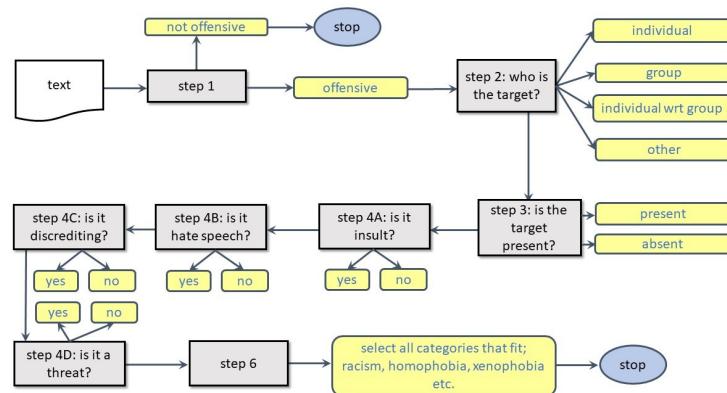


Figure 8: The annotation process

The annotators were also asked to mark the entire text as a scope. In the case of several offensive instances in one scope, they were instructed to mark all categories that apply to any of the instances (see the Appendix for details). An example of offensive text annotation is given in Figure 9.

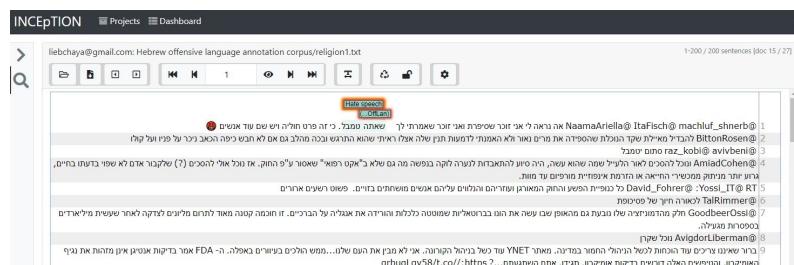


Figure 9: An example of annotation in the INCEption tool

5.3 Annotation results

The annotation of Hebrew texts based on the simplified taxonomy was conducted on a sample of 450 sentences annotated by two independent annotators. We report the Kappa values in Table 4. When working with categorical or ordinal data, the Cohen's Kappa coefficient (Cohen 1960) is a statistical metric used to evaluate the level of agreement between two raters or annotators. It provides a measure of inter-rater reliability, taking into account the agreement beyond what would be expected by chance.

Table 4: Agreement between annotators

Annotation	Type	Agreement (Cohen's Kappa)
Target		
Target 01	individual/group	0.84
Target 02	present/absent	0.49
Target 03	vulgar or not	0.63
Target 04	hate speech/insult	0.73
Target 04A	discredit	0.66
Target 04B	threat	0.97
Aspects 05	-	0.68

As can be seen, our annotators agreed on most cases in all categories, except one— Target 02 (present/absent). This category was indeed very difficult for objective labelling because it is not always very clear whether a target is taking part in the heated conversation because they annotate isolated comments without a context of the entire conversation. For example, someone can address a target via third parties such as “He is ***” or “I want to *** him”, which makes an impression that the target is absent; however, it is also possible that more than two people argue and two of them are talking about the third person in his/her presence.

5.4 Error analysis

Target	Sentence	Translation
01	תורה.האמת לא-aicפת פה ולא מה ואיך שעשנ' לדי' ובאייה און' ובווען צערן אלל' שלדי סדרני זאל בענין בוראה. כי, מה שאלא טבריט וטיפיה ווועה שעוגה גא אט להללויל לע 2000 נפערט... וואם כויהו גו וווער די' עעל מה קורה להפוגה באיראן או בכל מדינה מוסלמית אחרת.אלל' מה המשנה	Look, the truth is I don't care what and who and why and where my children do and in what way and in no way. But a child who grows up straight grows up in a natural and healthy way, because what is not straight is a deviation from what is normal and what is customary even if you roll over it 2000 times... and even an intellectual like you knows what happens to gays in Iran or any other Muslim country, but what does it matter?
02	מעשה שיי פין, היה מעוניין חום ליטסן קווקסנלי	Make a sex change, it will be interesting Haim Levinson a shemale
03	שוד מאנץ': אני שונא כסיש.חוושב שהומוואים זאת טעות בעניבטל. כהנא בזק, ננסי נציבות דלעת יהוי לבשל.אלל' כלב בימיורזיאת.ה'יכ' חזרה מ'ה איסכין, והמשותה פה אונק של מהבלם	More from Gantz: I hate Negroes, I think that gays are a mistake of nature, by and large... Kahana was right, women should know how to cook better, I ate a dog in Micronesia, the ultra-Orthodox MKs are a yikes, and the joint is a collection of terrorists
04	איך אפשר להקשיב לנכלה שבצעה שטע וגבבה קולות כדי להקם ממשלה אנט'יז'ינית	How can you listen to a crook who committed a crime and stole votes to establish an anti-Zionist government
04A	כשייר' הכסות מתנהגה כמ' עברניISM'ג'ץ מטערב. הפל של' אנט'יז'ינית' בעיניך שעסם בעסלאה כתבי' ישות מזורם מגולאת המדינה שלך	When the Speaker of the Knesset acts like a criminal who serves a criminal, the High Court intervenes. It's too bad you don't find it scandalous that someone accused of three serious charges is running your country
04B	מה הוא קורא לך זונה תdkiri אוטו	What is he calling you whore, stab him
05	גראה לך שם מנספּן לסתות בחורות? זה בטח נגד הקוראן לסביות	Do you think they are trying to seduce girls? It must be against the Koran lesbians

Figure 10: Annotation disagreement examples

Almost every disagreement case can be intuitively explained, as we did in the following paragraph for the examples in Figure 10.

For Target 01 (individual/group/etc.), the text is addressed to a particular individual (opponent in the argument), but at the same time, it describes an “issue” related to the entire gay community. We assume that it was a source of confusion and disagreement between two annotators.

For Target 02, the text is directed at someone named Haim Levinson. However, it is not clear if that person is participating in the conversation.

For Target 03, the issue is the word “ew” (or “eww”) that can be considered to be vulgar, but one annotator has decided against it.

In the example belonging to Target 04, one insults some female politician. But, because she represents some particular political party/wing, one of the annotators probably saw a generalization and the reference to a group via discriminatory group stereotypes.

The given comment from Target 04A is confusing because it discredits the Israeli PM, but at the same time, it is addressed to an individual opponent without explicit discredit.

In the case of Target 04B, there is a call to harm someone, but it is difficult to determine if this call is a direct threat.

In the text of Target 05 both the Quran and lesbianism are mentioned, which led to one annotator marking it as profanity and another to mark it as homophobia; upon closer examination, this comment is more likely to be homophobic.

6 Conclusions

This paper introduces a simplified yet comprehensive taxonomy that effectively classifies the varied offensive expressions in Hebrew. The taxonomy follows the combined schema for explicit and implicit offensive language that previously was tested using English datasets via the INCEption annotation platform (Lewandowska-Tomaszczyk et al., 2022). The taxonomy also addresses both straightforward and implied forms of offensive language, as introduced (Lewandowska-Tomaszczyk, 2023).

The accompanying dataset, one of the first of its kind focusing on Hebrew, contains a large number of examples including derogatory terms, culturally specific offenses, and other verbal aggressions in social media. Analysing this dataset has confirmed the validity and relevance of our proposed taxonomy, while simultaneously highlighting the complex interplay of linguistic and historical factors in the construction of offensive language. For most levels of our taxonomy, the agreement was substantial or even high, except for the target's presence/absence which was a more difficult case achieving moderate agreement only. It shows that offensive language annotation can be effective in Hebrew.

We hope that the provided taxonomy and dataset serve as foundational resources for future studies and interventions in the field of the Hebrew offensive language detection (Litvak et al. 2021).

Acknowledgments

The authors express their appreciation to Rivka Sheiner, Yam Biton, and Yovel Efraim for their valuable help with data collection and annotation.

Appendix A: Annotation instructions

The annotation instructions are divided into general and target-specific, where the latter refers to specific levels of the offensive language taxonomy. General instructions are as follows.

- (1) Be consistent! You cannot label comments with the same characteristic (that your decision is based on) with different labels.
- (2) If you have labeled a comment as NON-OFFENSIVE, do not label it with anything else (targets 01-04 must not be labeled in any way).
- (3) If you have labeled a comment as OFFENSIVE, proceed to targets 01-04.
- (4) Label the entire sentence, even if it is very long, from the first symbol to the last (including punctuation marks).

Target 01 corresponds to level 2 of our taxonomy which specifies the target of an offensive speech. These are the instructions given for annotating target 01:

– Target 01

- (1) A comment must be labeled with the GROUP label if it contains an offense targeting a group of people and does not address any specific person.
- (2) A comment must be labeled with the INDIVIDUAL label if it contains an offense targeted at a specific person and do not generalize it towards any associated group.
- (3) A comment must be labeled with the IND. wrt. GR./GR. wrt. IND. label if it simultaneously addresses a person and an associated community/group of people. For example, an offense is addressed to a group because it is led by a specific person or because this is a group of fans/followers of this person; an offense is addressed to a person who is associated with a group (“she is stupid because she is vegetarian”, “he is weak because he is homosexual”, etc.). If an offense is addressed to a group that is not associated with a specific person, then it must be labeled by the GROUP label. If an offense is addressed to a person who is not associated with a specific group, then it must be labeled by the INDIVIDUAL label.
- (4) A comment must be labeled by the NON-TARGETED label if an offense does not address any specific person or group (“this is a stupid country/shitty situation”, etc.)

Target 02 in the system indicates level 3 of our taxonomy specifying whether or not the target of an offensive speech is explicitly addressed in the text. The instructions given for annotating target 02 are:

Target 02

- (1) Label a comment as PRESENT if a target of an offense is a conversation participant. For example, “You are ***” or “I will do *** to you”. In this case, a second person (“you”) is present.
- (2) Label a comment as ABSENT if a target is not taking part in the heated conversation. For example, “Bibi is ***” should be labeled “absent” because obviously, the PM is not taking part in this conversation. The same goes for third parties such as “He is ***” or “I want to *** him”.
- (3) If a comment is long and contains several offensive instances, then label it as PRESENT if at least one of the instances is PRESENT.

Target 03 indicates whether the offensive language is vulgar and represents level 4 of our taxonomy. Its instructions appear below.

Target 03

- (1) A comment should be labeled as VULGAR if there is a use of words and phrases that refer to scatology (feces and excretion), effluvia (an unpleasant smell or exhalation), death, or sexuality, especially different sexual practices or parts of the body that are used to offend or shame. For example, “shit!” or “Pee on you” or “Smelly mouth” or “to fuck” or “scoundrel” are vulgar expressions.
- (2) A comment containing offensive words and curses that do not belong to the domains listed in the previous rule will be labeled as NON-VULGAR.

Target 04 indicates the strength of the offense in the text and represents the first two items in level 5 of our taxonomy. The guidelines provided for annotating target 04 are listed below.

– Target 04

- (1) A comment should be labeled as INSULT if there is an individual or group offense with no reference to any group stereotypes (generalized belief about a particular category of people). For example, “hypocrite” or “demagogue”.
- (2) A comment should be labeled as HATE SPEECH if its discriminating property is precisely the reference to a group or individual via discriminatory group stereotypes. For example, “(politics) leftist”, “Coccinelle”, or “bibist”.

Targets 04A and 04B represent the remaining two items in level 5 of our taxonomy - discrediting offensive language and threats. The guidelines are as follows for annotating these targets.

– Target 04A

- (1) Label a comment as DISCREDIT *only* if it describes a person/group as lying, cheating, immorality, unprofessionalism, or unfairness. For example: “he is a liar”, “he is a dump”, “he does not know how things work”, etc.
- (2) Label a comment as NON-DISCREDIT if it is OFFENSIVE but does not discredit any person/group.

– Target 04B

- (1) A comment should be labeled as THREAT if there is an active direct threat expressed in it. For example, “I will kill X” or “X should be killed” are threats.
- (2) Wishing is not threatening, and such comments should not be labeled as threats. For example, “I wish someone would kill X” or “I wish X would die” are not threats.
- (3) A threat should contain a mention of physical or other harm. For example, “Do not leave her alone till X resigns” or “We will file complaints against X” are intentions and not threats.
- (4) If a comment is long and contains several offensive instances, then label it as THREAT if at least one of the instances is THREAT.
- (5) Label a comment as NON-THREAT if it is OFFENSIVE but there is no active direct threat expressed in it.

Aspects of target 05 in the system stand for level 6 of our taxonomy which describes various aspects of explicit offensive language. In this category, the annotators were asked to select all aspects that apply to a given text (in alphabetical order). The guidelines referring to these aspects appear below.

Target 05

- (1) Racism: involves exhibiting discrimination and bias against individuals due to their racial or ethnic background.
- (2) Homophobia: encompasses a spectrum of unfavorable sentiments and emotions directed towards homosexuality or individuals who identify as lesbian, gay, or bisexual, or are perceived as such.

- (3) Sexism: pertains to bias or belittlement rooted in an individual's sex or gender, particularly such bias targeted at women.
- (4) Profane: characterized by irreverence or contempt for God (religious) or sacred principles or things.
- (5) Ageism: the unjust treatment of elderly individuals due to unfavorable and inaccurate stereotypes.
- (6) Ableism: the act of discriminating against and harboring societal biases towards individuals with disabilities, driven by the notion that conventional abilities are of higher value.
- (7) Classism: the complex combination of institutional, cultural, and personal behaviors and beliefs that allocate varying worth to individuals based on their socioeconomic class.
- (8) Ideologism: dealing with a system of ideas.
- (9) Other: You should label a comment as OTHER *only* if it does not fit any of the 8 aspects mentioned above.

Hebrew examples of texts for some of the targets are given in Figure 11.

Target	Type	Hebrew example	Translation
01	GROUP	שמאליים אנשים אינטלקטואליים מודע דודדים שהם חושבים הינם חכמים עם בקען מחשבתי ובודת להיכין נשים עם טיפשות אחות ורבה יותר כחמים וערכאים יותר ובעל היגיון ברא אמשיך לשים לهم מראה מזו הפראזף	Leftists are very shallow people who think they are smart with a fixed mindset. It is my duty to prove them stupid. We are much smarter and more valuable and have common sense. I will continue to put a mirror in front of their faces.
01	INDIVIDUAL	פופוליזם המבוסס על האדרטה קבוצה מדומיננטית המכונה "העם", וsonian ל' י' שמהומה לה כליטה בוגדייה המא"מ'ת על שלמותה. זה לא... ח...	Right-wing populism is based on glorifying an imaginary group known as "the people", and branding everyone outside of it as a treacherous elite that threatens its integrity. It's not h...
04A	NON-DISCREDIT	ארץ זבל חלב וחרא	A land of milk and shit

Figure 11: Hebrew examples for targets 01, 02, and 04A

References

- Belkina, Anna C, Christopher O. Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen & Jennifer E Snyder-Cappione. 2019. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications* 10(1). 5415.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5. 135–146.
- Bright, J. 2022. History under attack: Holocaust denial and distortion on social media. Supporting Data. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France, and the United Nations Department of Global Communications, United Nations, New York, USA.

- Caselli, Tommaso, Valerio Basile, Jelena Mitrovic, Inga Kartoziya & Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, 6193–6202. The European Language Resources Association (ELRA), Marseille, France.
- Chiril, Patricia, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully & Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le traitement automatique des langues naturelles* (TALN-PFIA 2019), 351–360. Toulouse, France, ATALA.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1). 37–46.
- Çöltekin, Çağrı. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the twelfth language resources and evaluation conference*, 6174–6184. The European Language Resources Association (ELRA), Marseille, France.
- Davidson, Thomas, Dana Warmsley, Michael Macy & Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, vol. 11, 512–515. San Francisco, California USA, AAAI Press.
- Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the first workshop on abusive language online*, 46–51. Long Beach, California, USA, Curran Associates, Inc.
- Fortuna, Paula, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically labeled Portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, 94–104. Florence, Italy, ACL.
- Grice, Herbert Paul. 1990 [1975]. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech acts*, 41–58. New York: Academic Press.
- Hamad, Nagham, Mustafa Jarar, Mohammad Khalilia & Nadim Nashif. 2023. Offensive Hebrew corpus and detection using bert. arXiv preprint arXiv:2309.02724.
- Haugh, Michael & Valeria Sinkeviciute. 2019. Offence and conflict talk. In Matthew Evans, Lesley Jeffries & Jim O'Driscoll (eds.), *The Routledge handbook of language in conflict*, 196–214. London: Routledge.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho & Iryna Gurevych. 2018. The inception platform: machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, 5–9. Santa Fe, New Mexico, USA, ACL.
- Koglavani, SV, S Malliga, KR Jaibabinaya, M. Malini & M. Manisha Kokila. 2023. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, vol. 81, part 2, 630–633, Elsevier.
- Lakoff, George & Mark Johnson. 1980. *Metaphors We Live By*. Chicago: Chicago University Press.
- Lewandowska-Tomaszczyk, Barbara. 2023. A simplified taxonomy of offensive language (sol) for computational applications. *Koninklijke Brill's Linguistics Series* 10(3). 213–227.
- Lewandowska-Tomaszczyk, Barbara, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene & Slavko Žitnik. 2023. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 19(1). 7–48.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrovic & Giedre Valunaite Oleškeviciene. 2021a. Lod-connected offensive language ontology and tagset enrichment. In Shubert R. Carvalho and Renato R. Souza (eds.), proceedings of the

- workshops and tutorials held at ldk 2021 co-located with the 3rd language, data and knowledge conference, vol. 3064, 135–150. CEUR Workshop Proceedings.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović, and Giedrė Valūnaitė Oleškevičienė. 2021b. Lod-connected offensive language ontology and tagset enrichment. In CEUR workshop proceedings, vol. 3064.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojszczak et al. 2023. Annotation Scheme and Evaluation: The Case of OFFENSIVE Language. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovje 49(1), 155–175.
- Liebeskind, Chaya & Shmuel Liebeskind. 2018. Identifying abusive comments in Hebrew Facebook. In *2018 IEEE international conference on the science of electrical engineering in Israel (ICSEEL)*, 1–5. IEEE, Eilat, Israel.
- Litvak, Marina, Natalia Vanetik, Chaya Liebeskind, Omar Hmida & Rizek Abu Madeghem. 2022. Offensive language detection in Hebrew: can other languages help? In *Proceedings of the thirteenth language resources and evaluation conference*, 3715–3723. Marseille, France: The European Language Resources Association (ELRA).
- Litvak, Marina, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout & Israel Beer-Sheba. 2021. Offensive language detection in Semitic languages. In *Multimodal hate speech workshop*, vol. 2021, 7–12. Düsseldorf, Germany: ACL.
- Liu, Ping, Wen Li & Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, 87–91. Minneapolis, Minnesota, USA: ACL.
- Mandl, Thomas, Sandip Modha, Anand Kumar M & Bharathi Raja Chakravarthi. 2020. Overview of the HASOC track at FIRE 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*, 29–32. Hyderabad India: Association for Computing Machinery (ACM).
- Mandl, Thomas, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia & Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation*, 14–17. Hyderabad India: Association for Computing Machinery (ACM).
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Mohaouchane, Hanane, Asmaa Mourhir & Nikola S Nikolov. 2019. Detecting offensive language on Arabic social media using deep learning. In 2019 sixth international conference on social networks analysis, management and security (SNAMS), 466–471. Granada, Spain: IEEE.
- Pan-European anti-racism network. 2022. ENAR Shadow Report 2006. <https://www.enareu.org/shadow-reports-on-racism-in-europe-203/>.
- Pitenis, Zeses, Marcos Zampieri & Tharindu Ramasinghe. 2020. Offensive language identification in Greek. arXiv preprint arXiv:2003.07459.
- Poletto, Fabio, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, Cristina Bosco, et al. 2017. Hate speech annotation: analysis of an Italian Twitter corpus. In CEUR workshop proceedings, vol. 2006, 1–6. Rome, Italy: CEUR-WS.

- Ranasinghe, Tharindu, Marcos Zampieri & Hansi Hettiarachchi. 2019. Brums at HASOC 2019: deep learning models for multilingual hate speech and offensive language identification. In *FIRE 2019* proceedings, 199–207. Kolkata, India: CEUR-WS.
- Schütze, Hinrich & Jan O Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management* 33(3). 307–318.
- Shlens, Jonathon. 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.
- Sigurbergsson, Gudbjartur Ingi & Leon Derczynski. 2019. Offensive language and hate speech detection for Danish. arXiv preprint arXiv:1908.04531.
- Smadja, Frank, Kathleen R McKeown & Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational linguistics* 22(1). 1–38.
- Technologies, Mindpool. 2023. Mindpool Technologies. Available at: <http://www.mindpool.com> (accessed 6 September 2023).
- Tova Hartman. 2022. The challenges of multiculturalism in Israel’s shared society – opinion. *Jerusalem Post*. Available at: <https://www.jpost.com/opinion/article-705192> (accessed 10 September 2023).
- Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven & Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. arXiv preprint arXiv:1608.08738.
- WALLA! TECH. 2022. Social media plays large role in fomenting online hate - report. *Jerusalem Post*. Available at: <https://www.jpost.com/international/article712070> (accessed 6 September 2023).
- Wine, M. 2016. National monitoring of hate crime in Europe: the case for a European level policy. In Jennifer Schweppe and Mark Austin Walters (eds.) *The Globalization of Hate The Globalization of Hate: Internationalizing Hate Crime?* 213–32. New York: Oxford University Press.
- Yasaswini, Konthala, Karthik Puranik, Adeep Hande, Ruba Priyadarshini, Sajeetha Thavareesan & Bharathi Raja Chakravarthi. 2021. IIIT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, 187–194. Online, ACL.
- Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.
- Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019b. SemEval-2019 task 6: identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.

About the Authors

Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and data mining. Especially, her scientific interests include Semantic Similarity, Language Technology for Cultural Heritage, Morphologically rich languages (MRL), Multi-word Expressions (MWEs), Information Retrieval (IR), and Text Classification (TC). Much of her recent work has been focusing on analysing offensive language. She has published a variety of studies and a few of her articles are under review or in preparation. She is a member of several international research actions funded by the EU.

Address

Jerusalem College of Technology, Department of Computer Science
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel

e-mail: liebchaya@gmail.com
ORCID: <http://orcid.org/0000-0003-0476-3796>

Natalia Vanetik is a senior lecturer and researcher in the Department of Software Engineering at the Shamoon College of Engineering. Her research interests include Natural Language Processing, text mining, and optimization. Specifically, her research covers diverse range of topics in NLP and machine learning, including social media analysis, job vacancy ranking, and the development of evaluation systems for summarization tasks. Her research also extends to graph theory applications in data mining and cross-lingual transfer learning.

Address

Shamoon College of Engineering, Department of Software Engineering
56 Bialik st.
84100 Beer Sheva, Israel

e-mail: natalyav@sce.ac.il
ORCID: <http://orcid.org/0000-0002-4939-1415>

Marina Litvak is a Senior Lecturer at Shamoon College of Engineering, Department of Software Engineering. Marina's research focuses mainly on Multilingual Text Analysis, Social Networks, Knowledge Extraction from Text, and Summarization. Marina published over 90 academic papers, including journal and top-level conference publications. She constantly serves on the program committees and editorial boards in multiple journals and conferences and collaborates on different research projects in Israel and abroad. She is a co-organizer of the MultiLing, FNP, Text2Story, and IACT workshops, collocated with top-level conferences.

Address

Shamoon College of Engineering, Department of Software Engineering
56 Bialik st.
84100 Beer Sheva, Israel

e-mail: marinal@ac.sce.ac.il
ORCID: <http://orcid.org/0000-0003-3044-3681>

Anna Bączkowska
University of Gdańsk

IMPLICIT OFFENSIVENESS FROM LINGUISTIC AND COMPUTATIONAL PERSPECTIVES: A STUDY OF IRONY AND SARCASM

Abstract.

The aim of this paper is to shed some light on the linguistic concept of implicit offensiveness. On the one hand, implicitness will be juxtaposed with indirectness as the two concepts are not conceived of here as synonymous. On the other hand, a typology of offensiveness (vs offensive language and vs offendedness) will be proposed, as well as the overarching term ‘covert meaning’ that will span figurative implicitness and non-figurative implicitness. The gradability of various forms of covert meaning and its overlap with overt meaning (subsuming explicit literal meaning and implicit literal meaning) will also be discussed. In the analysis, two sample implicit concepts will be examined (irony vs sarcasm) based on corpus data (of general English and dedicated offensiveness corpus) and using non-contextual embeddings. Theory-wise, the paper demonstrates that implicitness is a complex term which is fuzzy and gradable; methodology-wise, it shows how computational tools can be used to attest theoretical assumptions related to offensive covert terms.

Keywords

implicitness, offensiveness, irony, sarcasm, embeddings, corpus of offensive language

1 Introduction

The aim of this paper is to shed some light on the linguistic concept of implicit offensiveness. To better understand the term, the concept of offensiveness needs elaboration (Section 1) on the one hand and of implicitness per se (Section 2) on the other. It is assumed here that implicitness is not a notion tantamount to indirectness, as many scholars seem to think by treating them as synonyms; hence, indirectness will also receive due attention (Section 3). These three sections will provide a detailed discussion of how implicitness and indirectness differ and will manifest the complexity of the key notion of implicitness. This will be shown by resorting to two other concepts that underlie the notion of implicitness, namely literalness and figurativeness, their gradability and non-discreteness. The possible subcategories of implicitness will be mentioned, two of which (*irony* and *sarcasm*) will receive further empirical investigation (Section 4). This will serve only as an illustration of how the hazy category of implicitness may be further

detailed by resorting to computational tools. The analysis will rely on tools used in computational linguistics/science, one of which has already been widely employed by linguists, which is a corpus-assisted analysis, and the other popular in computational science and relatively new to linguistic investigations, which are lexical embeddings. The main objective of the paper, however, is to offer a gradable understanding of the concept of implicitness anchored in a tripartite model of the notion of offensiveness.

2 Offensiveness, offendedness and offensive language

Offensiveness is a topic that is increasingly popular both in pragmatic scholarship and in computer science studies. In linguistics, the interest in offensive language that is currently so widespread has grown out of previous studies on impoliteness (Culpeper 1996; Bousfield 2008), which in turn appeared as a reaction to politeness theories (Brown and Levinson 1978; Leech 1983; Lakoff 1973).

The terms *offensiveness*, *offence*, *offendedness*, and *offensive language* are not seen here as synonyms, even though there is some degree of overlap among them, and they are often used interchangeably by a number of authors. Generally speaking, offensive language spans prototypically negative, hurtful, and often vulgar words. An excellent example of offensive language is swearing. Swearwords are explicit language means (devices) used as carriers of offensive meaning. Offensiveness may be explicit or implicit; it is the latter that is of interest here. Implicit offensive contexts rely mostly on rhetorical devices (also dubbed tropes, figures of speech, figures of thought, stylistic figures), and thus they are, unlike language devices, figurative. Naturally, rhetorical devices, such as irony or sarcasm, while still being instantiations of implicit figures of speech, may contain vulgar words (offensive language) which are explicitly offensive. Offensiveness is conceived of here as the most general term which engulfs both straightforwardly expressed desire to hurt the addressee/target, that is, in an overt way, and purposefully obscured expressions voiced with the intention to hurt, that is, covert cases. Put differently, offensiveness spans overt and covert forms of expression. Along with explicit, offensive language and explicit/implicit offensiveness, the third term to be described in this study is offendedness, which requires that the target be or feel offended; in other words, it triggers offence. These notions are presented in Fig. 1.

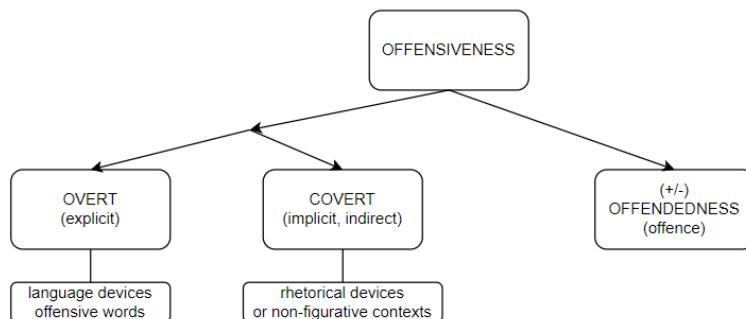


Figure 1: Offensiveness, offensive language and offendedness

To summarise, the schema illustrated by Fig. 1. shows the general term *offensiveness*, which can be further specified as either (1) overt, that is, explicit, encoded by offensive language (viz. language devices), or (2) covert, that is implicit or indirect, and then they can be realised through rhetorical devices in the case of implicitness or via non-figurative contexts (which will be further elaborated in the next section). Offensive language is usually enacted through single words (“slut”, “stupid”) that are vulgar or hurtful in some way or through phrases/utterances (“fuck off”, “mother fucker”). The explicit contexts rely on literal readings. While offensiveness is usually divided into explicit and implicit (e.g., Lewandowska-Tomaszczyk et al. 2021), it is proposed here to include indirectness along with implicitness as non-explicit cases and thus to use the term *covert* to encompass both options of inexplicitness. The main difference between them boils down to whether a given context uses figurative language (implicitness) or non-figurative yet non-literal (indirectness). These notions are discussed in detail in the following sections. The covert contexts are usually realised through phrases or sentences/utterances, as the use of, for example, rhetorical devices require that, yet occasionally single-word propositions can also function as covert cases, for example, when one says *Genius!* in an ironic way. The next (optional) step when discussing offensiveness is to consider whether the offensive contexts under investigation are analysed at the level of locution (and seen from the perspective of a speaker) or whether they carry illocutionary or perlocutionary force in a given context (and then they are perceived from the perspective of the hearer-target). Offendedness may but need not depend on the use of offensive language (hence “+/-” in the diagram), that is, explicit tokens of offensiveness, because, as already alluded to, offensive words may be perceived as non-offensive and non-offensive words as offensive. In other words, the explicit cases of offensiveness are juxtaposed with inexplicit cases, while offence/offendedness is a separate category which can but does not have to be considered in an analysis.

The offence is conceptualised here as the result of an offensive comment directed at a target (whether a ratified or non-ratified addressee). It is thus the illocutionary or

perlocutionary effect of (prototypically) an offensive comment; put differently, it is the target's reaction to an offensive remark who takes offence. The term offendedness was already used by Jay (1992: 161) to denote "a reaction to a word by a person who hears or reads" something that offends him or her, which fits perfectly to our understanding of the term. An offensive comment may contain evidently offensive words, such as swearwords, i.e., negatively emotion-laden words (explicit, offensive comment, e.g., "you are a fucking cunt") or common words which denotationally are not offensive but may become such in a specific context (implicit offensive comment, e.g., ironic one like: "You are a genius"). Interestingly, words denotationally burdened as offensive ("cunt", "fuck"), even though they are explicit and thus are examples of offensive language, may trigger reactions other than offence; they may be funny, flirtatious, etc., if uttered among close friends, in a specific sub-culture group slang or peer circle. The very presence of offensive words (such as vulgarisms) thus does not necessarily entail the speaker's offensive intent and/or the effect of taking offence in the target's emotional response. On the other hand, non-vulgar words may still be the source of distress and/or offence in the target (cynical, sarcastic or ironic comments in particular) even though they do not contain linguistic evidence manifesting the speaker-intended will to offence. In this respect, the term *offensiveness* used by Jay (1992) differs from what is proposed here as for Jay "Offensiveness is a term used to denote the degree to which a certain word or concept possesses negative or aversive properties". For Jay, offensiveness refers to (1) the semantics of a word and (2) negative words only. In this paper, offensiveness spans potentially harmful words, so neutral words that become negative context-wise and negative words that are neutralised context-wise are taken into account. Therefore, offensiveness encompasses both the (conceptual-)semantic and, unlike in Jay's proposal, the pragmatic meaning. What we converge on with Jay (1992: 161) is that Offendedness is a personal reaction to a stimulus; it is not a quality of a word per se. Offendedness in this study is a term subsumed by a more general one of offensiveness.

Another possibility is when the intention to hurt may be linguistically encoded but is left unnoticed by the hearer, who may fail to recover the ill intention. All these cases are spanned by an overarching concept of offensiveness, and thus, the term involves using or not using offensive language (explicitly vulgar words) and triggering or not triggering offence in the target. Swearwords and expletives are words which seem to be offensive due to the strong emotional charge they convey, yet their primary offensive content may be easily cancelled by situational context and the sender's and the receiver's attitude to such words or interactants, as "the use of swear words is heavily context-sensitive" (Goddard 2015: 190). The present analysis will deal with implicit offensiveness, that is, offensiveness expressed by covert meaning, yet not necessarily with offence (offendedness).

3 Implicit offensiveness

While much ink has been spilt on the reading of indirectness (elaborated in the next section), somewhat less popular is the topic of implicitness in pragmatics, yet an increase in interest can be observed. Implicitness has been defined as “a linguistic choice that the speaker in principle can deny that s/he has made”, which “goes beyond literal meaning”, covers “presuppositions and implicatures” and “goes beyond the field that these concepts denote”, it also “covers speakers’ attitudes and emotions” and the effect utterances have on the target (Östman 1986: 25–26). This definition does not seem to differ much from what is defined as indirectness; hence, it needs further elaboration.

Being offensive implicitly means expressing hurtful opinions, judgements or evaluations in an obscure, veiled way that is not straightforward. The motivation behind the avoidance of straightforward expression of offensiveness may be twofold. Some scholars are of the opinion that being impolite in a direct way is blunter and has a stronger hurtful effect on the target than when one expresses criticism in a veiled manner (for example, Leech 1983; Searle 1975; Brown and Levinson 1987). This early approach to (im)politeness and implicitness has been recently challenged by other scholars who claim the reverse, that is that what is hidden and requires some processing time to recover the meaning generates a stronger, wanton effect inasmuch as the interpretation one arrives at is contrary to expectations, and thus it is more surprising and eventually painful and debasing (for example, Martí 2006; Terkourafi 2013; Culpeper and Haugh 2014). Still others (Sanchez and Vogel 2013; Benikova et al. 2018) maintain that the problem remains unresolved.

Interestingly, Benikova et al. (2018: 177) notice that survey respondents perceive tweets aimed at insulting (rather than threatening) as more hurtful. The authors see the reason for that in the fact that explicit offensiveness is more often honest while implicit one is deceitful and that untruthful and dishonest comments cause greater pain and carry stronger hateful messages than when one is confronted with a truthful opinion. Be that as it may, implicitness seems to play an essential role in encoding and decoding offensiveness and thus deserves further elaboration.

In this study, it is proposed to treat implicitness as a concept associated not only with propositional messages and linguistic encoding (typically discussed under the umbrella term of indirectness) but also with multimodal ones, which warrant making contextual inferences of conceptual or conceptual coupled with linguistic, facets that typically constitute the (non-linguistic) contextual information. These clues are used to infer implicit meaning, which can encompass physical, cultural, social situational or experience-based features. This multimodal approach to implicitness is germane to what Kecske (2017: 208) holds that “linguistic data must be supplemented by non-linguistic, contextual interpretation processes”. Incidentally, high context dependence, according to Yus (1999), is the factor that increases implicitness, while Kurzon (2017) gives an

extreme case of implicitness, that of thematic silence whereby topic avoidance creates confusion as to what the speaker thinks of something in question.

In the simplified model of implicit offensiveness recently offered (Bączkowska et al. 2022), implicit offensiveness relies primarily on figures of speech, particularly irony, metaphor and overstatement or understatement. These four notions constitute the core subtypes of implicitness in the model proposed by Grice (1989). Metaphor is anchored in the idea of comparison, which is also the case of simile, yet the latter is made more evident through the use of “as” or “like”. Both have a similar underlying concept of talking of somebody by referencing another entity, yet they differ in the degree of implicitness. The recent model views the simile as a separate category within implicit offensiveness. Irony, in turn, is typically rooted in the idea of contrast, yet there are other options that, while being classified as irony, do not resort to contrasting entities.

The concept of indirectness is also subsumed by implicitness in this model, which will be elaborated on in the next section as a separate notion. The rationale behind the model is rooted in the Gricean theory, particularly the figures of speech that flout the first maxim of Quality, which reads, “Do not say what you believe to be false” (1989: 27). Figurative language in ironic utterances resorts to obviously false statements; for example, when used ironically, the metaphor “you are an angel” assumes that the entity referred to is not an angel but has features opposite to those ascribed to an angel. Resorting to untruthful statements is typical of figurativeness, which creates implicit meaning. The four figures of speech mentioned by Grice in connection with flouting the first maxim of Quality, that is, irony, metaphor, hyperbole and meiosis, were used as the core categories of implicitness in the model mentioned above, which was enriched by simile, indirectness and rhetorical questions.

Along with the Gricean theory, theorising on implicitness may benefit from the neo-Gricean and post-Gricean approaches to meaning, which notice the gradability of explicit/implicit meaning (Bączkowska 2022), in particular, the notions of implicature (Bach 1994b) and explicature (Sperber and Wilson 1986) are relevant in explaining degrees of implicit meaning. In both cases, meaning interpretation stems from some insertion of the missing, i.e., inexplicit meaning (e.g., through completion or expansion in Bach’s theory) to obtain a complete proposition, be that conceptual strengthening (for Bach, it may pivot on the expansion that involves lexical strengthening) of the logical form and/or sentence enrichments (for Sperber and Wilson) as well as contextual inferences based on them. The need to strengthen/enrich meaning at the level of *what is said* through implicatures or explicatures is the result of semantic indeterminacy (or semantic underdeterminacy, to use Bach’s parlance). It is claimed that sentences tend to be semantically nonspecific as the logical form need not determine a proposition, and thus, the semantic meaning alone often fails to make the inferential process successful. Implicature, in turn, is reserved for what is implicated (by *what is said*), which is a stage occurring after the enrichment process.

According to Bach (1987: 10), “Where expansion is involved, what is meant is closely related to what is said, since the former proposition is derived from the latter by the insertion of conceptual material, but is not identical to it.” Expansion bridges literalness and non-literalness, and thus, it strikes the middle ground between *what is said* and *what is meant* (implicated), which makes the notion of *non-figurative explicitness* and *figurative implicitness* linked by the intermediate stage of non-figurative implicitness (Fig. 2).

Non-figurative explicitness-----non-figurative implicitness-----figurative implicitness

Figure 2: Three-stage cline of explicitness-implicitness

However, there are contexts such as “You’re a pig”, “You’re an angel”, or “This is not my cup of tea”, which invite misgivings about the above cline. They go beyond this gradability scale inasmuch as they are not implicit, being perfectly understandable, yet they hinge on figurativeness. Their lucidity leans on high usage frequency, making them semantically unobscured even if, prototypically, they were inexplicit. Accordingly, another slot should be wedged in this scale between non-figurative explicitness and non-figurative implicitness, which can be labelled figurative explicitness (Fig. 3).

Non-figurative explicitness ---**figurative explicitness** ---non-figurative implicitness---figurative implicitness

Figure 3: Four-stage cline of explicitness-implicitness

On this view, the example “You’re an angel” (or “You’re a genius”) often mentioned in literature as good instantiations of irony (which by definition is rooted in implicitness) raises doubts about its implicitness status. In actual fact, it is so recurrent both in ironic and non-ironic comments (to pay a compliment) that it is probably no longer perceived as opaque (figurative), and its comprehension is most likely immediate. Due to the recurrence of usage, they have become fully transparent. Similar observations were voiced by Kecske (2017) regarding various types of fixed phrases (including idiomatic expressions), like *not my cup of tea, welcome abroad, piece of cake*, which are situation-bound, i.e., they derive their meaning in “a well-definable situation” and tend to occur in a particular social context. They are “coherent units” that are like “frozen implicatures”, which “are a part of a kind of collectively shared social background knowledge” acquired through social practices one partakes intrinsic to socialisation. Stemming from this assumption, situation-bound units are usually easily understood by the L1 speech community and are rarely misunderstood (Kecske 2017: 202), irrespective of their fixed or figurative/idiomatic meaning. As Kecske (2017: 203) asserts, saying, “not my cup of tea”, “No L1 speaker of American English will understand this expression literally. Its most salient meaning for them is its figurative meaning”. While Östman (1986: 34),

Blakemore (1989: 31), Carston (1988: 67), Yus (1999), Kurzon (2017: 220), Decock and Depraete (2018), among others, hold that explicitness and implicitness constitute a gradable cline and are thus non-binary concepts, Kecske (2017) also argues for the possibility for a proposition to be explicit and implicit at the same time. For example, the idiom “it’s not my cup of tea” is non-literal (idiomatic) yet at the same time explicit.

In a similar vein, Gibbs and Colston (2012: 7) maintain that some forms of metaphoric language, in particular, conventional metaphors, “may in some people’s view operate much like literal language and may be interpreted as readily as most non-figurative discourse”. Gibbs and Colston (2012: 26) are also in favour of the opinion that some instances of figurative meanings, such as “a novel, poetic metaphor (e.g., “Her kiss is the sting of a bee”) seem *more non-literal* than highly conventionalised phrases that appear to express something much closer to ordinary, perhaps literal, meanings”, such as *kick the bucket* [italics mine]. This stand is suggestive of the literal meaning being somewhat fluid and continuous itself, as it can be more literal or less literal. The fluidity of notions is thus not only applicable to the explicitness-implicitness typology or literal-figurative distinction, spanning all categories of offensiveness illustrated by Fig. 4, but to the constitutive elements on this continuous cline as well. Put differently, the notions of literalness and figurativeness are internally non-homogeneous¹ and manifest gradability. Gibbs (1993) asserts that, for example, on the literalness end, various forms of literary meaning may be identified, and their degree of literality may be evaluated differently by various evaluators depending on which dimension and salient feature they will focus on in the evaluation process. Thereby, the continuous line spanning the two extremes of literal vs figurative does not do full justice to the problem of literalness-figurativeness gradability, according to Gibbs and Colston (2012: 27).

For this reason, it is proposed here to conceive of contexts as “kick the bucket” or “not my cup of tea” as cases illustrating figurative explicitness rather than the default option of figurative implicitness, despite the fact that technically speaking, they incarnate implicitness due to their non-transparency. Aligned with this assumption, the interpretation of the following example should be seen in terms of explicit figurativeness. By saying, “You’re a pig”, one resorts to metaphor, yet hurling this insult at a target, at all likelihood, does not require complex processing (analysing the insult at a literal and metaphorical level, disregarding the sequence of these potential processes), making thus the adjustment process of re-conceptualisation highly unlikely. This metaphor is heard sufficiently frequently to be grasped immediately without dissecting it into separate layers of meaning; therefore, it exemplifies figurative explicitness. Another insult illustrates an interesting case expressed through “You (are a) slut”. Truth-conditionality is the decisive factor here, determining whether this utterance is figurative (when said to somebody who is not a prostitute, possibly for humorous purposes by resorting to metaphor) or not (when said to somebody who is a prostitute). In both cases, this

¹ Ariel (2002), for example, distinguishes three subtypes of literal meaning (linguistic, psychological and interactional).

utterance may be used in an offensive manner; the former can be an example of a jocular insult, and its understanding is obvious, and the latter as a slur (wherein the target is described through the prism of a social or professional group to which the target belongs). From this discussion, it transpires that inconsistent with the default expectations, figurative contexts can be explicit; still, they are not literal.

Aside from metaphors, another case of hazy categories comprises irony, including metaphorical irony and understatement inasmuch as they may cause doubts in assessing its figurativeness. Whilst the classic examples of irony (being based on contrasts) rely on figurativeness² (“You are an Einstein”), wherein the reading is based on the reversed meaning, there are contexts where no meaning opposition is at work, and thus, the meaning is less figurative/more literal. The oft-quoted example of the mother scolding her child for being a slob (“I love children who tidy up their room”; Gibbs and O’Brien 1991: 525) is a case in point. The utterance said by the mother does not seem figurative (e.g., it is not a metaphor and is not contrast-driven), yet it can be classified as a subtype of irony (verisimilar irony, to be precise), i.e., under the rubric of figurativeness. The reading of the mother’s judgement is literal (the sentence meaning), yet it remains somewhat non-literal (inviting further interpretations) at the what-is-communicated end. The reading here departs from the literal interpretation, with meaning remaining implicit, requiring from the receiver some mental re-conceptualisation. It feels more literal than opposition-based irony (e.g., “Beautiful weather” said on a rainy day creates an apparent contradiction) yet more figurative than dead metaphors (as in “Time is running out”) and routinised idiomatic expressions (as in “not my cup of tea”).

Another example, given by Trosborg (1995: 315–372), brings similar observations regarding the gradability of (non)figurativeness of ironic remarks. The utterance “There was nothing wrong with my car yesterday” may be read as ironic (it is an example of a complaint in Trosborg’s paper, and thus it is assumed to be ironic), which rests upon a truthful statement devoid of word or phrase level contradiction-based interpretations. It is the clash of this assertion with the current facts where the divergence from the expected is housed. This dissimilarity between the expected and the actual situation creates a meaning that is not entirely literal, although it seems so on the surface, yet it is not completely figurative either. To compare, this fuzziness is not present in a more obvious figurative irony expressed by the utterance “You did really well” upon a failed examination. Here, the word-based contradiction (“well” to be read as “awful”) leaves no doubt about the meaning expected to be interpreted as it is reversed (with probable intonation suggestive of sheer criticism rather than an honest compliment). The car example from Trosborg (1995) opens more than one potential interpretation and thus is not built on a binary opposition. Secondly, the content of the car utterance is cancellable to a greater degree than in the examination example, i.e., the speaker still has an option amenable to possible withdrawal from the complaint by saying that he or she did not

² Croft and Cruse (2004: 217) classify irony, understatement and hyperbole in terms of metonymy.

mean to attack the hearer, whereas in the exam case, the criticism is evident, and it is not subject to easily manageable cancellation. An even weaker opposition is sensed in ironic understatements (technically classified as figurative; see Croft and Cruse 2004), whereby the contradiction is realised only partially, as in “You are a bit tipsy” said to somebody completely drunk (Kapogianni 2014 dubs it scalar reversal). The implied meaning is mitigated here, which makes figurativeness somewhat attenuated. The discussion presented so far is illustrated by the diagram in Fig. 4.

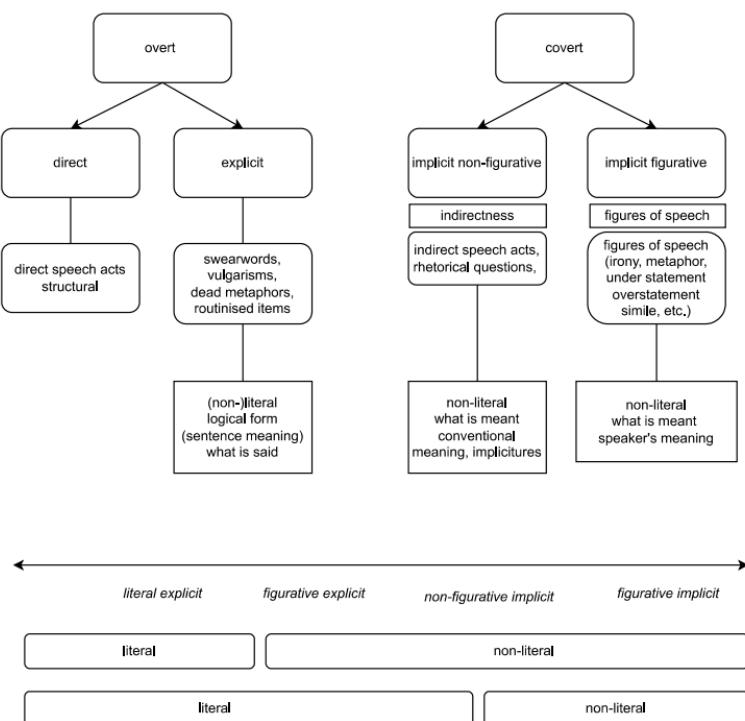


Figure 4: Overt vs covert offensiveness

In the diagram of offensiveness above, the categories are roughly divided into literal (rhetorical questions, indirectness) and non-literal (e.g., irony, metaphor). By and large, literal meaning is non-figurative and is engrained in the expressed assertion in the sentence or, roughly speaking, in *what is said*. Technically speaking, however, literal meaning and *what is said* are not identical concepts; the latter derives from the former as

a result of meaning disambiguation (e.g., by elaborating indexicals). In other words, the literal meaning is the underlying layer contributing to what is said. As already alluded to, in accord with Bach (1984a), meaning disambiguation already belongs to implicature. *What is said* corresponds to the elements of a sentence/utterance and its structure, yet it “does not mean that what is said must be made fully explicit” (Bach 1999: 335). For Bach, *what is said* includes language conventions, disambiguations and indexicals but excludes references to demonstratives, pronouns and proper names (Devitt 2021: 8). For Grice (1989; Devitt 2021: 52), *what is said* involves the contents encoded by an utterance (the logical form) as well as the speaker’s intentions.

Whilst literal meaning is considered context-invariant (for detailed definitions, see Ariel, 2002), *what is said* is the literal meaning after linguistic completion or the speaker’s aims. However, before the completion, the transparency of the meaning of a proposition may vary from entirely understandable to utterly obscure, even when it is not figurative; therefore, what it needs is semantic disambiguation. To give an example of disambiguation, in “She has taken enough from you”, we do not know who the pronouns refer to, what is the synonym of “take” in this context (and it could be “suffer”, “tolerate”, “receive”, etc.) and what the whole situation is about. A complete proposition is thus not achieved by virtue of the linguistic meaning only as the sentence is semantically underdetermined (Bach 1994b: 125). Likewise, by saying “I haven’t eaten breakfast” (meaning “today”), one does not specify it, so some lexical (or rather conceptual) completion must be enacted (Bach 1994b: 135). In “Have you eaten”, analysed by Bach (1987: 72–82), on the other hand, the possible completion may involve “today”, “at all”, “breakfast”, etc. To compare, these propositions are more opaque (inexplicit, non-literal) than “Dogs are bigger than cats”, which is literal and explicit.

On the other hand, “Dogs are dearer than cats”, while being literal, introduces some ambiguity relative to the polysemous meaning of “dear”, which may stand for either “expensive” or “cherished” (Bach 1987: 72). Bach classified this example as literal, not explicit. Another case in point illustrating the haziness of the term literal and highlighting its gradability is seen in “I don’t like Mary” (meaning “I love Mary”). Bach (1987: 72) views this context as literal on the surface (lexical) level yet inexplicit at the utterance level, wherein the whole sentence is not to be treated literally. The inexplicitness is achieved by negating a weaker statement (Bach 1994b). If we transformed the sentence slightly by implementing scalar implicature into “I don’t dislike you” (meaning “I detest/hate you”), we would obtain an instantiation of ironic understatement (that is implicit context); specifically, a scalar reversal irony (Kapogianni 2014), wherein the offensive content is only partially contradictory. In sum, Bach proposes several stages of literalness, of which three are mentioned here: literal explicit, literal inexplicit, and non-literal inexplicit.

According to Bach (1994b: 135), “it is not necessary for the hearer to identify the exact words the speaker has in mind but only what those words would contribute if they were used”. Depending on how much must be replenished and how much ambiguity the

missing words entail, the cases that require expansion (as Bach dubbed this process) may be positioned in Fig. 1 somewhere between explicit and implicit categories, either as examples of indirectness, somewhere midway between indirectness and explicitness, or possibly even between indirectness and implicitness. The last case is occasioned by the following example (cited after Bach 1987). Therefore, it is impossible to assign all potential propositions to a single category on the literal-figurative scale or explicit-implicit.

From this discussion, it transpires that these terms are gradable phenomena. Gradability is not a new concept, as it has already been mentioned by a number of scholars (Östman 1987: 156; Yus 1999: 491; Haugh 2015: 78; Kecses 2017; Bączkowska 2022). Östman (1987: 156) speaks of “necessarily gradient (...) dichotomy” with reference to implicitness-explicitness. Bach (1994a) proved that a sentence might sound ambiguous (i.e., semantically underdetermined), i.e., inexplicit, even without resorting to figurativeness. The classical account of gradability relies on the neo-Gricean notion of implicature (Bach 1994) on the one hand and the post-Gricean notion of explicature (Sperber and Wilson 1986) on the other. Both are somewhere in between ‘what is said’ and ‘what is implicated’. They are positioned in this study between the explicit and the implicit, in the category dubbed indirectness/indirection.

Cases covered by enrichment are treated as part of *what is meant* (i.e., implicature) by Grice, as part of *what is said* (i.e., explicature) by post-Gricean scholars and contextualism (e.g., by Recanati 2004), and as an area in between *what is said* and *what is meant* (i.e., impliciture) by Bach (1994). For Sperber and Wilson (1986: 182), explicatures are “the development of the logical form”, and thus they are explicit, whereas, for Bach (1994), implicatures are essentially implicit rather than explicit even though they flow from *what is said*. The three theories argue in favour of a shift in how the concept of implicitness has evolved over decades: from *what is meant*, i.e., implicature proposed by Grice, through the middle ground disconnected from both *what is said* and *what is meant* propounded by Bach, to its full integration with *what is said* advocated by Sperber and Wilson as well as Recanati.

From the discussion thus far, it can be concluded that an essential part of the concept of implicitness boils down to the discussions on the *what is said* vs *what is meant* distinction. Following Bach (1994a), the enriched meaning of *what is said* (implicatures) is already at the implicit level of interpretation, even though it is beyond figurative reading. The discussion shows that the notion of implicitness is not just non-literal meaning, nor is it any meaning that is figurative, as there are stages between the two extremes. While in this study, and in Bączkowska et al. (2022), implicitness is deemed figurative meaning, what neo-Gricean and post-Gricean theories sensitise us to is that the notions of literal and implicit meaning are not discrete and clear-cut but gradable. Furthermore, the literal and figurative constitute two extremes on a continuum cline (as proposed by Giora 2003), with the middle part coaching overlapping cases. This is in line with Giora’s (1997; 2003) graded salience hypothesis, which assumes that an

utterance may be labelled literal, figurative or both. As not all figurative meanings are implicit (but still are non-literal), what is proposed in Fig. 1 is to use the blanket term ‘covert’ meaning, which engulfs both figurative readings, which are seen here as implicit and non-figurative interpretations that are not necessarily literal.

4 (In)directness

Indirectness is a term that has long attracted scholarly attention and has been widely used in pragmatics, which, however, has not been understood unanimously. Scholars attach disparate meanings to this phenomenon and are not consistent in its definitions and taxonomies alike (Wierzbicka 1985: 175; Haugh 2015: 39; Grainger and Mills 2016: 34; Decock and Depraetere 2018), which are not only “much too general” and vague (Wierzbicka 1985: 175) but sometimes even contradictory (Grainger and Mills 2016: 34).

Blum-Kulka (1987: 141) defines indirectness somewhat imprecisely as a proposition that conveys “something more or different from their literal meaning”³. Pinker (2007: 437) adds the hearer’s perspective to it and emphasises its non-literal reading by saying that indirectness is “the phenomenon in which a speaker says something he does not literally mean, knowing that the hearer will interpret it as he intended”. In a similar vein, Haugh (2015: 15) defines indirectness as a proposition that pivots on non-literary meanings, yet they are “contextually unambiguous” and thus are inferable. Sbisà (2007: 3) details that indirectness contains “presupposed and implied contents”. On the other hand, Lampert (2012: 181) focuses on its subtypes and defines it very broadly by claiming that indirectness “border[s] on (...) many topics, including euphemism, circumlocution, metaphor, irony, repression, parapraxis”.

Indirectness is claimed to be marked compared to direct speech (Terkourafi 2013: 202). This is so because indirectness results from some mental effort, strategic thinking, and encoding deniability options (Haugh 2015: 15), which in turn require making roundabout utterances that drift away from the Gricean maxim of clarity (Terkourafi 2013: 202). Strangely enough, despite the apparent uneconomical measures involved in indirectness, it may be more efficient than direct speech because indirectness creates “a shorthand for a conjunction of inferences” (Terkourafi 2013: 203). While it is generally believed that indirectness entails the recognition of implicatures to be inferred by the hearer from the speaker’s utterance, Terkourafi (2013) convincingly proves that one-to-one correspondence between the speaker’s (illocutionary) intent and the receiver’s inference is rarely the case as several possible readings may be often inferred. The essential role of the hearer in the process of meaning inference, or even marginalisation

³ Literal meaning has been defined by the classical definitions as context-free, context-invariant, direct, and truth-conditional (Ariel 2002).

of the speaker and his/her intent, is stressed by a number of scholars, particularly those working within the discursive paradigm (e.g., Haugh 2014; Grander and Mills 2016).

Furthermore, the discussion on indirectness may revolve around three research strands. First, the classical definition captured by Searle (1975) will be presented; second, the impoliteness take will receive some attention; third, it will be elaborated from the point of view of terms of address. These three perspectives are the possible ways of understanding indirectness which seem to dominate the literature.

Indirectness usually immediately triggers associations with Indirect Speech Acts (ISA), as proposed by Searle (1975). What Searle propounds is that the meaning of an utterance differs from what might be entailed by its form, and thus, the illocutionary force of an utterance is only indicated, that is implied and must be inferentially retrieved by the addressee (1975: 59). The example given by Searle (1975) to illustrate ISA is the oft-quoted *Can you pass me the salt*, wherein an interrogative takes the function of a request, and thus, the original function of questioning is assigned to the imperative. Bach (1994: 71) concedes that such indirect illocutionary acts draw on literal and non-literal meanings.

Another take of indirectness problematises it on the scale of (im)politeness (House and Kasper 1981; Trosborg 1995; Haugh 2015; Grainger and Mills 2016; Decock and Depraetere 2018: 5). It captures the concept of (in)directness in terms of implicitness vs explicitness, which they usually understand as scalable face-threats (that are wedded to perlocutionary acts), and thus they embed it in the concept of (im)politeness. In line with this reasoning, the more explicit an utterance, the more direct it is, and the greater the possibility of face-threat. Thereby, by saying, “There was nothing wrong with my car yesterday” or “There is a horrible dent in my car” (Trosborg 1995), that is by resorting to an implicit complaint, the perceived face-threat is weaker than by uttering the explicit and direct “How on earth did you manage to be so stupid” (and dent my car) (Trosborg 1995). According to Grainger and Mills (2016: 2), “Directness is (...) considered to be in essence impolite”. It is generally acknowledged in classical theories that indirectness increases the “degree of optionality” and thus may entail diminishing the tentative force of a proposition (Leech 1983: 108). Therefore, it may have a strong mitigating function and result in softening criticism, complaints, requests, compliments, etc., and so be more polite (Searle 1975: 36; Leech 1983: 108–109; Leech 1983: 169; Brown and Levinson 1987: 20). However, there are also voices, primarily in more recent discussions, that it may not always hold (Leech 1983: 171; Wierzbicka 1985; Marti 2006: 1850; Culpeper and Haugh 2014; Haugh 2015: 15–16; Kadar 2017: 194; Bączkowska 2022). For example, indirectness may be subject to culture specificity (Blum-Kulka and Olshtain 1994; Wierzbicka 1985; Spencer-Oatey 1992: 17; Ogiemann 2009) or, contrary to what is commonly assumed, may also function at “low-stake” situation, such as intimate talk scenarios (i.e., when not much risk is taken of being impolite), which is contrary to the “high-stake” interactions usually expected in indirect speech (Terkourafi 2013: 202). In ritual offence, on the other hand, the implicit offence may be tantamount to rudeness

(Kadar 2017: 194). The degree of (im)politeness is also contingent upon language conventions. As observed by Blum-Kulka (1987: 132), conventionally indirect requests (e.g., “Could you please close the window”) are usually correlated with politeness, yet non-conventionally indirect ones are not always polite (as in “Why is the window open?”).

Decock and Depraetere (2018: 8) further add, while analysing complaints, that what is expressed directly, i.e., on-record (straightforwardly) in the (im)politeness theories parlance, may still be enacted at varying levels of explicitness. On this view, the term indirectness relates to the linguistic realisation governed by the speaker’s linguistic choices on the explicit-implicit cline. Terkourafi (2013: 198, 205) holds that it is not always the case that the speaker has some (illocutionary) intention to be uncovered by the hearer and also that often there are more than one (collateral) implicatures affording inferences to be retrieved by the hearer. Culpeper and Haugh (2014: 199) lean more towards the receiver by stressing that (im) politeness “is in the eyes and ears of the beholder”, which is generally typical of most scholars representing the second (discursive) and the third wave in impoliteness studies. Decock and Depraetere (2018) hold that “The degree of linguistic (in)directness is (...) related to explicitness and implicitness.” By this claim, they seem to subscribe to the thesis that, firstly, (in)directness and (im)PLICITNESS are two separate concepts and, secondly, that (in)directness can be characterised in terms of explicitness and implicitness (which they see as gradable concepts).

Indirectness may be in tune with a concept of addressivity, that is, it relies on participation framework design (Boxer 1993; Bączkowska 2021), wherein directness entails addressing somebody straightforwardly, that is, by using the second-person singular (“you are stupid”, “you are beautiful”), whereas indirectness requires resorting to some other linguistic forms, such as third-person singular or plural (“some people can be rude” when addressing an individual) or citing others (“he says that you are stupid”). In studies on complaints, Boxer (1993) distinguishes direct complaints, wherein the person (or institution) accountable for the offence is directly addressed, and indirect complaints, wherein a third party is being addressed.

Finally, Kiesling and Johnson (2010: 296) propose four subtypes of indirection, which they dub stance, topic, participation, and production. Stance indirection occurs when one uses conventional forms in unconventional social situations, e.g., when “distancing or power-indicating forms are used to create solidarity”. Topic indirection requires a “shadow” topic to be parallel to the current topic. Participation indirection occurs when the recipient is not the addressee (he or she may be, for example, a ratified overhearer). Production indirection, in turn, assumes that the speaker changes his or her identity.

From the discussion thus far, it transpires that indirectness (sometimes also dubbed indirection) revolves around the question of im/politeness and/or addressivity, and therefore, it substantially departs from the notion of implicitness that is elaborated chiefly

in terms of semantic/pragmatic meaning. Consequently, in this paper, indirectness/indirection is not conceived of as tantamount to implicitness; the latter will be further discussed in the remainder of this paper in the context of computational methods of language analysis.

5 Corpus-assisted and computational studies of implicit offensiveness

Corpus-assisted studies and computational studies (such as word embeddings) can be applied to an analysis of language to deepen subtle differences between meanings based on the context they typically occur in, as well as for verification of linguistic assumptions and taxonomy proposals. Corpus-assisted studies are widely used in English linguistics, yet word embeddings are hardly explored. In this section, both types of computational examinations will be employed in order to investigate two categories widely recognised as examples of implicitness, i.e. “irony” and “sarcasm”. Whilst the main objective of this paper is to detail the notion of implicitness and other terms akin to it and, stemming from this, to offer a gradable cline-based typology thereof, the empirical study presented below is elaborated for illustrative purposes, with the aim of showing possible computational tools that can be employed to further investigate the notion of implicitness in offensive contexts.

First, the notion of embeddings will be defined, and its relation to a linguistic theory will be described. Lexical, i.e., non-contextual embeddings, are vector representations of word meanings. They show semantic cosine⁴ similarity existing between words classified as exemplifications of the pre-defined categories (sometimes called keywords). The numerical values of similarity computed by neural networks contribute to the generation of embeddings that deploy various models (e.g., RoBERT, HateBERT) with the use of some NLP techniques (e.g., word2vec, GloVe or FastText). The techniques resort to one of the two model architectures (bag-of-words or skip-gram⁵) and envisage word meaning as vector-space word representations, i.e., numerical *vectors of meaning* that are based on mathematical calculations. Thereby, words are converted into vectors (hence the name of the word2vec technique). These real-valued vectors are next used as input data processed by neural networks that generate output data in the form of embeddings. Word embeddings are claimed to be very efficient in capturing language

⁴ Cosine is one of the possible measures of similarity that show the angle between two vectors, it can have a value between 0 and 1, the closer the value is to 1 the higher similarity.

⁵ In the continuous bag-of-words (CBOW) architecture, the target word is predicted on the basis of contextual words which surround the word that are treated as input data, whereas in the skip-gram architecture the word at hand predicts the neighbouring words. CBOW is more efficient and works better for frequent words while skip-grams are better with the prediction of rare words and operate well on smaller datasets.

regularities, particularly at the syntactic and semantic levels (Mikolov et al. 2013). Whilst a lot depends on the specific type of data available in the dataset used to generate embeddings, general trends can be observed. Non-contextual word embeddings rely on distributional information about a word, and they predicate on the assumption that words which frequently occur in the same neighbourhood tend to be semantically similar; thus, a word sense is dependent upon other word senses. The higher the dependence between words, the more similarity they display. Put differently, “similar words are likely to have similar vectors” (Mikolov et al. 2013).

This computational hypothesis resonates with the linguistic hypothesis of distributional analysis (Harris 1954) and distributional semantics (Lenci 2008). Harris (1954: 43) claims that “difference in meaning correlates with difference in distribution” and that words are “determined relatively to each other, and on the basis of the distributional relations among them (Harris 1951: 7). Lenci (2008: 16) contends in turn that “the distributional behaviour of a word in contexts is not only taken as a way to get at its semantic behavior, but indeed as a way to explain its semantic content at the cognitive level”. In corpus linguistics, the distributional approach to semantics is known as the neo-Firthian approach, and it hinges on the theory propounded by Firth (1957) regarding phraseological units, of which collocations constitute a frequent object of research in the corpus era. Firth’s (1957) well-known definition of collocations is that they are defined through the linguistic environment, which was expressed by Firth’s (1957: 11) oft-cited statement that “you shall know a word by the company it keeps”. The neo-Firthian approach to language analysis is emergent and empirical, as opposed to system-based and introspective studies (more in Bączkowska 2018). Emergent meaning surfaces from contextualised examples, mainly from authentic language, such as one gathered in a corpus. Corpus-assisted studies thus capitalise on a usage-based approach to language analysis (i.e., one coming from empirical data) and on a bottom-up direction of analysis, specifically on the statistics of co-occurrence.

In corpus linguistics, semantic meaning identification pivoting on distributional calculations is widely used and easily available, for example, through the word sketch tool built into the Sketch Engine used in this study. The semantic meaning evidenced via statistical analysis extracted from language corpora adheres to an alternative understanding of word sense (i.e., other than the traditional lexicographic definitions), which can be seen as a “cluster of citations for a word”, wherein clusters are formed when words “exhibit similar patterning and meaning” (Kilgariff 2003: 1-2). In the case of polysemous words, in turn, lexical senses may be “disambiguated by representing its context as a vector and identifying which cluster centroid the vector is closest to” (Kilgariff 1997). The key thing in word sense identification is how the word relates to various contexts in which it appears as word sense derives from its contexts. Hence, word meaning is dynamic; its meaning, or rather meaning approximation, may undergo some modifications depending on the source texts used to extract statistical information about

the word, which may highlight its distinct salient features. In this view, word sense is a derivative of its combinatorial behaviour, i.e., dependence on other words.

Correspondingly, in non-contextual word embeddings, meaning emerges from real value vectors fed into neural networks, which learn on the basis of historical data and predict future distribution. The computation is distributed and parallel, that is, non-linear, as already voiced by Rumelhart and McClelland (1986) in their widely-known *Parallel Distributed Processing*. As a result, the output of the computation eventually generates vector space, which is a distributed, continuous word representation. The computation rests on various algorithms that a given network arrives at in the so-called hidden layer, which is between the input and output layer, and stores the results as node weights (expressed numerically). The algorithms used for a calculation depend on the adopted model, which allows predictions of word distribution obtained when the model is trained. The output of the trained model is a word embedding that shows distribution-based word clusters abstracted from a repository of word usage. Embeddings thus converge with current corpus-linguistic methods of word analysis insofar as both approaches to language analysis are computational and are conducted on the strength of word (distributed) similarities. By the same token, words are not treated as discrete, atomic and unrelated entities but rather as tightly connected webs of semantic relations that create distributed word meaning space.

This approach to semantic meaning is robust and objective, yet, as with any methodology, it has its drawbacks. The biggest weakness of NLP-based linguistic analyses is the fact that they ignore extralinguistic factors, at least to some extent, as the availability of contextual and situational clues is limited when pursued only in the written mode, viz. textual data (Lenci 2008). The distributional meaning is inevitably language intrinsic and statistical, yet, on a positive note, NLP-based investigations allow generalisations flowing from a significant amount of data and observations from a bird's eye view. Thereby, they may positively contribute to other (subjective, semantic-conceptual/cognitive, fully culture-embedded and situated) studies, and, as such, they are not at odds with the cognitive (Langacker 1987) approach to meaning but rather complementary.

6 Irony and sarcasm

From the linguistic research, we know that sarcasm is conceived of in various ways, for example as a subtype of irony (e.g., Gibbs 2000; Nunberg 2001; Kihara 2005; Alba-Juez and Attardo 2014: 100; Pexman et al. 2019), a superordinate term relative to irony (Camp 2012), a notion independent of irony (Fowler 1965: 535; Caucci and Kreuz 2012: 1; Garmedia 2020) or a single notion encompassing both terms (Attardo et al. 2003; Kruger et al. 2005). The concept of "irony" is described as an expression (a figure of speech) that encodes contempt and derogatory comments (Grice 1989[1978]: 53); thus, it

expresses the speaker's evaluative opinion, often based on meaning reversal and/or a clash with hearer's expectations (Partington 2007: 1548; Cutler 1974: 118; Kapogianni 2011: 56), either targeted at a particular receiver or an inanimate event or situation (Dews and Winner 1999: 1580). Contrary to "irony", "sarcasm" is claimed to be always targeted at a person (Kreuz 2020: 147), and it is believed to be much stronger than ironic comments, as the aim of resorting to a sarcastic remark is to hurt the target, ridicule them, manifest hostility via aggressive, acrimonious and pointed comment (Garmendia 2018: 129; Filik et al. 2019). Additionally, both terms may be used for humour (Roberts and Kreuz 1994; Partington 2006: 182).

Given these widely accepted general definitions of the terms in question, the borderline between an ironic and sarcastic comment is not easy to draw, and for this reason, the two notions were selected for further scrutiny. In the analysis which follows, some more light will be cast on the semantic meaning of both concepts by relying on features retrievable from corpus data and revealed by word embeddings to uncover some more details regarding facets either shared by the two notions or displaying clear contrasts.

7 Corpus data and embeddings in implicit offensiveness studies

7.1 Research aims

The aim of the study which follows is to (1) investigate potential similarities or differences in meaning and possibly a semantic overlap between and gradability of two terms, "irony" and "sarcasm", both of which are instantiations of implicit linguistic categories, and for this purpose a corpus-assisted study will be discussed; (2) check the relationship between the two notions immersed in a broader context of other implicit/covert terms as well as words that typically co-occur with them (primarily adverbs), and this will be shown by examining lexical embeddings. It is hypothesised that (a) given the discussions in relevant publications, the differences between *irony* and *sarcasm* will not be conspicuous, yet some distinctive facets are expected to be uncovered, particularly in a dedicated corpus of offensive language; (b) the embeddings should display the apparent close affinity of the two terms; (c) the words which often pattern with the two terms mentioned in relevant publications (e.g., in Seto 1998) will show close distance in the embeddings.

7.2 Data and methods

The data employed for the corpus-assisted analysis are of two types. Firstly, a general English corpus, 36 billion words in size, is used as a reference corpus, which makes it a sufficiently big corpus to observe the combinatorial behaviour of single words in the

general language. This is currently the biggest corpus of English available on the Sketch Engine (SE), which will be used in the study. The Sketch Engine is a commercial corpus management system where a number of existing corpora in many languages are available. The other corpus has been created for the purpose of this study, and it consists of existing datasets of various types of offensive language in English garnered from social media (Twitter, Reddit, forums, among others). The Corpus of Offensiveness (COFF) consists of 18 open-access datasets available online that deal with various types of offensiveness (dubbed by the authors as cyberbullying, hate speech, harassment, toxic language, etc.) compiled by computational scientists. They were manually annotated by the scholars who created them, but before the human evaluation, they were retrieved through automatic data extraction to use them as train and test data for machine learning experiments. For the purpose of this study, the datasets were uploaded to the SE and converted into a corpus; that is, they were tagged, lemmatised and parsed, and eventually, 18 constitutive subcorpora were created. The same datasets of offensive language were used to generate non-contextual (lexical) embeddings.

The methods used in this study are twofold. The computational (neural networks) experiment that generates word embeddings (already described in the previous section) and an analysis of corpus data that relied on the tools available in the SE platform that involve: the Word Sketch (WS) and visualisations based on the output, the SE Thesaurus, which displays automatically retrieved synonyms, i.e., words that are recognised by the system as being potentially synonymous (yet need not be such according to linguistic classifications), as well as similar words.

The two terms have received considerable interest in linguistic research, yet no decisive and unanimous definitions have been provided, illustrating clear differences between them.

7.3 A corpus-assisted study

A corpus-assisted sample analysis of implicit offensiveness presented below manifests the differences in the distributional meaning between two types of implicitness, which cause confusion in scholarly discussions, namely irony and sarcasm. The semantic meanings of “irony” and “sarcasm” presented below demonstrate the distributional behaviour of these terms in structures where they pattern with verbs, nouns and adjectives as well as with and/or phrases. The diagrams in Fig. 5 and Fig. 6, retrieved from the Sketch Engine (SE) tool Word Sketch, visualise the terms “irony” and “sarcasm” in the following structures:

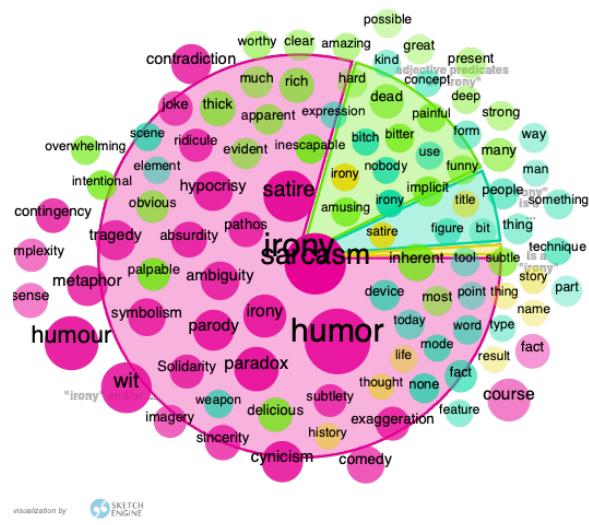


Figure 5: Word sketch of irony

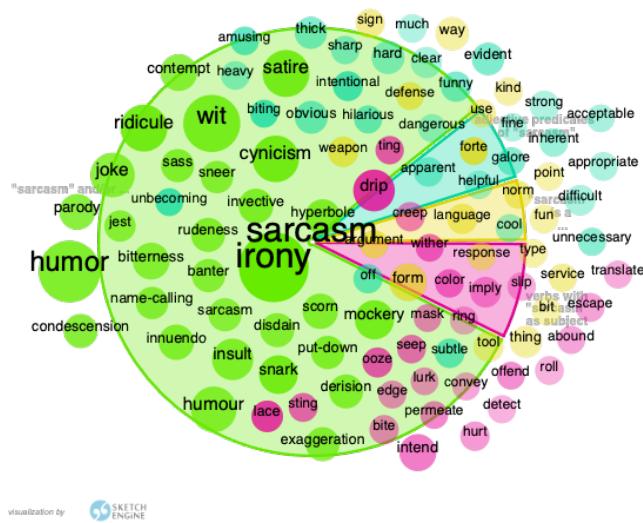


Figure 6: Word sketch of sarcasm

Is an irony: life, title, name, history, thing, story, thought

Is a sarcasm: post, paragraph, statement, sentence

Adjective predicates of irony: delicious, palpable, bitter, inescapable, amusing, funny, overwhelming, rich, intentional, subtle, apparent, painful, implicit, dead

Adjective predicates of sarcasm: evident, hard, thick, intentional, apparent, subtle, clear

Verbs with irony as subject: note, appreciate, miss, lose, recognise, enjoy, see, understand
Verbs with sarcasm as subject: drip, lace, ooze, sting, seep, hurt, creep, imply, bite.

Verbs with irony as object: about, escape, strike, surround, occur, lie, consider, seem, be

Verbs with sarcasm as object: drip, wither, detect, miss, intend, understand, ignore, avoid, note, love

“Irony” refers to abstract concepts, such as life or thought, while sarcasm refers to written discourse and social media (e.g., “post”, “sentence”). It is used in contexts showing both amusing meaning and painful readings, whereas “sarcasm” is reserved for clearly harsh cases. Finally, “irony” is expected to be enjoyed and appreciated, whereas “sarcasm” creeps and bites, stings, withers and hurts. Overall, in the general language data, sarcasm is used to encode stronger emotions and is expected to cause more hurtful results, which supports the definitions provided by linguistic publications.

Along with the Word Sketch option, the Thesaurus tool available on the SE allows noticing another interesting feature when “ironic” and “sarcastic” are examined in the Corpus of Offensiveness (COFF). Fig. 7 and Fig. 8 display the Thesaurus-based information gleaned from the COFF. To compare the results with the general corpus, Fig. 9 and Fig. 10 show the Thesaurus output from the enTenTen20 corpus.

The data sourced from the Thesaurus tool available on the SE essentially confirm what we know about the two adjectives, “ironic” and “sarcastic”, from relevant publications mentioned above, namely that “ironic” is much milder than “sarcastic”. The former pertains to words that belong to the categories indicating the following features encoded by adjectives: “cumbersome”, “amusing”, “tiresome”, “lax”, “speculative”, “questionable”, and “suspicious”. In more general terms, it can be said that they revolve around humorous features, clumsiness and strangeness. On the other hand, “sarcastic” triggers strong emotions and clear negativity, as the following categories suggest: “arrogant”, “aggressive”, “derogatory”, “rude”, “condescending”. Put differently, COFF demonstrates that “sarcastic” and words related to adjectives are understood as strongly offensive, while “ironic” is conceived of in a lighter sense and with jocular overtones. This study confirms the first hypothesis that the differences between the two terms are rather subtle yet discernable. However, the two terms certainly do not represent two separate linguistically verifiable notions; instead, a considerable overlap between them is clearly visible.

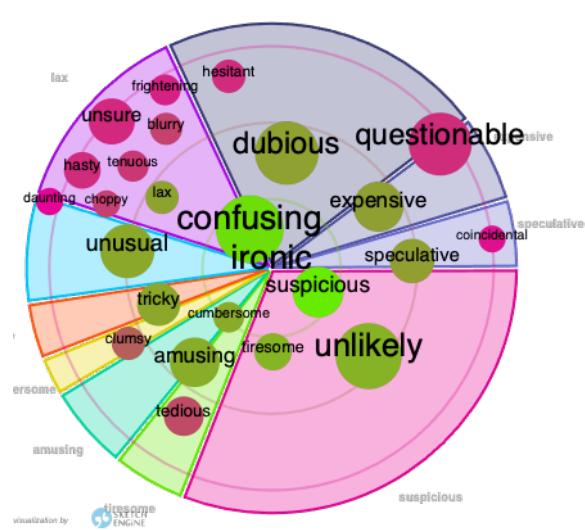


Figure 7: Synonyms of ironic in COFF (SE-Thesaurus)

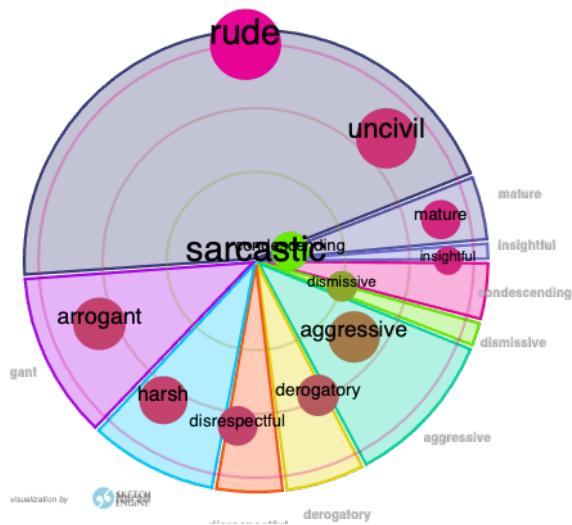


Figure 8: Synonyms of sarcastic in COFF (SE-Thesaurus)

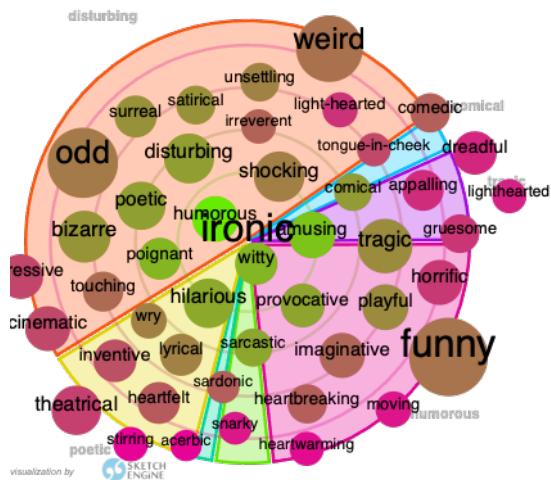


Figure 9: Synonym of ironic in enTenTen20 (SE-Thesaurus)

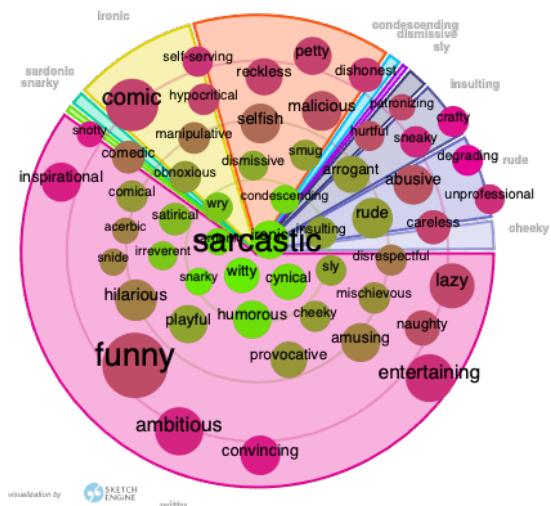


Figure 10: Synonyms of sarcastic in enTenTen20 (SE-Thesaurus)

Adjectives that are similar to “ironic” in the enTenTen20 corpus are, in turn, humour-oriented, with the groups congruent with “comical” (“comedic”, “light-hearted”) and “humorous” (“playful”, “funny”, “amusing”, “hilarious”), as well as words similar to “disturbing” (“surreal”, “satirical”, “shocking”, “weird”, “odd”, “poignant”) and “tragic” (“appalling”, “gruesome”, “dreadful”). Unlike in the COFF, there is additionally a group related to literary usage (poetic), which signals resorting to “irony” as an artistic device. Likewise, words identified as relevant for “sarcastic” in the enTenTen20 corpus are grouped mostly around predominant notions of “cynical” (“malicious”, “arrogant”, “obnoxious”) and witty/entertaining (“funny”, “hilarious”, “playful”). It is hard to notice a striking contrast between the two terms in this corpus of general English. In sum, the adjective *ironic* seems to carry stronger negative connotations in the general corpus (enTenTen20) than in the COFF data. Secondly, both “ironic” and “sarcastic” appear to have associations with wit and humour on the one hand and cynicism and arrogance on the other. The dissimilarity between the two terms is more clearly visible in the case of COFF data, wherein terms akin to “sarcasm” are more often reserved for hurting the target, while in the enTenTen20 data, the differences are less noticeable. From this, it transpires that social media users tend to polarise the two terms more clearly relative to general English.

7.3.1 Non-contextual embeddings

Whilst corpus data can prove helpful in discerning subtle differences in meanings, embedding can provide a more general perspective (a bird’s eye view) on the whole group of terms subsumed, in this case, by the category of implicitness. The non-contextual embeddings in Fig. 11 (Bączkowska et al. 2022) capture the distributed meaning of several implicit offensiveness categories and the relationships held among them. FastText [PCA⁶] shows that *irony* and *sarcasm* are moderately intertwined (contrary to what some scholars opt for), a stronger dependence is observable between “irony” and “metaphor”, and “metaphor” and “simile” are tightly interlocked. The concept of exaggeration is interconnected with two implicitness terms, i.e., “understatement” and “overstatement”, thus suggesting that it is an underlying notion in which the implicit terms are embedded. Furthermore, whilst “irony” tends to be merged with “metaphor”, which has been observed by a number of scholars (e.g., Partington 2006; Bączkowska 2022), sarcasm tends to combine with “understatement” and “overstatement”. A type of “overstatement” is hyperbole, which, as noticed by some scholars (e.g., Partington 2006: 12, 209; Burgers 2013: 15; Claridge 2011: 82), tends to conflate with irony. This feature is not shown by the embeddings, however, unless one assumes that social media users conceptualise “irony” in terms of “sarcasm”, which, in fact, is a claim put forward by a number of scholars (e.g., Attardo 2013: 40; Kapogianni

⁶ PCA (Principle Component Analysis) and t-SNE (t-distributed Stochastic Neighbour Embedding) are dimensionality reduction algorithms (van der Maaten and Hinton 2008).

2014: 635; Garmendia 2018: 128). As already mentioned above, the embeddings draw on social media datasets, and social media users are claimed to use the word “sarcasm” to mean “irony” and reserve “irony” only to one of its subtypes, namely situational irony (Attardo 2013: 40). This trend is evident in American English.

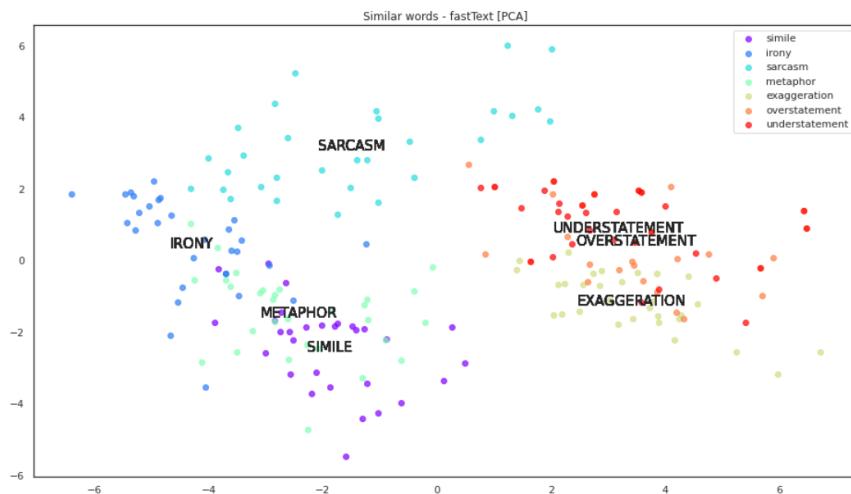


Figure 11: FastText [PCA]

Some types of implicit categories tend to pattern with specific nouns or modifiers. For example, “irony” is often premodified by “truly”, “sure”, “indeed”, “absolute”, etc., as observed by Seto (1998: 241–246). Fig. 12. (Word2Vec [PCA]) is a sample visualisation of select premodifiers and two nouns that tend to be used in ironic comments or metaphors. The embedding generated by Word2Vec [PCA] (Fig. 12) and other embeddings (e.g., FastText [t-SNE]) are suggestive of using “rather”, “never”, “always”, “truly”, and “absolute” in ironic comments. However, as evidenced by FastText[t-SNE], “rather” is a term most detached from both other modifiers and the covert categories. The metaphor goes with “genius”, “miracle”, and premodifiers “brilliant”, “absolute”, and “really”. Sarcasm patterns with “indeed”, “truly”, “certainly”, “evidently”, and “clearly”, while simile co-occurs with “real” and “indeed”.

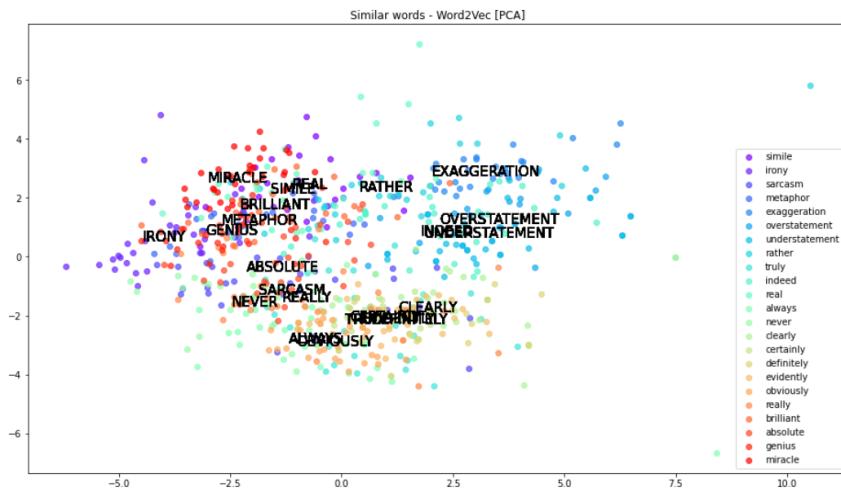


Figure 12: Word2Vec [PCA]

The analysis above shows that the concepts under scrutiny are far from discrete; in fact, they are blurry and tend to overlap, thus showing the gradability of their semantic meaning.

As already alluded to, embeddings are generated based on some data; thus, the output of the algorithms calculation by neural networks depends entirely on the type of data on the input end. The embeddings presented here draw on the social media dataset, and as such, they show the dependencies of the covert meaning terms in this specific type of discourse only. Notwithstanding the fact that obtaining different relationships among the terms in embeddings based on other types of data is highly probable, the embeddings presented above already manage to positively verify several features captured by the theoretical model proposed earlier (Bączkowska et al. 2022) as well as confirm some observations already voiced by scholars on conflated forms of implicit language and regarding the modifiers which tend to co-occur with the implicit/covert terms, and on the co-occurrence of certain premodifiers, nouns and adverbs.

8 Conclusions

Linguistically, the theoretical considerations presented in this paper have demonstrated that the concept of implicitness is not easy to grasp and define. The main difficulty in distinguishing implicit terms resides in the fact that they display gradability and are thus far from discrete terms. Consequently, the linguistic categories illustrating various forms

of opaque meaning (be it implicitness, covert meaning or non-literalness) tend to be fuzzy and are thus impossible to be neatly aligned in a clear-cut taxonomy. To reconcile the problem of fuzziness and gradability, it has been proposed in this paper to replace the concept of implicitness with the overarching notion of covert meaning, as not all inexplicit meanings are encoded by implicit forms (that is, figurative, rhetorical devices); some of them can be expressed by non-figurative forms that are, however, not literal. These are indirect forms that constitute a category staying between the literal and figurative. Moreover, figurative meanings can be explicit, even though prototypically, they are implicit.

Methods of analysis of textual data relying on corpus-assisted and computational (non-contextual embeddings) tools proved supportive in disentangling the complex meanings of non-explicit categories and words as well as visualising the relations held among them. The analyses based on corpora of general and offensive language have demonstrated that while jocularity, wit and contempt are typical of the notions of “irony” and “sarcasm” in both types of corpora, in the case of offensive data, the dissimilarity between “irony” and “sarcasm” is more conspicuous, with negativity being expressed more strongly by the latter. This study thus highlights the importance of dataset selection and its possible influence on the output data and their interpretation. Subtle insights into single concepts observable in corpus data can be supported by non-contextual embeddings that show a broader perspective and allow noticing relationships held among a number of clustering terms subsumed by a more general category of varying degrees of opacity. Lexical embeddings of implicit terms presented in this study have demonstrated that some terms tend to co-occur, which supports linguistic claims regarding the frequent incidence of merged forms, such as a combination of irony and understatement, irony and metaphor, etc. Modifiers (mostly intensifiers), which are claimed to be typical of some implicit terms (such as “totally”, “a bit”, etc.), have also occurred in our offensive data visualisations.

Nonetheless, based on the investigation presented in this paper, neither corpus data examination nor lexical embeddings managed to support the claim made by some scholars that sarcasm and irony should be conceived of as two independent concepts. On the contrary, they appear to manifest gradability and fuzzy borders, tend to overlap semantically and shift their meaning depending on the types of data used; in this case, the shift was observed between general English data and social media offensiveness data in English.

Acknowledgement

This study was conducted as part of a project on offensive language within the NexusLinguarum Cost Action (CA 18209). I want to thank Dr Slavko Žitnik for his kind permission to use the word embeddings cited in this paper.

References

- Ariel, Mira. 2002. The demise of a unique concept of literal meaning. *Journal of Pragmatics* 34. 361–402.
- Bach, Kent. 1987. *Thought and Reference*. Oxford: Oxford University Press.
- Bach, Kent. 1994a. Semantic Slack: What is Said and More. In Savas L. Tsohatzidis (ed.), *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, 267–291. London: Routledge.
- Bach, Kent. 1994b. Conversational implicature. *Mind and Language* 9. 124–162.
- Bach, Kent. 1999. The myth of conventional implicature. *Linguistics and Philosophy* 22. 327–366.
- Bączkowska, Anna. 2019. Funkcjonalna analiza zbitek leksykalnych w dyskursie medycznym. *Conversatoria Linguistica* XII. 7–24.
- Bączkowska, Anna. 2021. “You’re too thick to change the station” – Impoliteness, insults and responses to insults on Twitter. *Topics in Linguistics* 22(2). 62–84. DOI: <https://doi.org/10.2478/topling-2021-0011>.
- Bączkowska, Anna. 2022. Explicit and implicit offensiveness in dialogical film discourse in *Bridgit Jones* films. *International Review of Pragmatics* 14(2). 198–225.
- Bączkowska, Anna, Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedre Oleskeviciene Valunaite & Marcin Trojsczak. 2022. Implicit offensive language taxonomy and its application to automatic extraction and ontology. Paper presented at LLOD Approaches to Language Data Research and Management, Mykolas Romeris University in Vilnius, 21–22 September.
- Benikova, Darina, Michael Wojatzki & Torsten Zesch. (2018). What does this imply? Examining the impact of implicitness on the perception of hate speech. In Georg Rehm & Thierry Declerck (eds.): *Language Technologies for the Challenges of the Digital Age*, GSCL 2017, 171–179. DOI: https://doi.org/10.1007/978-3-319-73706-5_1.
- Blum-Kulka, Shoshana. 1987. Indirectness and Politeness in Requests: Same or Different? *Journal of Pragmatics* 11(2). 131–146.
- Bousfield, Derek. 2008. *Impoliteness in Interaction*. Amsterdam: John Benjamins.
- Brown, Penelope & Stephen Levinson. 1987. Politeness: Some universals in language usage. Cambridge: Cambridge University Press.
- Carston, Robyn. 1988. Implicature, Explicature and Truth-theoretic Semantics. In Ruth M. Kempson (ed.), *Mental Representations: The Interface between Language and Reality*, 155–181. Cambridge: Cambridge University Press.
- Croft, William & Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press.
- Culpeper, Jonathan. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics* 25(3), 349–367.
- Culpeper, Jonathan & Michael Haugh. 2014. *Pragmatics and the English Language*. London: Palgrave.
- Decock, Sofie & Ilse Depraetere. 2018. (In)directness and complaints: A reassessment. *Journal of Pragmatics* 132. 33–46.
- Devitt, Michael. 2021. *Overlooking Conventions*. Cham: Springer.
- Jay, Timothy. 1992. *Cursing in America*. Amsterdam: John Benjamins.
- Gibbs, Raymond. 1993. Process and products in making sense of tropes. In Andrew Ortony (ed.), *Metaphor and thought* (2nd eds.), 252–276. New York: Cambridge University Press.

- Gibbs, Raymond & Herbert Colston. 2012. *Interpreting figurative meaning*. Cambridge: Cambridge University Press.
- Gibbs, Raymond & Jennifer O'Brien. 1991. Psychological aspects of irony understanding. *Journal of Pragmatics* 16. 523–530.
- Giora, Raymond. 1997. Understanding Figurative and Literal Language: The Graded Salience Hypothesis. *Cognitive Linguistics* 8. 183–206. DOI:10.1515/cogl.1997.8.3.183
- Giora, Raymond. 2003. *On our mind: Salience, context, and figurative language*. New York: Oxford University Press.
- Goddard, Cliff. 2015. “Swear words” and “curse words” in Australian (and American) English. At the crosswords of pragmatics, semantics and sociolinguistics. *Intercultural Pragmatics* 12(2). 189–218. DOI: <https://doi.org/10.1515/ip-2015-0010>.
- Grainger, Karen & Sara Mills. 2016. *Directness and Indirectness Across Cultures*. London: Palgrave.
- Grice, Paul H. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press
- Haugh, Michael. 2014. *Im/politeness Implicatures*. Berlin: Mouton de Gruyter.
- Kapogianni, Eleni. 2014. Differences in use and function of verbal irony between real and fictional discourse: (Mis)interpretation and irony blindness. *Humor* 27. 597–618.
- Kecskes, Istvan. 2017. Implicitness in the use of situation-bound utterances. In Piotr Cap & Marta Dynel (eds.), *Implicitness: From Lexis to Discourse*, 201–215. Amsterdam: John Benjamins.
- Kihara, Yuka. 2005. The mental space structure of verbal irony. *Cognitive Linguistics* 16(3). 513–530.
- Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31. 91–113
- Kurzon, Dennis. 2017. Thematic silence as a speech act. In Piotr Cap & Marta Dynel (eds.), *Implicitness: From lexis to discourse*, 217–234, Amsterdam: John Benjamins.
- Leech, Geoffrey. 1983. *Principles of Pragmatics*. London: Routledge.
- Lakoff, Robin. 1973. The logic of politeness; or minding your p's and q's. In Claudia Corum, Thomas Cedric Smith-Stark & Ann Weiser (eds.), *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, 292–305. Chicago: Chicago Linguistic Society.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović & Giedre Valunaite Oleskeviciene. 2021. LOD-connected offensive language ontology and tagset enrichment. In Sara Carvalho & Renato Rocha Souza (eds.), *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*, 135–150. Zaragoza, Spain: CEUR Workshop Proceedings.
- Marti, Leyla. 2006. Indirectness and politeness in Turkish-German bilingual and Turkish monolingual requests. *Journal of Pragmatics* 38. 1836–1869.
- Mikolov, Tomas, Wen-tau Yih, & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT 2013*, 746–751, Atlanta, Georgia, 9–14 June 2013.
- Östman, Jan-Ola. 1986. *Pragmatics as implicitness: an analysis of question particles in Söf Swedish, with implications for the study of passive clauses and the language of persuasion*. Unpublished PhD Dissertation: UC Berkeley
- Pinker, Steven. 2007. The evolutionary social psychology of off-record indirect speech acts. *Intercultural Pragmatics* 4(4). 437–461.
- Rumelhart, David E. & James McClelland. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press.

- Sanchez, Liliana M. & Carl Vogel. 2013. IMHO: an exploratory study of hedging in web forums. In *Proceedings of the SIGDIAL 2013 Conference, Metz, France*, 309–313. Association for Computational Linguistics.
- Searle, John R. 1975. Indirect speech acts. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics*, Volume 3, *Speech Acts*, 59–82. New York, NY: Academic Press.
- Sperber, Dan & Deidre Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.
- Terkourafi, Marina. 2013. Re-assessing the speech act schema: twenty-first century reflections. *International Review of Pragmatics* 5. 197–216.
- Trosborg, Anna. 1995. *Interlanguage Pragmatics. Requests, complaints and apologies*. Berlin: Mouton de Gruyter. DOI: <http://dx.doi.org/10.1515/9783110885286>.
- Wierzbicka, Anna. 1985. Different cultures, different languages, different speech acts: Polish vs. English. *Journal of Pragmatics* 9(1–2). 145–178.
- Yus, Francisco. 1999. Misunderstanding and implicit/explicit communication. *Pragmatics* (9)4. 487–517.

About the Author

Anna Baczkowska, Dr Habil. Prof. UG, holds MA in English Philology, which she received from Adam Mickiewicz University in Poznan, as well as PhD in linguistics and D.Litt. in English Linguistics, which she received from the University of Lodz. Her research interests revolve around translation studies (film subtitles), cognitive semantics, corpus and computational linguistics, and discourse studies (media discourse). She has guest lectures in Italy, Spain, Portugal, UK, Norway, Kazakhstan and Slovakia, and she has also conducted her research during her scientific stays in Ireland, Iceland, Norway, Austria and Luxembourg.

Address

Institute of English and American Studies, University of Gdańsk
Wita Stwosza 51
80-308 Gdańsk, Poland

e-mail: anna.baczkowska@ug.edu.pl
ORCID: 0000-0002-0147-2718

Kristina Š. Despot
Institute for the Croatian Language

Ana Ostroški Anić
Institute for the Croatian Language

Tony Veale
UCD School of Computer Science, University College Dublin

“SOMEWHERE ALONG YOUR PEDIGREE,
A BITCH GOT OVER THE WALL!”

A PROPOSAL OF IMPLICITLY OFFENSIVE LANGUAGE TYPOLOGY

Abstract

The automatic detection of implicitly offensive language is a challenge for NLP, as such language is subtle, contextual, and plausibly deniable, but it is becoming increasingly important with the wider use of large language models to generate human-quality texts. This study argues that current difficulties in detecting implicit offence are exacerbated by multiple factors: (a) inadequate definitions of implicit and explicit offense; (b) an insufficient typology of implicit offence; and (c) a dearth of detailed analysis of implicitly offensive linguistic data. In this study, based on a qualitative analysis of an implicitly offensive dataset, a new typology of implicitly offensive language is proposed along with a detailed, example-led account of the new typology, an operational definition of implicitly offensive language, and a thorough analysis of the role of figurative language and humour in each type. Our analyses identify three main issues with previous datasets and typologies used in NLP approaches: (a) conflating content and form in the annotation; (b) treating figurativeness, particularly metaphor, as the main device of implicitness, while ignoring its equally important role in the explicit offence; and (c) an over-focus on form-specific datasets (e.g. focusing only on offensive comparisons), which fails to reflect the full complexity of offensive language use.

Keywords

implicit offence, implicitly offensive language typology, metaphor, humour, impoliteness triggers

1 Introduction

In our digitally-connected societies, we now have greater access to communication platforms that offer faster and more pervasive modes of linguistic expression. This increased freedom of expression allows marginalized groups to voice their concerns and

opinions, but it has also fostered an increase in public displays of aggressive and harmful language, such as through verbal aggression, taboo-breaking, hate speech, bullying, harassment, and badgering (Patrick and Hollis 2018). These egregious forms of speech can be seen as a violation of social norms of communication, particularly of the Gricean cooperative principle (Grice 1975). Although these behaviours are scarcely new to modern culture, and have always existed in various guises throughout history, the widespread availability of social media technologies expands their reach and magnifies their effects, which has led to increased efforts to mitigate their worst outcomes. Offensive language, whether explicit or implicit – through euphemism, metaphor, or “dog whistles” – can cause emotional distress, incite hatred, damage relationships, and escalate conflicts to the point of physical violence.

Offence takes root in many forms, and so computational approaches to its automated detection speak of different forms in different terms, whether *incivility* (Sadeque et al. 2109; Stoll et al. 2020), *toxic language* (Kunupudi et al. 2020), *abusive language* (Caselli et al. 2020; Waseem et al. 2017; Wiegand et al. 2021), *offensive language* (Zampieri et al. 2019) or *hate speech* (Gao et al. 2017; ElSherief et al. 2021; Schmidt and Wiegand 2017). Nonetheless, there is general agreement that offence through language has both explicit and implicit forms, and that its target may be quantifiable or amorphous, as either a specific individual or a group. Lexical (and lexicon-based) approaches to detecting explicitly offensive language have made headway by focusing on the specific words, and they are a reliable feature of many NLP (natural language processing) systems in many languages (e.g. Anand et al. 2023; Kogilavani et al. 2023; Roy et al. 2022; Shi et al. 2022; Sigurðarsson and Derczynski 2023; Trandabăt et al. 2022, etc.). However, the detection of implicit offence has proven far more challenging (see Lewandowska-Tomaszczyk et al. 2023b; Plaza-del-Arco et al. 2022; Waseem et al. 2017, etc.) not least because this kind of offence is designed to toy with our own internal censors, never mind those of machines. Explicit offence wears its intentions openly; implicit offence requires a typology of forms and methods to better understand its hidden intentions.

The main goal of this paper is to provide a detailed qualitative analysis of the types of implicitly offensive language (IOL) in real language use, following (Wiegand et al. 2021: 584) who stated that “although generic supervised methods may always represent a good baseline, the community should also be open that other more linguistically informed approaches can be more effective for particular subtasks in the detection of implicitly abusive language.” We first survey how related work has defined explicit and implicit offensive language, and examine existing typologies for classifying the latter for computational applications (Section 2). In Section 3, we suggest an operational definition of implicit offence and propose a new typology to organize its various types. Each type of offence in the typology is illustrated with examples from our dataset, the analysis of which will highlight the role of figurative language and humour when conveying offence. We draw attention to aspects that prove challenging for machine-learning approaches to detection, and suggest ways in which these challenges can be mitigated with finer-grained

typologies and datasets. Finally, in Section 4, we probe one of the public-facing LLMs (ChatGPT) and we analyse its output to check whether it can detect and grasp the full richness of the ways these devices are exploited in our data.

2 Related work

2.1 An overview of the definitions of implicitly offensive language in NLP research

The relation of explicitness and implicitness in analysing offensive language has been discussed so far from computational linguistic (e.g. ElSherief et al. 2021; Ocampo et al. 2023; Waseem et al. 2017) and from pragmatic perspectives (e.g. Assimakopoulos et al. 2017; Baczkowska 2022 among the most recent ones). Despite the fact that both approaches recognize the importance of figurative language, metaphor and simile in particular, a thorough figurative language analysis of the phenomenon has not been previously performed.

Waseem et al. (2017) were among the first to call for more work to be done on the detection of implicit abuse in language, which they define as “that which does not immediately imply or denote abuse” (Waseem et al. 2017: 79). In contrast, explicit language is “unambiguous in its potential to be abusive” (*ibid*), as is evident from e.g. racial or homophobic slurs. Kumar et al. (2018) prefer the terms *overt* and *covert aggression*. Their approach is one of many that makes a formal distinction between both types of abuse by reference to a lexicon of offensive words. If a text contains “more directly offensive words than potentially offensive words” (Kumar et al. 2018: 7), it is marked as overtly aggressive, while texts containing some potentially offensive words are marked as covert aggression. Gao et al. (2017) identify circumlocution, metaphor, and stereotypes as specific linguistic devices for conveying ill will through implicit hate speech, “in which hatefulness can be captured by understanding its overall compositional meanings” (2017: 775).

Caselli et al. (2020) aim for a subtler differentiation between explicit and implicit abuse and define implicit abuse as that which can only be “suggested or inferred” because “it does not have any surface evidence” (2020: 6197). For these authors, explicit abuse is manifested by means of “profanity, performative constructions, imperatives, idioms, adjectives, or nouns with a clear negative connotation.” In contrast, the devices used to veil abuse in implicitness include “sarcasm, metonymy, irony, litotes, euphemism, and inside jokes” (*ibid*). Although *negative connotation* is not otherwise defined, we infer from the given examples that it relates to words and expressions that convey a critical opinion or attitude about a target.

ElSherief et al. (2021) define implicit hate speech as “a subclass of hate speech defined by the use of coded or indirect language such as sarcasm, metaphor, and circumlocution.” Similarly, in their typology of implicit abusive language, Wiegand et al. (2021) also aim

for a practical definition by understanding implicit abusive language as an offence that is “not conveyed by (unambiguously) abusive words (e.g. dumbass, bimbo, scum)” (2021: 576). By lexicalizing explicitness, they consign critical statements whose abuse is only apparent at the phrase level to the class of implicit abuse, such as “I haven’t had an intelligent conversation with a woman in my whole life.” (Wiegand et al. 2021: 576), “Women should all stay at home in an apron, chained to the oven!” and “Get back in the kitchen” (Wiegand et al. 2021: 578).

From a computational perspective, explicit abuse is therefore the kind of offence that easily can be captured with a “blacklist” of undesirable terms. (We note that the term *blacklist* is now itself a proscribed term, with *blocklist* becoming the preferred alternative.) But the fragility of this perspective has long been understood by comedians. In his routine “The seven words you can’t say on TV”, George Carlin famously highlighted the existence of “two-way words” that have both acceptable and unacceptable uses. As he put it in his characteristically salty manner, “You can prick your finger … Just don’t finger your prick” (see Veale 2021).

2.2 Existing typologies of implicitly offensive language in NLP

From a computational perspective, two recent typologies of implicit offence are particularly relevant for our effort of defining the phenomenon. The first is by Wiegand et al. (2021), who analysed eight English datasets as part of a binary classification of micro-posts as abusive or otherwise. These authors calculated the proportion of explicit abuse by detecting the presence of abusive words (which is the term they use) from the lexicon of Wiegand et al. (2018). After manually annotating a random sample of 500 implicitly abusive micro-posts from each dataset, they proposed eight subtypes to characterize the implicit abuse in the data: 1) stereotypes; 2) perpetrators, a “morally-contemptible” sibling of stereotypes to which egregious and often criminal behaviours are commonly (if unfairly) attributed (Wiegand et al. 2021: 578); 3) abusive comparisons; 4) dehumanization, or the “act of perceiving or treating people as less than human” (2021: 579) by means of unfavourable comparisons with animals and other non-human tropes; 5) euphemistic constructions, often expressed using negation, such as “I wouldn’t touch [target] with a ten-foot pole;” 6) calls for action, where an author appeals for violence against the target); 7) multimodal abuse, i.e. abusive content presented via non-textual elements (e.g., visual memes, emoji); and 8) phenomena requiring world knowledge and inferences, which uses inference over shared knowledge in jokes, sarcasm, rhetorical questions, topical allusions and so on to communicate a criticism that is implied by a speaker and inferred by the reader.

When building their Twitter dataset of implicit hate speech, ElSherief et al. (2021) applied both fine-grained labels and, significantly, free-text descriptions of the implied aspects in each post. Their taxonomy of labels defines six classes of implicit hate speech that are “grounded in the social science literature” (ElSherief et al. 2021: 346). They define the class *White Grievance*, as evidenced in the tweet “Black lives matter and white lives don’t? Sounds racist,” as “frustration over a minority group’s perceived privilege and

casting majority groups as the real victims of racism” (2021: 347). The class *Incitement to Violence* includes displays of in-group unity and out-group rejection in the service of hateful ideologies, such as Nazism. The class *Inferiority Language* relates to acts of dehumanization and toxic behaviour in the furtherance of power disparities, while the class *Irony* unites sarcasm, satire, and any form of humour that is used to “attack or demean a protected class or individual” (2021: 347). The class *Stereotypes and Misinformation* labels instances of hate speech that exploit and reinforce the negative qualities unfairly ascribed to a protected class, while *Threatening and Intimidation* denotes the speaker’s willingness to embrace violence or other, subtler forms of intimidation.

These classes are used to label tweets in a second pass of annotation, after an initial pass has determined whether a given text is an example of explicit hate, implicit hate, or no hate at all. The methodology, which uses crowd-sourced annotations for this two-stage process, can be replicated on other datasets, and its work products can be used in the supervised training of machine learning models. The six classes, which are not mutually exclusive, are chosen to model the hate ideologies that are most prevalent in the US, and must likely be expanded to cover the kinds of implicit speech one might encounter in other cultural contexts and datasets.

3 A qualitative analysis and a typology of implicitly offensive language

The above-mentioned typologies emerge from a mix of top-down and bottom-up approaches, in which classes of implicit offence are postulated on the basis of their utility in previous work, or on their observed frequency in the datasets being analysed (e.g. ElSherief et al. 2021; van Aken et al. 2018; Wiegand et al. 2021). The classes are typically considered self-evident, and detailed origin stories for each class are often omitted. Moreover, even though some classes are shared by different approaches, the typologies of these approaches can still differ considerably.

3.1 An operational definition of IOL

This brief survey shows that researchers can also differ widely in their working definitions of implicit offence, and obviously, underspecified or unclear definitions lead to confused and unreliable annotations.

Moreover, the concept of *offence* itself poses definitional problems. There is no clear definition of this concept which would be undebatable. Since our annotation task was a part of a larger task of building offensive language taxonomy within COST Action *Nexus Lingarum* (see Section 3.2. and Lewandowska-Tomaszczyk et al. 2021), we have inherited the approach to (explicitly) offensive language defined within this task. Following the interpretation of offence as is presented in (Lewandowska-Tomaszczyk et al. 2023a), constructed according to the definition of impoliteness by (Culpeper 2005: 38), we

understand offence as an act of the language user (speaker) communicating face-attack intentionally, or the addressee (hearer) perceiving and/or constructing behaviour as intentionally face-attacking, or “else when a combination of intentional face attack and its perception by the hearer (audience) as such occur at the same time (Lewandowska-Tomaszczyk et al. 2023a: 9)”. Such behaviours, most often accompanied by the use of derogatory language, always have, or are presumed to have, cognitive and emotional consequences for at least one participant, frequently for a group, that is, they cause or are presumed to cause offence (Lewandowska-Tomaszczyk et al. 2023a).

Earlier definitions of IOL fall broadly into two groups: those driven by emotional-valence (is the negative valence unambiguous and explicit, or deniably implicit?), and vocabulary-driven (the overt presence of offensive words makes an offensive intent explicit, while an offensive intent that avoids any such words remains implicit, e.g. Wiegand et. al. 2021).

The valence of a text is not always self-evident, and it can be hard to achieve a satisfying level of agreement between different annotators. We thus rely on a vocabulary-driven definition that is more amenable to manual annotation and to machine-learning tasks. In this view, an explicitly offensive example is one that contains words designated as offensive in a dictionary, and that are undoubtedly aimed at the target. Conversely, if a text does not use a single lexical item that unambiguously denotes hatred or malicious intent towards a target, then any offence must be implicit. A basic threshold must be passed, then: a text must have at least one lexical item that, according to the relevant dictionary, has at least one sense marked as offensive, insulting, vulgar, or similar, and this exact sense must be used in the example.¹ Implicitness in this sense is not to be equated with indirectness (Searle 1979), but as will be evident from the analysis of various examples, they are tightly connected ideas since, in many cases, implicitly offensive language makes use of indirect speech acts, presuppositions and implicatures.

3.2 The dataset and methods

We begin by annotating texts that have already been pre-labelled as implicitly offensive. We then classify these short texts into different types of offensiveness via an in-depth qualitative analysis of each one. The qualitative analysis is performed within the framework of figurative language analysis used in Dancygier and Sweetser (2014) and Despot et al. (2019), taking into consideration various theories of humour surveyed in Veale (2021). We occasionally make references to the silencing theory (Caponetto 2021), and to the impoliteness theory (Culpeper 2011, 2016) when the analysed examples of what we consider to be implicit offence are structurally similar to various triggers of impoliteness (i.e. conventionalized impoliteness formulae). The resulting typology proposal emerges

¹ We are thankful to Luka Terčon for his help with this definition.

bottom-up from the types of implicit offence found in the examples when they are viewed through these various lenses.

We propose a two-level typology that differentiates between the content of an offence (aggressive speech, insulting speech, discrediting/condescending speech, dehumanization, derogation, and the discriminatory use of stereotypes, racism, sexism, etc.) and its form, i.e. the linguistic devices used to achieve it (metaphor, metonymy, simile, irony, hyperbole, euphemism, repetition, rhetorical questions, circumlocution, name-calling, contrasts, graphic devices, etc.). Unlike other typologies used so far in NLP (ch. 2.2.), we do not conflate the device used in a text with its ultimate intent. We believe this to be important because devices such as metaphor and simile, and even irony, can be used with or without offensive intent. Indeed, even aggressive speech can be used for self-mockery or harmless parody. Metaphor, for instance, can be used to elevate or to diminish.² We use it to personify the non-human and to dehumanize our fellow humans. Its myriad uses may be subtle or crass, implicit or explicit. It would be inappropriate, then, to consider dehumanization and metaphor as sibling classes in any taxonomy. Rather, metaphor is just one means by which dehumanization is achieved, and we tried to be careful to separate the means from the ends in our proposed typology of offence. This was not always an easy task, as is usually the case with any attempt at categorization – clear delineations are, of course, never to be expected considering the prototype theory (e.g., Lakoff and Johnson 1999).

For this study, we have used the data that had been pre-annotated within the WG 4 of the COST Action CA 18209 – *European network for Web-centered linguistic data science Nexus Linguarum* (Use Case 4.1.1. *Incivility in Media and Social Media*). Within this use case, 30 previously annotated offensive language datasets were used to annotate randomly extracted documents by applying a more detailed typology for offensive language annotation (Lewandowska-Tomaszczyk et al. 2021). A total of 331 documents, containing 1264 sentences, were annotated at multiple levels by two annotators and one curator, including for the offence target (whether the offence is aimed at an interlocutor, at a third party, or at a generalized other or group). The first annotation level marks whether the intent of an offensive utterance appears to be expressed explicitly or implicitly. We then collected 200 sentences that had been marked as implicitly offensive in this way by at least one annotator. A qualitative analysis of these sentences was then performed, and the dataset was then annotated by two other annotators for the nature of the offence and for the linguistic device that was used to convey it³. Both annotators compared their results, discussed points of disagreement, and agreed on common labels for examples that were similar in content. The complete dataset is available from <https://tinyurl.com/ymaff9pu> to view and download. As the dataset is limited in size, the study does not contain a

² For an overview of metaphor research in the field of cognitive linguistics see Dancygier and Sweetser 2014.

³ Both annotators are experienced in metaphor annotation tasks (see Despot et al. 2019).

quantitative aspect, but the focus is rather on an in-depth qualitative analysis the results of which can yield insights for future NLP tasks.

In the next sections, we propose a typology of implicitly offensive language, created on the basis of a qualitative analysis of the annotated data. Its application is currently being tested in the process of manual annotation of the LGBT related discourse within the FRENK dataset (Ljubešić et al. 2021).

Finally, we probed one of the public-facing LLMs (ChatGPT 3.5) to check whether it can detect and grasp the full richness of the ways these devices are exploited in our data.

3.3 Types of implicitly offensive linguistic units derived from the data

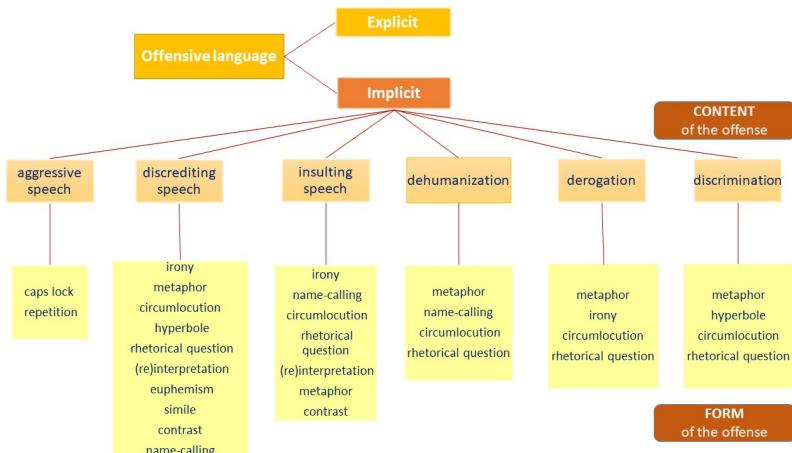


Figure 1: Types of implicitly offensive linguistic units derived from the data

3.3.1 Aggressive speech

In its implicitly offensive form, aggressive speech is the textual equivalent of yelling, hissing and snarling. It is marked by its confrontational tone and its intent to intimidate. This type of language is often used to assert dominance or to express anger, frustration, or dissatisfaction.

In our dataset, aggressive speech is realized through linguistic features such as repetition (example (1) below), all-caps spelling (examples (3), (4), and (5), below), and ill-tempered rhetorical questions that often carry veiled threats (example (1), (3) and (4)). A sense of intensity and urgency is created using punctuation (example (2)) and repetition (example (1)) – particularly of vowels (example (2)), in ways that can be interpreted as

impatient and impolite. Conversely, the use of all-caps is the textual equivalent of shouting to loudly and aggressively emphasize a point.

- (1) Is humiliation or hurt a likely response to that, Malleus? Huh? Huh?
- (2) Noooo!!! What a shock!
- (3) Should I dismiss EVERYTHING because Hitler bad?
- (4) Does any OTHER admin want to come and delete this idea for the sake of some shallow excuse like a banned user?
- (5) Dodo never, in detail, responded to my repeated and DETAILED complaints of his behavior.

Example (3) aims to counter a reference to Hitler with an aggressive form of prototype-based reasoning (Lakoff 1987). In example (4), in addition to aggressive speech, the metaphorical expression *some shallow excuse* achieves a discrediting effect through the conceptual metaphor UNDERSTANDING IS DEPTH.

3.3.2 Discrediting/condescending speech

As a form of implicit offence, discrediting speech aimed at an interlocutor/target attempts to undermine the credibility, competence, or integrity of their actions or words by implying that they are not intelligent, reasonable, well-mannered, trustworthy, reliable, consistent and so on. The effect can be manifested in various linguistic modes, from irony and sarcasm to rhetorical questions that challenge the interlocutor's judgment and motives, or through similes and metaphors that emphasize the interlocutor's inferiority or lack of credentials. Discrediting speech can erode trust, damage relationships, and devalue the interlocutor's contributions. It can also be used as a power play by a speaker who seeks to establish their dominance or authority over the interlocutor. Discrediting speech can also be intended to silence others, as when a speaker mockingly questions another's sincerity or seriousness (Caponetto 2021) with utterances in the vein of “surely you can't be serious” and “are you f**king kidding me?”

Discrediting speech is so commonplace it has given rise to formulaic discrediting expressions, as evidenced in examples (6) and (7) from our dataset. The fixed nature of these expressions also make them attractive (and easily learnt) from a computational point of view.

- (6) Anyway, that's off the point.
- (7) You have got to be kidding.

The fixed expression “off the point” in example (6) uses the conceptual metaphor IMPORTANCE IS CENTRALITY to discredit the interlocutor's words as unimportant or irrelevant to the discussion. The form [that][is][off the point] follows Culpeper's (2016) conventionalized impoliteness formula of pointed criticism, where the third element uses

metaphor to satisfy the model’s expectation of a negatively valenced claim. Example (7) is the typical case of sarcasm, where the speaker pretends that the interlocutor is joking so as to portray their words as risible. This use of discrediting speech is dubbed *seriousness silencing* by Caponetto (2021). The same expression in other context could indicate surprise or disbelief at something said without any intended offence to the interlocutor.

- (8) I wonder if he'd have the same position for a riot at a Linda Sarsour speaking event...
- (9) Well, if he's SO BUSY, why did he have enough time to repeatedly delete my posts?
- (10) You wouldn't know prosperity if you fell into a pile of 100's.
- (11) It's obvious (((David))) has never read one word written by poor old sweet-tempered Chuck He's a Lover Not A Fighter Darwin.
- (12) Nazi bad, Trudeau nazi therefore Trudeau bad. Is that all the thinking that's required?
- (13) Their egos are based in quicksand.

Examples (8) and (9) can be viewed as further instances of seriousness silencing insofar as they force the inference that the interlocutor is untrustworthy, inconsistent, dishonest, or all of these things. Example (10) packages implicit offence in the form of personalized negative assertion (Culpeper 2016). Here the desired inference is not so much forced as explicitly stated: “You do not know what real prosperity is.” Since the interlocutor is claimed to be using words that they do not understand, they should be ignored, making this another case of seriousness silencing (Caponetto 2021). Example (11) is especially complex, and draws on a number of discrediting stratagems. The speaker not only employs metonymic hyperbole (“never read one word”), but also over-familiarity (“Chuck for Charles”), frivolous metaphors (describing Darwin as “He’s a Lover Not A Fighter”), and long-winded circumlocution (“poor old sweet-tempered Chuck … Darwin”) to imply that the interlocutor is unqualified and ill-informed on the subject. These stratagems unite to emphasize the gulf between the knowledgeable speaker and an ignorant interlocutor. In example (12), the speaker employs a Me Tarzan You Jane speaking style to mock the interlocutor’s lack of sophistication. The interlocutor may not speak this way, but the point – metaphorically made – is that they certainly think this way. The point is hammered home by a rhetorical question that cements the speaker’s dim view of their thought processes. Example (13) taps into the conceptual metaphor COMPLEX SYSTEMS ARE BUILDINGS (BASE IS THE FOUNDATION) to suggest that the interlocutor lacks the intellectual foundations on which to build such a grandiose self-image. The use of the source concept *quicksand* also implies a threat, by suggesting that the interlocutor is not smart enough to see the danger they are in.

Irony often uses the pretence of extreme politeness, intimacy or compassion to discredit another, by implying that an interlocutor needs to be treated with great delicacy, as one might treat a child or a self-regarding fool. The pretence must be exaggerated so as to be penetrated by the audience. One may be over-kind to be cruel, and one can be so polite as to be impolite (see Culpeper 2016 on the impoliteness meta-strategy of sarcasm or mock politeness).

- (14) I like that notion. My dearest sir, you are so correct!
- (15) He is so sweet.
- (16) How very sad.

Therefore, in examples (14)–(16), the speaker’s true intent is at odds with the sentiments they express.

Discrediting can also be achieved by a forced reinterpretation of other’s words or actions:

- (17) I thought we were here to work on an encyclopedia, but you just seem to be here to stir up as much drama as possible.
- (18) Less than 100 edits to articles in 2.5 years, yet you have no problem chiming in and rudely making bad situations worse.

Examples (17) and (18) are, following from the framework of Culpeper (2016), personalized negative assertions. Each uses a similar structure to convey a clash of perspectives: the main clause expresses the speaker’s view of reality as a self-evident truth, before a contrastive conjunction introduces a critical view of the interlocutor’s actions, which is also offered up as a truth. Each example also relies on a metaphorical expression (“stir up drama,” “chime in”) to discredit the interlocutor’s motives (or credentials) for doing what so displeases the speaker.

To discredit another person and their ideas, a speaker might also offer a forced choice between a self-evidently positive worldview and a self-evidently negative worldview. In this way, the speaker makes it abundantly clear which side they, and the interlocutor, are to be found on:

- (19) We as a country need to decide if we want individual liberty more than an overbearing centralized government telling us how to live and wasting all our money on bureaucracy.

In example (19), the speaker brings the entire “country” into the debate, making it clear that the interlocutor is not just a discredit to themselves, but to the nation; in other words, they are “a national disgrace.” Moreover, the metaphorical terms “overbearing,” “centralized,” “telling us,” and “wasting money” only reinforce the speaker’s critical view of the interlocutor’s views.

In the following examples, another person is discredited as being rude and impolite:

- (20) I don’t like when people are biased and rude.
- (21) Such crass, rude, and low-class behavior.
- (22) As said by Sir Thomas More: They’ll think that somewhere along your pedigree a bitch got over the wall!
- (23) Seems ghetto manners have rubbed on Harry big time.

Example (20) is offensive if, like that of (21), it is intended to target the interlocutor. In that case, the implication employs the GENERIC IS SPECIFIC schema to place that specific interlocutor in a generic class of people that are deserving of criticism. The implied offence of example (21) is perhaps simpler, even if it also relies on circumlocution, because it offers a cumulative sequence of negatively valenced words: “crass,” “rude,” and “low-class.” Example (22) is somewhat more complicated, for us as analysts and for machines as automated detectors, as it quotes a memorable line attributed to Thomas More in the film *A Man for All Seasons*. The line draws on the conceptual metaphor CLASS IS BREEDING to describe a target’s parentage in animal terms. The key term is not so much “bitch” – which has a negative valence when used of humans – but “over the wall,” which implies that the interlocutor’s family line has been compromised by undesirable outside influences. As in examples (21) and (22), it is left to us to infer that the line refers to the interlocutor, to create what Culpeper (2016) calls a personalized negative reference, and this ambiguity adds to the implicitness of the offense. Nonetheless, there is a sexist dimension to the slight, since it is implied that the baleful influence is female. Example (23) smacks of both sexism and racism in its use of the expression “ghetto manners,” which metaphorically equates the target, Harry, with the typical denizens of the typical ghetto. If the target is Prince Harry, the racism is magnified, insofar as it implies these manners were acquired from his wife, an actress of mixed race who is a frequent target of racist jibes.

Because it is often difficult to distinguish condescending speech from discrediting speech, we view them as a single composite category. Condescending speech uses or evokes a patronizing or superior tone to imply that the interlocutor is somehow inferior or inadequate, and the effect is always to discredit its targets. From their superior position of presumed authority, the condescending speaker may offer unsolicited advice, deliver lectures, judgments or sermons, or identify specific failings in the interlocutor. This power imbalance can contribute to feelings of frustration and resentment on the part of the intended target.

- (24) This is the worst attempt at a dog whistle that I have ever seen, we all know what she really meant: the only people you should be arresting are poor, black, men.
- (25) Don't justify murder by saying it was never alive or it's my right.
- (26) In other words, you intend to reward unproductive or irresponsible parents and penalize good parents.
- (27) You confuse science with censorship.
- (28) Allow me to spell it out for you.
- (29) But he's not an Australian, so fix it up and get your facts right.

Examples (24) through (29) each lecture their targets from a position of assumed authority. Example (24) implicitly accuses its target of deceitful behaviour, by suggesting that their views are more discreditable than they would care to admit. Just as dog whistles are heard only by dogs, a speaker who engages in “dog whistle” communications uses coded language to mask the worst aspects of their social or political views, which may be sexist,

racist or reactionary. The dog whistle itself is a violation of the Gricean maxim of quality (Grice 1975), but here the target is further accused of incompetence, insofar as their efforts are so nakedly transparent. It sharpens the implied offence somewhat to imagine the target's intended audience as "dogs."

Example (25) reinterprets the target's views as a morally unjustifiable defence of murder. That the speaker feels qualified to make this judgment, which effectively casts the target as an apologist for crime, puts them in a self-appointed position of moral superiority. To achieve a greater impact, the speaker uses two heavily valenced metaphors: "justify," which activates the frames of *justice* and *guilt*, and the more hyperbolic "murder," which activates the frames of *evil*, *crime*, and *immorality*. The act of reinterpretation is more explicit in example (26), since the speaker openly admits to rephrasing the interlocutor's thoughts "in other words." These other words recast the disputed ideas in a form that is likely oversimplified, so as to discredit the target for defending such a self-evidently illogical position. Example (27) also expresses the interlocutor's position in other words, but as in (26), these words are unlikely to be ones that a target would agree with. We can see that implied offence by reinterpretation operates by replacing the interlocutor's own words with much more disobliging alternatives. The condescending presumption, which is clearly articulated in example (28) as "allow me to spell it out for you," is that the target cannot express themselves clearly without the help of a superior thinker. The conceptual metaphor at work in example (28), INTELLIGENCE IS LITERACY, adds to this implied offence by suggesting that the target even needs help with their spelling. By overriding a target's own words, these offences may also be seen as attempts at silencing, which can in turn be interpreted as an impoliteness trigger in the mould of (Culpeper 2016). Example (29) is a more obvious example of discrediting speech with a condescending tone, in which the interlocutor is chastised as an errant child that should go and fix their own mistakes.

3.3.3 Insulting speech

Insulting speech shares many similarities with discrediting speech, insofar as the end-result is the diminishment of its target. However, insulting speech is also highly personal; it aims to do more than discredit an arbitrary target as a fellow disputant, but aims to undermine them as people. The speaker is specific in their insults, using personal knowledge of the target in their attack. An insult of the form "even a moron like you realizes that" ... may even acknowledge a target's correctness on a particular issue. In contrast, discrediting speech always aims to undermine a target's credibility or competence, but it is not always personal. This personal dimension can add to the offensiveness of insulting speech when it cuts closer and deeper.

Our view on insulting speech gels with Culpeper's (2016) impoliteness trigger of insult, which identifies personalized negative vocatives, personalized negative assertions, personalized negative references, and personalized third-person negative references as common devices for implicitly insulting language. The personal dimension may also be

exploited by using names, especially nicknames and diminutives, since these are often used by intimate acquaintances. However, the names used in insulting speech – so-called name-calling – are crafted using metaphors, stereotypes, and metonymies to foster greater distance from a target, not intimacy.

- (30) Thanks to the dad Pierre Trudeau & his Communist Islamic loving son of course.
- (31) Jihadi Justin always needs to be mentioned in his Post Nation as the Hero.
- (32) You think any of those women would look at you?
- (33) It's not me who looks like an idiot in this discussion.
- (34) I guess that makes you a genius by donkey standards.
- (35) You are just pretending like your life has some meaning.
- (36) You obviously don't get out much.
- (37) Think I hurt his feelings with his easy vanity degree in Literature, his public sector PC parasitic non-job, and his wife and four mistakes.

Example (30) begins with the sarcastic formula (“Thanks to”), and indirectly targets Justin Trudeau via the epithet “Communist Islamic loving son” [of Pierre Trudeau].⁴ Although the words “Communist” and “Islamic” are not derogatory in themselves, in the political language games that are played online they carry a highly-charged negative valence, especially when used in the construction [X]-loving. The combination of sarcasm and contextual negativity means that this clear insult is still an example of implicit offense. The alliterative nickname “Jihadi Justin” in example (31) strengthens the connection between Trudeau and a label with a heavily negative valence. The personal attacks in examples (32)–(36) are heightened by being so directly targeted. In (32), the target (*you*) is assumed to be an unattractive male, perhaps an incel, that is unworthy of the attention of the women he seems to scorn. In Culpeper’s (2016) terminology, the unpalatable question undermines the target’s self-serving assumptions about who is deserving of attention and who is deserving of scorn while, as suggested by Leech (1983), this indirectness may also sharpen the perception of impoliteness. Example (33) is a prototypical case of indirectness, and one cannot help but read it with a heavy stress on *me*. To preserve another’s feelings, it is common to say “It’s not you, it’s me.” Impoliteness, however, allows us to imply that “It’s not me, it’s you.” Still, the offence remains plausibly deniable here, since formal indirectness often decreases offensiveness. As shown in Panzeri and Carrus (2016), derogatory content can survive in conditionals and questions but tends to be diminished by indirect reports, while negated slurs are frequently rated as inoffensive.

Example (34) is reminiscent of Gore Vidal’s memorable quip about Andy Warhol, whom he described as “the only genius I’ve ever known with an IQ of 60.” Like Vidal’s quip, example (34) has the rug-pulling structure of a classic ironic putdown. First the interlocutor is built up (“makes you a genius”) and then torn down (“by donkey standards”). The incongruity that is sparked by this “switch of scripts” (Raskin 1984) or “shift of frames”

⁴ The strict father – nurturant parent morality and the differences in worldview have an impact on understanding language, and connecting frames to certain moral categories (Lakoff 1996).

(Coulson 2001) is the source of the humour that arises from this “internal mismatch” (Culpeper 2016) of viewpoints. The preamble to example (35), “You are just pretending,” is unnecessary here, and is often left unsaid in an implicit offence. The remainder of the statement, “like your life has some meaning,” would work on its own as a sarcastic putdown, and the accusation of pretence weakens the wit by making the putdown less implicit than it could be. Example (36) is more formulaic as a case of sarcasm, but still requires inferencing to understand. Those who “don’t get out much” may be presumed to lack the world knowledge needed to engage in serious debate, so this can be seen as another instance of seriousness silencing. Example (37) is more complex, and is characterized by heavy uses of circumlocution and figurative language. There is a note of mockery in the opening, “Think I hurt his feelings,” which implies that the target is insecure and over-sensitive. As to what they might be insecure about, the speaker goes on to dismiss, with equally mocking figurative language, the target’s education, work and family. The expressions “parasitic non-job” and “four mistakes” are particularly offensive cases of dehumanization, and it is left to the reader to understand the metonymy PROCESS FOR RESULT as referring to the target’s children. This accumulation of negatively valenced terms has a compounding effect. Not only are the target’s children mistakes, they are mistakes made by a vain and lazy parasite.

3.3.4 Dehumanization

Dehumanization has a very specific reading in the context of implicit offence. It is not enough that a conceptual metaphor, simile or analogy causes the target to be viewed as a nonhuman entity, such as an animal or a thing, but the mapping must prompt us to think less of the target. Shakespeare did not dehumanize Juliet in this sense by comparing her to the sun, any more than one is dehumanized by a comparison to an oil painting, an angel, or a god. Henri Bergson argued in his essays on laughter that “we laugh each time a person gives us the impression of being a thing” (Bergson 2011; Veale 2021), and it is this inclination to laugh that is exploited by implicit offence to make a target risible in the eyes of others. Animalistic and mechanical metaphors are often effective in this regard, as are diminutives, or any exclusionary language that denies the target’s humanity or agency. Dehumanization has traditionally been understood as an “extreme form of prejudice that enables violence and cruelty” (Haslam and Loughnan 2016), but recent work has revealed subtle and everyday forms of the denial of humanness. Haslam and Loughnan (2016) distinguish two forms of dehumanization: one in which people are denied uniquely human attributes, as when they are likened to animals to imply a lack of refinement, self-control, intelligence and rationality, and another in which they are denied equally essential human attributes such as free will, as when they are likened to machines or tools. We define dehumanization as any conceptualization of a human as a nonhuman entity for purposes of causing offence. The metaphors we use – especially those used unconsciously or non-deliberately – can have a significant impact on our reasoning and shape our social and

political attitudes (see Grady 2017 for an overview). Dehumanizing conceptualizations can thus promote a disregard for the human rights of certain groups (Markowitz and Slovic 2020).

In the following examples from our dataset, source concepts such as *commodity*, *cesspool*, *trash*, *dog*, *flood*, and *puppet* are used to achieve dehumanization of the speaker’s targets.

- (38) Trudeau and the #Liberals have the ability to stop it but choose not to as it serves their agenda of turning Canada into a third world cesspool.
- (39) Unfortunately by the time she seeks the good guys for commitment, her value is enormously lowered due to some combination of depreciating looks, a promiscuous past, alpha widowhood, and kids needing provision.'
- (40) Better yet; that she-dog. Here girl, here girl! Good girl! Here, I will give you biscuit!
- (41) You will never be anything more than a replaceable component to be put to work.
- (42) Trailer park trash in the comments still crying.
- (43) I wouldn't be surprised if the government took more taxes off our paychecks so that Trudeau could feed and house more rapeugees that are flooding our country.
- (44) Here's some food for thought, have you ever wondered if some of the malicious sock puppets that attack you were operated by a wikipedian admin who didn't want to lose credibility because of slander?

Example (38) is both dehumanizing and racist. The literal content of a “cesspool” is not people but toxic waste, but “third world” implicitly equates this waste – via the metonymy PLACE FOR INHABITANTS – with the inhabitants of economically underdeveloped countries that are mostly non-white. Example (39) is another personalized third-person negative reference in the terms of Culpeper (2016), but its discriminatory metaphors are sexist rather than racist. A female target is objectified as a “depreciating” commodity whose “promiscuous past” has placed a burden on the state in the form of “kids needing provision.” Example (40) employs an extreme form of dehumanization to conceptualize its female target as a biddable dog that can be enticed with a biscuit. The term “she-dog” seems to be a circumlocution that avoids (yet still primes) the more overt sexism of “bitch.” By placing its target in the submissive position of a dog eager for treats, the speaker projects themselves into a position of power and dominance. The pretence has a performative aspect which fills out what might otherwise be a partial metaphorical mapping (Lakoff and Johnson 1980). By literally acting out their metaphor (“Here girl, here girl! Good girl!”), the speaker is using what we will henceforth label a *metaphor literalization strategy*.

Examples (41) and (42) do not employ this strategy, but rather use dehumanizing metaphors for simple name-calling. Example (41) uses a mechanization metaphor that views its target as a replaceable cog in a machine. The offence is heightened with the condescending claim that “You will never be anything more,” and the speaker’s sense of superiority is heightened by the assumption that they are somehow raging against the same machine, making this an example of both dehumanization and discrediting/condescending speech. Example (42) also implies the social superiority of the speaker by labelling its

targets as “Trailer park trash,” which denotes the class of poor, low-class people (so-called “trailer trash”) who live in trailer parks. While these people are often white (so-called “white trash”), the metaphor is not so much racist as it is classist, and the offence is compounded by suggesting its targets are not just unsophisticated but over-sensitive and baby-like (“still crying”).

Example (43) employs the neologism “rapeugees.” This is ambiguous, insofar as it might refer to the victims of rape or to the perpetrators of rape, but the larger metaphor of unwanted newcomers “flooding our country” plays to a common trope: that they are unwanted because they are criminals. It is commonplace for demagogues to stereotype a whole group based on a small number of very negative and highly salient examples (Lakoff 1987). To characterize them as perpetrators rather than victims betrays a lack of empathy and compassion, while the metaphor *flooding* further dehumanizes them as a mindless force that lacks individual agency. Example (44) also denies agency to some of its targets by using the metaphor of *sock puppets*. The example is conspiratorial in tone, but the metaphor itself is fascinating. Sock puppets are a special kind of puppet insofar as they are primitive (a hand in a sock) and unconvincing. It is their ease of detection that makes the offence implicit here, since the interlocutor is presumed to lack the basic intelligence needed to detect such a crude form of deception from others.

3.2.5 Derogation

Dehumanization diminishes its targets by treating them as non-human entities deserving of scorn, but it is possible to achieve the same end without denying their humanity. Derogatory speech, or derogation, diminishes the target by attacking the kind of human that they seem to be. A common derogation strategy is to view the target as a child, or as a person of unusual sensitivity, or one with special needs or physical or mental impairments. The implication that the target is under-developed as a human being can be achieved with adjectives and verbs that imply incompetence, dependency, or frailty, or with rhetorical questions that challenge the target's abilities and worth. In short, dehumanization views its targets as non-human, while derogation views them as sub-human. In the following examples, the target is variously seen as a child (examples (47)–(53)) or as a person with an illness or disability (examples (52)–(55)).

- (45) Atta boy!
- (46) Good boys!
- (47) Did the money your grandparents sent you in a singing card by the suped up fucktarded laptop you're rocking?
- (48) I like when the little guys stick up for themselves.
- (49) Snotty 17-year-olds projecting their daddy and mammy issues on the world.
- (50) 2 questions for you, my dude: Who hurt you, and how old are you?
- (51) Grow up, sometimes things that hurt your feelings get posted to the internet.

- (52) A person with a mental illness just got elected to a position where he has a direct influence over the education of children.
- (53) But continue to lick windows at your own pleasure.
- (54) That person has nothing to offer the school board other than the fact he lives a delusional life and wants to be a 'role model' to other children.
- (55) Anyone who thinks Trudeau has done anything to support women is delusional and engaging in some next-level cognitive dissonance.
- (56) Oh tut tut, such unproven accusations are so uncivil, but posturing from a nobody pretending to be a somebody still leaves a nobody.

Examples (45) and (46) are performative in much the same way as example (40). They literally treat the interlocutor as a child by acting out the consequences of this literal truth, and in doing so they intensify the derogatory offence. Examples (47) through (51) use a similarly derogatory conceptual metaphor, *INGÉNUE IS A CHILD*, but they differ greatly in the strategies that they employ. Example (47), the most complex one, takes the form of an unpalatable question (Culpeper 2016), and also makes extensive use of circumlocution. The metaphorical source domain of childhood is richly evoked to suggest arrested development, by imagining that the target still receives money from their grandparents, and “in a singing card” no less. Although the target is said to be “rocking” their laptop, in what is ostensibly a compliment, the derogatory neologism “fucktarded” betrays the pretence at the heart of this ironic praise. As in examples (40), (45) and (46), this pretence is part of a larger performative act, one that fleshes out a more complete mapping between the target and a stereotypical child. Examples (52) through (56) each invoke the conceptual metaphor *IGNORANCE IS A MENTAL ILLNESS* to portray the target as mentally lacking. Example (52) uses the vagueness of the term *mental illness* to hint at the most damning diagnosis, while (53) uses a derogatory stereotype (“lick windows”) that is more often heard in the school yard. Example (54) uses scare quotes to reject a positive view of its target (“role model”) in favour of a derogatory one (“delusional life”), while example (55) accuses its target of actively engaging in delusional behaviour. Example (56) is performative in its patronizing use of “tut tut,” while “nobody pretending to be a somebody” is also a clear attempt at seriousness silencing, one that seeks to discredit the target and diminish their credibility.

3.3.6 Discrimination (racism, sexism, ageism, etc.)

Discrimination is the unjust or prejudicial treatment of individuals or groups based on their personal characteristics, such as race, gender, religion, or sexual orientation. Discriminatory language achieves these ends with various linguistic devices, including slurs, stereotypes, and exclusionary labels. Such language, which can perpetuate harmful attitudes, reinforce power imbalances, and lead to social exclusion and violence, can be explicit or implicit, but we focus here on implicit cases. Even implicit discriminatory language use may be rooted in systemic and structural forms of discrimination, such as racism, sexism, homophobia, or

ableism, and can target marginalized groups by dismissing or undervaluing their particular perspectives.

- (57) Although, you are a Christian, so what more can I expect.
- (58) Wanting to live in a segregated tax-payer subsidized Islamic enclave after you immigrate to a multicultural pluralist society isn't culturally appropriate for Canada, so it isn't the right setting to do so.
- (59) What I do know is that I live on a small Island and it seems that a lot of people from various countries in the world are willing to desert their homelands and their families just to come here, in my mind that tells me a lot about said people and their home countries, most of all it highlights their weaknesses of how their own countries have let them down and how they feel the need to feed off somebody else's country, how should we English interpret this without sounding rude ???
- (60) I dunno if you're actually saying it should be that way, or just saying, it would be ideal, in a perfect world if those people didn't have kids.
- (61) Intelligent black teen does not like blacks.
- (62) Ever notice that the more intelligent a person is the more they do not like blacks?

Example (57) combines the insulting speech of an *ad hominem* attack with discrimination based on religious affiliation. It is also a case of discrediting/condescending speech, since it questions the target's qualifications to express a valid opinion. Example (58), which contrasts “a multicultural pluralist society” with “a segregated tax-payer subsidized Islamic enclave,” is discriminatory in a subtler fashion. The words comprising the latter may not exhibit an obviously negative valence, but they do appear negative when juxtaposed with the obvious positivity of the former (consider society vs. enclave, pluralist vs. segregated). Since “Islamic” is placed in the latter cluster, it too takes on a relatively negative valence. The result, which is made explicit with the clause beginning *so*, is intended to “other” and exclude Muslims. Example (59) is grounded in white grievance (ElSherif et al. 2021) and what Saul (2018: 365) identifies as a dog whistle working in tandem with racial resentment. The implicitness of the offence is served by the speaker’s attempts to appear reasonable, but this only makes their derogatory tropes about immigrants – such as “need to feed off,” which suggests parasitism – all the more offensive. Example (60) uses circumlocution to evoke a trope about over-breeding among minorities, and equates a world with eugenics to “a perfect world.” Examples (61) and (62) also make use of implicit racism. The first licenses this racism by appearing to agree with the smartest members of the targeted group, while the second generalizes this reasoning to the “more intelligent people” of all races. The offence in the latter is sharpened by the implication that the targeted group suffers from self-hate, making other-hate acceptable, or even logical.

4 The Subtle Knife

Why do speakers resort to implicit offence when explicit offence is so much more direct, and so much easier for an interlocutor to understand? A language like English gives us a surfeit of ways to express an offensive meaning openly and without ambiguity. However, we have seen in our data analyses that implicitness holds many attractions for a speaker who aims to offend. First, implicitness affords plausible deniability to a speaker, who may later have good reason to resile from their aggressive stance. When a speaker shares a view with others who are more extreme and less subtle in their articulation of a grievance, they may employ “dog whistles” to make their position more palatable to a broader audience. Recall that a speaker using social media is addressing more than a single interlocutor and a specific target, but is simultaneously playing to the gallery too. Implicitness that does not rely on formulaic constructions demands more creativity of a speaker, but it also affords more opportunities for play, for grandstanding, and for more eye-catching humour. A mocking linguistic turn is more cutting if its target feels that not just the speaker, but large parts of the audience, are laughing at them also. Explicit offence that relies on vulgarity and obvious epithets can be dramatic, but it rarely encourages linguistic play. In contrast, implicit offence that exploits metaphors, similes, tropes and other figurative gambits establishes a stage on which different kinds of play, from performative play acting and verbal showmanship to aggressive back-and-forth jousting, can take place.

This poses a certain danger for a speaker who finds that the interlocutor is better able to wield the figurative weapons that they have chosen for themselves. Consider a mode of figuratively adversarial humour that Veale, Feyaerts and Brône (2006) call “trumping”. Examples (40), (45) and (46) show how speakers can performatively treat their own metaphors as literal truths, to playfully twist the knife they have inserted into their targets. But this *metaphor literalization strategy*, as we name it here, is also available to the interlocutor, who may use the strategy to turn the metaphor against its user. It seems apt that a noteworthy user of the trumping strategy is Stormy Daniels, an adult film star who is a frequent target of former US president Donald Trump. It is alleged that after a brief affair with Daniels, Trump paid the actress for her silence and reported the payment as a legal expense. Trump denies the charges and frequently mocks the actress as “horse-face.” She, in turn, has taken ownership of Trump’s insult, and advertises her live events as “straight from the horse’s mouth.” Dynel and Poppi (2020) conducted an analysis of the linguistic strategies used by Daniels on her Twitter account to fend off abusive messages, many of which originate from Trump supporters. Their analysis shows how Daniels deftly turns abusive metaphors against their users, often by appealing to an obvious sexual meaning that has eluded her would-be tormentors. For example, when told she should “do a 180 before it’s too late” she replies ‘I do 180s all the time … cowgirl to reverse cowgirl!’ A recent attack by Trump on her looks inspired a supporter to tweet at the actress that “Trump wouldn’t touch you with a 10-foot pole.” She retorted by taking ownership of the figurative pole and turning it against its bearer and his master: “True. He used a 3-inch one”.

Trumping shows that metaphors are unstable, two-edged swords when one aims for implicit offense, so it is important to separate the form of the offence from the device that enables it. A question that has motivated us throughout this paper is whether modern NLP technologies can detect and grasp the full richness of the ways these devices are exploited in our data. The detection of implicitly offensive language is growing in importance with the widespread adoption of large language models (Li 2022), as e.g. used in the chatbot *ChatGPT* (Leiter et al. 2023). A language model (LM) is, at its core, a probability distribution over the strings of a language; that is, a mechanism for assigning probabilities to different sequences. From a finite number of observations and with a finite number of parameters, they derive their ability to say that one string is more likely than another. The piecewise fashion in which they finitely model the infinite makes them excellent predictors of language: they can fill in blanks, or predict the most probable continuation to strings that they (and we) have never seen before. When the occasion calls for explicit offence, LMs are quick to oblige. For instance, an LM will assign a high probability to the expletive that most obviously fills the gap in this masked utterance: “I thought I told you to ___ off!” LMs do not have different modules for syntax, semantics, and pragmatics, but conflate different factors in language production into a single end-to-end mapping of inputs (text prompts) to outputs (text continuations). LMs learn to play language games from the large datasets on which their parameters are tuned. This makes them savants at the language of small-talk, formal speeches, literary genres, computer coding, and math problems. Their ability to complete our thoughts by completing our sentences means that they can also cleverly veil offensiveness with implicitness, just by learning the tacit rules of the language games that we humans play whenever we wield words like hidden daggers.

Large language models (LLMs) have many more parameters, are trained on ever larger datasets, and use wider context windows for maintaining coherence across a discourse. They are also capable of greater abstraction and more powerful generalizations. There are credible fears that LLMs may be weaponized by bad actors to generate hateful content on an industrial scale. This makes the automated detection of offence, both explicit and implicit, a prime concern for public-facing LLMs such as ChatGPT. Consider the model’s analysis of the phrase in this paper’s title. It correctly interprets the word “bitch” to mean “an individual who was not from the same social class or racial background [who] managed to enter the family bloodline, perhaps through a romantic relationship or an affair”, and goes on to note that the word “is particularly derogatory, as it implies that the individual [...] was a woman who was seen as being of lower social status or moral character.” It concludes that “making derogatory comments about someone’s ancestry or heritage is not acceptable behavior.” Those who *talk the talk* of sensitivity and respect do not always *walk the walk*, but ChatGPT shows us that a productive facility for the former is a useful platform on which to build the latter.

Let us consider what ChatGPT 3.5 makes of a creative instance of trumping:

User: I told my boss that "I do the work of two people for this company!" And he agreed with me, saying "Yes, Laurel and Hardy." Was my boss being complimentary or offensive?

ChatGPT: It's difficult to say for certain without knowing your boss and the context of the conversation, but it's possible that your boss was using a form of trumping that could be interpreted as both complimentary and offensive.

On one hand, your boss may have been acknowledging that you're doing a lot of work by comparing you to two people instead of one. This could be seen as a compliment, as it suggests that you're working hard and doing a good job.

On the other hand, the reference to Laurel and Hardy could be interpreted as a subtle insult. Laurel and Hardy were a comedy duo known for their bumbling antics, so your boss may have been implying that you're not as competent as you think you are. Alternatively, your boss may have been using humor to diffuse the situation and acknowledge that you're doing a lot of work without actually commenting on your competence.

ChatGPT's response is quite impressive. It starts by diagnosing our example – a confrontation between a worker and their employer – as a case of trumping. In an earlier interaction, we had defined trumping for the model as a linguistic strategy in which “one speaker undermines an idiom or metaphor used by another speaker by treating it as a literal statement.” ChatGPT now uses the term fluently, to diagnose the implicit offence at the heart of the interaction. It sees how the speaker's idiom, “the work of two people,” has been reinterpreted by the interlocutor so that it no longer signifies twice the average productivity, but now means “the work of two fools.” But it also appreciates the subtlety that comes from implicitness, noting that the reference to the bumbling duo may also serve other, more pragmatic and less biting conversational goals.

ChatGPT also has an ear for sexual euphemism; it is trained on internet content, after all. Here is how we paraphrased the Stormy Daniels “pole” interaction for the model's consideration:

User: A politician has an affair with an adult film star, but then denies it. He says "I wouldn't touch her with a 10-foot pole." She replies "True. You used a 3-inch one." What does she mean?

The model sees the interaction for what it is: an idiom-driven denial, and a sarcastic retort that plays on the meaning of “pole.” It is even shrewd enough to place this word in scare quotes:

ChatGPT: The adult film star is implying that the politician did have a sexual relationship with her, despite his denial. When the politician says "I wouldn't touch her with a 10-foot pole," he is trying to distance himself from the adult film star and suggest that he would never be involved with someone like her. However, the film star's response suggests that the politician did have a sexual relationship with her, but used a much smaller "pole" than he claims. In other words, he did have a sexual encounter with her, but it was much shorter and less impressive than he wants people to believe.

While ChatGPT seems well-attuned to the nuances of implicit offense, it sometimes fails to appreciate the speaker's larger figurative ploy. Consider how it handles the following prompt:

User: The journalist Christopher Hitchens once said this about a controversial figure: "If you gave Jerry Falwell an enema, you could bury his remains in a matchbox." Was he being offensive? And if so, how?

Hitchens' quip is analyzed in Veale (2012) as a complex, highly-figurative circumlocution whose offence is heightened by its implicitness. Here is how ChatGPT understands the quip:

ChatGPT: Yes, Christopher Hitchens' comment about Jerry Falwell was intended to be offensive. The statement is a form of hyperbolic insult, which exaggerates the negative qualities of the person being criticized in an extreme and provocative way. The use of the enema metaphor is particularly crude and graphic, and is intended to paint Falwell in an extremely negative light. The implication is that Falwell's views and actions are so odious and repellent that they could be reduced to waste matter that could fit in a matchbox. The statement is also intended to be humorous, in a dark and satirical way. By reducing Falwell to such a base and ridiculous image, Hitchens is mocking him and undermining his credibility and influence.

ChatGPT is quick to detect the satirical insult in the quip, and it does not equivocate about its meaning. Hitchens offers and expects no quarter, so there is no plausible deniability here. The model also grasps the figurative intent of *enema*, even if its circumlocutionary function masks much of its vulgarity. Yet, while it also recognizes the hyperbolic nature of the metaphor, it fails to see the figurative significance of the "matchbox," which is inversely proportional to its size. Ultimately, the insult is a novel variation on the explicit insult "You are full of shit." By inferring that an enema would kill Falwell and leave so little behind, we can also read it as a variation on the explicit insult "You are a total shit." By unpacking what Hitchens *could* have said, but chose not to, we appreciate his ingenuity all the more. Still, we should give credit where it is due. While ChatGPT fails to see the reason for the circumlocution, it does grasp the speaker's overall goal of discrediting Falwell by undermining his credibility.

5 Conclusions

Our present analysis highlights the main issues with previous typologies and datasets: (a) an inadequate definition of implicit vs. explicit offense; (b) a form-oriented typology of implicit offense, and therefore (c) the conflation of content and form in modes of offense; (d) treating figurativeness, particularly metaphor, as the principal player, while ignoring this

player's other leading role in explicit offense; and (e) an over-focus on specific forms that are not broadly representative of the phenomenon, such as on offensive similes.

This paper has provided a qualitative analysis of more than a hundred examples of IOL, an operational definition and proposed a new typology of implicitly offensive language. The typology distinguishes between the content of an offense, where speech can be insulting, discrediting or condescending, dehumanizing, derogatory, or discriminatory, and the various forms used to express it, including metaphor, metonymy, simile, irony, hyperbole, euphemism, repetition, rhetorical questions, circumlocution, name-calling, generalization, contrastive statements, and the use of graphic devices such as all-caps text. As we have seen, many examples combine several of these devices to heighten their offensive impact, as when metaphor, irony, circumlocution, and euphemism all work together to deliver a brutal payload.

This new typology is intended to serve as the foundation for identifying Implicitly Offensive Language (IOL) in 'real-world' texts and interactions. We have already conducted some promising initial experiments applying this typology in the context of LGBT discourse, and the preliminary results are satisfactory considering the complexity of the phenomenon, especially regarding difficulties with clearly defining the concept of 'offense' as such and the frequent conflation between offence and socially unacceptable ideology. It remains questionable, taking the annotation campaigns into account, whether all of the instances we treated as offensive can be unambiguously classified as such. Given that achieving a satisfactory level of inter-annotator agreement, and even intra-annotator agreement, with this intricate phenomenon, has proven to be exceptionally challenging (Lewandowska-Tomaszczyk et al. 2023), we have aimed to keep the typology as straightforward as possible, while still carefully considering its linguistic intricacies. We acknowledge that striking a perfect balance between operability and the depth of linguistic analysis is likely an unattainable goal. It may be valuable for NLP tasks to develop another typology that more comprehensively addresses pragmatic theories and factors, such as interlocutors, intent, uptake, and both discursive and social context. Additionally, some of the categories we propose could be more carefully considered in the context of issues such as free speech (see O'Driscoll 2020), socio-historical norms and values, and frameworks of social control.

In earlier research, implicitness was often explored as a matter of linguistic form, due to ease of annotation and the popularity of certain devices such as metaphor and irony. However, a focus on form alone can be misleading. While offensive similes clearly use form to be unkind, such comparisons prove extremely rare in our dataset, where they constitute just one example. When an audience has to expend substantial cognitive effort to unpack an implicit offense, the effort draws others in, as with a joke, and makes them complicit in the offence (Veale 2012). Some of this effort is needed to negotiate the conceptual devices that are used, and some is needed to grasp the pragmatics of their use, such as whether the speaker wants us to see their target as "sub-human" or even non-human, or whether they want to silence them for their views.

Explicit and implicit offensiveness are related but ultimately very different language games. In the former, obviousness is a virtue; in the latter, it is a vice. The former wears its offence on its sleeve, while the latter makes us work to perceive it, often by masking negative intent with words that have little or no negative valence in the mental lexicon. Large language models, or LLMs, can learn to play these language games well if they are trained on very large datasets. Mechanisms of self-attention allow LLMs to focus on different aspects of a word’s meaning in different contexts, and so learn how to use these words as game pieces in different games. We have shown that public-facing LLMs such as ChatGPT are already rather good at grasping the offence implicit in the mischievous use of idioms, metaphors and even sarcasm and irony. These models need no further bouts of supervised learning over large datasets, as they come *pretrained* – this is the “P” in GPT (Radford et al. 2019) – on a wealth of contextual language use. What is needed now is not a vast dataset with annotations painted in broad strokes, but smaller, targeted datasets that are annotated in a much finer and discriminating fashion, to support the fine-tuning of LLMs. Large pretrained language models do not need copious amounts of additional training data, but can learn from a few targeted examples, in what is called *few-shot learning* (Brown et al. 2020). Anyone with a web browser can see for themselves how far LLMs have already come. It will take finesse and insight, not scale and generality, to take them the rest of the way.

Acknowledgements

The present study has been conducted within the Use Case WG 4.1.1. Incivility in Media and Social Media, COST Action CA 18209 *European network for Web-centred linguistic data science Nexus Linguarum*.

References

- Assimakopoulos, Stavros, Fabienne Baider & Sharon Millar (eds.). 2017. *Online Hate Speech in the European Union: A Discourse-Analytic Perspective*. Cham: Springer.
- Anand, Manish, Kishan Bhushan Sahay, Mohammed Altaf Ahmed, Daniyar Sultan, Radha Raman Chandan & Bharat Singh. 2023. Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science* 943. 203–218. DOI: <https://doi.org/10.1016/j.tcs.2022.06.020>.
- Bączkowska, Anna. 2022. Explicit and implicit offensiveness in dialogical film discourse in Bridget Jones films. *International Review of Pragmatics* 14. 198–225.
- Barnden, John A. 2016. Communicating Flexibly with Metaphor: A Complex of Strengthening, Elaboration, Replacement, Compounding and Unrealism. *Review of Cognitive Linguistics* 14(2). 442–473.
- Bergson, Henri. 1911. *Laughter: An Essay on the Meaning of the Comic*. Translated by Cloutesley Brereton and Fred Rothwell. New York: The Macmillan company.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

- Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th NeurIPS Conference on Neural Information Processing Systems*, volume 33, 1877–1901. Vancouver.
- Caponetto, Laura. 2021. A Comprehensive Definition of Illocutionary Silencing. *Topoi* 40. 191–202.
- Caselli, Tommaso, Valerio Basile, Jelena Mitrović, Inga Kartozija & Michael Granitzer. 2020. I Feel Offended, Don’t Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Ishara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6193–6202. Marseille: European Language Resources Association (ELRA).
- Coulson, Seana. 2001. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge, UK: Cambridge University Press.
- Culpeper, Jonathan. 2005. Impoliteness and entertainment in the television quiz show: The weakest link. *Journal of Politeness Research* 1(1). 35–72. DOI: 10.1515/jplr.2005.1.1.35
- Culpeper, Jonathan. 2011. *Impoliteness: Using Language to Cause Offense*. Cambridge: Cambridge University Press.
- Culpeper, Jonathan. 2016. Impoliteness Strategies. In Alessandro Capone & Jacob Mey (eds.), *Interdisciplinary Studies in Pragmatics, Culture and Society. Perspectives in Pragmatics, Philosophy & Psychology*, vol 4, 421–445. Cham: Springer.
- Dancygier, Barbara & Eve Sweetser. 2014. *Figurative language*. Cambridge University Press.
- Despot, Kristina, Mirjana Tonković, Mario Brdar, Mario Essert, Benedikt Perak, Ana Ostroški Anić, Bruno Nahod & Ivan Pandžić. 2019. MetaNet.HR: Croatian Metaphor Repository. In Marianna Bolognesi, Mario Brdar & Kristina Despot (eds.), *Metaphor and Metonymy in the Digital Age: Theory and methods for building repositories of figurative language*, 123–146. Amsterdam: John Benjamins Publishing Company.
- Dynel, Marta & Fabio I. M. Poppi. 2020. Arcana Imperii: The power of humorous retorts to insults on Twitter. *Journal of Language Aggression and Conflict* 8(1). 57–87.
- ElSherief, Mai, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury & Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Fortuna, Paula, Juan Soler-Company & Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* 58(3). DOI: 102524. <https://doi.org/10.1016/j.ipm.2021.102524>.
- Gao, Lei, Alexis Kuppersmith & Ruihong Huang. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In Greg Kondrak & Taro Watanabe (eds.), *Proceedings of the 8th International Joint Conference on Natural Language Processing*, 774–782. Taipei, Taiwan: AFNLP.
- Gibbs, Raymond W. & Herbert L. Colston. 2012. *Interpreting Figurative Meaning*. New York: Cambridge University Press.

- Grady, Joseph. 2017. Using metaphor to influence public perceptions and policy: Or, how metaphors can save the world. In Elena Semino & Zsófia Demjén (eds.), *The Routledge Handbook of Metaphor and Language*, 443–454. London/New York: Routledge. Taylor and Francis Group.
- Grice, Paul H. 1975. Logic and Conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, 41–58. New York: Academic Press.
- Haslam, Nick, Stephen Loughnan, Catherine Reynolds & Samuel Wilson. 2007. Dehumanization: A new perspective. *Social and Personality Psychology Compass* 1(1). 409–422. DOI: <https://doi.org/10.1111/j.1751-9004.2007.00030>.
- Haugh, Michael & Jonathan Culpeper. 2018. Integrative pragmatics and (im)politeness theory. *Pragmatics and its interfaces*. 213–239. Amsterdam: John Benjamins. DOI: [10.1075/pbns.294.10hau](https://doi.org/10.1075/pbns.294.10hau).
- Haugh, Michael & Valerie Sinkeviciute. 2019. Offence and conflict talk. In Matthew Evans, Lesley Jeffries & Jim O'Driscoll (eds.), *The Routledge Handbook of Language in Conflict*, 196–214. London: Routledge. DOI: [10.4324/9780429058011-12](https://doi.org/10.4324/9780429058011-12).
- Kogilavani, S. V., Senthil Malliga, K. R. Jaiabinaya, Malini, M. & Manisha Kokila, M. 2023. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings* 81. 630–633. DOI: <https://doi.org/10.1016/j.matpr.2021.04.102>.
- Kumar, Ritesh, Atul Kr. Ojha, Shervin Malmasi & Marcos Zampieri 2018. Benchmarking Aggression Identification in Social Media. In Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri & Shervin Malmasi (eds.), *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, 1–11. Santa Fe.
- Kunupudi, Deepthi, Shantanu Godbole, Pankaj Kumar & Suhas Pai. 2020. Toxic language detection using robust filters. *SMU Data Science Review* 3(2). Available at: <https://scholar.smu.edu/datasciencereview/vol3/iss2/12> (accessed 5 September 2023).
- Landau, Marc J., Lucas A. Keefer & Trevor James Swanson. 2017. ‘Undoing’ a Rhetorical Metaphor: Testing the Metaphor Extension Strategy. *Metaphor and Symbol* 32(2). 63–83. DOI: [10.1080/10926488.2017.1297619](https://doi.org/10.1080/10926488.2017.1297619).
- Leiter, Christoph, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen & Steffen Eger. 2023. ChatGPT: A Meta-Analysis after 2.5 Months. *ArXiv* 2302.13795.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović & Giedre Valunaite Oleskeviciene. 2021. LOD-connected offensive language ontology and tagset enrichment. In Sara Carvalho & Renato Rocha Souza (eds.), *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*, 135–150. CEUR Workshop Proceedings. Zaragoza.
- Lewandowska-Tomaszczyk, Barbara, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene & Slavko Žitnik. 2023a. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 19(1). 7–48. DOI: <https://doi.org/10.1515/lpp-2023-0002>.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Chaya Liebeskind, Giedre Valunaite Oleskevicienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojsczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, Olga Dontcheva-Navratilova, Agnieszka Borowiak, Kristina Despot & Jelena Mitrović. 2023b. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovje* 49(1). DOI: <https://doi.org/10.31724/rihjj.49.1.8>.
- Li, Hang. 2022. Language models: past, present, and future. *Communications of the ACM*. 65(7). 56–63.

- Ljubešić, Nikola, Darja Fišer & Tomaž Erjavec. 2021. Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.0. *Slovenian language resource repository CLARIN.SI*. ISSN 2820-4042. Available at: <http://hdl.handle.net/11356/1433> (accessed 5 September 2023).
- Markowitz, David M. & Paul Slovic. 2020. Social, psychological, and demographic characteristics of dehumanization toward immigrants. *PNAS* 117(17). 9260–9269.
- Ocampo, Nicolas, Ekaterina Sviridova, Elena Cabrio & Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1997–2013. May 2-6, 2023 ©2023 Association for Computational Linguistics
- Panzeri, Francesca & Simone Carrus. 2016. Slurs and negation. *Phenomenology and Mind* 11. 170–180. DOI: https://doi.org/10.13128/Phe_Mi-20117.
- Park, Namkee, Kyungeun Jang, Seonggyeol Cho & Jinyoung Choi. 2021. Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior* 121. 106795. DOI: <https://doi.org/10.1016/j.chb.2021.106795>.
- Patrick, Clea & Leah Hollis. 2018. Contentious Cloud Chatter: A Comparative Analysis of Aggressive Speech. *Comparative Civilizations Review* 78(8).
- Plaza-del-Arco, Flor Miriam, M. Dolores Molina-González, L. Alfonso Ureña-López & María Teresa Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems* 258. 109965. DOI: <https://doi.org/10.1016/j.knosys.2022.109965>.
- Qadir, Ashequl, Ellen Riloff & Marilyn Walker. 2015. Learning to Recognize Affective Polarity in Similes. In Lluís Màrquez, Chris Callison-Burch & Jian Su (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 190–200. Lisbon: Association for Computational Linguistics.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). p. 9.
- Raskin, Victor. 1984. *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel.
- Roy, Pradeep Kumar, Snehaan Bhawal & Cn Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language* 75. 101386. DOI: <https://doi.org/10.1016/j.csl.2022.101386>.
- Sadeque, Farig, Stephen Rains, Yotam Shmargad, Kate Kenski, Kevin Coe & Steven Bethard. 2019. Incivility Detection in Online Comments. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang & Soujanya Poria (eds.), *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 283–291. Minneapolis, Minnesota: Association for Computational Linguistics.
- Saul, Jennifer M. 2018. Dog whistles, Political Manipulation, and Philosophy of Language. In Daniel W. Harris, Daniel Fogal & Matt Moss (eds.), *New Works on Speech Acts*, 360–383. Oxford: Oxford University Press.
- Schmidt, Anna & Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In Lun-Wei Ku, Cheng-Te Li (eds.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. Valencia: Association for Computational Linguistics.
- Searle, John. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press. DOI:10.1017/CBO9780511609213.

- Shi, Xiayang, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen & Shaolin Zhu. 2022. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering* 101. 108005. DOI: <https://doi.org/10.1016/j.compeleceng.2022.108005>.
- Sigurbærsson, Gudbjartur Ingi & Leon Derczynski. 2023. Offensive language and hate speech detection for Danish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 498–3508). Marseille: European Language Resources Association (ELRA). DOI: <https://doi.org/10.48550/arXiv.1908.04531>.
- Stoll, Anke, Marc Ziegele, Oliver Quiring. 2020. Detecting impoliteness and incivility in online discussions. Classification approaches for German user comments. *Computational Communication Research* 2(1). 109–134. DOI: 10.5117/CCR2020.1.005.KATH.
- Trandabăt, Diana, Daniela Gifu & Adrian Pleșescu. 2022. Detecting offensive language in Romanian social media. *Procedia Computer Science* 207. 2883–2890. DOI: <https://doi.org/10.1016/j.procs.2022.09.346>.
- van Aken, Betty, Julian Risch, Ralf Krestel & Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem & Jacqueline Wernimont (eds.), *Proceedings of the Workshop on Abusive Language Online (ALW)*, 33–42. Brussels: Association for Computational Linguistics.
- Veale, Tony, Kurt Feyaerts & Geert Brône. 2006. The cognitive mechanisms of adversarial humor. *HUMOR: The International journal of humor research* 19(3). 305–339.
- Veale, Tony. 2012. *Exploding the Creativity Myth: The computational foundations of linguistic creativity*. London: Bloomsbury Academic.
- Veale, Tony. 2021. *Your Wit is My Command. Building AIs with a Sense of Humor*. Cambridge, Massachusetts: The MIT Press.
- Waseem, Zeerak, Thomas Davidson, Dana Warmsley & Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy & Joel Tetrault (eds.), *Proceedings of the first workshop on abusive language online*, 78–84. Vancouver: Association for Computational Linguistics.
- Wiegand, Michael, Maja Geulig & Josef Ruppenhofer. 2021b. Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 358–368. Stroudsburg, Pennsylvania: European Association for Computational Linguistics.
- Wiegand, Michael, Josef Ruppenhofer & Elisabeth Eder. 2021a. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty & Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587. Strasbourg: Association for Computational Linguistics.

About the Authors

Kristina Štrkalj Despot is a Senior Research Fellow and a Vice Director at the Institute for the Croatian Language. Her research interests center around cognitive linguistics, figurative language, semantics, pragmatics, conceptualization, the relationship between mind and language, historical linguistics, e-lexicography, and language resources and technologies. She is the author of five books and more than 50 papers, and she has received the National Science Award of the Republic of Croatia (2012) and Fulbright Postdoctoral Award and the subsequent postdoctoral fellow position at UC Berkeley (2011/2012), as well as two additional national awards.

Address

Institute for the Croatian Language
Republike Austrije 16
10000 Zagreb, Croatia

e-mail: kdespot@ihjj.hr
ORCID: 0000-0001-9004-5103

Ana Ostroški Anić is a Senior Research Associate at the Institute for the Croatian Language. Her research expertise involves terminology, specialized knowledge representation and cognitive linguistics (particularly metaphor theory). She is the author of the book *Metafora u terminologiji* [Metaphor in Terminology], a co-author of 6 additional books and edited volumes, and over 30 research papers.

Address

Institute for the Croatian Language
Republike Austrije 16
10000 Zagreb, Croatia

e-mail: aostrosk@ihjj.hr
ORCID: 0000-0001-9999-0750

Tony Veale is an Associate Professor in the School of Computer Science at University College Dublin (UCD), Ireland. He has been a researcher in the areas of Computational Linguistics, Cognitive Science, Cognitive Linguistics and Artificial Intelligence since 1988, both in industry and in academia. He received his Ph.D in Computer Science from Trinity College, Dublin in 1996. He is the author of *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity* (Bloomsbury Academic, 2012), and 4 other books as well as more than 50 research papers. He has developed text-understanding and machine translation systems, natural-language-processing tools, and patented web-based question-answering technology for a number of EU and US companies and projects.

Address

University College Dublin, School of Computer Science
Comp Science and Informatics
Belfield Dublin 4, Ireland

e-mail: tony.veale@ucd.ie
ORCID: 0000-0003-2375-1811

Milana Bolatbek
al-Farabi Kazakh National University

Shynar Mussiraliyeva
al-Farabi Kazakh National University

DETECTION OF EXTREMIST MESSAGES IN WEB RESOURCES IN THE KAZAKH LANGUAGE

Abstract

Currently, the Internet information and communication network has become an integral part of human life. People use social networks such as Twitter, VKontakte, Facebook, etc., to establish global contacts, exchange opinions, gain knowledge, etc. The active participation of not only individual users, but also information organizations in the entire world space makes it necessary to develop measures that correspond to modern trends in the development of information and communication technologies to ensure national security, in particular, the organization of events related to countering the strengthening of ideas of extremism and terrorism.

Countering the spread of aggressive information on the global network is an urgent problem of society and government agencies, this task is solved by filtering unwanted Internet resources. However, terrorist and extremist groups rationally use web technologies to perform various functions, including information dissemination, propaganda, fundraising and extremist missions. In such a situation, the Internet poses a threat to national security.

In this paper, we investigate the issue of creating semantic analysis models to identify extremist messages in the Kazakh language. For the study, a proprietary text corpus was assembled and models based on bigrams and word input methods were proposed. According to the results of experiments, the proposed model shows the highest indicators for evaluating machine learning methods.

Keywords

web-resources, social network, cybersecurity, extremism, text classification

1 Introduction

Nowadays the Internet is actively used to post extremist materials. The Internet provides extremists with effective means to spread their ideologies. Social networks, blogs, forums and other platforms allow them to easily reach a wide audience by anonymously

or pseudonymously expressing their views and attracting new supporters. They can use social media to find people who share their views and convince them to join extremist organizations. Using the global Internet network and computer communication capabilities, ideologists of extremist movements and groups actively influence the consciousness of citizens, especially young people. The problem is of a global nature and is very relevant for the Republic of Kazakhstan, as one of the main participants in the global political process. The fight against online extremism requires the combined efforts of governments, technology companies and society as a whole.

Taking into account the above conclusions, it can be concluded that the task of building models to identify extremist texts in the Kazakh language on web resources is the most relevant.

The purpose of the paper is a comprehensive study and creation of a model of semantic analysis to identify extremist texts in the Kazakh language on web resources.

Research objectives. To achieve this goal, the following tasks are provided:

- (1) creation of a corpus of extremist texts in the Kazakh language for training and testing machine and deep learning methods for identifying extremist texts in the Kazakh language on web resources;
- (2) creation of a semantic analysis model for identifying extremist texts in the Kazakh language on web resources;
- (3) creation and study of methods for identifying extremist texts in the Kazakh language on web resources;
- (4) creating a list of extremist keywords in the Kazakh language;

2 Literature review

The paper by Irani et al. (2021) addresses the challenge of early detection of online hate speech spreaders by framing it as a social media author profiling task. The primary goal of the study is to predict whether an author is likely to spread hate speech in the future based on a small collection of tweets. The authors approach this problem as a social media author profiling task.

To address the complexity of hate speech detection, Badjatiya et al. (2017) conduct extensive experiments using multiple deep learning architectures. The primary emphasis is on leveraging semantic word embeddings to capture nuanced linguistic patterns associated with hate speech. The experiments are carried out on a benchmark dataset comprising 16,000 annotated tweets.

The study by Khanday et al. (2022) addresses the impact of the COVID-19 pandemic on global social dynamics, emphasizing the role of social isolation as a major protective measure. It highlights the prevalence of people expressing themselves on platforms like Facebook and Twitter, with a specific focus on the dissemination of disinformation and hate speech on Twitter. The primary objective of the research is to detect hate speech on

Twitter during the COVID-19 pandemic, leveraging machine learning and ensemble learning techniques.

The research by Munasinghe et al. (2022) contributes to the broader field of hate speech detection by bridging the gap for low-resource languages and showcasing the applicability of advanced techniques, such as deep learning, to improve moderation capabilities in diverse linguistic contexts. The study concludes that the proposed deep learning-based approach for hate speech detection in the Sinhala language is a significant advancement over traditional machine learning models. The model's robust performance, demonstrated generalizability, and positive impact of extra features underscore its efficacy in addressing the challenges posed by hateful expressions on social media platforms.

Mansur et al. (2023) aim to provide a comprehensive resource outlining the various approaches, methods, and techniques employed to address the issue of Twitter hate speech. The intention is to assist researchers in developing more effective models for future studies. By presenting current perspectives and underscoring research opportunities, the paper contributes to the continuous improvement of hate speech detection systems. Ultimately, this advancement benefits social networking services by mitigating the risk of targeted harassment through the early detection of hate messages generated by users before they are posted.

Identifying Multiword Expressions (MWEs) in tweets is a challenging task due to the complex linguistic nature of MWEs combined with the non-standard language use in social networks. The article by Zampieri et al. (2022) explores joint experiments on MWE identification and its influence on Hate Speech Detection (HSD) tasks in English Twitter data. Two systems are compared: a lexicon-based system and a deep neural networks-based (DNN) system. The DNN system, based on recurrent networks and utilizing pre-trained contextual embeddings from BERT, is experimentally evaluated in seven configurations. A new DNN architecture is proposed for incorporating MWE features into the HSD task. The performance of the proposed DNN architecture is compared with previous MWE-based HSD systems on four Twitter HSD corpora.

The research by Mussiraliyeva et al. (2022) introduces a novel corpus designed for the detection of religious extremism in social networks and open sources, focusing specifically on the Kazakh language. The study aims to assess the feasibility of detecting extremist messages through binary classification and compares the performance of various machine learning algorithms using a pre-existing corpus.

Govers et al. (2023) propose and examine the research field of ERH (Extremism, Radicalization and Hate speech) context mining to unify disjoint studies. They evaluate the entire design process, including socio-technical definition-building, dataset collection strategies, technical algorithm design, and performance evaluation. The systematic literature review (SLR) spanning from 2015 to 2021 provides a cross-examination of textual, network, and visual approaches to detecting extremist affiliation, hateful content, and radicalisation towards groups and movements. The paper identifies consensus-

driven ERH definitions and suggests solutions to existing ideological and geographic biases, particularly addressing the lack of research in Oceania/Australasia. The hybridized investigation on Natural Language Processing, Community Detection, and visual-text models indicates the dominating performance of textual transformer-based algorithms in ERH context mining.

Another paper by Johansson et al. (2016) explores the feasibility of detecting violent extremism through the identification of warning behaviours in written text, referred to as linguistic markers. The approach involves leveraging computer-based analysis, specifically employing natural language processing techniques.

Gaikwad et al. (2021) present a comprehensive and comparative systematic literature review of datasets, classification techniques, validation methods, and online extremism detection tools. The review follows the PRISMA methodology, systematically collecting 64 studies on extremism research. This includes 31 studies from SCOPUS, Web of Science (WoS), ACM, IEEE, and 33 additional sources, encompassing theses, technical reports, and analytical reports using the Snowballing technique.

The paper by Okechukwu et al. (2023) explores the detection of hate and offensive speech using the Term Frequency-Inverse Document Frequency (TF-IDF) technique combined with a majority voting ensemble of machine learning classification models. The paper concludes that the employed approach, combining TF-IDF and a majority voting ensemble, is effective in detecting hate speech with a high level of accuracy. This research contributes to the ongoing efforts to address hate speech on social media platforms, providing a methodology that demonstrates promising results in terms of accuracy and F-Measure.

We applied a generalized offensive language detection methodology, partly based on the methodology referred to in this section.

3 Methodology

The methodology for extremist language detection in social media typically involves a series of steps to preprocess data, extract features, select a suitable model, train and evaluate the model, and post-process the results. Next chapters describe a generalized methodology for extremist language detection.

3.1 Data collection

The methodology for identifying extremist texts on the web includes a number of steps and approaches that help to automatically identify content that violates norms and laws related to extremism. It is important to note that identifying extremist content is a difficult task, as it requires a balance between freedom of expression and preventing the spread of malicious material. The general methodology for identifying extremist texts is

presented below: the first stage is data collection: data is collected from various sources, such as social networks, forums, blogs and news sites, then determined keywords, phrases, or topics related to extremism that can be used to filter content. The second stage is the preprocessing of the texts of the collected corpus, here algorithms for tokenization, removal of stop words, lemmatization and cleaning of unwanted characters are performed, where Natural Language Processing (NLP) methods are used. The third stage is the formation of signs, the characteristics of the text that can be used to identify extremist content are determined. This may include the use of n-grams, tonality analysis, the definition of keywords and phrases, as well as other textual features.

Machine learning algorithms such as classification or clustering are used to train the model based on annotated data, a suitable algorithm is selected depending on the nature of the task (for example, random forests, the support vector method, neural networks), the effectiveness of the model is evaluated by metrics such as accuracy, completeness, F1-measure.

Extremist texts in the Kazakh language were found mainly among comments on news portals, on social networks *Youtube*, *Vkontakte*. However, the extremist texts found in this way were not sufficient to train machine learning systems. About 400 texts of extremist content in the Kazakh language containing about 4,400 words have been found and included in the corpus in open sources.

In order to expand the corpus containing extremist texts, the task was set to choose the type of social network that is actively used among the Kazakh-speaking audience. The analysis of texts on the social networks *Vkontakte*, Facebook, Twitter was carried out, as a result of which the social network *Vkontakte* was selected. Next, it is necessary to provide lists of groups whose texts should be uploaded to the input of the parser module. The list of groups on which parsing is made in the social network was formed using key words defined using the TF-IDF method for extremist texts in open sources. In particular, the words “war”, “jihad”, “Sham”, “Syria” and others were used to identify social media groups. In the region field, “Kazakhstan” was selected (Fig.1).

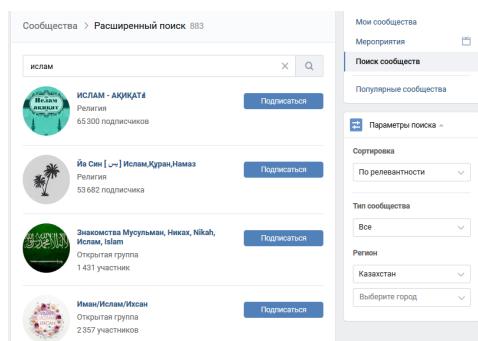


Figure 1: Example of searching for social network groups

In the present study, both extremist texts and texts of neutral content were collected in the corpus. Texts including calls to commit extremist activities, financing of extremist organizations, interest in weapons, training in counterfeiting weapons, etc. were chosen as the text of extremist content. For the texts of the neutral category, groups of general vocabulary were selected, widely known in the Kazakh audience, which do not contain religious content.

For the annotation of the data, the texts of extremist ideas from the social network Vkontakte were collected, and all messages were checked manually to make sure that they were correctly labelled. Annotation rules and message examples are given in Table 1.

Table 1: Annotation rule

Categories	Rules	Examples
Extremist message	(i) means extremist thoughts.	I think it's okay to do jihad to get a reward.
	(ii) includes extremist actions that may be performed	I came to jihad to Syria.
Neutral message	(iii) not related to extremism.	I really like this book.

The total number of extremist messages included in the corpus is about 1,200, and the total number of words in the corpus is about 140,000 words.

The compiled text is saved as a document in aCSV format. The document containing extremist texts consists of 4 columns: the serial number of the message, the message itself, the updated version of the message and the attribute "1/0", to which the corresponding entry belongs. Fig 2 – Example of messages of an extremist nature

The extremist texts in the Kazakh language are analysed, some characteristic features are revealed:

- (1) replacement of the original letters of the Kazakh language with Cyrillic;
- (2) frequent usage of bigrams;
- (3) frequent use of religious terms in Arabic;
- (4) the presence of several spellings of the same word;
- (5) a large number of spelling errors in the corpus (not grammatical, but typographical errors that occur when typing).

Extremist texts in the corpus contain the following information:

- (a) official reports of the Afghan Taliban movement ("Voice of Jihad");
- (b) reports of the Al-Qaeda organization;
- (c) extremist about actions interested in the Kazakh language, and in relation to texts (records, request for translation into Kazakh, rendering of Kazakh legislation);
- (d) false translation of the Book of the Quran;
- (e) records related to weapons training exercises;
- (f) drug records;

- (g) gender discrimination;
- (h) entries with a list of Caucasian jihadists;
- (i) records of users interested in visiting Syria (asking how to get to Syria, asking how jihad is happening, etc.) [103].

Word clouds. Word clouds were used to visually represent data. Messages from users with potentially extremist ideas are shown separately in Figures 2a and 2b.

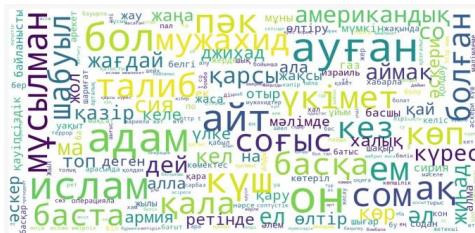


Figure 2a: Word cloud of messages related to extremism



Figure 2b: Word cloud of messages unrelated to extremism

The extremist part of the corpus is characterized by words such as *шаабуыл* ['attack'], *қарсы* ['against'], *соғыс* ['war'], *куи* ['power'], *өттір* ['kill'], and in the neutral part there are often words such as *адам* ['person'], *үйрен* ['train'], *қала* ['city'], *бақытты* ['happy'], *уақыт* ['time'].

3.2 Proposed architecture

This section provides a general architecture of the system for identifying extremist texts in Kazakh in social networks based on natural language processing methods using machine learning algorithms. In the process of learning machine and learning algorithms, the characteristics of the features used are given. The proposed architecture consists of several modules: a data extraction and preprocessing module, a tag extraction and data markup module, and an analysis module using deep learning models.

All words that semantically do not affect the classification process were excluded as stop words, such as “and”, “tomorrow”, “yesterday”, “it turns out”, etc. After clearing the data and deleting stop words, each sentence was tokenized into words based on empty space. Disabling a short word: short words whose length does not exceed three in the texts are excluded.

Letter conversion: after short words are erased, the letters in the social network texts are converted to lowercase. This is an important step, as the analysis algorithm is case-sensitive. Probabilistic models, for example, consider the words “Bad” and “bad” to be different words, and they count the appearance of each word separately. If words are not converted to lowercase, this can reduce the effectiveness of the classifier. Low register is a common method of pre-processing text.

Stemming: stemming is the process of removing affixes from words and bringing words into their root form. Removing suffixes helps reduce the complexity of labels and improves the readability of classifiers.

Delete @ and other punctuation marks: another common method of pre-editing text is to remove punctuation marks from text data. This process involves standardizing text, ensuring that phrases like “Hurray” and “Hurray!” are recognized as a single expression.

There are many ways to represent text data in vector format, such as a Bag of Words (BOW), TF-IDF. These vectors are vectors based solely on lexicon, usually denoting the number of words in a document or its relative weight. Word embedding is a complex of different approaches to learning visualizations aimed at comparing small-sized vectors with words in a dictionary when modelling language and processing natural language.

4 Results

In this paper, the problem of classifying the text into extremist and neutral categories was solved using several machine learning methods. The result of the experiment is presented in the following Table 2:

Table 2: The result of classifying text into extremist and neutral categories using machine learning methods

Machine learning methods	Accuracy	F1-Score	AUC-ROC
Logistic regression	0.77	0.63	0.86
k-nearest neighbors	0.78	0.66	0.5
Decision Tree	0.77	0.64	0.73
Random Forest	0.73	0.52	0.68
Gradient Boosting	0.75	0.58	0.71
SVM	0.73	0.53	0.68
Naïve Bayes	0.86	0.84	0.86
Proposed model (TF-IDF bigram LSTM)	0.9	0.88	0.89

In conclusion, three different models have been created to identify extremist texts. A number of experiments were carried out on the proposed models. As a result of the experiments carried out, a model is proposed to identify texts of extremist content on web resources, which, along with the symbols of TF-IDF+bigrams, will be based on deep learning algorithms using the stemming algorithm. The combination of these features showed high results in all evaluation parameters during the experiments.

5 Conclusion

The creation of a semantic model for identifying extremist texts is an important aspect in the development of effective filtering and detection systems for such content. The creation of a semantic model for identifying extremist texts is an important aspect in the development of effective filtering and detection systems for such content. Semantic models are able to take into account not only individual words, but also their interrelationships and context, which makes it possible to more accurately understand the meaning of the ideas expressed. This is especially important when analysing extremist texts, where often reinterpreted and ironic constructions are used.

In this work, research was carried out to create a model of semantic analysis of the identification of extremist texts in the Kazakh language on web resources and the following results were achieved: for the first time, a corpus of extremist texts in Kazakh has been created for teaching and testing machine learning methods to identify extremist texts in Kazakh; for the first time, a semantic analysis model has been developed, characterized by the application of the TF-IDF method to bigrams, in which a pre-stemming algorithm is implemented at the level of embedding words of the LSTM network, taking into account the peculiarities of the Kazakh language and increasing the accuracy of determining extremist texts; methods of forming a set of features and methods based on combining n-grams and word input and improving the quality of classification of extremist texts have been developed; for the first time in the Kazakh language, a list of extremist keywords has been formed; as a result of the developed models and methods, software has been created to identify extremist texts in the Kazakh language.

Taking into account these factors, the semantic model becomes an important tool in the development of systems aimed at identifying extremist content on the web. It also highlights the need for constant updating and development of the model so that it can effectively cope with emerging new forms of extremist expressions.

Acknowledgements

This work was supported by the project “Development of models and methods to identify youth extremism and ensure the safety of youth in the modern information space” funded by the grant of young scientists for research and/or research- technical projects for 2023-2025 granted by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP19576868. Supervisor of the project is Milana Bolatbek, email: bolatbek.milana@gmail.com.

References

- Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta & Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (WWW ‘17 Companion), 759–760. International World Wide Web Conferences Steering Committee. DOI: <https://doi.org/10.1145/3041021.3054223>.
- Gaikwad, Mayur, Swati Ahirrao, Shraddha Phansalkar & Ketan Kotecha. 2021. Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools. *IEEE Access* 9. 48364–48404. DOI: [10.1109/ACCESS.2021.3068313](https://doi.org/10.1109/ACCESS.2021.3068313).
- Govers, Jarod, Philip Feldman, Aaron Dant & Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Computing Surveys* 55(145). 1–35. DOI: <https://doi.org/10.1145/3583067>.
- Irani, Darius, Avyakta Wrat & Silvio Amir. 2021. Early Detection of Online Hate Speech Spreaders with Learned User Representations. *Conference and Labs of the Evaluation Forum*. Available at: <https://api.semanticscholar.org/CorpusID:237298964> (accessed 30 September 2022).
- Johansson, Fredrik, Lisa Kaati & Magnus Sahlgren. 2016. Detecting Linguistic Markers of Violent Extremism in Online Environments. *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications*. IGI Global. DOI: [10.4018/978-1-5225-1759-7.ch118](https://doi.org/10.4018/978-1-5225-1759-7.ch118).
- Khanday Akib Mohi Ud Din, Syed Tanzeel Rabani, Qamar Rayees Khan & Showkat Hassan Malik. 2022. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques, *International Journal of Information Management Data Insights* 2(2). 100120. DOI: <https://doi.org/10.1016/j.ijimi.2022.100120>.
- Mansur, Zainab, Nazlia Omar & Sabrina Tiun. 2023. Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities. *IEEE Access* 11. 16226–16249. DOI: [10.1109/ACCESS.2023.3239375](https://doi.org/10.1109/ACCESS.2023.3239375).
- Munasinghe Sidath & Uthayasaner Thayasivam. 2022. A Deep Learning Ensemble Hate Speech Detection Approach for Sinhala Tweets. *Moratuwa Engineering Research Conference (MERCon)*, 1–6. DOI: [10.1109/MERCon55799.2022.9906232](https://doi.org/10.1109/MERCon55799.2022.9906232).
- Mussiraliyeva, Shynar, Batyrkhan Omarov, Paul Yoo & Milana Bolatbek. 2022. Applying machine learning techniques for religious extremism detection on online user contents, *Computers, Materials & Continua* 70(1). 915–934. DOI: <https://doi.org/10.32604/cmc.2022.019189>.
- Okechukwu, Chukwuemeka, Idris Ismaila, Joseph Ojeniyi, Morufu Olalere & Olawale Surajudeen Adebayo. 2023. Hate and Offensive Speech Detection Using Term Frequency-Inverse Document Frequency (TF-IDF) and Majority Voting Ensemble Machine Learning Algorithms. 4th International Engineering Conference (IEC 2023), Federal University of Technology, Minna,

- Nigeria. Available at: <http://repository.futminna.edu.ng:8080/jspui/handle/123456789/18492> (accessed 20 September 2023).
- Zampieri, Nicolas, Carlos Ramisch, Irina Illina & Dominique Fohr. 2022. Identification of Multiword Expressions in Tweets for Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 202–210, Marseille, France: European Language Resources Association.

About the Authors

Milana Bolatbek holds PhD in Information security systems. She is senior lecturer at Al-Farabi Kazakh national university in Almaty (Kazakhstan). She is also supervisor of the project named “Development of models and methods to identify youth extremism and ensure the safety of youth in the modern information space”. Her research interests include information security, natural language processing, semantic analysis, social media analysis.

Address

Department of Information systems, Al-Farabi Kazakh national university
Al-Farabi av., 71, 050040, Almaty, Kazakhstan

e-mail: Bolatbek.Milana@kaznu.kz
ORCID: 0000-0002-2153-180X

Shynar Mussiraliyeva is candidate of physical and mathematical sciences. She is head of the department of Information systems at Al-Farabi Kazakh national university in Almaty (Kazakhstan). She is also supervisor of the project named “Multi-ideology Cyber Extremism Classification in the Kazakh language using Artificial Intelligence”. Her research interests include information security, cryptography, semantic analysis, social media analysis.

Address

Department of Information systems, Al-Farabi Kazakh national university
Al-Farabi av., 71, 050040, Almaty, Kazakhstan

e-mail: Shynar.Musiraliyeva@kaznu.kz
ORCID: 0000-0001-5794-3649

Talya Natanya
Jerusalem College of Technology

Chaya Liebeskind
Jerusalem College of Technology

CLICKBAIT DETECTION IN HEBREW

Abstract

The prevalence of sensationalized headlines and deceptive narratives in online content has prompted the need for effective clickbait detection methods. This study delves into the nuances of clickbait in Hebrew, scrutinizing diverse features such as linguistic and structural features, and exploring various types of clickbait in Hebrew, a language that has received relatively limited attention in this context. Utilizing a range of machine learning models, this research aims to identify linguistic features that are instrumental in accurately classifying Hebrew headlines as either clickbait or non-clickbait. The findings underscore the critical role of linguistic attributes in enhancing the performance of the classification model. Notably, the employment of a machine learning model resulted in an impressive accuracy of 0.87 in clickbait detection. Moving forward, our research plan encompasses dataset expansion through the best machine learning model assisted labelling, with the objective of optimizing deep learning models for even more robust outcomes. This study not only advances clickbait detection in the realm of Hebrew but also emphasizes the fundamental importance of linguistic features in the accurate classification of clickbait.

Keywords

clickbait detection, machine learning, weak supervision, Hebrew

1 Introduction

In today's digital landscape, the internet has become a primary source of information and news consumption for many individuals. However, the abundance of online content has also given rise to a significant challenge – the proliferation of clickbait. Clickbait is detrimental to the online experience due to its misleading and sensational nature; it creates a false sense of importance or urgency, exploiting readers' curiosity for the sole purpose of generating website traffic or ad revenue. This not only wastes users' time but also contributes to the spread of misinformation and devalues the credibility of online content. It undermines trust between publishers and readers, hinders informed decision-making, and erodes the overall quality of online information.

By developing effective clickbait detection methods, we aim to empower internet users to discern between authentic, high-quality content and clickbait articles. Detecting

and flagging clickbait can help users save time, avoid misinformation, and foster a more transparent and reliable online environment. Furthermore, clickbait detection can assist content creators, news outlets and platforms in maintaining their credibility and building trust with their audiences.

As evidenced by previous research, linguistic features play a crucial role in clickbait detection. For instance, Prokofeva and Akulovich (2021) posit that clickbait is the result of selecting and combining linguistic mean. Their investigation into the Russian language highlights the mechanisms of creating a clickbait in the semantic field of "funny" from the positions of the addressee, the subject of laughter, the object and the matter of laughter. This confirms the possibility of developing a model for identifying clickbaits. Likewise, in his research on Indonesian, Fakhruzzaman et al. (2021) demonstrate that the distinction between clickbait and non-clickbait hinges on the usage of informal words, named entities, and various parts of speech. In Arabic, Al-Sarem et al. (2021) discovered that employing linguistic features led to the optimal performance of machine learning classifiers. Drawing upon these findings, we hypothesize that these linguistic components will also prove significant in detecting clickbait in Hebrew. Employing deep learning models, we aim to quantify the significance of these components and demonstrate their contribution to clickbait identification.

Our methodology for clickbait detection includes several stages as described in Figure 1.

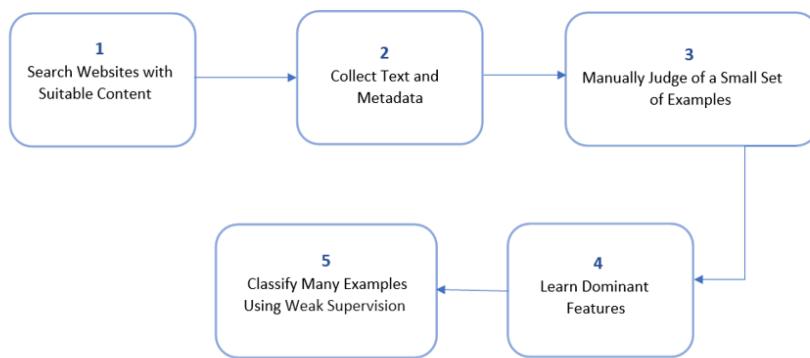


Figure 1: The methodology's stages

In the first step, we identified websites that were likely to contain clickbait content, characterized by provocative headlines and emotionally charged language. Subsequently, texts and metadata were meticulously gathered from the selected websites, encompassing relevant fields. Following this, a subset of examples designated for the machine learning models underwent manual assessment. In the fourth step, leveraging

machine learning models, such as Linear Regression, Support Vector Machine (SVM) and Naive Bayes, our aim was to identify dominant features that yielded optimal model performance. By reducing the features list to encompass solely linguistic features, headline structural features, and URL structural features, we attained an outstanding accuracy of 0.87. This highlights the effectiveness of these feature selection methods.

Looking ahead to the fifth step, our future plan entails enhancing the dataset's comprehensiveness through the use of weak supervision annotations to avoid manual labelling of large quantities. This approach involves automatically labelling a range of examples by the machine learning model developed in step four. This effort will result in the creation of an expansive dataset, which will serve as the foundation for upcoming in-depth investigations. These investigations will involve advanced deep learning models characterized by intricate neural networks and a multitude of parameters, reinforcing the need for a substantial dataset to achieve accurate and reliable outcomes.

2 What is clickbait?

According to the Oxford Dictionary, “Clickbait is a content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page”. Dictionary.com defines clickbait as “A sensationalized headline or piece of text on the Internet, designed to entice people to follow a link to an article on another web page”. These two dictionary definitions refer to clickbait more as an attractive headline. Yet, Wikipedia, a collaborative encyclopaedia, portrays clickbait more as intentionally misleading. It describes clickbait as “A text or a thumbnail link that is designed to attract attention and to entice users to follow a link and read, view, or listen to the linked piece of online content, being typically deceptive, sensationalized, or otherwise misleading”.

It seems there is no universally agreed-upon definition of clickbait; it has a variety of definitions, depending upon its attributed purpose, whether it is an intentionally misleading headline or just an attractive one. The usage of clickbait capitalizes on the human nature of curiosity; it presents teaser messages in headlines, signals new information to readers, provokes their curiosity, and ultimately leads them to click on the headline.

Previous studies on clickbait (Chakraborty et al. 2016) have considered BuzzFeed, an American Internet media, news, and entertainment company with a focus on digital media, a source of clickbait. Its editor, like Wikipedia, portrays clickbait more as intentionally misleading, considering it "a headline that is dishonest about the content of the article". However, he claims that a headline can be eye-catching but accurately describe the article. Facebook, an online social media, in its efforts to reduce clickbait exposure to users, defines it as a headline that encourages clicks without providing a clear indication of the content users will encounter. This definition refers to clickbait more as ambiguous, than misleading.

Within the Natural Language Processing (NLP) research community as well, some researchers (Chakraborty et al. 2016; Anand, Chakraborty and Park 2017) regard clickbait to be primarily an attractive headline as defined by Potthast et al. (2016), “Clickbait refers to a certain kind of web content advertisement that is designed to entice its readers into clicking an accompanying link”, whereas others (Cao, Le, et al. 2017) view clickbait as intentionally misleading. For instance, Agrawal (2016) defines clickbait as follows: “Clickbaits, in social media, are exaggerated headlines whose main motive is to mislead the reader to ‘click’ on them”.

In the following paper, the accepted definition of clickbait is the following: clickbait refers to a headline that entices readers to seek more information through a link by manipulation, exaggeration, or dishonesty about the content of the article, while a non-clickbait headline delivers straightforward key information.

3 Clickbait detection

Clickbait detection has been a topic of interest in previous studies, where researchers have investigated different approaches to address this pervasive issue. In these studies, both classical machine learning methods and deep learning methods have been employed to detect clickbait content.

Zuhroh and Rakhmawati (2020) conducted a literature review on clickbait detection studies. The review analyzed a total of 21 works, which were presented in English or Indonesian. In the study, Zuhroh and Rakhmawati (2020) examined various aspects, including the sources from which the data were collected, the data preprocessing techniques employed, the features utilized in each study, and the classification methods employed. Among the 21 works, ten utilized deep learning techniques, while the remaining works employed machine learning methods.

3.1 Classical machine learning methods

Machine learning is a field of artificial intelligence that involves developing algorithms and models capable of learning from data, recognizing patterns, and making predictions or decisions without explicit programming. The initial and crucial step in machine learning process, as described in Figure 2, is feature extraction which encompasses identifying and extracting relevant information or features from the raw data, that can be numerical or categorical in nature. The goal is to transform the data into a suitable representation that captures important patterns and relationships. After the feature extraction, the data is divided into two sets: a training set and a test set. The training set is used to train the machine learning models by feeding the features and their corresponding target variables into the models. The models learn patterns and relationships between the features and the target variable, enabling them to make

predictions or decisions on new, unseen data. The test set is used to evaluate the performance of the trained models. By applying the trained models to the test set, their ability to generalize and make accurate predictions on unseen data can be assessed. This evaluation helps measure the effectiveness of the machine learning algorithms and guides further improvements, if necessary.

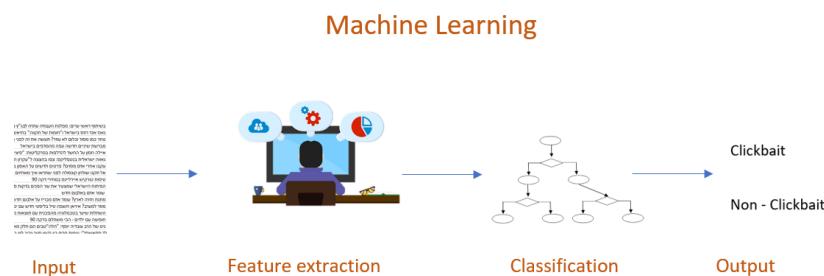


Figure 2: Machine learning process

Previous studies in the field of clickbait detection using machine learning have used a range of features, including (1) basic statistical features, such as headline's length, average word length, length of the longest word and more (Potthast et al. 2016; Cao, Le, et al. 2017; Chakraborty et al. 2016; Biyani, Tsoutsouliklis and Blackmer 2016; Marreddy et al. 2021), (2) structural features, for example, the number of uppercase words, presence of superlative adverbs and adjectives or whether the headline contains third personal pronouns (Potthast et al. 2016; Cao, Le, et al. 2017; Biyani, Tsoutsouliklis and Blackmer 2016; Marreddy et al. 2021) and (3) semantic features like sentiment polarity and the count of negative/positive sentiment words (Potthast et al. 2016; Cao, Le, et al. 2017; Chakraborty et al. 2016; Biyani, Tsoutsouliklis and Blackmer 2016).

While some articles (Chakraborty et al. 2016; Marreddy et al. 2021) focused solely on analyzing the headline without considering the accompanying text, others (Potthast et al. 2016; Cao, Le, et al. 2017; Biyani, Tsoutsouliklis and Blackmer 2016; Dam, Panday and Thapa 2021) incorporated features related to the text itself, for instance, the similarity between the headline and the first five sentences of the text, as well as the ratio of stop words present in the text.

To represent the words in the headline and the content of the URL, the bag-of-words (BoW) approach has been employed in some previous studies (Potthast et al. 2016; Marreddy et al. 2021). The BoW method involves creating a collection or “bag” of words from the given text, disregarding grammar and word order. Each word is then assigned a numerical value based on its frequency in the text, allowing for a quantitative representation of the textual data into a format suitable for machine learning algorithms.

3.2 Deep learning methods

Deep learning is a subfield of machine learning that focuses on training artificial neural networks with different topologies to learn and extract high-level representations from unstructured and complex data such as images, speech, and text. Unlike classical machine learning, which relies on manual feature engineering, deep learning algorithms automatically learn hierarchical representations from raw input data, enabling them to discover intricate patterns and relationships. Figure 3 depicts this process of deep learning.

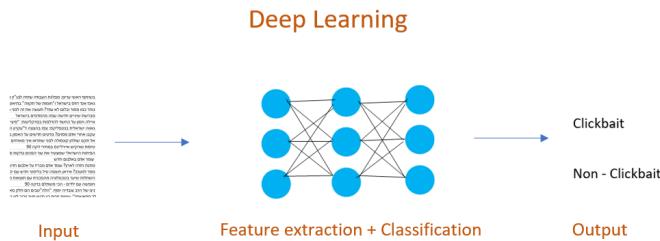


Figure 3: Deep learning process

In previous studies in the field of clickbait detection using deep learning, different topologies were used. Some studies have used shallow topology, such as Word Embedding (Anand, Chakraborty and Park 2017; Agrawal 2016), which addresses the information sparsity problem of bag-of-words by representing text in a reduced dimension. It captures semantic relationships between words by mapping them to vectors based on their contextual usage, while others have utilized deeper networks, where the Word Embedding acts as input to the hidden layer, like Convolutional Neural Networks (CNN) (Zheng et al. 2018; Fu et al. 2017) and BERT (Bidirectional Encoder Representations from Transformers) (Fakhruzzaman et al. 2021; Marreddy et al. 2021).

In this paper, our focus on linguistic interpretability necessitates the prioritization of machine learning models over deep learning models. Despite the impressive capabilities and increased accuracy demonstrated by deep learning models, their inherent lack of interpretation is a significant challenge. The complex workings of these models often remain obscure and incomprehensible to humans, making it challenging to pinpoint the specific features essential to their classification decisions. In contrast, machine learning models offer a valuable advantage in terms of interpretation. Their internal mechanisms are more transparent and explainable, providing a clearer view of the rationale behind

their output. This transparency allows us to discern the specific words within a title that have the most significant impact on the model's classification decisions, enabling a deeper exploration of linguistic features in the clickbait detection process.

4 Clickbait detection in Hebrew

While most of the research in this area has focused on English-language content (Anand, Chakraborty and Park 2017; Potthast et al. 2016; Dong et al. 2019; Kumar et al. 2018; Cao, Le, et al. 2017), there have also been efforts to extend this research to other languages, including Turkish (Geçkil et al. 2018), Thai (Klairith and Tanachutiwat 2018), Telugu (Marreddy et al. 2021) and Chinese (Zheng et al. 2018), where the challenge in dealing with other languages lies in the absence of certain features, such as word semantics and recognition of superlative words. However, one language that has not yet been extensively studied in this context is Hebrew. While there has been exploration into explicitly offensive language detection in Hebrew, the exploration of implicitly offensive language detection (Liebeskind and Liebeskind 2018; Litvak et al. 2022), including the specific case of clickbait, remains unexplored. To fill this gap in the literature, we propose a methodology for clickbait detection in Hebrew, with the aim of providing a framework that can be useful for other languages as well.

4.1. Data collection

The data for this study were collected from Hebrew news websites: Mako¹, N12², Ma'ariv³ and Walla⁴ spanning diverse domains such as politics, health, sports, finance, entertainment, and science. To ensure a comprehensive clickbait sample, initial content collection, which includes around 1050 articles, focused on a region with abundant sponsored content, while an additional 230 articles were meticulously chosen out of 350 articles from a standard region, to create dataset equilibrium. To collect the data, web scraping techniques were employed, enabling the automated extraction of relevant information from these identified sources. The resulting dataset comprises around 700 clickbait and 700 non-clickbait examples, providing a balanced and representative foundation for the study.

Each headline was manually labelled as either clickbait or non-clickbait. For this task we predefined a set of categories based on our definition of clickbait: “clickbait refers to a headline that entices readers to seek more information through a link by manipulation, exaggeration, or dishonesty about the content of the article”.

¹ <https://www.mako.co.il/>

² <https://www.n12.co.il/>

³ <https://www.maariv.co.il/>

⁴ <https://www.walla.co.il/>

- (1) A headline that piques curiosity but fails to provide substantial information.

(Hebrew reading direction: right to left)

a. **נִגְמַלּוּ מָסֻכָּר וּפְחָמִימֹת יוֹרְדוּ בָּמִשְׁקָל :** *כַּךְ עֲוֹבְדַת הַשִּׁיטָה שֶׁל מַכְוֵן אַבְרָהָם-סּוּן*
Ngmlw mswkr wph. mmwt wyrdw bmql: kk 'wbdt hšt .h šl mkwn 'brhmswn⁵
[‘Freed from an addiction to sugar and carbohydrates and lost weight: this is how the method of the Abramson Institute works’] (there is no description of the method in the link)

b. **שְׁנֵי אוֹיְבֵי הַגְּדוּלִים שֶׁל הַגָּב :** *חוֹסֵר בְּפִעְולֹת גּוֹפְנִית וַיִּשְׁבַּת מְמוֹשְׁכָת*
šn 'wbw hgdwlm šl hgb: h.wsr bp'lwt gwpnt wyšbh mmmwškt
[‘The two biggest enemies of the back: lack of exercise and prolonged sitting’] (in the link there is an advertisement about a therapeutic swimming class)

- (2) A headline that presents a fact in an exaggerated and arguable manner.

a. **הַקִּיץ בְּפֶתַח / שִׁיפְלָנִיקְיָה מִזְגְּנָן שְׁכּוֹלָט צִוְיכִים לְדֹעַת!**
hqṣ bpṭh! typm lnqw mzgn škwlm ṣrk̠m ld't!
[‘Summer is coming! Tips for cleaning air conditioners that everyone should know!?’]

b. **הַטְּרִיקָה לְנַאֲגַשְׁת הַכִּי טְעִימִים בָּאָרֶץ**
Hṭrq ln'gts hk t'mm b'rṣ
[‘The trick to the tastiest nuggets in the country’]

- (3) A headline that references an unidentified subject or entity.

a. **לֹא דָקְכָּבִים :** *הַתְּפִקֵּיד הַחֲדִישׁ שֶׁל גַּיְא פְּלֶג וְעַמְלִיה דּוֹאָק*
l' rq ktbm: htqpd hh. dš šl gy' plg w'mlh dw'q
[‘Not just reporters: the new role of Guy Peleg and Amelia Douek’]

b. **עֲבָשִׂיו וְהַרְשָׁמִי :** *מְגִישַׁת חֲדִשׁוֹת 12 עֹזְבָּת לְמַתְחָרִים מַעֲרִיז 13*
'ks̠w zh rsm: mgšt h. dšwt 12 'wbrt lmth. rm m'rws. 13
[‘Now it is official: News anchor 12 is moving to the competitors from channel 13’]

- (4) A headline that attempts to evoke emotional reactions from readers.

a. **מִשְׁרָדָה הַבְּרִיאָה מַזְהִיר :** *לֹא תָּאמְנוּ אַיִּזָּה מְדִכְבִּים נִמְצָאים בְּחֵזֶק הַרוֹלְ-אַפְּס*

⁵ ISO259-2 Hebrew transliteration

Mšrd hbr'wt mzhr: l' t'mnw 'yzh mrkbn nms. 'm btwk hrwl-'ps
[‘The Ministry of Health warns: You won't believe what ingredients are inside
the roll-ups’]

b. **תופתעו לגלות כמה עולה קיטרינג לאירועים קטנים בבית**
twpt'w lglwt kmh 'wlh qyt. rmg l'rw'm qt .nm bbt
[‘You will be surprised to find out how much catering for small events
at home costs’]

(5) A headline that attracts readers to seek more information in a manipulative way.

a. **עינב בובליל : "זה חטוף שעור לי לחוור לשגרה"**
ynb bwblil: "zh htwsp š'zr ly lh. zwr lgzrh"
[‘Einav Bobailil: “This is the supplement that helped me get back in
shape”’]

b. **איו טעה : זה שלב שאתה מצלג עליי בנקי השירותים - צפי**
yzw t'wt: zh hšlb š'tm mdlgm 'lw bnqw hšrwtm | spw
[‘What a mistake: the step you skip in cleaning the toilets | watch’]

This category is more general and there are headlines that also meet its criteria (like 3a). However, if a headline satisfied any of the previous categories, we prioritize to associate it to the more appropriate category.

(6) Non-clickbait

בעלי עסקים? קבלו במתנה פגישה פיצוח עסקים עם ניר דובדבני
B'l 'sqm? Qblw bmtnh pgst pys.wh. 'sq mnr dwbdbn
[‘Business owners? Get a free business coaching meeting from Nir Duvdevani’]

The use of terms like “gift” or “free” suggests an exaggerated or potentially misleading claim, but in the link, you can actually include your details and receive a free meeting.

מאות מקרים חתלוות : מעצר הunganת החשודה הוארך
m'wt mqqr ht' llwt: m's. r hgnnt hh. šwdh hw'rк
[‘Hundreds of abuse cases: the detention of the suspected nursery schoolteacher
was extended’]

While the specific name of the nursery schoolteacher is not explicitly mentioned in the headline, it can be inferred from the context and wording used. This makes the headline not qualify as clickbait since it indirectly provides information about the nursery schoolteacher in question.

4.2 Data annotation

The annotation task was performed by two annotators and unfolded over four rounds, strategically aimed at enhancing the categories precision to magnify the model's capacity for adept classification. In every round, each annotator meticulously assessed one hundred headlines. The results were then compared, and any discrepancies were thoroughly discussed and resolved to establish a shared understanding. In each round, the kappa index was calculated to measure the level of agreement between the annotators. The kappa index provides a measure of agreement beyond what would be expected by chance alone. Higher kappa values suggest a stronger level of agreement between annotators, indicating a more reliable and consistent annotation process.

In the first round, the agreement percentage between the annotators was 76.84%, indicating a moderate level of agreement. To further assess the reliability of the annotations, the kappa index was calculated, resulting in a value of 0.517. A kappa index between 0.41 and 0.60 indicates moderate agreement.

The second round demonstrated a notable improvement in agreement due to the discussions and resolutions. The agreement percentage increased to 90.81%, and the kappa index rose to 0.763. A kappa index between 0.61 and 0.80 suggests substantial agreement.

In the third round, although there were more examples of political issues that caused slight confusion, they contributed to refining the understanding. The agreement percentage was 87.5%, with a corresponding kappa index of 0.608.

By the fourth round, a high level of agreement was achieved, with the agreement percentage reaching 94.62%. The kappa index for this round was 0.883, indicating almost perfect agreement.

Throughout the four rounds, each example was carefully categorized according to its corresponding category (one of the five categories or not clickbait). Figure 4 summarizes the distribution of our data after the annotation task. The most prominent percentages are 27% of the data satisfied category 2, 22% satisfied category 3, and 15% satisfied category 4. It is worth noting that the ratio between clickbait and non-clickbait is not equal at this point. This is because we deliberately focused on the areas with more clickbait, as it is usually easier to identify what is certainly not clickbait and add those examples later.

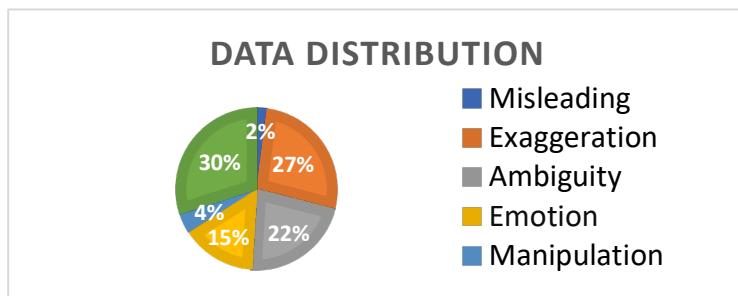


Figure 4: Data distribution

The remaining samples were labelled by one annotator, with consultation from the second annotator only in cases of doubt or uncertainty. This collaborative approach ensured the final quality and accuracy of the entire dataset.

4.3 Machine learning

The process of detecting clickbait using machine learning involves three key stages, as discussed in section 3.1. In this chapter, we delve into the implementation of each stage:

- (1) Feature extraction (Section 4.3.1): we meticulously extract diverse features from the raw data, aiming to capture essential aspects of the clickbait content.
- (2) Training classification models (Section 4.3.2): this stage involves exploring various machine learning algorithms and fine-tuning hyperparameters to create an accurate clickbait detection model.
- (3) Evaluation (Section 4.3.3): we assess the model's performance using standard metrics and a dedicated test set, comparing it against state-of-the-art approaches to validate its effectiveness.

4.3.1. Features set

For our machine learning model, we harnessed an extensive set of features, categorized into four distinct groups: statistical features based on frequency, structural features on the headline, semantic features, and structural features on the URL. Each category captures different aspects of the headlines, providing valuable insights for clickbait detection.

- (1) Statistical features based on frequency. This category includes features that capture basic statistical information about the headlines: the number of words (the headline length) and the average word length.
- (2) Headline structural features. This category focuses on the structural characteristics of the headlines: the number of stop words, the number of

acronyms, the number of exclamation marks, the number of question marks, the number of quoted words, the number of all kinds of punctuations, whether the headline contains a number, whether the headline contains an adverb, whether the headline starts with a number, whether superlative adverbs and adjectives are present, whether the headline contains demonstratives and whether the headline contains third person personal pronouns. The POS tags were extracted by using the YAP tool (More and Tsarfaty 2016).

- (3) Semantic features. This category encompasses features related to the semantic aspects of the headlines: the number of negative/positive sentiment words, negative/positive/natural sentiment score of the entire headline and the similarity between the headline and the first five sentences of the URL content. The semantic scores were calculated by using the HeBERT model (Chriqui and Yahav 2022).
- (4) URL structural features. This category focuses on features extracted from the associated URL of the headlines: frequencies of dashes, ampersands, upper case letters, commas, periods, equal-to signs, percentage signs, pluses, underscores, digits, and URL depth (number of forward slashes in the URL).

In addition to the aforementioned features, an essential aspect of our approach involves the representation of the textual content itself. We adopt a Bag-of-Words (BOW) approach, wherein the words in each headline are first stemmed to their root forms. This step aids in capturing the inherent essence of the language while reducing inflections and variations.

In the statistic group, there are two features, while the headline structural group involves 13 features, the semantic group entails ten features, and the URL structural group encompasses 11 features. Additionally, the Bag of Words (BOW) representation comprises 3531 words. In total, there are 3,567 features.

4.3.2 Methods

For the clickbait detection task, we employed five distinct machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, and Bagging. Each model offers a unique perspective on the task, allowing us to uncover intricate relationships within the features and their impact on clickbait classification.

- (1) Logistic Regression (Yu, Huang and Lin 2011): Logistic Regression is a binary classification technique that calculates the probability of an instance belonging to a certain class. It is a straightforward yet powerful method for estimating class probabilities based on the input features.
- (2) Support Vector Machine (SVM) (Chang and Lin 2011): Support Vector Machine is a robust classification algorithm that seeks optimal hyperplanes to separate classes, maximizing the margin between them. SVM is adept at handling high-

dimensional data and can handle both linear and non-linear separation through kernel functions.

- (3) Random Forest (Breiman 2001): Random Forest is an ensemble learning technique that constructs multiple decision trees to collectively make predictions. By combining these individual trees, Random Forest can mitigate overfitting and provide more reliable results through aggregation.
- (4) Naïve Bayes (Christopher, Prabhakar, Hinrich et al. 2008): Naïve Bayes is a probabilistic algorithm based on Bayes' theorem that leverages conditional probabilities to classify data. Despite its simplistic assumptions, it can be remarkably effective in various classification tasks.
- (5) Bagging (Breiman 1996): Bagging, short for Bootstrap Aggregating, employs multiple instances of a base model to train on different subsets of the data. These individual models' predictions are then aggregated to produce the final outcome, enhancing model performance and reducing variance.

For each model, we embraced a feature selection algorithm known as Recursive Feature Elimination with Cross-Validation (RFECV) (Guyon et al. 2002). This method holds the promise of identifying the most pertinent subset of features that would yield optimal results for each individual model. By iteratively excluding less informative features and assessing performance using cross-validation, RFECV enabled us to fine-tune the models to their peak capabilities and to ensure that each model harnessed only the most influential features, thus enhancing their predictive accuracy and interpretability.

Furthermore, our exploration encompassed the fine-tuning of a deep learning model known as HeBERT (Chriqui and Yahav 2022), an innovative adaptation of the BERT architecture tailored for the Hebrew language. This endeavor aimed to assess the effectiveness of employing advanced deep learning techniques within our dataset context. By adapting the HeBERT model to our data, we sought to unveil the potential advantages and insights that can be gained from leveraging more intricate neural network architectures for clickbait detection.

The evaluation of each model was executed on a designated test set, employing a data split approach that allocated 80% of the dataset for training the model, and reserved the remaining 20% for testing. To assess the performance, we used basic evaluation measures, including 1. accuracy, the percentage of predictions that were correct 2. precision, the percentage of positive predictions that were actually positive 3. recall, the percentage of positive cases that were correctly identified and 4. F1 score, the harmonic mean between precision and recall.

4.3.3 Results

Table 1 shows the optimal number of features chosen by Recursive Feature Elimination with Cross-Validation (RFECV), along with the corresponding results for each model. As we can see, the SVM model demonstrated superior scores across all the evaluation

metrics. This finding was further supported by the McNemar test (McNemar 1947), which showed that the results were statistically significant for all models except linear regression (LR). In the context of the bagging model, it is important to note that all the features were used. This is because bagging-based ensemble models inherently incorporate a form of feature selection by utilizing random subsets of features during training. Each individual model within the ensemble is trained on a distinct subset of features, which contributes to mitigating the ensemble's overall variance.

Table 1: Machine learning results

Model	Number of features	Accuracy	Precision	Recall	F1
SVM	1711	0.87	0.88	0.87	0.87
Logistic Regression	2317	0.85	0.86	0.85	0.85
Random Forest	2762	0.82	0.82	0.82	0.82
Bagging	All(3568)	0.79	0.79	0.79	0.79
Naive Bayes	1878	0.76	0.81	0.76	0.75

Given SVM's superior performance, we proceeded with a detailed analysis by groups, starting with checking the remaining features across the four feature groups and the BOW. The results were as follows: in the statistical group, no features were retained. The headline structural group saw a reduction from 13 features to ten (NumQuote, Has3PersPron, HasAdverb are not included). The semantic group did not yield any influential features. In the URL structural group, the number of features was trimmed from 11 to three (NumAmpersand(&), NumComma(,), NumPeriod(.) were left). Lastly, the BOW group exhibited a substantial reduction from 3,567 features to 1,698. We then proceeded to assess the impact of each group of reduced features on the model's performance. The outcomes, depicted in Table 2, reveal that the BOW group yields the greatest influence. This observation was further validated by a McNemar test (with a significance level of 0.05), which demonstrated a statistically significant difference between the results obtained from the three sets of features and the results obtained without BOW. This analysis underscores the influence of the headline words on the classification.

Figure 5 illustrates the confusion matrix of the SVM model, a visualization that provides insights into the model's performance by showcasing the distribution of actual and predicted class labels. This confusion matrix shows that the model correctly predicted 117 instances of clickbait and 130 instances of non-clickbait. However, it predicted 24 false negatives (misclassifying clickbait as non-clickbait) and 11 false positives (misclassifying non-clickbait as clickbait).

Table 2: SVM – results per group

Features Groups	Accuracy	Precision	Recall	F1
Structural	0.74	0.7	0.83	0.61
URL	0.74	0.68	0.87	0.56
BOW	0.82	0.81	0.86	0.76
Structural, URL	0.8	0.77	0.87	0.69
Structural, BOW	0.85	0.84	0.9	0.78
URL, BOW	0.84	0.83	0.89	0.77
Structural, URL, BOW	0.87	0.87	0.9	0.83

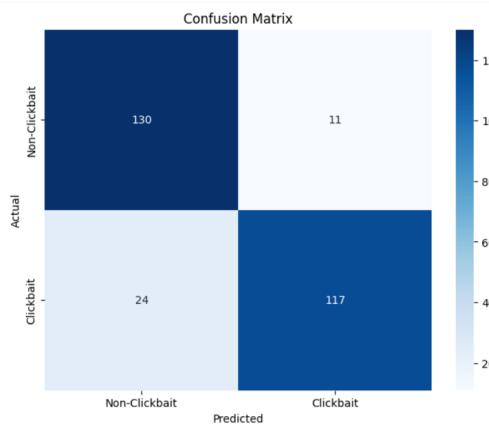


Figure 5: Confusion matrix

To obtain a more comprehensive understanding of the linguistic characteristics underlying clickbait classification, we employed word clouds, a graphical representation of text data, where the size of each word reflects its frequency or importance within a given dataset, for both clickbait and non-clickbait content. It offers a visual snapshot of the most prevalent words, allowing for quick insights into the themes and trends present in a corpus of text. For clickbait headlines (Figure 6), the dominant words include terms like **ז** ['That way'], **נ** ['what'], **ה** ['how'], **ן** ['end'], **מ** ['method'], **ל** ['before'] and **ה** ['believe']. These words suggest a focus on intriguing, curiosity-inducing language often associated with clickbait content. Conversely, in non-clickbait headlines (Figure 7), we observe the prevalence of political concepts and figures, including **נתניהו** ['Netanyahu'], **ממשלה** ['government'], **נ** ['against'], **חוק** ['law'], **כנסת** ['parliament'], **אישור** ['approved'] and **לפיד** ['Lapid']. This indicates a strong emphasis on political discourse and legislative matters in non-clickbait headlines.



Figure 6: Clickbait word cloud



Figure 7: Non-clickbait word cloud

In our attempt to fine-tune a deep learning model called HeBERT, we conducted two experiments: one using only the headlines, and another combining headlines with the content of the URLs. The results, as presented in Table 3, were also subjected to a McNemar test with a significance level of 0.05. The test showed that there was no statistically significant difference between the results obtained from the SVM model and our deep learning model. This means that, at this time, we cannot say that deep learning models are superior to machine learning models for clickbait detection in Hebrew. This may be because deep learning models tend to exhibit superior performance when presented with a substantial volume of data.

Table 3: Deep learning results

	Accuracy	Precision	Recall	F1
Headlines	0.87	0.87	0.87	0.87
Headlines + Text	0.88	0.88	0.88	0.88

5 Conclusion and future work

In summary, this study has delved into the significance of linguistic features in clickbait detection, revealing their crucial role in differentiating clickbait from non-clickbait in Hebrew. By employing machine learning models, we were able to quantify the importance of these features and demonstrate their effectiveness in clickbait identification. Our findings also highlight the potential of machine learning and deep learning models for robust clickbait detection in Hebrew. The impressive accuracy of 0.87 achieved by the SVM model and the comparable performance of our fine-tuned HeBERT model further validate the feasibility of our approach.

While this study has shed light on the effectiveness of machine learning and deep learning models for clickbait detection in Hebrew, there are several avenues for further exploration and enhancement. First, the expansion of the dataset through weak supervision annotations holds potential for improving the performance of deep learning models. This dataset enrichment could involve diverse examples that encompass various linguistic styles, ensuring better generalization across different contexts. Additionally, the incorporation of more sophisticated deep learning architectures, such as recurrent neural networks (RNNs) or transformer-based models, can offer improved capabilities for capturing intricate linguistic patterns within clickbait headlines. These architectures can potentially unlock higher accuracy rates by effectively modelling the sequential nature of language and contextual relationships. Lastly, as the landscape of online content continues to evolve, ongoing model adaptation and regular updates are essential to maintain the models' relevance and efficacy. By exploring these avenues, we aim to continually refine and advance our clickbait detection methodologies, contributing to a safer and more informed digital environment for users consuming content in the Hebrew language.

Data availability statement

The data generated and analysed during this study are available in the Git-Hub repository: <https://github.com/Talya-Natania/Clickbait-Data>.

References

- Agrawal, Amol. 2016. Clickbait detection using deep learning. In *2016 2nd international conference on next generation computing technologies (ngct)*, 268–272. Dehradun, India: IEEE.
- Al-Sarem, Mohammed, Faisal Saeed, Zeyad Ghaleb Al-Mekhlafi, Badiea Abdulkareem Mohammed, Mohammed Hadwan, Tawfik Al-Hadhrami, Mohammad T Alshammari, Abdulrahman Alreshidi & Talal Sarheed Alshammari. 2021. An improved multiple features and machine learning-based approach for detecting clickbait news on social networks. *Applied Sciences* 11(20). 9487.

- Anand, Ankesh, Tanmoy Chakraborty & Noseong Park. 2017. We used neural networks to detect clickbaits: you won't believe what happened next! In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings* 39, 541–547. Aberdeen, UK: Springer International Publishing.
- Biyani, Prakhar, Kostas Tsoutsouliklis & John Blackmer. 2016. "8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30. DOI: <https://doi.org/10.1609/aaai.v30i.1.9966>
- Breiman, Leo. 1996. Bagging predictors. *Machine learning* 24. 123–140.
- Breiman, Leo. 2001. Random forests. *Machine learning* 45. 5–32.
- Cao, Xinyue, Thai Le, et al. 2017. Machine learning based detection of clickbait posts in social media. *arXiv preprint arXiv:1710.01977*.
- Chakraborty, Abhijnan, Bhargavi Paranjape, Sourya Kakarla & Niloy Ganguly. 2016. Stop clickbait: detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 9–16.
- Chang, Chih-Chung & Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2(3). 1–27.
- Chriqui, Avihay & Inbal Yahav. 2022. HeBERT & HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition. *INFORMS Journal on Data Science* 1(1). 81–95.
- Christopher, D Manning, Raghavan Prabhakar, Schütze Hinrich, et al. 2008. Introduction to information retrieval. *An Introduction To Information Retrieval* 151(177). 5.
- Dam, Shiva Ram, Sanjeeb Prasad Panday & Tara Bahadur Thapa. 2021. Detecting clickbaits on nepali news using SVM and RF. In *Proceedings of 9th IOE Graduate Conference*, vol. 9, 140–146. Nepal.
- Dong, Manqing, Lina Yao, Xianzhi Wang, Boualem Benatallah & Chaoran Huang. 2019. Similarity-aware deep attentive model for clickbait detection. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II* 23, 56–69. Macau, China: Springer International Publishing.
- Fakhruzzaman, Muhammad N, Saidah Z Jannah, Ratih A. Ningrum & Indah Fahmiyah. 2021. Clickbait headline detection in indonesian news sites using multilingual bidirectional encoder representations from transformers (m-bert). *arXiv preprint arXiv:2102.01497*.
- Fu, Junfeng, Liang Liang, Xin Zhou & Jinkun Zheng. 2017. A convolutional neural network for clickbait detection. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 6–10. Changsha, China: IEEE.
- Geçkil, Ayse, Ahmet Anil Müngen, Esra Gündogan & Mehmet Kaya. 2018. A clickbait detection method on news sites. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 932–937. Barcelona, Spain: IEEE.
- Guyon, Isabelle, Jason Weston, Stephen Barnhill & Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46. 389–422.
- Klairith, Praphan & Sansiri Tanachutiwat. 2018. Thai clickbait detection algorithms using natural language processing with machine learning techniques. In *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 1–4. Phuket, Thailand: IEEE.
- Kumar, Vaibhav, Dhruv Khattar, Siddhartha Gairola, Yash Kumar Lal & Vasudeva Varma. 2018. Identifying clickbait: A multi-strategy approach using neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1225–1228. MI, Ann Arbor, USA. DOI: <https://doi.org/10.1145/3209978.3210144>.

- Liebeskind, Chaya & Shmuel Liebeskind. 2018. Identifying abusive comments in Hebrew Facebook. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, 1–5. Eilat, Israel: IEEE.
- Litvak, Marina, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia & Rizek Abu Madeghem. 2022. Offensive language detection in hebrew: can other languages help?. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3715–3723. Marseille, France: European Language Resources Association.
- Marreddy, Mounika, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni & Radhika Mamidi. 2021. Clickbait detection in Telugu: Overcoming NLP challenges in resource-poor languages using benchmarked techniques. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Shenzhen, China: IEEE.
- McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2). 153–157.
- More, Amir & Reut Tsarfaty. 2016. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 337–348. Osaka, Japan: The COLING 2016 Organizing Committee.
- Potthast, Martin, Sebastian Köpsel, Benno Stein & Matthias Hagen. 2016. Clickbait detection. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings* 38, 810–817. Padua, Italy: Springer International Publishing.
- Prokofeva, Natalia A & Irina A Akulovich. 2021. The language means of comicality in clickbait headings. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Seriya 2, Lazkoznanie* 20(3). 151–165.
- Yu, Hsiang-Fu, Fang-Lan Huang & Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning* 85. 41–75.
- Zheng, Hai-Tao, Jin-Yuan Chen, Xin Yao, Arun Kumar Sangaiah, Yong Jiang & Cong-Zhi Zhao. 2018. Clickbait convolutional neural network. *Symmetry* 10(5). 138.
- Zuhroh, Nurrida Aini & Nur Aini Rakhmawati. 2020. Clickbait detection: a literature review of the methods used. Register: *Jurnal Ilmiah Teknologi Sistem Informasi* 6(1). 1–10.

About the Authors

Talya Natanya graduated with honors a bachelor's degree in computer science. Currently pursuing a master's degree in data mining at the Jerusalem College of Technology and conducting research in the field of Natural Language Processing (NLP) under the supervision of Dr Chaya Liebeskind.

Address

Jerusalem College of Technology, Department of Computer Science
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel

e-mail: tyazdi@g.jct.ac.il
ORCID: 0009-0006-1150-8573

Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and data mining. Especially, her scientific interests include Semantic Similarity, Language Technology for Cultural Heritage, Morphologically rich languages (MRL), Multi-word Expressions (MWEs), Information Retrieval (IR), and Text Classification (TC). Much of her recent work has been focusing on analysing offensive language. She has published a variety of studies and a few of her articles are under review or in preparation. She is a member of several international research actions funded by the EU.

Address

Jerusalem College of Technology, Department of Computer Science
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel

e-mail: liebchaya@gmail.com
ORCID: <http://orcid.org/0000-0003-0476-3796>

Barbara Lewandowska-Tomaszczyk, University of Applied Sciences in Konin

Chaya Liebeskind, Jerusalem College of Technology

Anna Bączkowska, University of Gdańsk

Jurate Ruzaitė, Vytautas Magnus University

Ardita Dylgjeri, “Aleksandër Xhuvani” University of Elbasan

Ledia Kazazi, “Aleksandër Xhuvani” University of Elbasan

Erika Lombart, Université de Louvain-la-Neuve, Belgium

OPINION EVENTS: TYPES AND OPINION MARKERS IN ENGLISH SOCIAL MEDIA DISCOURSE¹

Abstract

The paper investigates various definitions of the concept of opinion as opposed to factual or evidence-based statements and proposes a taxonomy of opinions expressed in English as identified in selected social media. A discussion situates opinions in the realm of pragmatics and reaches to philosophy of language and cognitive science. The research methodology combines a thorough linguistic analysis of opinions, proposing their multifaceted taxonomy with the automatically generated lexical embeddings of positive and negative lexicon acquired from the analysed opinionated texts. As proposed, the definition of the concept of opinion is best apprehended when looked upon in terms of an *opinion event*, with a number of necessary conditions on the one hand, and those that are characteristic of an explicit opinion prototype on the other. Lists of opinion discourse markers show their preferential uses either in positive or negative opinionated texts; however, no sets of necessary and/or sufficient opinion markers properties have been acquired from the analysed texts. The conclusions indicate a polysemous understanding of naturally occurring social media opinionated texts and a definitional flexibility of the boundaries around lexical positive and negative types of opinion markers.

Keywords

belief, , evidence, negative, opinion (-event, -marker, -SCHEMA), persuasion, positive, prototype

¹ The authors wish to express their gratitude for the insightful comments from the anonymous referees of the paper, which provided ground for some new reflections on the theme.

1 Focus of the paper and research question

The present paper focuses on the identification of various approaches to the definition and scope of the concept of opinion, providing an argument toward accepting the definition of opinion embedded in social context in terms of an event rather than as a set of necessary and sufficient markers. A particular emphasis is laid here on similarities and contrasts in positive and negative opinion linguistic markers, opinion taxonomies and schemas.

The second part of the paper refers to the identification of opinion markers in the category of positive and negative opinions, and presents relevant examples with the identification of their relevant parameters.

The main research questions asked in the present study refer to the presence of unique or, minimally, preferred language markers of texts expressing opinion as opposed to those stating facts. The second issue is an attempt to identify unique or preferred contexts and/or language markers for positive and negative verbal opinion expression.

2 Multiple outlooks on *opinion* concept: a literature survey

The field of opinion studies is not a subject of one research domain. Apart from linguistics, which provides insight from phonetics, particularly prosody, as well as lexicon, the field profits from construction grammar and cognitive semantics with the notion of mental models, meaning co-constructions, semantic properties and pragmatically characterised contexts, as well as discourse analysis illuminating the phenomenon of speech events, implicatures and politeness studies. Resort is also made here to some philosophical investigation of truth and judgement, logic of argumentation, and persuasive appeals, social sciences, culture studies, political science, media studies, emotion studies, theory of science, law and other professional fields.

2.1 Polysemic character of *opinion*

Although opinion is a polysemous concept, with a number of senses, which are characterised by varying properties, all of them emphasise subjectivity and agency as characteristic features. Nevertheless, they can differ in their relation to *truth* (in its Aristotelian, pragmatic, or consensual senses), or to *fact*, rendered in its diverse ways (scientific, social, etc.).

We propose that there is more to the opinion concept than one prototypical category member. We rather see it as a *Lakovian* (Lakoff 1987) *radial category*, in which there are a few prototypical members, all of which are interlinked by family resemblance (Wittgenstein 1953) ties. The family resemblance relies on two criterial properties.

One of the opinion criterial properties is the degree of opinion basis on *scientific factual evidence*, which provides identification conditions differentiating *opinions* from *factual knowledge*.

The other opinion category criterial property refers to *the degree of the opinion holder's (Agent/speaker's) conviction with regard to the truth/falsity of the proposition expounded*. This property aims at differentiating between telling what the speaker considers *true* and what would be his/her (*intentional*) *lie*. The typology of opinions can thus be carried on according to these two different types of differentiation criteria, e.g., scientific/professional hypotheses are typically both based on evidence and their authors believe what they propose to be true, i.e., conforming to facts.

Although immersed in social context and based on *social facts* on the other hand, fake news, are either not based on scientific proof/evidence or rely on biases or else, on intentionally manipulated data (e.g., by using faulty logic, insufficient evidence, emotive persuasion, etc., cf. Hoehn et al 2023).

A prototypical context of an opinion expression involves a face-to-face interaction and an opinion expressed orally. In the present study, we give a range of opinion texts as used in selected social media website sources (itemized in Appendix 2). While some scholars (e.g., Kumar and Gupta 2021) propose that opinionated texts are restricted to precisely such texts acquired from blogs, social networking sites or any other online portals in which the users have expressed their disposition and point of view towards any particular product or service, we analyse such and similar data in the present study, and we consider opinions in rather wider – general lexicographic – terms as – immersed in a *social-cultural context* – views or judgements formed about a person, object, property, or an event, not necessarily based on proven fact or evidence.

We do not enter the discussion whether particular opinions are *right* or *wrong* as, in general, opinions cannot be classified as such because, apart from the criterial properties discussed above, they are based on personal perspectives and experiences. However, as mentioned before, opinions can be based on factual information or evidence as well, which can make them more or less credible. It is important to note that while opinions cannot be right or wrong, they can obviously be either pleasing, praising (example 1), or else harmful or offensive (example 2).

- (1) EXTREMELY fair teacher. Nice man too. Is enjoyable. Favors weekly online assignments and quizzes (<https://www.ratemyprofessors.com/professor/1270844>)
- (2) Evil personified!!! Should be strung up and left!! (Fb)

Thus, as proposed in our two criterial properties and the radial category membership of opinion concepts, one can find definitions of opinions as subjective interpretations not necessarily evidence-based, or else facts which are evidence-based or logically derived from evidence. They can be considered statements of belief, attitude, value, judgement, or feeling, as opposed to facts which are statements that can be confirmed by proof.

Subjective statements are often seen in terms of evaluations, which, in simplest terms, are value-laden, *personal opinions* that something is good or bad (Hunston 2000: 5; Bednarek 2006: 19; Van Linden 2012: 43). The good-bad cline in evaluative language developed in an axiological approach to language and has replaced the long-standing true-

false approach typical of truth-conditional semantics. Generally, two strands in the axiological approach can be distinguished. One of them relies on *nonveridicality* and it revolves around the concept of *modality* (Taboada and Trnávka 2013: 10), wherein an opinion hinges on the degree of *probability* when describing an entity. It is often grammar-oriented in the sense that the choices of grammatical categories, such as specific modal verbs, are decisive in verbalizing personal evaluations:

- (3) He's *pretty* nice, but he's *definitely* one of the worst teachers I've ever had. He's always *extremely* vague & unhelpful about his assignments. *Personally*, he never revised my drafts for my essays even though we were required to turn them in, & then I got bad grades on most of my essays. I emailed him about it several times & he never replied.

The other standpoint is attitudinal, and it encodes evaluations through the choice of adjectives or morphological affixes (usually diminutives to signal appreciation or augmentatives to suggest contempt); thus it is both lexis- and grammar-oriented. There is also an *integrative approach*, elaborated *inter alia* by Biber and Finegan (1989) or Thompson and Hunston (2000), which takes into account both options.

Evaluative language is tightly connected and sometimes used interchangeably with *stance*, yet they are not exactly the same concepts. While evaluative language is associated primarily with implicit meaning, stance is regarded as explicitly expressed lexico-grammatical structures (Biber and Zhang 2018). As a result, evaluative language tends to be examined by describing the contextual information and connotations, whereas stance focuses on recurrent structures appearing in larger datasets, hence corpus-assisted analyses are typically employed to investigate stance (Biber and Zhang 2018). It has been noticed, based on a qualitative analysis (Biber and Zhang 2018), that opinionated texts (unlike persuasive or informative) are marked for stance. In our study, both the implicit and the explicit, the conceptual and the lexico-grammatical aspects will be explored as they are seen as constitutive elements of the prototype of opinions. In example (4), for instance, the stance is expressed very directly by explicitly specifying the speaker's relation to the object:

- (4) *I totally agree!* Britain to be 'proper' Britain!

In example (2), meanwhile, the negative emotions expressed can be considered as an example of affective stance, but the speaker provides these strongly attitudinal generalisations without explicitly stating that he/she is taking a stance.

Some other outlooks consider opinions to be subjective conclusions, which might contain preferred (even biased) options, with prototypical opinions that need no verification. On the one hand, opinions can concern choices from logical premises (professional opinions) and on the other, they can be judgments not necessarily based on facts or evidence (e.g., on intuition, on biased preconditions e.g., conspiracy theories). Linguistic bias is referred to by Beukeboom and Burgers (2017) as "a systematic asymmetry in word choice that reflects the social-category cognitions that are applied to

the described group or individual(s)". Bias is often referred to as some vague ideas in which some confidence is placed. Instead, professional opinions are to be considered within the category of Objective opinions - based on facts and evidence, without personal biases or emotions and Expert opinions - based on the knowledge and expertise of a particular field or profession. On the other hand, there should be a distinction made between fact-based professional opinions and language-based professional opinions. Experimental sciences are clearly fact-based, while most humanities, as well as to go further, legal opinions, are all supported by linguistic and discourse analyses.

Opinions can reside in beliefs or sentiments shared by most people, expressing the voice of the people (public opinion), messages expressing a belief about something that can be held with confidence but not substantiated by positive knowledge or proof:

- (5) Drug dealers, rapists, murderous, people who organise grooming gangs when caught and found guilty should be hanged or shot no mercy the only way to stop these hideous crimes.

Such instances as example (5) not only express biased and prejudiced beliefs that lack evidence or grounding, but can also qualify as hate speech in some jurisdictions.

On the other hand, professional opinions indeed indicate (partly) subjective – preferences, albeit immersed again in a larger socio-cultural context, based on or concluded from evidence as in the legal opinion stating the reasons for a judicial decision (Król 2015). Nonetheless, even if a judicial decision is based upon the bounding procedure, the legal argumentation and the legal discourse – it is also given in view of the existing *leeways*, i.e., freedom to act within certain limits, in the choice of the legal norms to be applied in a given case and moreover the normative leeways inherent in the legal rules. Furthermore, as any judgment expressed in language, is also constrained by the semantic (interpretative) options existing in the determination of the meaning of the chosen rules. All of this may lead to the stance of the "rule-scepticism", which is coupled with the "fact-scepticism" based on uncertainty of the evidence (compare e.g., Wróblewski 1969; Król 1987).

Considering opinions as legal documents stating the reasons for a judicial decision might somewhat confuse the previous definitions given about opinions as just mere vague ideas and beliefs. These kinds of opinions are to be included into another broad category, that of *Informed Opinions* – based on research and analysis of available information, and obviously stemming from particular social circumstances.

From all the above-mentioned definitions, it might be concluded that since it is both challenging and interesting to correctly identify and analyse opinions, and since it is important to consider the type of opinion when evaluating its credibility and relevance to a particular topic, a classification of opinions in some rather broad categories might come to the scholars' help. As such opinions might be broadly classified as: Personal Opinions,

Expert/Informed Opinions, and Public Opinions.² A distinction between Biased Opinions and Objective Opinions makes use of an amount of verified background knowledge, which serves as a logical basis for a particular judgement.

2.2 Status of truth/fact/evidence/proof and opinion

The act of expressing one's opinion on a subject is often linked to a discussion of the status of its opposite, i.e., fact, or truth, of a particular assertion. Assertion is considered to be linked to truth, while in expressing an opinion the status of fact or truth seems to be suspended. As we wish to restrain from an ontological discussion of the concept of truth, we resort rather to the idea of evidence and will use the concept of *evidence* as *proofs* in discussing opinions.

This approach is connected with a type of semantics that is not synonymous with a truth-conditional specification of a proposition (Lewandowska-Tomaszczyk 1997). As Seuren (1985: 27–29), argues semantics “must primarily define a proposition in terms of what it does to any given discourse domain”. In other words, the meaning of a linguistic unit should be characterized first of all in terms of *changes* it brings about to a given *discourse domain* i.e., “a systematic modification, or increment, which it brings about whenever it is added to an appropriate given discourse domain”. The speaker constructs a discourse domain, i.e., in Seuren's wording, s/he builds “a picture of a partial world” (Seuren 1988: 213) in a *semantic space* (Langacker 1987: 147).

The understanding of facts and opinions can be influenced by the social and cultural context in which language is used. Social constructivist theories assert that language and meaning are products of social interaction. Societal norms, beliefs and values can shape the categorization of statements as either factual or opinionated. As Searle suggests language is not an automatic system correlating itself to the external world (Searle 1998). More specifically, names do not designate external objects by themselves, and sentences do not describe external states of affairs or facts by themselves. It is human beings, who use a language, that build the bridge connecting a language and the world, and that create the referring (or predication) relation of names (or sentences) to the corresponding objects (or states of affairs).

As a result, a comment that at first blush seems negative may in fact express a positive opinion and the other way around, which is the case of irony (see below for details). Opinionated statements where an insult is a complimenting remark (as in “You are awful” said to somebody who has passed an examination with the highest mark) or where an apparently positive comment is used ironically, that is to encode criticism (as in “You are an Albert Einstein”), are frequent forms of expressing one's opinion, in particular when

² One needs to consider the fact that *both personal opinions as well as expert opinions may be public if published e.g., in social media*. Here we used the term in a constrained sense as “an aggregate of the individual views, attitudes, and beliefs about a particular topic, expressed by a significant proportion of a community” as authored by W. Phillips Davison in an entry in *Encyclopaedia Britannica* (2023).

they are burdened with some emotional load. A common practice on social media, in some informal contexts for example, such as X (formerly Twitter), is to resort to vulgarisms in vocatives to signal bonding and in-group identity, as in “How are you bitches”. This is a particularly popular way of communicating among teenagers, and it is not seen as expressing a negative opinion but rather as a compliment or a form of ingratiation, as the social norms valid for social media are much more relaxed and allow one to resort to vulgar language while engaging in positive opinion forming. Expressing positive emotions through swearwords is not uncommon as swearwords have a wide array of functions, even when they sound vulgar and apparently negative they may encode social bonding, group solidarity, flirting, banter, jocularity, etc. (Bączkowska and Gromann 2023). Thus, a negatively-charged lexical marker alone may not do justice to the speaker-intended meaning while signalling opinions implicitly.

Thus, the context in which language is used plays a vital role in discerning between facts and opinions. A series of pragmatic elements such as implicatures, presuppositions and speech acts influence the interpretation and classification of statements as factual or opinionated. Grice’s *conversational implicature theory* (1975), especially, helps explain how people infer meaning in everyday conversation and how context and cooperation play essential roles in understanding indirect or implied messages. According to Grice, effective communication relies on a cooperative principle, which assumes according to the Grice’s *Cooperative Principle* that participants in a conversation generally cooperate and make their contributions in a way that is relevant, informative, and conducive to the overall purpose of the conversation (Grice 1975).

Apart from the Cooperative Principle, two other opinion discourse analysis relevant principles to be considered are the Politeness Principle and the Cognitive Change one. The politeness theory follows and enriches to some extent the Cooperative Principle firstly proposed by Grice (1975). Lakoff (1973) and Leech (1983: 132) discussed that politeness and truth are often mutually incompatible (the fact acknowledged as a *white lie*) and so are politeness and brevity. It seems that the cooperative and politeness principles and the tension between them reflect a dual purpose in human discourse: to act efficiently with other people and maintain social relationships.

The third principle is the principle of a *Cognitive Change*. Some discourse can be best interpreted as though it followed a maxim “change the receiver”. This fulfils the need to rearrange mental representations, a process that can be best affected in the absence of pressing practical and social constraints. Such can be the case of expressing opinions, personal beliefs and attitudes. This objective is clearly linked to the presence of a *persuasive force* as part and parcel of opinion defining properties and surface in some cases in the form of *linguistic manipulation*. Persuasive language often blurs the lines between facts and opinions to influence beliefs and attitudes. The wide and somewhat blurred semantic field of the term “manipulation” includes such key elements as “negative” intention of the speaker (Akopova 2013). Manipulation is a pragmatic aspect that achieves its goals without evident detection of communicative intention: the speaker purposefully chooses such a

form of utterance that lacks direct signals of his/her true intentions. It makes use of mechanisms that lead the listener to perceive verbal messages uncritically.

AI and digitalization have given rise to an unprecedented amount of information, some of which may be misleading or false which, on their own, affect our understanding of facts and opinions.

Opinions are often influenced by personal biases, cultural beliefs and emotions. Linguistic bias is defined as a systematic asymmetry in word choice that reflects the social-category cognitions that are applied to the described group or individual(s). Relying on evidence can help mitigate these biases and move towards a more objective understanding of reality. For instance, example (6) expresses biased attitudes towards Muslim migrants through such linguistic/rhetorical choices as the (rhetorical) question, which serves here as an indirect expression of indignation; insult/offensive labelling (“evil cretins”), and the strategy of polarisation between “us” vs. “them” (“they” come to “our” country). Simultaneously, the utterance interrogative form allows the speaker to withdraw from the opinion conveyed, thus functioning as a face-saving device in case of irrefutable verbal attacks from other interactants:

- (6) Why are these evil cretins allowed to come into our country

In a recent paper, Kaiser and Wang (2020: 116) identify an aim of their study as shedding light “on why people struggle with the seemingly easy task of distinguishing facts and opinions.” The authors investigate whether a statement being perceived as fact or opinion is modulated by its, as the authors call it, *linguistic packaging*. Language in fact provides these multiple ‘packaging options’ for expressing the same basic information (e.g., Chafe 1976; Lambrecht 1996). The authors try to show to what extent the ‘linguistic packaging’ which contains similar basic information can give us clues as to what extent the statement might be treated as a judgement or belief and where it is more likely to be supported by evidence (facts). They use subjective adjectives (e.g., “important”, “amazing”, “frustrating”, “impressive”, “hideous”, “unclear”, “impossible”) in three positions (as prenominal modifiers, in predicative position, and in appositive roles) to test for the potential effects of linguistic packaging on people’s perception of subjectivity versus factual information. They also consider a lexically based hypothesis, according to which the presence of a subjective adjective should trigger comparable ratings of subjectivity in any structural position. Results suggest that sentences conveying the same core information, using the same words, can receive different subjectivity ratings depending on how those words are put together, thus suggesting that subjective adjectives in different syntactic positions influence people’s subjectivity ratings. Also, when subjective adjectives are in the positions that can be associated with new information (predicative or appositive functions), the text receives higher subjectivity ratings than when the same adjectives are in a position associated with old/already known information (prenominal modifiers of definite nouns).

Overall, it is suggested that linguistic packaging choices can blur the distinction between fact and opinion. If a speaker/writer wishes to present an opinion as an objective fact, they should present the relevant subjective information in prenominal modifier position and not in predicative position. Furthermore, as argued in Hoehn et al. (2023) in an exhaustive analysis of different forms of bias, both lexical choices, grammatical structures, as well as logical argumentation and emotional persuasion can play a vital role in the persuasive strategies identified in opinions.

Apart from subjective adjectives, mentioned before, there also are other linguistic devices that can help distinguish between facts and opinions. Facts are not normally expressed with qualifiers such as: “seems”, “looks like”, “probably”, “possibly”, etc. as is the case of opinions. However, this distinction is not exhaustive since speakers can omit qualifiers when stating an opinion so it resembles a fact and as a result detection remains difficult (Kaiser and Wang 2021).

A good example of such cases are structures typically used to signal negative opinions. In English, personalised negative opinions can be expressed by means of several structures (Culpeper 2011: 135). The personalised negative assertion involves the following elements (words in parentheses are optional elements of a sentence): (you)(be)(so/such a) + Noun (e.g. “You are such a disappointment”). Without the qualifier, the sentence is still an assertion while at the same time being far from an objective fact. This is not valid, however, where the addressee is actually the representative of the name-labelling category, as in “You are (such a) Jew” wherein the recipient can have a Jewish origin, and then it is a fact, or can only have features typically ascribed to Jews (e.g., being mean), and then it is a subjective evaluation/opinion.

Along with personalised negative assertions, vocatives are often employed to express personalised negative opinions by resorting to the following structure: (you) + Adjective + Noun. This structure is typically used in emotional language, e.g. “You (rotten) dickhead”. However, this option need not rely on strong emotional language as apparently neutral statements can also encode negative opinions, as in the vocative “Mr President” when said to somebody who lost in a presidential re-election. This is an obvious irony, and thus a subjective opinion, when the audience knows about the failure in the campaign, or an objective fact if it is a piece of news not yet revealed. Another type of structure proposed by Culpeper (2011) to signal subjective negative opinion is personalised negative reference, which involves (your) + Adjective + Noun, as in “(your) little ass”. Here, the qualifier is required, and the statement is subjective. Finally, personalised third-person negative references are also examples of expressing opinions, even though in a less direct way, as in “(the) (daft) bimbo or prostitute”, especially when uttered in the presence of the target. Theoretically, the “prostitute” example can function as either a subjective negative opinion (if the target is thus compared to a prostitute) or an objective fact (if based on truth). When deprived of modifiers, stand-alone nouns can function as exclamatory expressions, and when offensive language is used, they can have the status of taboo-based secondary interjections (for more details on interjections see the next section).

Reasssuming the discussion on the status of (fact) evidence thus, one can try and summarize the evidence-presence scale according to the following criteria:

- (a) No evidence - gossip, hearsay
- (b) Uncertain evidence – beliefs
- (c) Logical fallacies
- (d) Conspiracy theories, fake news; myths, stereotypes, ideology, etc.
- (e) Evidence present
 - (i) in terms of outside proven facts
 - (ii) in terms of the logic of discourse [e.g., legal opinions]
 - (iii) degrees of certainty, though allowing ambiguity in scientific /professional opinions

3 Opinion schema

Our knowledge is organised into what Lakoff dubs *Idealised Cognitive Models* (ICMs, Lakoff 1987)), which are referred to as knowledge *frames* in computational literature (Schank and Abelson 1977) and framing in other subject domains (e.g., media studies). ICMs or frames are relatively stable knowledge structures that represent a given experience. In Cognitive linguistics, they are considered “idealised” because, rather than relating to any specific experience, they incorporate knowledge from various related experiences in order to form a more generalised (or sometimes over-simplified) abstract representation (Lakoff, 1990). Stores of knowledge are built up over time and through repeated exposure to experience. When shaping opinions individuals rely on their existing *opinion schemas*.

An opinion schema is a cognitive framework that people use to organise and interpret information related, in some part, to external evidence, but primarily to their attitudes, beliefs, and values. It is a mental structure that helps individuals process new information and form judgements about a particular topic or issue. What needs to be added at this point is the role and impact of properties which function as constitutive factors in opinion formation. They include external evidence and logical argumentation on the one hand but also emotional appeals, appeals to authority and other possible shades of Aristotelian persuasive appeals.

Except for those and various other factors, such as personal experiences, cultural background, social norms, and media exposure, opinion schemas can also be shaped by cognitive biases and heuristics, which can affect how people perceive and interpret information.

Opinion schemas can be changed or modified through exposure to new information, persuasive messages, or social influence. However, they can also be resistant to change, especially if they are deeply ingrained or associated with one's identity or self-concept. These features and characteristics of opinion schemas are to be based on the Cooperative, Politeness and Cognitive Change Principles.

Schematic knowledge can be subverted, exploited, altered or violated in order to create particular effects (Emmott et al. 2014). Schemas can be altered through the process of Accretion, in which new information is added to existing schemas, Tuning in which information within a schema is modified in some way and Restructuring, in which new schemas are created.

Creating opinion schemas is crucial to both Discourse Analysis and AI. AI has demonstrated how schemata are essential to text processing, and this idea has been accepted in discourse analysis as a partial explanation of coherence.

However, at times, Schemata are also considered as a potential barrier to understanding. Since one of the main functions of language is to manipulate the environment and the interlocutors as well as to establish and correctly maintain human and social relationships, the human mind must build new schemata and adjust existing ones if it is to adapt to new experiences it faces.

Agent /Author [Opinion holder] (professional/not, authoritative/not; subjectivity) – Evaluative statement (polarity, emotionality, persuasive appeals) –

Agent: present physically, online, or default

Goal - Object + Channel – Addressee (Receiver) - preconceived beliefs, ideology, preferences, – Effects (persuasion) – Feedback (adopting - rejecting)

Precondition: Agent's biased/subjective beliefs/judgments/preferences/framing

Channel: acoustic/auditory - visual; direct (face-to-face) – indirect (mail, online)

Addressee: present [synchronous or asynchronous] or default

Receiver: audiences, non-addressed participants

Channel: spoken, written/visual

Expression: explicitly or implicitly evaluative, persuasive force

Explicit: (a) containing mental verbs ‘think/suppose/consider/demonstrate, etc (b) containing evaluative adjective and phrases (c) containing comparative and superlative degrees of adjectives and adverbs, (d) containing comparative numerical values (d) affective/emotional language

Implicit: metaphoric & metonymic language, similes, under- and over-statement, irony, sarcasm, rhetorical question, humorous/jocular, ambiguous, vague

Several researchers have applied the schema theory in the study of the workings of humorous discourse. Raskin (1985) proposes that discourses can be characterized as humorous if they are compatible with two different schemas that are opposed in some way. Thus, a joke typically describes a certain “real” situation and evokes another “unreal” situation which does not take place and is fully or partially incompatible with the former (Raskin 1985: 108). Semino (1997) also suggests that humour can be created by the switching between schemas involved in jokes. Implicit meaning can be conveyed by means of *inter alia* the following means: (a) similes: e.g., “avoid this guy like the plague”, (b) metaphors: e.g., migrants or refugees represented as venomous animals (“their poison”), as a natural disaster (“another excuse to flood the country with the 3rd world”), or a burden (“we have to pay out to bring them here”), (c) over- and understatement (as in “tons of busy

work in the class") or (d) irony (or its stronger, acerbic form, i.e., sarcasm): e.g. "We should take no more immigrants of any sort. And as for 'rescuing' people they mostly went where the money was".

Irony is hard to identify, though in writing it is sometimes signalled by adding inverted commas, which can serve as a surface-level indicator in automated text analysis.

As already signalled above, negative opinions can predicate on ironic comments. These can be built on the concept of contrast and opposition, also known as the reversal of meaning, which is the most common subtype of irony. A good example to illustrate this type of irony is when the utterance "You are very smart" is said to somebody who has done something evidently stupid.

However, there are more options where irony can be used in either an exaggerated manner (as in the case of the so-called surreal irony) or in a more subtle fashion (as in verisimilar irony). By resorting to absurd ideas, the surreal irony is involved, as in – "I am beautiful, aren't I?" – "Yes, and I'm Brad Pitt". The self-praising comment expressed by the first speaker is followed by an obviously unreal situation wherein the interlocutor seems to identify himself with a handsome actor. On the other hand, no contradictory information is built-in a statement in verisimilar irony inasmuch as the speaker expresses a true opinion s/he holds on something, yet it is uttered in a situation where the external circumstances clash with this opinion. The oft-quoted examples to illustrate this option elaborate the so-called "mother scenario", where the mother says, for example, "I love children who tidy up their rooms" to her son while looking at his room that is in a total mess. Here, the truthful opinion that is generally on the positive note and can be a context-free general opinion indicates a problem existing in a real-life situation at the moment of uttering this opinion.

Needless to say, all the above types of irony can be used for jocular purposes, and perceived as such in a favourable context by the addressee; for example, in a situation where previous utterances involved a sequence of humorous comments and thus a non-offensive meaning is expected or when it can be safely assumed that the interlocutor is sufficiently aware of the speaker's jocular personality and/or his/her jocular intentions. Opinions can thus be expressed non-explicitly and often by resorting to the opposite of what they manifest on the surface level.

While exaggeration is typical of surreal irony (and occurs in what is dubbed hyperbolic irony), it is not limited to this trope. Non-ironic cases of exaggeration are also common, particularly in statements focusing on minimising the effect (negative exaggeration), and often involving humour. For example, when the captain of a sinking ship informs the passengers about the hopeless situation by saying "We have a bit of a problem", the opinion he states is an obvious understatement (although it is also an evidence-based fact). The exaggeration aiming at enlarging the situation (positive exaggeration) in turn, as in the case of hyperbolic irony, instantiates overstatement. A comparison, such as a simile or metaphor, may also rely on extreme references, and then they are also illustrative of an overstatement, as in "*You are white as a ghost*" or "*You are an angel*".

(7) Dr Jeter is the bomb.

The implicit forms show different degrees of vagueness or indeterminateness in revealing the speaker's opinion that require a cognitive effort on the part of the addressee in order to re-create the speaker-intended conceptual meaning. In other words, understatement, overstatements, as well as other figures of thoughts (such as similes, metaphors) trigger some conceptual-semantic reformulations, i.e., a reconceptualisation (Lewandowska-Tomaszczyk 2010; Bączkowska, 2022). However, a grammatical (-semantic) reformulation is also possible. For example, in the case of indirect speech acts (in the Searlian sense) the structure of a sentence must be transformed from a question into a statement (or request, etc.), as in the rhetorical question "Are you stupid?" Here, the speaker employs an interrogative form while in fact meaning a statement that encodes a strong, negative opinion about the target.

Another group of devices often deployed to signal opinions in a non-explicit way comprises interjections. These involve primarily the so-called secondary, that is word-like interjections, such as the semi-taboo "ghee" (which is a euphemistic form of *Jesus*) or non-taboo items such as "yuck" and "ugh":

- (8) Ugh. Skip him.

On the other hand, unless they are used purposefully, cases like *ouch* are generally not taken into account as means of opinion-forming inasmuch as they are stimulus-bound and, essentially, uncontrolled and involatile vocalizations, i.e., primary interjections. Taboos and semi-taboo interjections, that is expletives to use Biber et al.'s (1999) parlance, largely appear in contexts exhibiting negative opinions (e.g., "Jesus, what are you doing"). Syntactically, they tend to be non-clausal, stand-alone items or loosely attached to larger syntactic units. Due to the laconicism of interjections (as in, say, "wow"), they are potentially subject to multiple interpretations being strongly context-dependent; hence, they epitomise the non-explicit aspect of opinion-forming. To give an example, "wow" can be perceived as an expression of delight such as the following example:

- (9) Wow wow wow wow! What an amazing place. Stunning location, fabulous, attentive and wonderful staff ...

or contempt, if used ironically; in both instances they reveal the speaker's opinion.

"Ouch" in turn typically functions as evidence-based statement signalling pain, e.g., when one touches something hot, like a kitchen hob; thus, the vocalization testifies this fact. However, one might easily imagine a situation where "ouch" can be deployed to manifest the speaker's opinion, i.e., when it is not an uncontrolled reaction to some external pain-inducing stimulus generating involuntary vocal reaction, but a vocal expression triggered by a negative conceptualisation of somebody's verbal or non-verbal behaviour, uttered with full awareness and will.

Finally, a similar reversal of meaning is observable in im/politeness studies, particularly in what is known in Im/politeness Theories as negative politeness, which "consists in

minimizing the impoliteness of impolite illocutions” (Bousfield 2008: 53). An apparently polite structure can in fact signal distance, depersonalisation or disrespect. This negativity is not, however, manifested straightforwardly but in a non-explicit fashion, i.e., by being impolite but in a subtle way inasmuch as it redresses the threat of negative face to the addressee. In case the receiver actually feels a threat to his face and in consequence takes offence, negative politeness, similar to irony, allows immediate cancellability of the actually ill-intended expression of negative opinion, which is not valid for explicitly stated negative opinions. To illustrate this case of mitigated negative evaluation (opinion), one can easily imagine the addressee’s feeling of disappointment, coldness or even exclusion (from a social milieu) when a friend or a colleague the speaker knows well suddenly uses formal language, as in – “Hi, Mary” – “Good morning, Mr Brown”. The more formal the term of address in the reply (e.g., “Mr Brown”, “Mr President”, etc.), the stronger the effect of negative evaluation of the addressee, even though on the surface level the utterance seems perfectly polite. Naturally, depending on non-verbal signals that accompany such a scenario, it cannot be excluded that the person who resorts to the more formal language has the intention to express deference and subordination rather than aloofness, detachment and/or unfriendliness. Alternatively, it can encapsulate ironic jocularity in the case of honorifics (if the addressee is not the President) or non-ironic jocularity (if the person is the President and is being addressed by a befriended colleague). Personal, subjective evaluations mingle here with objective facts. One may really be the President (evidence-based fact) and can be addressed by professional honorifics (such as “Mr President”), by somebody close, e.g., his wife (subjective evaluation/opinion) for humorous purposes, in which case “Mr President” captures a fact and a subjective evaluation at the same time. The myriad of potential options for interpretation an analyst/observer must tackle brings us to the conclusion that analysing communication in its entirety, that is allowing for the interplay of all multimodal resources (gestures, facial expression, eye contact, body posture and movement, etc. and prosodic features), is a crucial step in interpreting opinions expressed by others.

All these types of statements mentioned above (whether involving irony, simile, metaphor, over- or understatement, hyperbole, indirect questions or interjections) represent some form of non-explicit opinion forming. They also highlight the blurred borderline and the gradability existing on the opinion-fact cline: while an opinion is typically subjective and has a strong evaluative function, a fact tends to be more objective and evidence-based, yet the interpretation is context-dependent, and their co-existence cannot be fully excluded. There are a number of cases, as demonstrated above, whose status is not so clear-cut as statements seem to share the features of both categories or have the potential to be deployed in either way.

4 Opinion prototype (necessary, typical and characteristic properties) of opinion and opinion peripheral senses (family resemblance types)

PROTOTYPE:

Necessary properties:

Individual judgement or Community judgement

Typical properties:

Reference to proven facts/evidence

Agent's (speaker's) [opinion holder's] conviction as to the truth expressed

Goal: evaluation, persuasion by logic (arguments), emotions, appeal to authority (references)
(Aristotelian persuasive appeals)

Characteristic properties: face effects [face maintenance - praising (positive opinion), face threat/losing – criticising, blaming, etc. (negative opinion)]

The process of communication among humans is an intriguing process which allows us to express our opinions, sometimes based on generalisations, which might affect the degree of persuasiveness of an argument or judgement. According to the *prototype theory* some of the cases that belong to a certain category are more typical members of that category than other cases (Rosch 1975; Lakoff 1987; Langacker 1987; Gärdenfors 2000). Opinions based on uncertain or non-existent evidence are distortions from the prototype.

An *opinion prototype*, defined as a most representative mental representation of a typical or idealized cognitive opinion model (ICOM) about a particular topic or issue, can be useful in guiding individuals' opinions and attitudes based on their similarity to and departures from the prototype.

However, given the fact that opinions are based on limited or biased information, they can also lead to oversimplification or stereotyping of different facts.

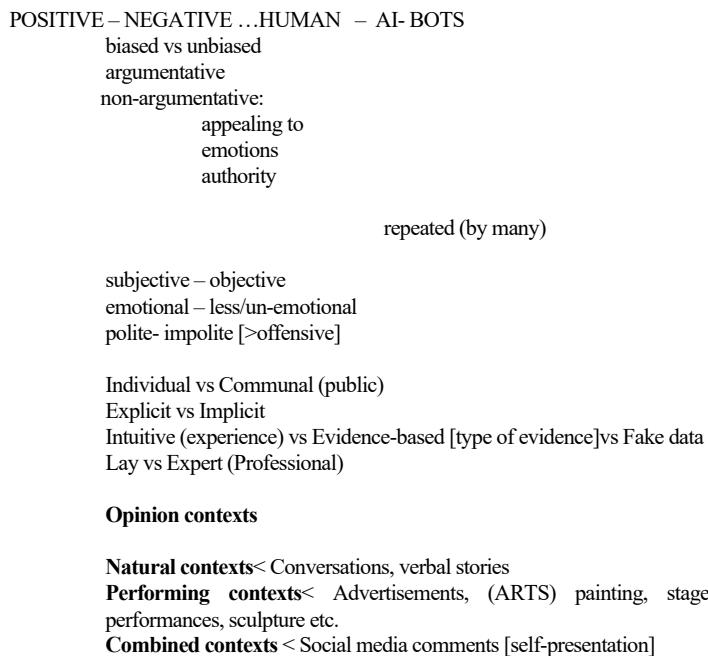
Different contexts in which opinionated discourse takes place, different agents engaged in the discourse and the different individual judgements they make, lead to substantial variations of the characteristics of the typical opinion prototypes. For example, a typical opinion prototype about politics may include beliefs such as "government should provide better basic services to all citizens" or "people should have the right to express their opinions and ideas freely." As stated before, cultural and societal factors play a significant role in shaping the nature of opinionated discourse. These factors can influence the language used, the topics discussed, and the way in which opinions are expressed. Due to the fact that opinions are based on different sources and engage distinct persuasive appeals (Aristotle), logical arguments, emotions, appeal to authority or else are based on rumours or hearsay, they form rather a set of prototypes, i.e., a radial category of prototypical concepts (Lakoff 1987), linked by a family resemblance (Wittgenstein 1956) of the presence of (incomplete) knowledge, and varying framing, defined as "the process of culling a few elements of

perceived reality and assembling a narrative that highlights connections among them to promote a particular interpretation” (Entman 2007: 4).

For instance, in cultures that value collectivism over individualism (cf. Hofstede 2011), opinionated discourse may prioritise the group's interests and values rather than individual perspectives. On the other hand, in cultures that prioritise individualism, opinionated discourse may place more emphasis on personal opinions and experiences.

A study by Dahlman et al. (2016) is concerned with situations where a generalisation is used in argumentation to make an audience agree with a certain judgement. The authors investigated to what extent the agreement is influenced by prototype effects. They conducted two experiments that investigated how the activation of the prototype affects the persuasiveness of the argument. This prototype effect increases the persuasiveness of the argument in situations in which the audience finds the judgement more warranted for the prototype than for the actual case (positive prototype effect) but decreases persuasiveness in situations where the audience finds the judgement less warranted for the prototype than for the actual case (negative prototype effect).

5 Proposed typology of opinions



6 Opinion research methodologies

6.1 Opinion linguistic expression

Opinion expression falls into categories of diverse types. It refers both to multi-genre and multi-disciplinary types. It can vary depending on culture types and contexts. The expression of opinion engages different modalities and distinct senses. It can be verbal, visual, tactile, olfactory, gustatory, or else it can engage more than one or all of the senses, making the opinion expression multimodal to different extent.

The language level and both its corresponding linguistic units, as well as particular framing, make opinion expression an element of still another categorial taxonomy, relative to:

Pragmatic framing identifies contextual knowledge and evidence:

e.g., *We here eat more apples than anybody in the world* [may be true upon evidence]

Syntactic framing imposes the order of linguistic elements used an opinion and – together with *Semantic framing* identifies degrees of certainty and conviction by particular Agents: e.g., *My/Our opinion is..../According to me...*

Lexical framing is marked by relevant lexical items, as e.g., cognitive verbs (e.g., *I think, I believe, I feel*), adjectives that express evaluation or judgement (e.g., *good/bad, worthy, valuable*), and expressions that convey personal feelings or experiences (e.g., *I love, I hate, I enjoy*).

Opinions may be reinforced with persuasive language, such as rhetorical questions, appeals to authority, and emotional appeals as in:

(10) Why are these evil cretins allowed to come into our country? (Fb)

(11) Not long enough punishment for the devastation they cause, let them go back to the country of Origin and peddle their poison there.
(<https://www.facebook.com/britishnationalism>)

Those opinionated texts which are introduced by means of unambiguous opinion markers such as “I think/I don’t think/I do not think”, “in my opinion” or “according to me” are Explicit Opinionated Texts, as contrasted with those which are Implicit Opinionated Texts, unaccompanied by any such markers, e.g., “This hall is more crowded than the other”.

In the spoken subcorpus of the British National Corpus (BNC), for example, there are 23,309 utterances introduced by “I think”, identified in an index containing 1,027,432

utterances (http://pelcra.clarin-pl.eu/SpokesBNC/#search/pl/I%20think/-1/0/20/-1/-1/-1/-1/1000/NN.*/-1,1/4/true/0/-1/-1/-1/-1/-1/-1), e.g.,

- (12) Er, ah, I think he was intending to come in

No data was found in this dataset for the phrase “I don’t think”. And yet, the search for the frequency of the unabridged marker “I do not think” returns 12 utterances, in which the negative opinion element “not” can be regarded as particularly salient and emphasised for their negative force:

- (13) I do not think that that evidence second and third hand as it is, is really enough for me to be confident that the conversion will answer. (http://pelcra.clarin-pl.eu/SpokesBNC/#search/pl/I%20do%20not%20think/-1/0/20/-1/-1/-1/1000/NN.*/-1,1/4/true/0/-1/-1/-1/-1/-1/-1)
- (14) I cannot prove that but I do not think that the tests which have been carried out by the County Council refute that possibility.

In example (14) in particular, the complete phrase “I cannot prove that but I do not think” is an explicit opinion marker, which, additionally, excludes the factual reading (“I cannot prove that”) of the opinionated utterance.

A number of other explicit opinion markers can also be mentioned, some of them identified in the present paper, as in the below example of the negative opinion marker it’s not clear to me that, emphasised by the intrusive say:

- (15) Now, it’s not clear to me at the moment that that is going to be the direction, say, of the Roman Catholic Church, because we seem to have a much more conservative Pope (http://pelcra.clarin-pl.eu/SpokesBNC/#search/pl/there%20are/-1/20/20/-1/-1/-1/1000/NN.*/-1,1/4/true/0/-1/-1/-1/-1/-1/-1)

Implicit opinionated texts are much harder to identify. In spoken datasets in particular, the utterances devoid of any explicit opinion clue, can be considered ambiguous between opinionated and factual readings, as e.g.,

- (16) So he is actually taking some action, and he will come back to us fairly soon he said about the results of the survey (http://pelcra.clarin-pl.eu/SpokesBNC/#search/pl/he%20will%20come/-1/0/20/-1/-1/1/-1/1000/NN.*/-1,1/4/true/0/-1/-1/-1/-1/-1/-1)

The part of the utterance in (16) invoking the evidence in the form of the ‘saying’ reference is typically probabilistic with reference to the person’s coming back fairly soon, but can be considered factual information regarding the act of ‘saying’.

Such examples as the ones discussed in the present study indicate a communicative Speech Event status of the opinionated texts. It assumes a structure of the scenario in which a communicator (opinion holder) conveys a message (opinion on a Theme) to an Addressee in a particular context and in terms of a given cultural model.

6.2 Computational opinion identification – Vector-space models – Distributed representation

The automatic identification and distinction of opinions as opposed to factual knowledge is not easy. The polysemy of opinion definitions, their multilevel and multidimensional typologies, ambiguities particularly in the case of implicit opinionated texts, make such a goal a genuine challenge. The utterances produced in a context e.g.,:

- (17) He is sick.
- (18) My cat/this program/He is very clever.
- (19) They are ready to attack their neighbours.
- (20) It's raining outside

can be considered statements of facts or opinions. In other words, without *explicit linguistic opinion markers*, the judgement concerning the status of these utterances can be rather hard or impossible to make. The knowledge of the outside and internal discourse contexts will be necessary in most of such cases.

And yet, some of the computational methods that have been proposed are conducive in attempts to perform automatic opinion identification (Wang et al. 2019).

Word embedding is one of the most common methods for representing a document's vocabulary. It can determine the context of a word in a document, its semantic and syntactic similarity, its relationship to other words, and so on. Typically, the word embedding representation is a real-valued vector that encapsulates the meaning of the word in such a way that words that are closer in the vector space are anticipated to be similar in meaning.

Word2Vec (Mikolov et al. 2013) is a well-known neural network prediction model that efficiently computes word embeddings by learning from textual data. The two architectures included are the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model (SG). The Continuous Bag-of-Words (CBOW) model is designed to make predictions about the target word based on its surrounding context words. For example, given the sentence “the girl is _ a banana”, where the underscore represents the target word, the CBOW model aims to forecast the target word “eating”. On the other hand, the Skip-Gram (SG) model operates in the other direction, predicting the context words based on the provided target word. Word2Vec is one of the effective methods that represent aspects of word meaning and aid in enhancing opinion mining and, in particular, sentiment classification accuracy (Al-Saqqa and Awajan 2019).

To generate word embedding using the Word2Vec SG method, we employed a large corpus of 25 datasets focusing on various forms of offensive language and collected by Lewandowska-Tomaszczyk et al. (2023).

We learned 50-dimensional embeddings utilising 5 epochs, a window of 5 words, and a minimum frequency of 2. We used the cosine similarity measure, which employs the cosine of the angle between the two vectors to determine the degree of similarity between two words.

For our word embedding analysis, we used two visualisation methods: T-SNE (t-Distributed Stochastic Neighbour Embedding) and heatMap. T-SNE is a dimensionality reduction technique commonly used for visualising high-dimensional data in a lower-dimensional space. Its primary purpose is to reveal patterns and structures in complex data sets that may not be immediately apparent in the original high-dimensional space. Data points that cluster together in the t-SNE plot are considered similar or related, while those that are farther apart are considered dissimilar or unrelated. For each of the investigated terms, we extracted the top 30 most similar words, excluding words whose substrings are the investigated term, its lemma, or its stem. Then, we applied the t-SNE algorithm to the embeddings of the investigated terms and their top 30 most similar words (with complexity set to 15).

A heatmap word embedding visualisation is a graphical representation that uses colours to display the similarity or relatedness between words in a word embedding space. It calculates pairwise similarity scores between words, assigns colours based on these scores, and creates a matrix where cells represent word similarities. High similarity scores are depicted with warm colours, while low scores use cool colours. The heatmap helps users identify clusters or patterns of similarity among words, making it a valuable tool for understanding the semantic relationships between words in natural language processing tasks.

To clarify the rationale for embeddings application, we first present them in a lower number of words. Four frequent synonymy – antonymy examples were excerpted from opinion texts: intuitive – insightful, stupid – ignorant and their embeddings were generated (Figures 1–2)

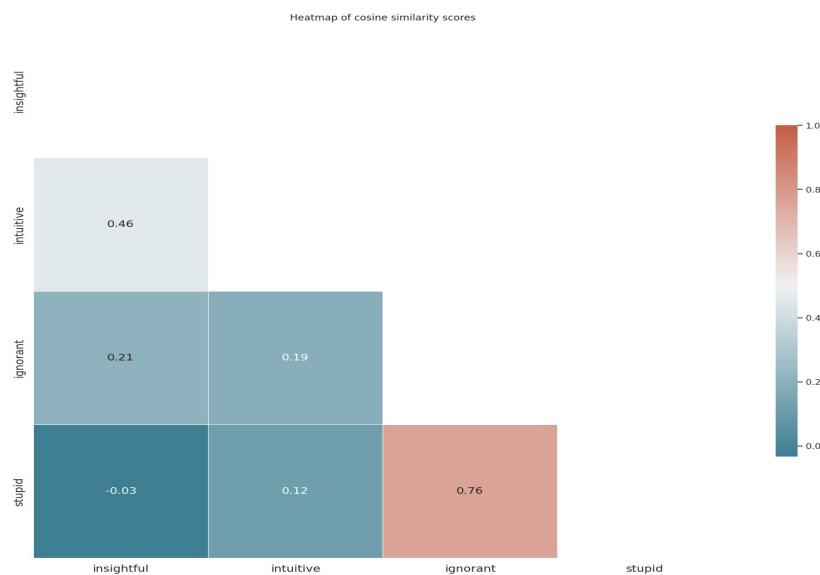


Figure 1: Word2Vec cosine similarity heatmap: *intuitive – insightful, stupid – ignorant*

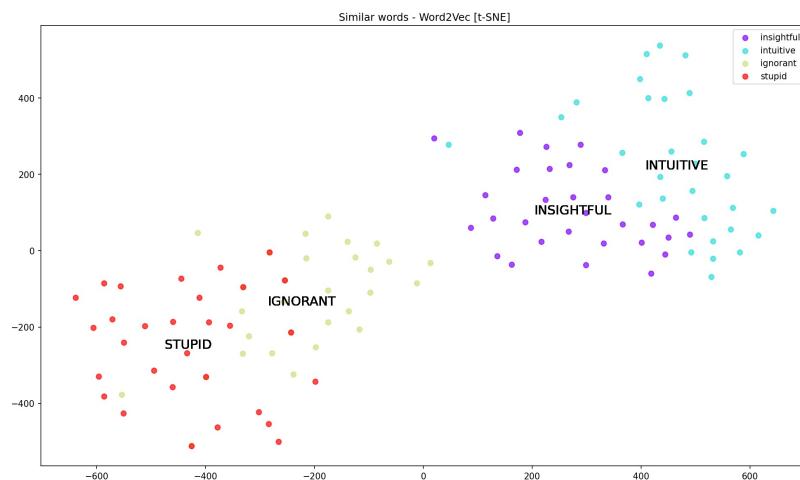


Figure 2: Word2Vec top 30 neighbouring vectors visualization using t-SNE: *intuitive-
insightful, stupid - ignorant*

The embedding results confirm psychological and cognitive linguistic assumptions concerning the close semantic, conceptual, and neural positions not only of synonymous but also antonymous, lexical pairs (for a cognitive-linguistic interpretation of these phenomena consult Lewandowska-Tomaszczyk 1996).

To provide more compelling mental closeness in linguistic results, in the next round we applied the embedding to 5, equally frequent, headwords in opinions, linked by the conventional synonymy - antonymy links: excellent-good-bad, rich-poor (Figures 3–4).

Both the correlation between synonymy good-excellent reported in the heatmap and that between next synonymous pairs good-bad as well as rich-poor, but also the correlations between the antonymic pairs good-bad, rich-poor in a polysemous (financial and mental) senses are precisely equally high or higher (compare the values for good-excellent 0.35 and for good-bad 0.86). This shows a close conceptual similarity between the two types of lexical semantic relations - synonymy as well as antonymy in particular.

As an example of embeddings, twenty words from the opinionated texts (15 negative and 5 positive) were excerpted as opinion discourse markers and their embeddings presented in Figures 5–6.

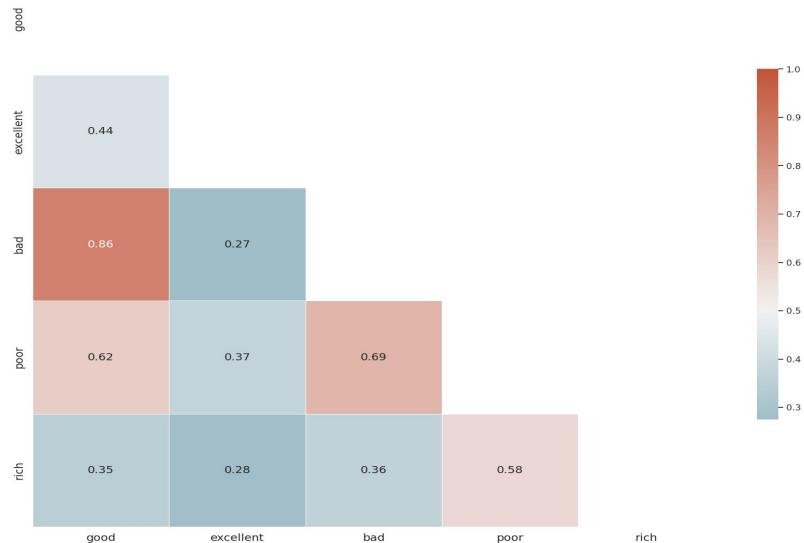


Figure 3: Word2Vec cosine similarity heatmap: *excellent-good-bad, rich-poor*

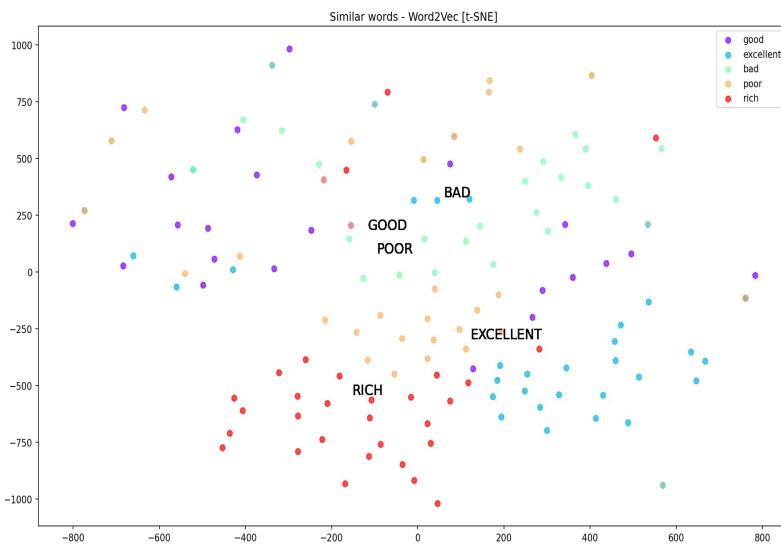


Figure 4: Word2Vec top 30 neighbouring vectors visualization using t-SNE: *excellent-good-bad, rich-poor*

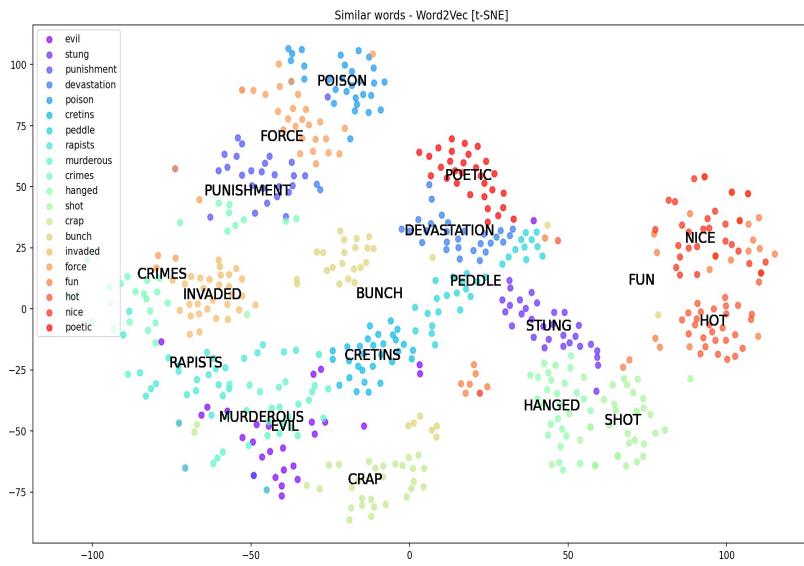


Figure 5: Word2Vec cosine similarity heatmap: twenty words from the opinionated texts (15 negative and 5 positive)

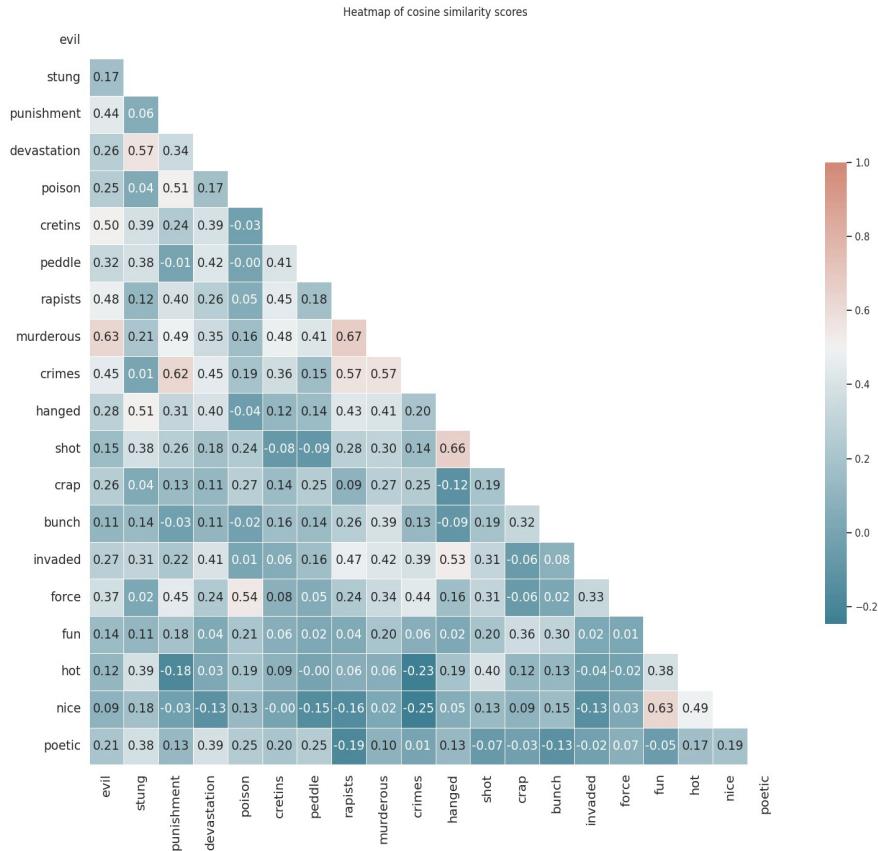


Figure 6: Word2Vec cosine similarity heatmap: twenty words from the opinionated texts (15 negative and 5 positive)

Apart from some obvious examples of words semantically close are e.g., “crimes” and “punishment”, linked by the causative relation, or “shot” and “hanged”, considered as co-hyponymic forms of the same superordinate category, there appear interesting cases of actual ambiguity in this respect e.g., “nice”, “fun” and “hot”, as well as “poetic”, which are all positive in most of the contexts on the one hand, and on the other “hot” has the highest correlation with “shot”, and “nice” correlates with “stung” in the highest degree. These relationships are also represented in the graph, in which the particular cluster shows the same relations visualized as distance representations.

7 Discussion and conclusions

Our linguistic and computational analysis of positive and negative opinions makes it possible to propose a general definition of an *Opinion Speech Event* as a semiotic act, embedded in a social-cultural context, of expressing one's judgement on a subject, person, property or event.

Opinions are generally identifiable through their characteristic lexical, syntactic, and semantic/pragmatic markers. However, single, separate markers cannot be considered necessary or sufficient. They are rather characteristic, sometimes, as in the case of idioms – typical, but, without additional outside world knowledge and context, they do not exploit a final unique set of opinion distinguishers.

As a semiotic act, opinion expressing can lend itself to a more effective identification in face-to-face contexts in which all signs of body language as well as other modalities can be more easily observed. From a linguistic viewpoint, the presence of paralinguistic signals such as prosody in particular can contribute to a more effective multimodal - verbal, visual, auditory etc., identification of opinionated as opposed to fully factual, evidence-based statements.

To reassume, we might propose that an opinion is a communicative Speech Event which can be defined in the following terms:

Source [Human/Thing/Property/Event] >
Opinion Author/Agent [Opinion Holder] > Affect (Interest/(Dis)pleasantness)) >
OPINION (mode) → Communicative Intent [sharing, (mode)→ Persuasive appeal →
→
Opinion Addressee (default)
Results/Consequences
(if Opinion Addressee > **Human Source** > [Pragmatics (Face maintenance (praising) &
Emotions (joy/satisfaction/happiness/encouragement)
or else
/Face threat (blaming) →
>>Emotion >(sadness/humiliation/anger/discouragement)
→
Change (for all Sources): individual/public opinion persuasive effects (yes/no effects of
varying force with the Addressee(s)

Symbols: >causality; → implication

In other words, an Opinion Speech Event assumes the presence of an Author, i.e., opinion holder, who expresses his/her (positive or negative) judgement/evaluation of a Theme to an Addressee. The Addressee may play either the role of an opinion receiver or an opinion Theme (subject), or else function in both the interactional roles. The term *speech event* is neutral as to the communicative medium – it can refer either to spoken or written communication. This type of speech event assumes the presence of both a Theme and an

Addressee, as well as a transfer medium, together with a persuasive force presence in the Opinion expression act. Effects of such a Speech Event are either positive or negative, embodied in raising the Human Theme, and the Addressee's emotionality and raising his/her potential of a particular evaluative judgement. Opinions thus are meant to exert a change: both pragmatically (face maintenance or else face threat or loss) and emotionally via their affective impact polarity.

In this study we analysed definitions of the concept of *opinion* as distinguished from factual or evidence-based statements. On the basis of the literature survey and our own social data excerpts analysis we proposed a taxonomy of opinions expressed in English as identified in selected social media. The taxonomy is based on two criterial properties – the relationship between the opinionated text and facts/evidence and the relationship between the proposition, and on the Agent/Author's (opinion holder's) conviction with regard to its truth/falsity to distinguish between truth and lying.

Through a linguistic analysis of opinionated texts that we acquired manually from various social media websites, we proposed lists of positive and negative opinion lexical discourse markers and their taxonomy. Our discussion was confronted with the automatically generated lexical embeddings of these positive and negative opinion markers.

The results did not support a thesis of unique specification of particular positive or negative opinion discourse markers in terms of a syntactic distribution or presence of a single positive or negative marker – as even what seems to be the most insulting lexical offensive markers can be negated in an opinion e.g., “This essay is not stupid, it is rather not completed”, which reverts their sense from the negative to a more positive one. The same initial message conveyed in the sentence though, can strengthen the negativity of the opinion, for example by the use of metalinguistic negation as in “This essay is not stupid, it is appalling”. This is a piece of evidence that even the use of negative markers with the negative opinion sentence, will not always convert it to a positive statement. Nevertheless, what can be obtained via our study is a set of *preferential lexical usage types in either positive or negative opinion contexts*, as was scrutinised in the analysed texts. The analysed linguistic markers can be proposed to refer not only to explicitly lexical (positive or negative senses), but the relevant polarity judgement is formulated on the basis of a complete collection of morphological, syntactic, and, last but not least, pragmatic markers. This will be particularly evidenced in some cases of implicitly positive or negative senses, in which the whole set of linguistic, paralinguistic and multimodal markers – let alone background knowledge concerning the communicative events and participants – matter. The latter are not considered in this study, especially that in social media communication some of these clues are not available.

Our present conclusions point to a less ambiguous understanding of opinionated texts which is established only by the types of the categorial boundaries of a more advanced syntactic and semantic analysis of the texts both in their pragmatic and cultural settings, but also considering individual identity characteristics of their authors. Furthermore,

irrespective of some of these properties, as the present data analysis shows, there exists a *definitional flexibility* of the boundaries around lexical positive and negative types of opinion expressing markers.

Acknowledgements

The study was prepared in the Linguistics group of COST Action CA 21129 *What are Opinions? Integrating Theory and Methods for Automatically Analyzing Opinionated Communication (OPINION)*.

Appendix 1

Preferred keywords to express opinions

- In opinions based on stronger evidence:
demonstrate, confirm, discover
- Based on weaker evidence
claim, argue, view, suspect
- Emotive/evaluative language in opinions
Praising –
very good, best, excellent, fantastic, superb
Blaming –
poor, weak, unconvincing, unsupported
strong blaming (offensive) - vulgar
- Descriptive:
Use of comparative lexis:
It could be better / It couldn't be better!
It's not so good, could be better [negative]
- Explicit/Implicit
It's awful! I couldn't look at it
- Ambiguous between positive and negative opinion
extremely, unusually, unexpectedly
I haven't seen anything like that in all my life

Appendix 2

Opinionated exemplar samples.

Underlined are proposed opinion lexical/phraseological markers

POSITIVE OPINIONS

- (1) Dr. Jeter is the bomb. He's not real strict, he works with you. He's definitely fair and treats you with respect. I would recommend him anytime. He is very knowledgeable about Criminal Justice issues. He's certainly a “go-to” guy.
<https://www.ratemyprofessors.com/professor/1244444>

- (2) They're just the cutest couple ever and you can tell he absolutely loves being a dad. I love watching his pancake videos in IG
https://www.reddit.com/r/popculturechat/comments/15yhvey/former_tennis_superstar_serenawilliams_and/
- (3) He is so hot. I would bottom for him in an instant
<https://twitter.com/jw291983/status/1694191915145855114>)
- (4) Wow wow wow wow! What an amazing place. Stunning location, fabulous, attentive and wonderful staff. And the food, oh my goodness - to die for!!! We went with our two young children (1&7) and they couldn't have been more accommodating. Best restaurant ever, can't recommend highly enough!
https://www.tripadvisor.com>ShowUserReviews-g4505725-d12658522-r912012364-The_Mussel_House_Restaurant-Ksamil_Saranda_Vlore_County.html#
- (5) i came into #Barbie with HIGH expectations and they still exceeded them. margot robbie was AMAZING as barbie. the whole film was so visually alluring, it was hilarious yet gut-wrenching. i'm definitely going to see it again this weekend
<https://twitter.com/cambeserious/status/1681847455141310465>
- (6) ... For the first time, Prince Harry tells his own story, chronicling his journey with raw, unflinching honesty. A landmark publication, Spare is full of insight, revelation, self-examination, and hard-won wisdom about the eternal power of love over grief.
<https://www.amazon.com/Spare-Prince-Harry-Duke-Sussex/dp/0593593804>
- (7) Kadare's writing is absolutely lyrical. Chronicle of Stone is the story of a city through one imaginative boy's musings. Physical place is such a large part of history that often goes unobserved. This is a magical book capturing the beauty and opposing rawness of life.
https://www.goodreads.com/book/show/708124.Chronicle_in_Stone?from_search=true&from_srp=true&qid=dSezT0IyQs&rank=2
- (8) The casting is also amazing. Cillian Murphy does an outstanding job playing Thomas Shelby. Cant wait for the next season. Definitely give it a try!
https://www.reddit.com/r/netflix/comments/auqzfs/is_peaky_binders_good/
- (9) Great prof, passionate about the subject, cares about student opinions, and really gets you thinking. There is a lot of reading, but that's just Russian literature for you. Actually, by the end, there were times I found myself wanting to read more.
<https://www.ratemyprofessors.com/search/professors?q=ani%20kokobobo>
- (10) Oh boy, this is obviously going to turn out great for everyone involved.
https://www.reddit.com/r/LinusTechTips/comments/15r1pfz/the_problem_with_linus_tech_tips_accuracy_ethics/)

- (11) Perfect jacket will not s e arrepender.ca SA perfect on the body! I ordered g but the M would be fairer! I still liked
https://www.aliexpress.us/item/3256805548166912.html?spm=a2g0o.productlist.main.1.28c1443fZTzymc&algo_pvrid=7d038dc0-fed6-43f8-ac26-fbb7c6fd9548&algo_exp_id=7d038dc0-fed6-43f8-ac26-fbb7c6fd9548-0&pdp_npi=4%40dis%21USD%2138.29%2111.85%21%21%21278.45%21%21%402101ea7116952907118767926e2a70%2112000034156325241%21sea%21US%210%21A BS&curPageLogUid=CuzBQAHIOS2X

NEGATIVE OPINIONS

- (1) the guy is impossible to understand and he doesn't understand any questions that you ask him. avoid this guy like the plague. he is also very unclear about what he wants done for the class (Teach Rev)
<https://www.ratemyprofessors.com/professor/1089225>
- (2) Terrible professor. Difficult to understand. For the projects he doesn't explain the directions until the day before it's due. If you ask him any question his response will likely be, "I don't understand" Would not recommend to anyone!!(Teach Rev)
<https://www.ratemyprofessors.com/professor/1089225>
- (3) I wish we could just let them sort their own s**t out, we are not the worlds parents. #Sudan
<https://www.facebook.com/britishnationalism>
- (4) He's pretty nice, but he's definitely one of the worst teachers I've ever had. He's always extremely vague & unhelpful about his assignments. Personally, he never revised my drafts for my essays even though we were required to turn them in, & then I got bad grades on most of my essays. I emailed him about it several times & he never replied. (Teach Rev)
<https://www.ratemyprofessors.com/professor/1089225>
- (5) Ugh. Skip him. (Teach Rev)
<https://www.ratemyprofessors.com/professor/1089225>
- (6) British politicians are destroying our nation. #Sudan
<https://www.facebook.com/britishnationalism>
I totally agree! Britain to be 'proper' Britain! (Fb; #Sudan)
<https://www.facebook.com/britishnationalism>
- (7) Vote Reform surely they are better than this bunch who seem to be invaded [i.e., invading] us with Islamics !!
<https://www.facebook.com/britishnationalism>
- (8) We should take no more immigrants of any sort . And as for "rescuing" people , they mostly went where the money was. Now the dream has crumbled we have to pay out to bring them here. And suppose some of these people leaving a conflict are the enemy?
<https://www.facebook.com/britishnationalism>
- (9) Delivered is cheap rag of very bad cotton. Worthless product, scam. Don't buy anything here
https://www.etsy.com/uk/listing/892209899/indian-cotton-handmade-summer-women?ga_order=most_relevant&ga_search_type=all&ga_view_type=gallery&ga_search_query=&ref=sc_gallery-7-1&pro=1&frs=1&referrer_page_guid=f6b2872cd42.27e8b29ca76fd982ab42.00&plkey=90ef4edb683c11e9300cbc3c77bdf3c50c4fe385%3A892209899

AMBIGUOUS OPINIONS

Conventional structural ambiguity, i.e., the ambiguity caused by the syntactic relations, is not particularly frequent in opinionated text. An example at from the <https://www.quora.com/What-are-some-good-examples-of-ambiguous-phrases> can be one such example:

- (1) My mother never made chocolate cake, which we all hated.

The sentence is ambiguous between 3 readings:

- (i) reading 1. (negative judgement) *We hated the fact that our mother never made chocolate cake for us*
- (ii) reading 2. (positive judgement wrt mother, negative wrt chocolate cake) *We hated chocolate cake and that was why mother never made it*
- (iii) reading 3. (positive judgement) *We loved the fact that mother never made chocolate cake that we would all hate (i.e., would not like)*.

BALANCED OPINIONS

- (1) Very very slow grader takes up to 4 weeks to return simple assignments. Not that hard to grade math, it's either right or wrong. That being said he is extremely fair but very dull class. I hate math and got a B but I wouldn't want to take another class with him
<https://www.ratemyprofessors.com/professor/1270844>
- (2) The fabric is very, very, VERY thin so I don't think I'll even be able to wear them just to be at home, maybe in the bedroom with my partner but that's about it. The fit is great and they make my butt look good but I just don't see myself exercising wearing them due to how thin the fabric is.
https://www.aliexpress.us/item/3256803758682307.html?spm=a2g0o.productlist.main.35.28c1443fZTzymc&algo_pvrid=7d038dc0-fed6-43f8-ac26-fbb7c6fd9548&algo_exp_id=7d038dc0-fed6-43f8-ac26-fbb7c6fd9548-17&pdp_npi=4%40dis%21USD%213.44%211.38%21%21%2125.00%21%21%21%402101e a7116952907118767926e2a70%2112000027528238423%21sea%21US%210%21ABS&curlPageLogUid=J6ibSZv9K63g

List of source websites of opinionated samples:

Rate my professor: <https://www.ratemyprofessors.com/>
Reddit: <https://www.reddit.com/>
Twitter: <https://twitter.com/>
Tripadvisor: <https://www.tripadvisor.com/>
Amazon: <https://www.amazon.com/>
Goodreads: <https://www.goodreads.com/>
Facebook: <https://www.facebook.com/>
Etsy: <https://www.etsy.com/>
Aliexpress: <https://www.aliexpress.us/?gatewayAdapt=glo2usa>
Quora: <https://www.quora.com/What-are-some-good-examples-of-ambiguous-phrases>

References

- Akopova, Asya. 2013. Linguistic manipulation: definition and types. *International Journal of Cognitive Research in Science, Engineering and Education* (online only). 1(2).
- Al-Saqlqa, Samar & Arafat Awajan. 2019. The use of word2vec model in sentiment analysis: A survey. In *Proceedings of the 2019 international conference on artificial intelligence, robotics and control*, 39–43. New York, NY: Association for Computing Machinery.
- Bączkowska, Anna. 2022. Explicit and implicit offensiveness in dialogical film discourse in Bridgit Jones films. *International Review of Pragmatics* 14(2). 198–225.
- Bączkowska, Anna & Dagmar Gromann. 2023. From Knobhead to Sex goddess: swear words in English subtitles, their functions and representation as Lingusitic Linked Data. *Rasprave* 491(1). DOI: <https://doi.org/10.31724/rihj.49.1.4>.
- Dahlman, Christian, Farhan Sarwar, Rasmus Bååth, Lena Wahlberg & Sverker Sikström. 2015. Prototype Effect and the Persuasiveness of Generalizations. *Review of Philosophy and Psychology* 7. 163–180.
- Davison, W. Phillips. “Public opinion”. *Encyclopedia Britannica*, 7 Nov. 2023. Available at: <https://www.britannica.com/topic/public-opinion> (accessed 12 November 2023).
- Emmot, Catherine, Marc Alexander & Agnes Marszałek. 2014. Schema theory in stylistics. In Michael Burke (ed.) *The Routledge Handbook of Stylistics*, 130–134. Abingdon: Routledge.
- Entman, Robert M. 2007. Framing bias: Media in the distribution of power. *Journal of Communication* 57(1). 163–173.
- Gärdenfors, Peter. 2000. *Conceptual spaces. The geometry of thought*. Cambridge: MIT Press.
- Grice, Herbert Paul. 1975. Logic and conversation. Reprinted in H. P. Grice. 1989. *Studies in the way of words*, 22–40. Harvard: Harvard University Press.
- Grice, Herbert Paul. 1978. Further notes on logic and conversation. Reprinted in H. P. Grice. 1989. *Studies in the way of words*, 41–57. Harvard: Harvard University Press.
- Grice, H. Paul. 1989. *Studies in the way of words*. Harvard: Harvard University Press.
- Hoehn, Sviatlana, Barbara Lewandowska-Tomaszczyk, Sjouke Mauw & Dove Gabbay. 2023. Dripping the poison: The instruments of bias. *Journal of Applied Logic* 10(4). 555–585.
- Hofstede, Geert (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture* 2(1). DOI: <https://doi.org/10.9707/2307-0919.1014>.
- Kaiser, Elsi & Catherine Wang. 2020. Distinguishing fact from opinion: Effects of linguistic packaging. *Papers of the Cognitive Science Society*. Available at: <https://cognitivesciencesociety.org/cogsci20/papers/0028/0028.pdf> (accessed September 11 2023).
- Kaiser, Elsi & Catherine Wang. 2021. Packaging Information as Fact versus Opinion: Consequence of the (Information-) Structural Position of Subjective Adjectives in *Discourse Processes* 58(7). 617–641.
- Król, Małgorzata (ed.). 2015. *Opinie prawne w praktyce* [Legal opinions in practice]. Warszawa: C. H. Beck.
- Król, Małgorzata. 1987. Concept of Normative Leeway in an Application of Law, *Studies in the Theory and Philosophy of Law* 3. 19–35.
- Kumar, Akshi & Divya Gupta. 2021. Sentiment Analysis as a Restricted NLP Problem. In Fatih Pinarbasi, Faith & M. Nurdan Taskiran (eds.), *Natural Language Processing for Global and Local Business*, 65–96. Hershey, PA: IGI Global.

- Available at: <https://www.igi-global.com/dictionary/sentiment-analysis-as-a-restricted-nlp-problem/89393> (accessed 28 October 2023).
- Lakoff, Robin. 1973. The Logic of Politeness, or Minding Your P's and Q's. *Chicago Linguistics Society* 9. 292–305.
- Lakoff, George. 1990. *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press.
- Langacker, Ronald. 1987. *Foundations of cognitive grammar*. Stanford: Stanford University Press.
- Leech, Geoffrey. 1983. *Principles of Pragmatics*. London: Longman.
- Lewandowska-Tomaszczyk, Barbara. 1996. *Depth of negation - A cognitive semantic study*. Lodz: University of Lodz Press.
- Lewandowska-Tomaszczyk, Barbara. 2010. Re-conceptualization and the emergence of discourse meaning as a theory of translation. In Barbara Lewandowska-Tomaszczyk & Marcel Thelen (eds.), *Meaning in translation*, 105–148. Frankfurt am Main: Peter Lang.
- Lewandowska-Tomaszczyk, Barbara, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene & Slavko Žitnik. 2023. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 19(1). 7–48.
- Mikolov, Tomas, Kai Chen, Greg Corrado, & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Preece, Jennifer, Yvonne Rogers & Helen, Sharp. 2002. *Interaction Design: Beyond Human-Computer Interaction*. New York: John Wiley & Sons.
- Raskin, Victor. 1985. *Semantic mechanisms of humor*. D. Reidel Publishing Company.
- Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology* 104. 192–233.
- Schank, Robert C. & Robert P. Abelson. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. New York: Lawrence Erlbaum.
- Searle, John. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Searle, John. 1998. *Mind, Language and Society*. New York: Basic Books.
- Semino, Elena. 1997. *Language and World Creation in Poems and Other Texts*. Abingdon: Routledge.
- Wang, Rui, Deyu Zhou, Mingmin Jiang, Jiasheng Si, & Yang Yang. 2019. A survey on opinion mining: From stance to product aspect. *IEEE Access* 7. 41101–41124.
- Wittgenstein, Ludwig. 2001 (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.
- Wróblewski, Jerzy. 1969. Legal reasoning in legal interpretation. *Logique et Analyse* 12(45). 3–31.

About the Authors

Barbara Lewandowska-Tomaszczyk is Professor Ordinarius Dr Habil. in Linguistics and English Language at the Department of Language and Communication at the University of Applied Sciences in Konin (Poland). Her research focuses on cognitive semantics and pragmatics of language contrasts, corpus linguistics and their applications in translation studies, lexicography and online discourse analysis. She is invited to read papers at international conferences and to lecture and conduct seminars at universities. She publishes extensively, supervises dissertations and also organizes international conferences and workshops.

Address

Department of Language and Communication, University of Applied Sciences in Konin
1, Przyjazni str.
62 510 Konin, Poland

e-mail: barbara.lewandowska-tomaszczyk@konin.edu.pl
ORCID: 0000-0002-6836-3321

Anna Baczkowska, Dr Habil. Prof. of University of Gdańsk, she holds MA in English Philology, which she received from Adam Mickiewicz University in Poznań, as well as PhD in linguistics and D.Litt. in English Linguistics, which she received from the University of Łódź. Her research interests revolve around translation studies (film subtitles), cognitive semantics, corpus and computational linguistics, and discourse studies (media discourse). She has guest lectures in Italy, Spain, Portugal, UK, Norway, Kazakhstan and Slovakia, and she has also conducted research during her scientific stays in Ireland, Iceland, Norway, Austria and Luxembourg.

Address

Institute of English and American Studies, University of Gdańsk
Wita Stwosza 51
80-308 Gdańsk, Poland

e-mail: anna.baczkowska@ug.edu.pl
ORCID: 0000-0002-0147-2718

Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and data mining. Especially, her scientific interests include Semantic Similarity, Language Technology for Cultural Heritage, Morphologically rich languages (MRL), Multi-word Expressions (MWEs), Information Retrieval (IR), and Text Classification (TC). Much of her recent work has been focusing on analysing offensive language. She has published a variety of studies and a few of her articles are under review or in preparation. She is a member of several international research actions funded by the EU.

Address

Jerusalem College of Technology, Department of Computer Science
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel

e-mail: liebchaya@gmail.com
ORCID: 0000-0003-0476-3796

Jūratė Ruzaitytė is Professor at the Department of Foreign Language, Literary and Translation Studies and a senior researcher at the Centre of Intercultural Communication and Multilingualism at Vytautas Magnus University, Kaunas, Lithuania. She holds a Doctor of Philosophy (PhD) focused in Linguistics from the University of Bergen, Norway. She has rich experience in (international) research projects, including a national project (Semantika-2, 2018-2019) in the framework of which a software for automated detection of offensive online comments in Lithuanian was created. She is also the Associate Editor of the Lithuanian Applied Linguistics Journal and a board member of the Lithuanian Association of Applied Linguistics. Her research interests include sociolinguistics, pragmatics, discourse analysis, language and ideology, hate speech, propaganda, and disinformation.

Address

Department of Foreign Language, Literary and Translation Studies
Vytautas Magnus University
K. Donelaičio g. 58
LT-44248 Kaunas, Lithuania

e-mail: jurate.ruzaite@vdu.lt
ORCID: 0000-0002-4519-8087

Ardita Dylgjieri is professor of Stylistics and Text linguistics at the Department of Foreign Languages, Faculty of Human Sciences, University of Elbasan ‘Aleksandër Xhuvani’, in Albania. She holds an MA in World Literature and a PhD in Linguistics (Pragmatics and Critical Discourse Analysis). In the framework of her PhD thesis and not only, she has shown deep and special interest in Political Discourse Analysis and all its main linguistic peculiarities. Her other research and academic interests include Psycholinguistics, Sociolinguistics, Linguistic Diversity, Second Language Acquisition and Cognitive Linguistics.

Address

Department of Foreign Languages, Faculty of Human Sciences, University of Elbasan ‘Aleksandër Xuvani’
Rruge ‘Ismail Zyma’
Elbasan, Albania

e-mail: ardita.dylgjeri@uniel.edu.al
ORCID: 0000-0003-4016-2701

Ledia Kazazi is a professor of English Language and Linguistics at the Department of Foreign Languages at the University of Elbasan “Aleksander Xhuvani” in Elbasan, Albania. She holds a PhD in Cognitive Linguistics. Her research focuses on all aspects of Cognitive Linguistics, especially Conceptual Metaphor and Conceptual Metonymy but also expands to Cognitive Semantics, Cognitive Narratology, Multimodal Discourse Analysis and Critical Discourse Analysis. She has published several articles and given several conference talks on topics related to the aforementioned disciplines.

Address

Department of Foreign Languages, Faculty of Human Sciences, University of Elbasan “Aleksander Xhuvani”
Rruge “Ismail Zyma”
Elbasan, Albania

e-mail: ledia.kazazi@uniel.edu.al
ORCID: 0000-0002-9516-3018

Erika Lombart is a research associate at UCLouvain's Language and Communication Institute. After spending five years at CENTAL, the Centre Traitement Automatique de la Langue, she defended her doctoral thesis in September 2001. Her thesis was centred around "the non conventional implicit in discussion forums". In addition, she authored a book titled "Entre les lignes des réseaux sociaux" (Between the lines of social networks) published by Editions L'Harmattan. Her research currently centres around investigating the application of metaphors in political discussion as well as examining the formal indicators of implication in discourses on social networks. She is actively participating in the COST Opinion Action.

Address

Institut Language et Communication
Collège Erasme
Place Cardinal Mercier 31, boîte L3.03.02,
1348, Louvain-la-Neuve
Belgium

e-mail: erika.lombart@uclouvain.be
ORCID: 0000-0001-7968-0413