Lodz Papers in Pragmatics 19.1 (2023): 7–48
Special issue on *Various dimensions of expressivity*
https://doi.org/10.1515/lpp-2023-0002

Barbara Lewandowska-Tomaszczyk
University of Applied Sciences in Konin

Anna Bączkowska
University of Gdańsk

Chaya Liebeskind
Jerusalem College of Technology

Giedre Valunaite Oleskeviciene
Mykolas Romeris University

Slavko Žitnik
University of Ljubljana

AN INTEGRATED EXPLICIT AND IMPLICIT OFFENSIVE LANGUAGE TAXONOMY

Abstract
The current study represents an integrated model of explicit and implicit offensive language taxonomy. First, it focuses on a definitional revision and enrichment of the explicit offensive language taxonomy by reviewing the collection of available corpora and comparing tagging schemas applied there. The study relies mainly on the categories originally proposed by Zampieri et al. (2019) in terms of offensive language categorization schemata. After the explanation of semantic differences between particular concepts used in the tagging systems and the analysis of theoretical frameworks, a finite set of classes is presented, which cover aspects of offensive language representation along with linguistically sound explanations (Lewandowska-Tomaszczyk et al. 2021). In the analytic procedure, offensive from non-offensive discourse is first distinguished, with the question of offence Target and the following categorization levels and sublevels. Based on the relevant data generated from Sketch Engine (https://www.sketchengine.eu/ententen-english-corpus/), we propose the concept of offensive language as a superordinate category in our system with a number of hierarchically arranged 17 subcategories. The categories are taxonomically structured into 4 levels and verified with the use of neural-based (lexical) embeddings. Together with a taxonomy of implicit offensive language and its subcategorization levels which has received little scholarly attention until now, the categorization is exemplified in samples of offensive discourses in selected English social media materials, i.e., publicly available 25 web-based hate speech datasets (consult Appendix 1 for a complete list). The offensive category *levels* (types of offence, targets, etc.) and *aspects* (offensive language property clusters) as well as

the categories of *explicitness* and *implicitness* are discussed in the study and the computationally verified integrated explicit and implicit offensive language taxonomy proposed in the study.

Keywords
(categorization) aspects, categorization schemata, corpus data(sets), explicit, implicit, language levels, offensive language, tagging, taxonomy, word (lexical) embeddings

1 Introduction

In the last decade, social media grew exponentially and offered everyone area to express themselves online. Enabling people to write on different online platforms, without even identifying themselves, led to a new era of freedom of speech. Despite this new technology establishing a new medium for communication which is a very positive thing, it also introduced many drawbacks. Social media have become a place where discussions and controversies that take place may result in insults and hatred. Having in mind the development of communication and media technologies, the research on offensive language recognition and identification could be useful in dealing with everyday variety of received texts which may contain abusive content. Offensive language research may lead to establishing algorithms spotting offensive content and allowing the intelligent software to automatically protect users from undesirable attacking messages.

The present study focuses on two objectives – (a) the first relates to the definitional revision and enrichment of the explicit offensive language taxonomy by reviewing the collection of available corpora and comparing tagging schemas applied there and (b) the second aim involves the focus on proposing an overarching model, where varying subtypes of implicitness used in the context of offensive language are conceptually linked relying on implicit offensiveness.

The verification of the offensive language taxonomic categorization was originally proposed in Lewandowska-Tomaszczyk et al. (2021) and extended in the present study, using publicly available annotated datasets and (contextual) word embeddings, along with deep neural approaches. The first part of the paper presents our 2021 model (Lewandowska-Tomaszczyk et al. 2021), followed by a critical discussion of the problem concerning conceptual and structural differentiation among 17 offensive language headwords we propose in the present study. With ample reference to contemporary linguistic literature, the definitions of the headwords are developed, and in the second part of the paper a description and results of relevant meaning discriminatory verification methods are provided. We apply approaches based on non-contextual word embeddings (i.e., Word2Vec and fastText), as well as more advanced computational linguistics methods (i.e., pre-trained transformer models). Analysis of the results reveals that discriminating among the related terms may require methodology combining a deeper analysis of language levels and a wider adoption of the contextual language use criteria.

In order to meet the research objectives, over 60 relevant offensive language corpora are identified and the concept of offence is discussed. Out of these datasets, 30 are English-based and 25 of them were available for research purposes. Typologies of offensive language current in the literature are presented and reviewed for their coherence with the sound semantic and contextual criteria. Based on their analysis, we develop a new taxonomy model of offensive language types and employ a computational linguistic exploratory analysis to examine selected offensive language datasets. This aims to lead towards uncovering clearer similarities and differences among existing categories, further discussed in terms of semantic and contextual criteria to be applied in a linguistically motivated annotation procedure.

This paper is organized as follows: Sections 2 and 3 provide linguistically sound interpretations of different offensive language categories and aspects. Each category is clearly identified, and the offensive language taxonomy is proposed. We also provide some insights on the implicit offensive language and its justification. In Section 5, we use computational linguistic techniques to validate the proposed offensive language taxonomy using word embeddings. We conclude the paper with a discussion of results and justification of the explicit part of taxonomy based on the results of the exploratory analyses.

2 Offence and offensive language

Offence is part and parcel of impolite and/or uncivil events. It involves either behaviour, or language, or else both language and behaviour that are the addressee's face-aggravating acts in a particular context (Goffman 1955) . Offence comes about when the language user (speaker) communicates face-attack intentionally, or the addressee (hearer) perceives and/or constructs 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 [Culpeper (2005); Haugh and Sinkeviciute (2019); Haugh and Culpeper (2018)]. 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*.

Offence is a central notion for research on impoliteness, aggression, and conflict talk (Culpeper and Haugh 2021). While the term *offence* (*offensive*, *offensiveness*) poses definitional problems (Culpeper 2011), some definition has been recently offered based on a corpus study of *offence*/*offensive* (Culpeper and Haugh 2021), which provides approximations of the term (and its synonyms) as well as stresses its variability (e.g., across British and Australian English). *Offence* is a multifaceted concept that assumes a verbal attack conducted in interpersonal interaction, which breaches social norms, and leads to moral evaluations and/or emotional disturbance in the addressee (Haugh and Sinkeviciute 2019). The adjective *offensive* labels a subjective evaluation of some behaviour that one judges as being insulting, abusive, hurtful, morally wrong, and socially inappropriate, and which can refer either to an individual or a specific group (religious, ethnic, social, etc.) (Culpeper and Haugh 2021). However, the struggle to define the term *offence* (*offensiveness* or *offensive* language) and its related terms (such as *insults*, *slurs*, *harassment*, *taboo*, etc.) could be minimised to a large extent or even reconciled, if, instead of treating them in terms of binary notions, we assume the gradability of these meanings as they belong to one superordinate schema. Thus, a cognitive linguistics approach (e.g., the Langackerian schematic networks in Langacker 1987) to (conceptual) meaning may, at least partially, solve the problem. In Langacker's terms, words typically have polysemic senses: "Most lexical items have a considerable array of interrelated senses, which define the range of their conventionally sanctioned usage. These alternate senses are conveniently represented in network form" (p.31). In other words, this is not the case that lexical units are characterised by a single, closed type of lexical structure, but are rather subject to contextual variation. In our study too, *offensiveness* is the most generic term, which is an abstraction emerging out of lower-level notions (schemas) that contribute to its meaning, understood in terms of a descriptive generalisation. Consequently, one superordinate context in the proposed taxonomy is characterized in our study by means of several other keywords (e.g., *toxic*, *offensive*, *harassment*, *cyberbullying*, *threat*).

In linguistic and cultural models, offensive language is part of incivility (Lewandowska-Tomaszczyk 2017) and impoliteness research (Culpeper 2011). Incivility is a broader concept than impoliteness and refers to deviant, aggressive, vulgar, face-threatening, and rude behaviour whose aim is to exert control over others in order to empower one's own position in a community of practice at the cost of others (Lewandowska-Tomaszczyk 2017). The impoliteness position, on the other hand, evolved as a reaction complementing politeness theories (Brown and Levinson 1987) that relied on and were criticised for their claims of cross-cultural universality. The notion of impoliteness has been largely studied from the point of view of the discourse analytic approach by e.g., Christie (2006) and Eelen (2001), who argue against the claim of considering impoliteness as a lacking-politeness act rather than a mostly intentional act of hurting an interactant. Hence, it can be considered subjective, context-dependent, and situated, and can be studied (*post-factum*) primarily from the internal point of view i.e., the perspective of the interlocutors, involved in a conversation, dubbed the *emic level*. Other scholars (Culpeper and Haugh 2014) hold that impoliteness should be analysed both from an internal (*emic*) and an external perspective (i.e., the researcher's, the *etic* level). In our study, an offensive language taxonomy is proposed for Twitter message language, assuming the *emic-etic* perspective, as well as the gradeability of the keyword taxonomy membership.

3 First taxonomy of offensive language

In this section we begin by presenting our first proposal of offensive language taxonomy (Lewandowska-Tomaszczyk et al. 2021) and then extend it to cover all aspects of offensive language categorization.

3.1 Initial offensive language taxonomy proposal

The main focus of our preliminary SALLD paper (Lewandowska-Tomaszczyk et al. 2021) was the definitional revision and enrichment of offensive language taxonomy making reference to publicly available offensive language datasets and testing them on available pretrained lexical embedding systems. We reviewed some of the available corpora and compared tagging schemas applied there while making an attempt to explain semantic differences between particular concepts of the category OFFENSIVE in English. A finite set of classes that cover aspects of offensive language representation along with linguistically sound explanations was presented, based on the categories originally proposed by Zampieri et al. (2019b) in terms of offensive language categorization schemata, and tested by means of Sketch Engine tools on a large web-based corpus. In our analytic procedure, we firstly distinguish offensive from non-offensive language. The non-offensive cases, beyond the scope of our research, are left unaccounted for. Secondly, the question of Target is taken up – if there is no identifiable addressee of offence, the language is considered an example of self-expression, having, e.g., an exclamatory function (e.g., swear words used to express anger, frustration, pain, etc.). An analytic procedure followed, which focused on a discussion of particular offensive language subcategories and setting up a taxonomy presented in Figure 2. The offensive language categorization schemata were next juxtaposed and discussed with reference to non-contextual word embeddings fastText, Word2Vec, and Glove. The methodology for mapping from existing corpora to a unified ontology as presented in this paper was provided (see Figure 1).

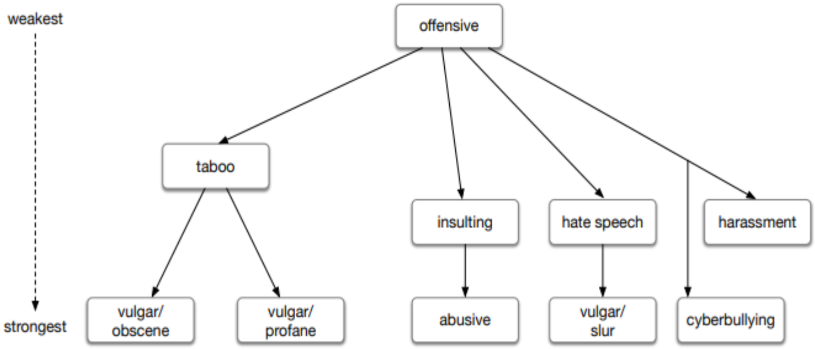


Figure 1: Initially proposed taxonomy of offensive language categories.

3.2 Headword categories and definitional problem

The lack of consensus among researchers concerning both the repertory of classificatory headwords as well as their definitions, observed in the literature and in the available tagset systems, posed classificatory and computational problems. Some authors, e.g., Poletto et al. (2020) or Founta et al. (2018a), consider *abusive language* as a general superordinate term involving hate speech, offensive language, and other subcategories. Some other researchers (Razavi et al. 2010) take offensive language as an overarching category for such terms as profanity and rudeness. We based our discussion on the Sketch Engine data and collocate frequencies and proposed offensive language to serve as a superordinate category in our system to cover instances of upsetting or embarrassing language because of its denigrating character. With reference to legal contexts (Bretschneider and Peters 2016), offensive language is defined as the term indicating hurtful or derogatory comments by one person or a group to another person or a group. In terms of socio-cultural standards, offensive language identifies a number of hierarchically arranged subclasses categorized according to (i) the type of offence addressee and (ii) reference to various matters of the taboo status. A number of subtypes were identified in the schema, with some of the categories used in the literature, such as toxic language and its levels (Kunupundi et al. 2020), left out in our first proposal due to its generally vague definitional character.

The model in our previous study (Lewandowska-Tomaszczyk et al. 2021) was closely related to the category system as an originally 3-level approach proposed by Zampieri et al. (2019b), with our modifications of retaining 2 categories and 4 sub-levels, tested by means of Sketch Engine tools on a large web-based corpus and juxtaposed to non-contextual word embeddings fastText, Word2Vec, and Glove on the relevant datasets. The repertory of offensive categories comprised 11 types of language:

(1) offensive, (2) taboo, (3) insulting, (4) hate speech, (5) harassment, (6) vulgar, (7) vulgar/obscene, (8) vulgar/profane, (9) abusive, (10) vulgar/slur and (11) cyberbullying.

Upon examining 60 of publicly available datasets, we conclude that only some of the datasets and annotation schemas publicly available are based on more clearly identifiable criteria e.g., datasets by Chung et al. (2019), Ousidhoum et al. (2019), Zampieri et al. (2019a) examined here, from the OLID resources (Zampieri et al. 2019b, Zampieri et al. 2020), and the dataset introduced in Bretschneider and Peters (2017), consisting of 3 sub-datasets, based on Facebook posts, including the comments published in response to them (Bretschneider and Peters 2017). Most of the others analysed in the present paper have been often annotated by crowdsourced volunteers and the offensive language categories in their annotation schemes have been fairly spontaneously attributed with no explicit diversification criteria identified or referred to. These considerations decided about extending our Proposal 1 to Proposal 2, whose aim is to model a more extended offensive language taxonomy on a set of more strictly identifiable criteria.

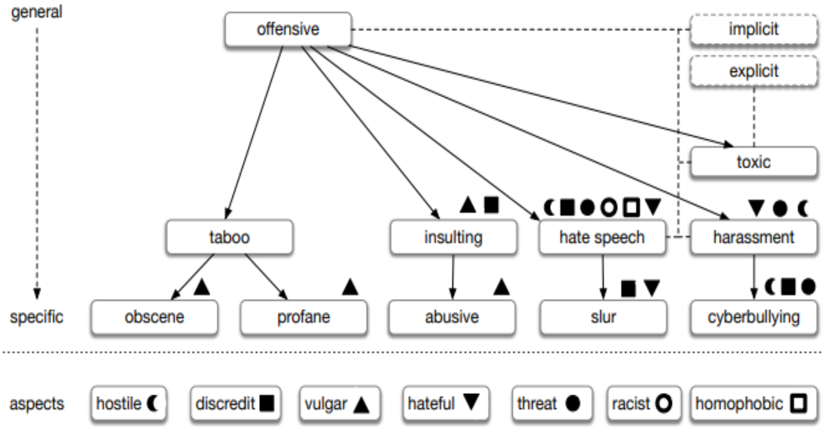


Figure 2: Offensive language taxonomy.

The upper part is hierarchically organized from general to specific, some categories are annotated with contexts, differentiated are also implicit and explicit offensive types. The bottom line identifies offensive language variants [aspects], marked by symbols (e.g., circles) to indicate semantic overlaps.

3.3 Extension of the initial proposal

The current proposal aims at an extension and enrichment of the initial taxonomy to provide in it more variance among offensive language types, as well as to put forward clearer linguistic and distributional criteria of their identification in actual language types. The computational instruments used for their verification in actual texts, as well as the results, follow in the forthcoming sections of the paper.

In the current study, we extend the number of proposed headwords to 17 categories, taxonomically structured similarly to that in our initial proposal, while making an attempt to provide linguistic and definitional justification for the extension, whose basic taxonomic structure is sketched in Figure 3.

As experienced with any types of classification attempts, the identification of categories of offensive language is not easy or straightforward. We propose the extended taxonomy, bearing in mind a fuzzy categorial nature of all concepts, first signaled by mathematicians and philosophers (Zadeh 1965), and widely recognized in cognitive studies and cognitive linguistics at present (Lakoff 1987). The categories of offensive language are no exception and they are subject to natural "leaking" as any taxonomic system referring to natural language. On the other hand, there are some reasons to assume that the current proposal might present a more explicitly delineated schema than the ones current in the offensive language identification systems.

The concept of *offence* is considered in the present study as a dominant category, which includes both more weakly and more strongly experienced offence, depending on the use of expressive means by the offender(s). The expressive means can represent a range of semiotic character – from offending behaviour, external attributes, such as garment or hairstyle, which can be visually offending to some addressee(s) through various types of verbal offence, of different degrees of intensity. Furthermore, the messages conveyed by the vocal communication channel are most direct, while in the present study we focus on the offensive language in the social media texts, in which expressiveness can be additionally marked by multiple repetitions, capitalisation, punctuation, or visual symbols, not considered in the present study.

The second category of offensive language considered here is toxic language of a fairly non-crisp definitional context. In our taxonomy, toxic texts, which express what is generally referred to as toxic content, represent another category of offensive language. However, while generally offensive language results in varying degrees of emotional offence, toxic language is clearly strongly hurting to the addressee, irrespective as to whether there is a direct interactional addressee or an absent one referred to, toxic language is both containing high intensity and can prototypically be contained in more than one sentence. Language toxicity is subject to a number of proposals, some of them further subdivide it into various types (Jigsaw and Google 2018), in which toxic is categorized as in terms of the types of toxicity which involve: toxic, severe_toxic, obscene, threat, insult, identity_hate. Although the subcategories used in this format can be considered toxic of different degrees, elements of toxicity are found in a number of other offensive types as well, therefore we propose to classify toxic as one of the dominant categories, subordinate only to offensive language.

While in the present study on the explicit category with OFFENSIVE proposed to function as the dominant class, overarching all others, we propose to consider abusive language to be viewed as a constituent of the superordinate category of offensive language, characterized in legal terms as harsh, violent, profane, or derogatory language which is directed to violate the dignity of an individual, including profanity and slurs of racial, ethnic, or sexist manner. According to Nobata et al. (2016), abusive language includes dominance, derailing, harassment, and threat; it may discredit race, religion, or gender, and may also include stereotypes and spam. Other offensive and abusive language categorization systems identify such subclasses as benevolent, derailing, discredit, dominance, harassment, hate, hostile, insult, obscene, profane, spam, and stereotype (Pamungkas and Patti 2019).

Offensive is furthermore considered as causing someone to feel resentful, upset or hurt, annoyed, or even insulted. Speech may be offensive because it contains personal insults or degrades others. It also may contain terms with a current or historical meaning relating to a particular gender, race, sexual orientation, or other person or groups (Sai and Sharma 2020).

The categories of *harassment*, *threat*, *cyberbullying*, and *toxicity* (*toxic*) do not only engage language – they are manifested in explicit and intentional acts of *behavioural* and *linguistic types*, with strong elements of aggression, of a consistent character in the case of *harassment* and *cyberbullying*, and of the varying frequency nature as far as *threat* is concerned. Both *harassment* and *threat* on the other hand can have an explicit or implicit character. It is only *cyberbullying* though which is solely constrained to digital communication, while *threat*, *toxicity*, and *harassment* can be performed via both offline as well as online communication modes.

In the current study (Fig. 2), the attribution markings that encode aspects (the square, triangle, circle, square, and half-moon) are added to select terms in the taxonomy, yet it must be borne in mind that the aspects may also be assigned to other terms. The current markings show only the intuitive and primary (most salient) facets associated with particular terms. For example, in the context of *hate speech*, the most salient aspects seem to be *hostile*, *discredit* and *threat*, whereas the attributive feature of *vulgar* appears to frequently pattern with *obscene*, *profane*, and *slur*.

Definitions

The analysis of the corpus language materials and their contextual variants, referring to relevant approaches available in the literature, serves as our point of reference in proposing the following offensive language categorization structure with the definitional distinctions of the category semantic interpretations:

Taboos are dysphemistic words referring to sacred and religious concepts or sexuality, for example to bodily waste, sex organs, or the act of having sex that are used in a non-literal and non-religious meaning, which cause embarrassment and offence (Andersson and Trudgill 1990). Examples of taboos are: "God!", "Goodness!". "Shit!", "Fuck!" These examples, however, are exclamations that, unless directed towards an addressee in some expressions (e.g., "You are shit", "Go fuck yourself"), would not pass muster as offensive language in our analysis, which assumes attacking some addressee (whether individual or group). There are authors (McEnery 2004), who do not agree that taboos are only non-literal (e.g., "You are fucked") as those used in the literal meaning ("Let's fuck") may also instantiate taboos. In our taxonomy, we accept all vulgarisms to be a subcategory of taboo, at a default level, leaving aside in the present analysis the contexts in which taboo words and phrases are used in non-insulting, non-offensive contexts as in some more friendly, slang exchanges. Example from the annotated material: "He will literally be able to walk into the enemy team and do whatever he wants being tanky as fk giving his target no room to escape and doing massive damage" (**taboo**/vulgar).

Profane language involves highly offensive expressions wherein religious words are used non-literally (Ljung 2010). Example from the annotated material: "Leviathan motherfucker!!!" (**profane**/vulgar).

Vulgarity is based on the use of words and phrases denoting scatology, effluvia, death, or sexuality, in particular various, usually deviant sexual practices or bodily sexual organs used to cause embarrassment or offence (Ljung 2010; Allan and Burridge 2006). Technical, scientific or childish terms (copulate, penis, poo) are excluded from this group (Ljung 2010). Example from the annotated material: "Always fucking bullying women with their disgusting stereotypes and slurs" (insulting/racist/**vulgar**/hostile).

Insults refer to language whose aim is to taunt or ridicule the addressee in order to cause emotional reactions. Such acts are typically expressive and not truth-conditional, i.e., not based on the relationship between language and reality but on the way language is used to achieve special and communicative goals, and they may be intentional or unintentional (Jucker 2000). Insults involve pejorative terms yet they do not primarily aim at oppressing the target, which is typical of slurs, so their denigrating force is weaker, i.e., "they are just mean and nasty" (Cousens 2020). In our taxonomy, they are thus treated as weaker forms of offence, as opposed to abusive language that is its stronger version. Moreover, insults focus on personal characteristics or behaviour (Jeshion 2013), unlike slurs, which are built around group membership (Cepollaro 2015). Example from the annotated material: "to everyone commenting said rude shit is a type of bullying im antibullying and if you dont like what she does unfollow"(**insulting**/vulgar/hostile).

Cyberbullying is an aggressive, violent, hostile, hurtful, and, importantly, repetitive verbal and/or nonverbal behaviour, often anonymous, private (e.g., through mails) or public (e.g., through social media), creating a sense of fear involving inter alia online harassment and denigration, leading to strong negative emotional responses of the addressee, such as nervousness, anxiety, depression, or terror (Alhujailli et al. 2020). In other words, it entails the use of technology, it is repeated, and it is a deliberative act of threatening and hurting, aimed at either an individual or a group (Ramirez et al. 2010). In our taxonomy, cyberbullying is considered in the major part of its manifestation to involve language, although the contextual factors mentioned above play a criterial role in its definition.

Hate speech (hateful) is the act of humiliating individuals or groups by resorting, usually explicitly, to offensive expressions addressed at communities that are weaker or inferior in terms of nationality, gender, sexuality, ethnicity, among other cultural/social properties (Lewandowska-Tomaszczyk 2020). Hate speech may take the form of the so-called hard hate speech, whereby prosecutable measures are involved or soft hate speech, which is legal but incites to discrimination and intolerance (Baider and Kopytowska 2018). Thus, hate speech, i.e., discourses of the discriminatory nature, resonate with hate crime, i.e., illegal, punishable acts. Hate speech is also considered by many scholars to operate at the illocutionary level, i.e., it conveys an intended message (Fraser 1998, Assimakopoulos 2020), as e.g., "Isn't she clever?" can have a locution of a question but be interpreted as a praise at the illocutionary level. We adopt this position as some of its contextual and expressive properties are not always immediately observed in less extensive text excerpts available to the analysis. The discriminatory feature in our taxonomy is that hate speech may be truth-conditional or not, whilst slurs are mostly expressive and evaluative, rather than typically non-truth-conditional. Example from the annotated material: "bullying asians is good, always justified" (**hate speech**/racist). The example contains a racist opinion against Asians, unsupported by evidence, based on hateful and xenophobic stereotypes.

Slurs are expressions that carry negative attitudes by attacking group addressees on the basis of demographics, such as nationality, race, religion, sexual orientation, etc., and punishable slanders, which, unlike slurs, are mostly truth-conditional. Slurs are counted as types of hate speech (see Hornsby 2001). Jucker (2000), in turn, maintains that the differences in meaning between slurs and insults are difficult to grasp and depend largely on the reaction of the target (the perlocutionary effect). Allan (2015) claims that slurs are deliberate and assume a perlocutionary effect, i.e., a (verbal) reaction from the addressee. In our taxonomy, following Hornsby (2001) and Croom (2011), slurs are viewed as subtypes of hate speech, yet, unlike Hornby, we do not take the essentialist view typical of the semantic approaches, but, rather, gravitate towards a pragmatic-cognitive approach, in which prototypical slurs can be distinguished as well as context-specific ones, subject to a more varied interpretation. Example from the annotated material: “hol up nigga who da fuck wuz u mufugging bullying nigga” (**slur**/hostile/vulgar).

Abuse is associated, apart from the basically verbal spoken offence, with the use of oral gesture or written language directed to a victim. It can include the act of harassing, labelling, insulting, scolding, rebuking, excessive yelling towards an individual or a group. It can also include the use of derogatory terms, the delivery of statements intended to frighten, humiliate, denigrate, or belittle a person (Koller and Darida 2020). Example from the annotated material: “I hate my fucking life and I fucking hate the people who are belligerently bullying me because now I have no friends thanks” (**abusive**/vulgar).

Obscene types are taken to refer to any utterance or act that strongly offends the prevalent morality of the time. Even non-image obscenity can be illegal for production, and its distribution has been judged by indecency laws applying both to real life and to the Internet. The interpretation of the meaning of obscenities is extensively discussed in legal domain in Hudson (2012). Example from the annotated material: “You have several times now said I should be blocked for calling somebody a cunt” (**obscene**/vulgar).

Homophobic language comprises many different forms, though typically it aims to refer to an individual’s sexual orientation and it can be used without harmful intent.
While some language is clearly homophobic, in some other cases it can be more nuanced and difficult to identify. The most common forms of homophobic language may include such expressions as “that’s so gay” and “you’re so gay”. Such comments may be directed towards people who are perceived to be gay on the one hand, while on the other such phrases are used to pass judgements that something is bad or nonsense – without the sense related to sexual orientation, e.g., the sentence “those trainers are so gay”, carrying the reference to “uncool” trainers. When homophobic language refers to abuse, particularly directed at gay people, they may include words like “fag”, “faggot”, “dyke”. Homophobic language used intentionally is a subdomain of abusive and offensive categories, typically addressed at people who are thought to be gay or simply different in some way from others (Adams et al. 2007). Example from the annotated material: “queue the faggot getting fagged out bullys bucker fags anti bullying at your service stop hitting yourself ya little bitch mods that s from south park don t you dare ban me for due to your uneducated ways” (hate speech/**homophobic**/discredit).

Threat in legal domain is a statement intended to frighten or intimidate a person or a group into believing in prospective harm they will experience. Real threats comprise such categories of speech as obscenity, child pornography, fighting words, and the intimidation of imminent lawless action (O’Neill 2001). Threat as a category is not directly a part of offence, although it can be used as a verbal element accompanying cyberbullying, hate speech, and its subclasses, homophobic, misogynic, and others of this character. Example from the annotated material: “bullying is good if justified” (offensive/hateful/hostile/**threat**).

Harassment is defined also in legal domain as a punishable act. Such speech is defined as *severe or pervasive* enough to create a *hostile or abusive work environment* based on race, religion, sex, nationality or citizenship, age, disability, military status. Also, in certain jurisdictions, harassment includes sexual orientation, marital status, political affiliation, criminal record, prior psychiatric treatment, occupation, citizenship status, personal appearance (Adams et al. 2007). Harassment, similarly to threat, is not directly associated with offence, although it is directly related to cyberbullying as one of its possible superordinate categories. Example from the annotated material: "If you delete it again you will suffer the consequences. This is your last warning" (**harassment**/threat/hostile).

Racist comments are considered a subcategory of hate speech. There are suggestions to treat comments as racist where speakers intend to humiliate or incite hatred against a class of people on the basis of race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin (Leets 2001). Example from the annotated material: "nothing hurts me worse than when I think be abt that nignog!" (insulting/**racist**/hateful/vulgar).

Hostile language is viewed as a subcategory of hate speech by characterizing hostile as antagonistic, opposed and unfriendly (Baider and Kopytowska 2018). Example from the annotated material: "You forgot to add someone with higher league bullying out that guy but i guess its a fine adaptation" (insulting/**hostile**).

Discrediting is a keyword related to hate speech and to other types of stronger offence and is defined as acts aiming to deprive of good reputation or cause disbelief in a person or a group (Erjavec and Kovacic 2012). Discrediting can also be performed as an illocutionary act, which will be considered a part of implicit offence, not discussed in detail in the present paper. Example from the annotated material: "Those wimps are the reason why we're losing more and more rights by the day" (insulting/racist/hostile/**discredit**).

Toxic is a general category and comprises speech acts, discursive practices, and behaviour which may inflict harm. It may include harms arising from speech devoid of slurs and epithets, which will be considered as implicit offence on the one hand, while it may contain deeply derogatory terms on the other (Tirrell 2018), directly linked to the category of offence. However, as toxic is primarily offence experiencer-oriented and requires the analysis of larger contexts, it is typically used as a tag with reference to larger stretches of texts. Example from the annotated material: "You are pathetic, screw you!*"* (**toxic**/discredit).

The classification proposed in the present study does not represent a uni-dimensional schema as can be concluded from the above definitions. It is based on a number of criteria and subcategories in terms of several (sub-)levels. The subcategories were built on the above definitions and are further validated by the Word2Vec heatmap that shows similarities between pairs of the terms at hand calculated on the basis of cosine distance (for details see Figure 3).

4 Implicit offensiveness

Implicit offensive language has so far received little scholarly attention. The current study focuses on proposing an overarching model where varying subtypes of implicitness used in the context of offensive language are conceptually linked relying on implicit (mostly figurative) offensiveness (e.g., irony, metaphor, hyperbole, etc.). Based on the analysis of existing typologies and definitions, a new model of implicit offence has been proposed. In the model, offensiveness is anchored in current approaches to offensive, as well as impolite language (Culpeper 2011, 2021; Haugh and Sinkeviciute 2019).

Whilst implicitness is a term that causes definitional problems, the common denominator for various approaches to this notion may be a description provided by Spisà (2007: 3), namely that it focuses on tacit communication and relies on meaning that is presupposed and/or implied. Put differently, implicit communication is one wherein meaning is not encoded directly; it is expected to be decoded by the addressee based on some pragmatic clues and by referring to background knowledge as well as contextual information. In other words, implicit meaning is actualized through contextual enrichment, detailing the reference and addressee's interpretation. Meaning is often vague or ambiguous as a result of resorting by the sender to either uncertainty and/or polysemy, often built on figurativeness or idiomaticity. Consequently, the addressee must put some cognitive effort in order to retrieve the intended meaning.

In our model of implicitness (Bączkowska 2022; Bączkowska et al. 2022), which derived mainly from Gricean and neo-Gricean approaches (1968) to implicit language, we distinguish four core categories. Following Grice (1989: 34), the main notions we use to describe implicitness are ones which flout the maxim of Quality, and they encompass: irony, metaphor, overstatement (Gricean hyperbole), and understatement (meiosis in Grice's parlance). **Irony** is understood here as expressing some derogatory evaluation by resorting to its contrary. This is the most common subtype of irony, yet all other, less typical forms of irony (such as surreal irony or verisimilar irony) are also encompassed by this category, including the concept of sarcasm. Thus, by saying "You are a genius" one would mean the opposite, that is "you are not a genius" or "you are stupid". Examples from the annotated material: "celebs simply can't handle bullying and endorse censorship so they can enjoy a safe space online. (offensive/hostile/implicit/irony); "Make bullying great again!" (offensive/hostile/implicit/sarcasm).

As for the term **metaphor,** we subscribe to the definition of conceptual metaphors offered by Cognitive Linguistics (Lakoff and Johnson, 1980) wherein one thing is seen in terms of another one (as in "argument is war" where the act of quarrelling is conceptualised in terms of military measures). Any intermediate forms between the metaphor and metonymy, that is **metaphtonymy** (as proposed by Goossens 1990) are subsumed by the overarching term *metaphor* in our model.

Example from the annotated material: "First... dude come back from hibernation". (offensive/hostile/implicit/metaphor). Some metaphors are moribund (even though relatively rare) or dead, and they do not trigger implicit inferences; therefore, dead metaphors will not be subsumed by implicitness in our model as they were included in the explicitness model. Similarly to metaphors, a comparison is also made in the case of **similes**, which are based on a structure containing "as" or "like". Similes are implicit figures of speech which are more transparent in meaning than metaphors as the comparison is straightforwardly stated, and thus less cognitive effort is needed to recover the speaker-intended meaning (e.g., "You are like Einstein"). Example from the annotated material: "Her win ratios were like a perfect 50% yea skilled players can" (insulting/discredit/implicit/simile). We could thus say that similes are more explicit than metaphors, even though they are still classified as implicit forms of offensiveness.

We use the term **overstatement** rather than a hyperbole as, while the former clearly indicates resorting to evaluations exceeding the state, object, or event observed in the reality, the latter may be understood commonly as an exaggeration in either direction, i.e., as ascribing too much or too little value to the object described. Devaluation is reserved in our model for the notion labelled as **understatement,** and it correlates positively with the Gricean concept of meiosis. The examples below illustrate overstatement and understatement as proposed here:

(1) I'd rather die than merry you wacko. (overstatement)
(2) We have a slight problem since you are a bit sloshed (said to somebody completely drunk)

Example from the annotated material: "And his cosmetic surgery will be done on the government dime no doubt" (offensive/hostile/implicit/understatement).
Example from the annotated material: "They should only impose this on China!" (offensive/racist/implicit/overstatement). In example (1) the implicit offence expressed through an overstatement contains explicit offensiveness encoded at lexical level by a form of address ("you wacko").

To these notions, we also added the Searlian concept of **indirectness** as an element staying somewhat midway between the four types of implicitness derived from Grice mentioned above and the subtypes of explicit forms of offensiveness elaborated in Section 3.3. To be more specific, indirectness is treated in our model as a subtype of implicitness which, however, shares some features of explicit language inasmuch as it does not encode meaning in a straightforward way (thus bearing similarity to implicitness) yet it eschews figurativeness that is so typical of implicitness (and thus bearing similarity to explicitness). The oft-quoted example provided by Searle (1969), "Can you pass the salt", being illustrative of signalling a request through an interrogative, epitomises our claim that, without resorting to figurativeness, one can still be non-direct in expressing meaning. This indirectness involves syntactic transformation; in the example above it is an interrogative (using the structure of inversion) used in the function of a request (rather than a question), without entailing any conceptual transformations typical of irony, metaphor, or over- and understatement, wherein meaning is elaborated on at both the literal and non-literal levels. Indirectness dwells on literalness, which is structurally encoded in a non-default manner, contrary to figurative implicitness that houses re-conceptualisations of non-literal meaning. Indirectness and figurative implicitness are thus reconciled as both instantiate non-explicitness in the more general sense and require cooperative reasoning on the part of the addressee. Put differently, they both rely on covert meaning yet to a different degree (Bączkowska 2022). Example from the annotated material: "Because Cassiopia ****s on half those examples you mentioned if she is played properly" (Hint: She is incredibly strong versus melee) (offensive/ vulgar/implicit).

An interesting feature of implicit messages is that they assume the possibility of being cancelled or denied. For example, following some clues, the addressee may come to conclusions that the sender has the intention to offend the addressee, yet once such an unfavourable interpretation becomes obvious to the sender, s/he may easily withdraw from this line of reasoning and deny his/her bad intentions as implicit message allows some degree of doubt. Cancellability has a risk-mitigation function and it is intrinsic to implicit language; it also contributes to vagueness or ambiguity mentioned above.

It is not difficult to notice that the example of irony "You are a genius" or, "You are beautiful like Mona Lisa" may be easily classified as an exaggeration (overstatement) and/or an ironic simile, which may at first glance introduce confusion and apparent definitional inconsistency. However, as observed by a number of scholars (e.g., Partington 2006, Popa 2010, Bączkowska 2022), these terms may also occur as conflated forms, and thus one may deal with ironic metaphor, hyperbolic irony, or others. Our model allows for such options.

Implicit language does not preclude the use of explicit offensive language. In other words, explicit lexemes may be enmeshed in utterances even if the overall meaning of the utterance is implicit. In (3), the sentence is obviously offensive if meant to be ironic, even though no explicit offensive language was employed, yet in (4) a swearword ("fucking") is juxtaposed with a label suggesting complimenting the addressee ("genius"), yet despite the apparent compliment the whole sentence may be read as offensive and implicit (ironic). Swearwords are often employed to strengthen pragmatic effects, such as ironic interpretations or humorous overtones, especially when used as premodifiers in the function of boosters.

(3) You're a genius. (used in an ironic way)
(4) You're a fucking genius. (used in an ironic way)

From the discussion above, it obviously transpires that the terms explicitness and implicitness cannot be treated as two mutually exclusive notions, as they may linearly co-occur (as in the ironic "fucking genius") and may show degrees of (non-) explicitness (as in frequent idioms or dead metaphors). An expression containing an explicit offensive word may only be a constitutive element of a larger unit of meaning that can be interpreted otherwise, i.e., as a non-offensive and/or implicit form of expression. On the other hand, what is formally treated as figurative/idiomatic may lose it semantic obscurity as a result of, for example, high frequency of use. This transition area links our model of explicit offensiveness with our model of implicitness.

5 Computational linguistic exploratory analysis

In this section, we employ the existing computational linguistic tools to examine the offensive language datasets and uncover clear differences/similarities among existing categories by creating our own word embeddings. At the stage of data import we perform some pre-processing steps, such as (1) whitespaces removal, (2) replacement of emojis with text, (3) specific texts removal (e.g., identification of questions, answers, retweets), (4) special symbols removal (e.g., &, $), (5) links removal, (6) html tags removal, and (7) change of text into lowercase.

5.1 Word embeddings analysis

In our analysis of the relationships between the categories of offensive language, we have adopted the methodology presented in the SALLD paper (Lewandowska-Tomaszczyk et al. 2021). We used different word embedding methods to represent the categories as vectors in a reduced dimension space and to calculate the distance between them. Word embeddings represent similarities between words in the form of numeric vectors, allowing similar words to have similar vector representations. Word embeddings can be learned either from pre-trained or non-pre-trained corpora. There are two fundamental differences between the analysis presented here and what has been done in the past. First, we have not used pre-trained vectors, but have learnt the word embedding from scratch. Second, the learning was done from a corpus of offensive language rather than a corpus that represents general language.

We have collected a large corpus consisting of 25 corpus datasets (Appendix 1.). Each of the datasets focus on different forms of offensive language. We believe that our corpus represents the researched categories better than a general corpus that may not fully cover all the categories.

We examined two types of word embedding methods, Word2Vec and fastText, the two techniques learning word embeddings from datasets. Both methods aim to create representations of words in the form of vectors. However, fastText, unlike Word2Vec, utilizes character n-grams, i.e., sequences of n items from a given sample of text instead of words to predict other words. It represents words as the sum of their character n-gram vectors. For both of the methods, we learned embeddings of 50 dimensions, running 5 epochs, taking into account a window of 5 words and minimum frequency of

2. We have also experimented with different parameters, but these gave the best results. We have also augmented input text with inclusion of offensive categories into text but it also did not improve the representations.

While there are a variety of metrics used to measure the similarity distance between two words, cosine similarity is more obvious and the most used of those measures in word embeddings. Cosine similarity is the normalized dot product of two vectors, i.e., the angle between them. The cosine similarity of two vectors whose orientations are the same is 1, 0 for vectors with a 90-degree angle, and -1 for two vectors that are diametrically opposite, regardless of their size. Separately, we performed the same analysis for explicit and implicit categories.

5.1.1 Explicit offensive analysis

First, we computed the pairwise cosine similarity between our categories. Figure 3 presents a heatmap of the cosine similarity scores. A heatmap is a color-coded visual representation of data, where the red colour region implies high similarity and the blue region implies low similarity.

The figure's blue rows (and columns) of similarity between categories discredit, threat, cyberbullying, and harassment, and others are very visible. Since words have a low cosine score when they do not share similar contexts, it may be inferred that these categories can be easily distinguished from the others, and they define separate categories of offensive language.

The most similar categories are homophobic and racist (0.87), and hateful is very close to both of them, with 0.85 and 0.83, respectively. These categories are probably the most prevalent ways in which hatred is expressed in the corpus. Profane and vulgar are also similar categories (0.83). Offensive, which was selected as the most general category, has a relatively high similarity with all the categories (>0.56) except for the four categories (discredit, threat, cyberbullying, and harassment) mentioned earlier.

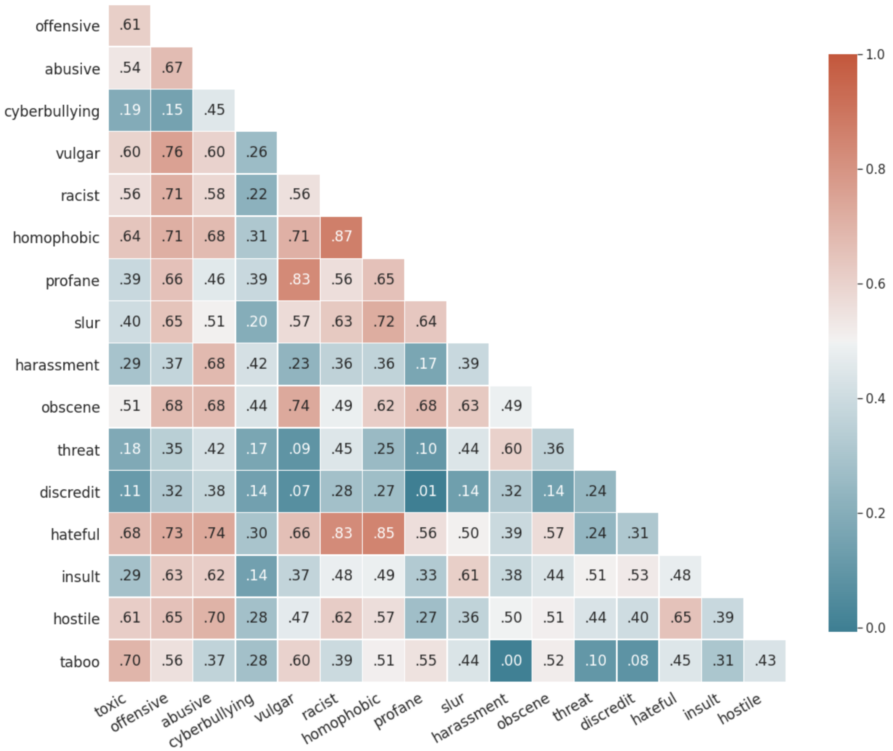


Figure 3: Word2Vec cosine similarity heatmap.

Next, we analysed the categories in their lemma forms. In linguistics and computational linguistics, a lemma refers to the base form or dictionary form of a word. For example, the lemma of "offensive" is "offence". Lemmatization is the process of reducing inflected forms of a word to their lemma, which can help with tasks such as language processing, text analysis, and machine translation.

For each of the categories, we extracted the top 30 most similar words, omitting words that the category, its lemma, or its stem are their substring. Then, we ran an t-SNE (t-distributed Stochastic Neighbor Embedding) method on the embeddings of the categories and their top 30 most similar words (with perplexity of 15). T-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm used for dimensionality reduction and data visualization. It maps high-dimensional data points into a lower-dimensional space (usually 2D or 3D) while preserving the pairwise distance between the data points as closely as possible. T-SNE works by identifying similar data points and placing them close together in the lower-dimensional space, while also ensuring that dissimilar points are far apart. This approach helps to reveal the underlying structure of the data that may be hidden in a high-dimensional space. T-SNE is particularly useful for visualizing complex and non-linear relationships among data points, which may be difficult to capture using other dimensionality reduction techniques. It has found applications in various fields, including biology, computer vision, and natural language processing.

T-SNE is advantageous because it further reduces the dimensionality of the vectors to two dimensions, allowing visualization of the vectors and their spatial arrangement or distribution. Figure 4 illustrates the t-SNE transformation of our embedding vectors from fifty to two dimensions.

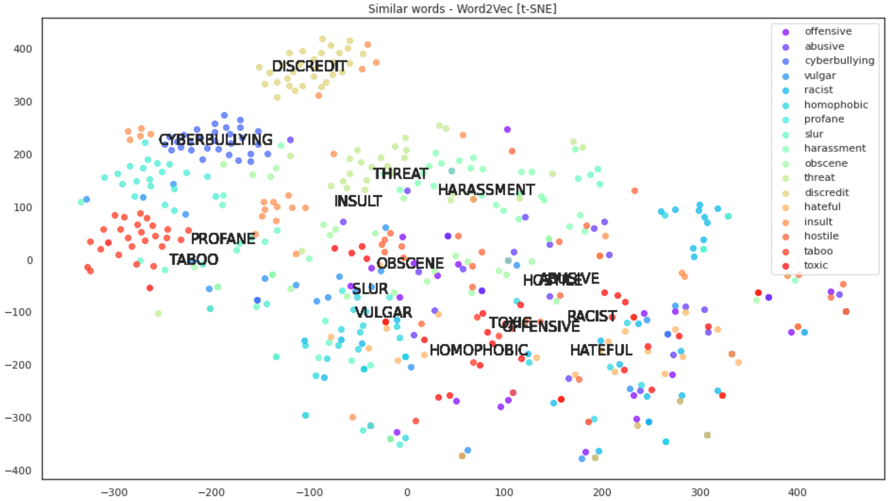
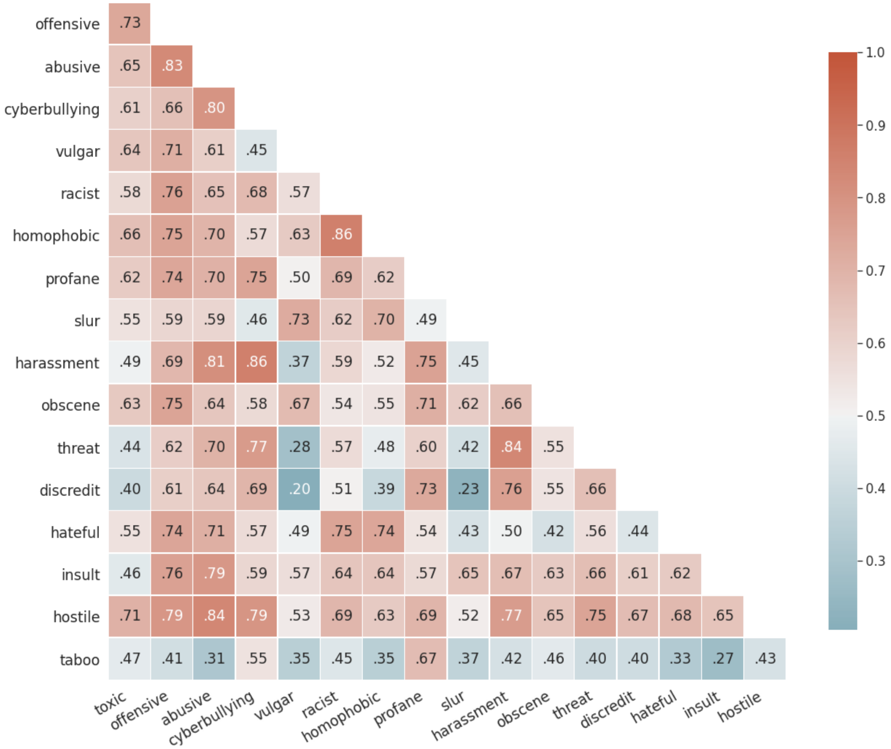


Figure 4: Word2Vec top 30 neighbouring vectors visualization using t-SNE.

We note that the distances between the categories in Figure 4 do not reflect the cosine similarity between their vectors. Whereas the cosine similarity considers all 50 dimensions of the vector, the t-SNE figure visualizes a reduced vector of two dimensions. Furthermore, the categories' labels are located at the mean point of the cluster. The mean is a measure that is significantly affected by the scatter and endpoints of the cluster.

The t-SNE figure shows that relatively well-delineated clusters are formed by the discredit, and cyberbullying. Threat-harassment-insult and profane-taboo form clustered categories on their own. However, the rest of the categories are spread widely and overlap.



	toxic	offensive	abusive	cyberbullying	vulgar	racist	homophobic	profane	slur	harassment	obscene	threat	discredit	hateful	insult	hostile
offensive	.73															
abusive	.65	.83														
cyberbullying	.61	.66	.80													
vulgar	.64	.71	.61	.45												
racist	.58	.76	.65	.68	.57											
homophobic	.66	.75	.70	.57	.63	.86										
profane	.62	.74	.70	.75	.50	.69	.62									
slur	.55	.59	.59	.46	.73	.62	.70	.49								
harassment	.49	.69	.81	.86	.37	.59	.52	.75	.45							
obscene	.63	.75	.64	.58	.67	.54	.55	.71	.62	.66						
threat	.44	.62	.70	.77	.28	.57	.48	.60	.42	.84	.55					
discredit	.40	.61	.64	.69	.20	.51	.39	.73	.23	.76	.55	.66				
hateful	.55	.74	.71	.57	.49	.75	.74	.54	.43	.50	.42	.56	.44			
insult	.46	.76	.79	.59	.57	.64	.64	.57	.65	.67	.63	.66	.61	.62		
hostile	.71	.79	.84	.79	.53	.69	.63	.69	.52	.77	.65	.75	.67	.68	.65	
taboo	.47	.41	.31	.55	.35	.45	.35	.67	.37	.42	.46	.40	.40	.33	.27	.43

Figure 5. fastText cosine similarity heatmap.

We performed the same analysis for the fastText embeddings. The heatmap of the fastText is shown in Figure 5. In general, the cosine similarity scores of the fastText are higher than the scores of the Word2Vec model. Most of the figure is painted in shades of red. The one and only category that can be easily distinguished from the others is taboo. Offensive, abusive, hostile, profane, racist, threat, and insult have a high similarity score with almost all the other categories.

The t-SNE figure of the fastText method is presented in Figure 6. The method's reduction of dimensions places an order in the data and better distinguishes between the categories. Relatively well-delineated clusters are formed by the discredit, cyberbullying, threat-harassment, profane, taboo, and toxic-vulgar categories. It is difficult to separate the hostile, insult, and the general offensive categories.

However, the distinctiveness is not the only factor deciding on the proposed taxonomy; there are also theoretical insights from the scientific literature analysed above, leading to establishing the proposed classification of the offensive language and identifying offensive language as an overarching taxonomical value which embraces other sub-categories of the taxonomy.

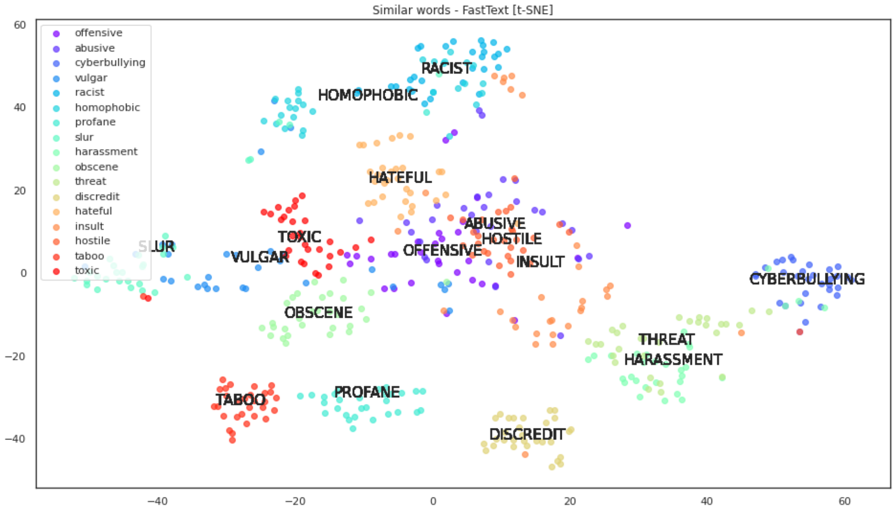


Figure 6: fastText top 30 neighbouring vectors visualization using t-SNE.

5.1.2 Implicit offensive analysis

First, the pairwise cosine similarity between our implicit categories was obtained. Figure 7 shows the cosine similarity score as a heatmap. To repeat, a heat map (or heatmap) is a data visualization technique that shows variation between categories. The figure's blue rows (and columns) of similarity between categories, such as sarcasm, irony, and others, are readily apparent. Since words with dissimilar contexts have a low cosine score, it may be deduced that these categories can be readily discriminated from the others and establish distinct categories of implicitly offensive language. Except for exaggeration, the simile is likewise extremely different from the rest.

The most similar categories are understatement and overstatement (0.79). The category of exaggeration is fairly similar in its meaning to both of them, with respective values of 0.79 and 0.76. The higher the values the more similar the indicated categories are.

Figure 8 illustrates the t-SNE transformation of our implicit embedding vectors from fifty to two dimensions. When compared to the intertwined nature of sarcasm and irony, simile and metaphor may be thought of as separate concepts. Metaphor, irony, and exaggeration all intersect to some extent. It seems that exaggeration encompasses both overstatement and understatement (respectively, in the sense of downplaying or playing up features).

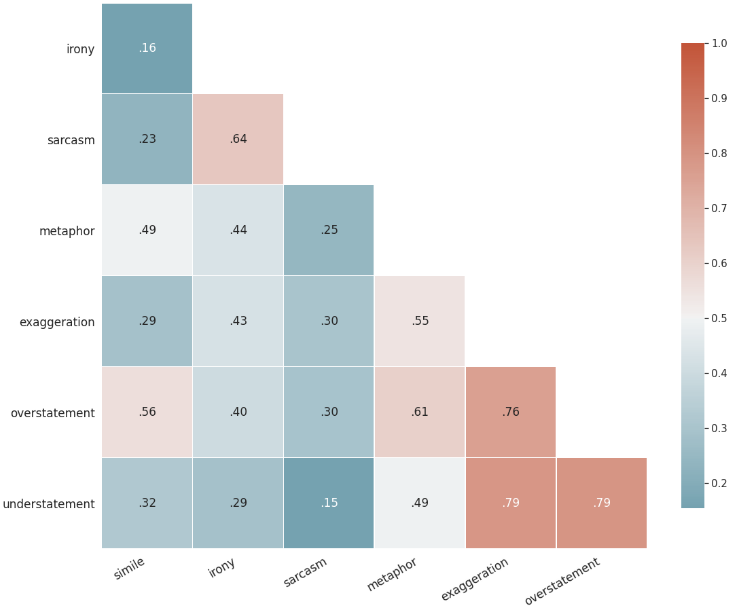


Figure 7: Word2Vec cosine similarity heatmap (implicit).

We performed the same analysis for the fastText embeddings. The heatmap of the fastText is shown in Figure 9. Understatement and overstatement are, as with the Word2Vec heatmap, the most similar categories (0.96). Exaggeration is somewhat close to both, with respective values of 0.83 and 0.87. However, there is far less blue area, which makes distinguishing between other categories difficult.

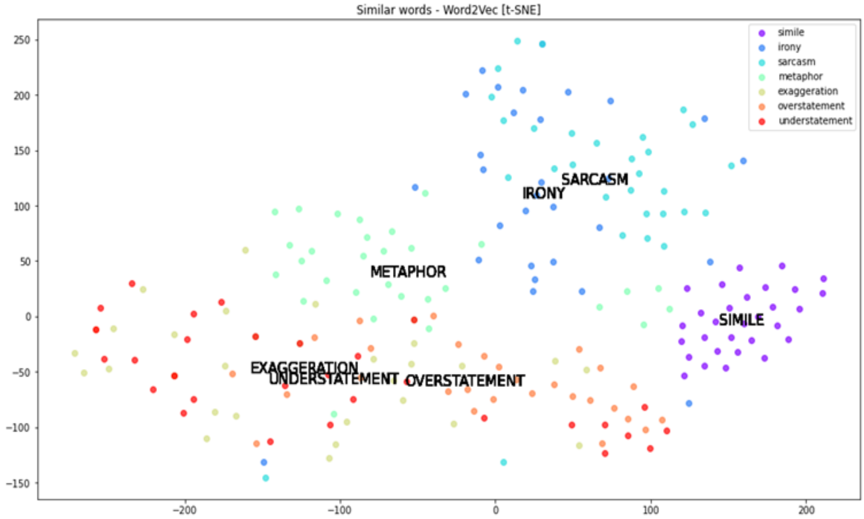


Figure 8: Word2Vec top 30 neighbouring vectors visualization using t-SNE (implicit).

Figure 10 shows the t-SNE graph produced by the fastText method. The method's reduction of dimensions organizes the data and improves the distinction between categories. The categories of simile, metaphor, sarcasm, and irony form clusters that are quite well-defined. The figure also demonstrates that exaggeration includes both overstatement and understatement (respectively, in the sense of downplaying or playing up features).

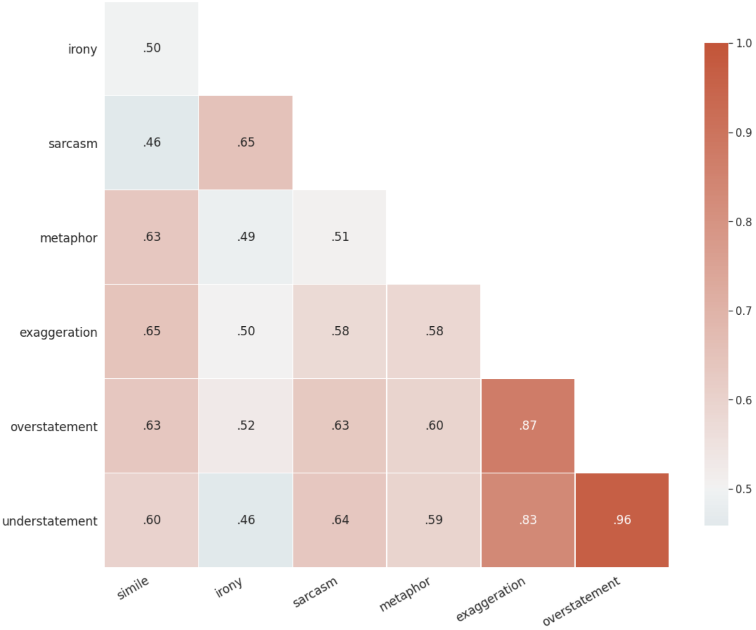


Figure 9: fastText cosine similarity heatmap (implicit).

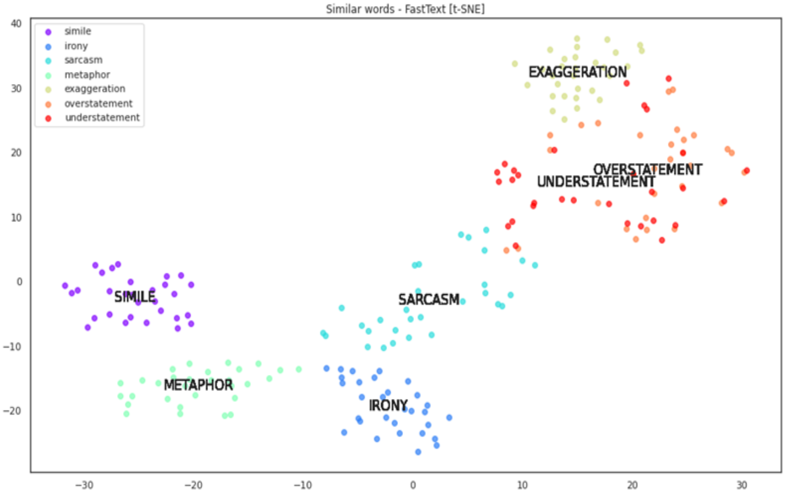


Figure 10. FastText top 30 neighbouring vectors visualization using t-SNE (implicit).

5.2 Results and discussion

We need to point out that the categories in this analysis do not exactly align with the ones presented by our taxonomy (Figure 2). The categories in this analysis are inferred directly from existing offensive language corpora. Therefore, some of the categories are not part of our taxonomy and some of the categories are missing as they were not covered by any dataset (e.g., *slur*).

It is also important to discuss the threats to the validity of the results. Apart from these decisions, the texts and sources from the datasets are different (e.g., Twitter posts, forum posts, news comments) which can highly influence the results. Still, taking these limitations into consideration, we believe the results may show us some rough relationships among offensive language categories to some extent. Note that most of these datasets did not declare their annotation criteria or justifications of specific categories. Therefore, comparison in this analysis can be based only on the *as-is* relation.

The heatmaps (Figure 5 and Figure 3) demonstrate that the taxonomy proposed in this paper fares better as the similarity is generally significantly higher for the majority of the categories. The results show differences for some of the categories as they form clearly separated clusters.

Some of the overlapping categories may be semantically similar (or vaguely defined by different datasets) for which we defined clear linguistic descriptions in order to build a comprehensive offensive language categorization. The heatmap supports a number of our specific claims, for example that *slur* is a strong offence as it displays a high correlation with *abusive*, *vulgar*, *racist*, *homophobic*, and *profane*. Therefore, *slur* is located at the bottom of the diagram, and *hate speech* is a more general term in our taxonomy than *slur* (hence it is above *slur* in the diagram).

Similarly, *cyberbullying* and *harassment* show a high correlation (0.64); thus, they are connected in the diagram. *Cyberbullying* is more specific than *harassment* (*cyberbullying* is one of possible types of bullying), hence it is subordinate to *harassment*.

The particular character and position of *threat*, *harassment*, and *cyberbullying* in the offensive language ontology, which are accounted for in the present study, have not been explored in such detail in previously available literature. The models explored by other researchers of *severe toxic*, *harassment*, *hate speech*, *profane*, *vulgar*, and *abusive* are the models proposed for the identification of certain domains of offensive language, which might make the recognition of the categories within the models, although they are limited to the most salient cases and demonstrates varying specificity. Thus, our proposed taxonomy is aimed at a broader account and at a more inclusive schema for offensive language identification.

6 Conclusions

The study offers a linguistic model of explicit and implicit offensive language categories by resorting to empirical corpus-based methods and a review of relevant literature and verified by means of computational methodology.

Following a critical analysis of over 60 offensive language datasets, 25 English datasets were selected and validated by means of exploratory computational analyses. The results achieved by the computational modelling support the linguistic proposal, although a certain specificity of the employed datasets, and some of the limitations of the used annotation systems, have to be considered. The relevance of the current contribution in regard to the presented taxonomic insights might appear to be crucial for further advancements in offensive language detection and a more adequate category identification. Nevertheless, and this can be considered a limitation of this study, we do not expect a full agreement (i.a., inter-annotator agreement) in the case such a taxonomy is applied to the annotation of large offensive language data sets in English or in other languages. As shown in the values acquired from embeddings and heatmaps, the semantic character of the linguistic category of offence does not conform, similarly to a number of other natural language categories (Lakoff 1987a), to the exceptionless type of hierarchically structured language taxonomies. Neither can all the instances be captured by a system of necessary and sufficient conditions in the definitions of categories and sub-categories of language, although attempts at such definitions have been presented here. The identification and interpretation issues that might arise in connection with the integrated offensive language taxonomic system discussed in this paper will be further tested both during the annotation campaign planned to be carried out and, if needed, in the further analytic steps regarding aspects of the taxonomic categorization.

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.*

Appendix 1. English datasets used in the present study

Types:
Level A (offensive vs. non-offensive)
Level B Offensive (subtypes)
Level C (implicit vs. explicit)
Level D (morphosyntactic features)
Size indicates the number of posts in a dataset

Project	**Source**	**Size**	**Tags**	**Reference**	**Type**
Automated Hate Speech Detection and the Problem of Offensive Language	Twitter	24 802	Hierarchy (Hate, Offensive, Neither)	Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. ArXiv,.	A, B
Hate Speech Dataset from a White Supremacy Forum	Stormfront (Forum)	9 916	Ternary (Hate, Relation, Not)	de Gibert, O., Perez, N.,García-Pablos, A., and Cuadros, M., 2018. Hate Speech Dataset from a White Supremacy Forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, pp.11-20.	A
Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter	Twitter	16 914	3-topic (Sexist, Racist, Not)	Waseem, Z. and Horvy, D., 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: Proceedings of the NAACL Student Research Workshop. San Diego, California: Association for Computational Linguistics, pp. 88-93.	A
Detecting Online Hate Speech Using Context Aware Models	FoxNews, posts	1 528	Binary (Hate / Not)	Gao, L. and Huang, R., 2018. Detecting Online Hate Speech Using Context Aware Models. ArXiv.	A
Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter	Twitter	4 033	Multi-topic (Sexist, Racist, Neither, Both)	Waseem, Z., 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In: Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science. Copenhagen, Denmark: Association for Computational Linguistics, pp. 138-142.	A
When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data	Twitter	712	Hierarchy of Sexism (Benevolent sexism, Hostile sexism, None)	Jha, A. and Mamidi, R., 2017. When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data. In: Proceedings of the Second Workshop on Natural Language Processing and	A

				Computational Social Science. Vancouver, Canada: Association for Computational Linguistics, pp. 7-16.	
Overview of the Task on Automatic Misogyny Identification at IberEval 2018	Twitter	3 977	Binary (misogyny / not), 5 categories (stereotype, dominance, derailing, sexual harassment, discredit), target of misogyny (active or passive)	Fersini, E., Rosso, P. and Anzovino, M., 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).	A
CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech	Synthetic / Facebook posts	1 288	Binary (Islamophobic / not), multi-topic (Culture, Economics, Crimes, Rapism, Terrorism, Women Oppression, History, Other/generic)	Chung, Y., Kuzmenko, E., Tekiroglu, S. and Guerini, M., 2019. CONAN - COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, pp. 2819-2829.	A
Characterizing and Detecting Hateful Users on Twitter	Twitter	4 972	Binary (Hateful/Not)	Ribeiro, M., Calais, P., Santos, Y., Almeida, V. and Meira, W., 2018. Characterizing and Detecting Hateful Users on Twitter. ArXiv	A
A Benchmark Dataset for Learning to Intervene in Online Hate Speech	Platform Gab, posts	33 776	Binary (Hateful/Not)	Qian, J., Bethke, A., Belding, E. and Yang Wang, W., 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. ArXiv	A
A Benchmark Dataset for Learning to Intervene in Online Hate Speech	Reddit	22 324	Binary (Hateful/Not)	Qian, J., Bethke, A., Belding, E. and Yang Wang, W., 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. ArXiv	A
Multilingual and Multi-Aspect Hate Speech Analysis	Twitter	5 647	Hostility, Directness, Target attribute and Target group	Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. and Yeung, D., 2019. Multilingual and Multi-Aspect Hate	A, B, C

				Speech Analysis. ArXiv	
Exploring Hate Speech Detection in Multimodal Publications	Twitter	149 823	Six primary categories (No attacks to any community, Racist, Sexist, Homophobic, Religion based attack, Attack to other community)	Gomez, R., Gibert, J., Gomez, L. and Karatzas, D., 2019. Exploring Hate Speech Detection in Multimodal Publications. ArXiv	A
Predicting the Type and Target of Offensive Posts in Social Media	Twitter	14 100	Branching structure of tasks: Binary (Offensive, Not), Within Offensive (Target, Not), Within Target (Individual, Group, Other)	Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R., 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). ArXiv,.	A, C
hatEval, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter	Twitter	13 000	Branching structure of tasks: Binary (Hate, Not), Within Hate (Group, Individual), Within Hate (Agressive, Not)	Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F., Rosso, P. and Sanguinetti, M., 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. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 54-63.	A, B, C
Peer to Peer Hate: Hate Speech Instigators and Their Targets	Twitter	27 330	Binary (Hate/Not), only for tweets which have both a Hate Instigator and Hate Target	ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G. and Belding, E., 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018). Santa Barbara, California: University of California, pp. 52-61.	A
Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages	Twitter and Facebook	7 005	Branching structure of tasks. A: Hate / Offensive or Neither, B: Hate Speech, Offensive, or Profane, C: Targeted or Untargeted	Modha, S., Mandl T., Majumder, P., Patel, D. 2019. Overview of the HASOC track at FIRE 2019. In: Proceedings of the 11th Forum for Information Retrieval Evaluation	A, B
Large Scale Crowdsourcing	Twitter	80 000	Multi-thematic (Abusive, Hateful, Normal, Spam)	Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I.,	A, B

and Characterization of Twitter Abusive Behavior				Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N., 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. ArXiv	
A Large Labeled Corpus for Online Harassment Research	Twitter	35 000	Binary (Harassment, Not)	Golbeck, J., Ashktorab, Z., Banjo, R., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A., Gergory, Q., Gnanasekaran, R., Gnanasekaran, R., Hoffman, K., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M., Naik, S., Nixon, H., Ramachandran, P., Rogers, K., Rogers, L., Sarin, M., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z. and Wu, D., 2017. A Large Labeled Corpus for Online Harassment Research. In: Proceedings of the 2017 ACM on Web Science Conference. New York: Association for Computing Machinery, pp. 229-233.	A
Ex Machina: Personal Attacks Seen at Scale, Personal attacks	Wikipedia posts	115 737	Binary (Personal attack, Not)	Wulczyn, E., Thain, N. and Dixon, L., 2017. Ex Machina: Personal Attacks Seen at Scale. ArXiv	A
Ex Machina: Personal Attacks Seen at Scale, Toxicity	Wikipedia posts	100 000	Toxicity/healthiness judgement (very toxic, neutral, very healthy)	Wulczyn, E., Thain, N. and Dixon, L., 2017. Ex Machina: Personal Attacks Seen at Scale. ArXiv	A
Detecting cyberbullying in online communities	World of Warcraft, posts	16 975	Binary (Harassment, Not)	Bretschneider, U. and Peters, R., 2016. Detecting Cyberbullying in Online Communities. Research Papers, 61.	A
Detecting cyberbullying in online communities	League of Legends, posts	17 354	Binary (Harassment, Not)	Bretschneider, U. and Peters, R., 2016. Detecting Cyberbullying in Online Communities. Research Papers, 61.	A
A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research	Twitter	24 189	Multi-topic harassment detection	Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. and Sheth, A., 2018. A Quality Type-aware Annotated Corpus and Lexicon for Harassment	A

				Research. ArXiv	
Ex Machina: Personal Attacks Seen at Scale, Aggression and Friendliness	Wikipedia posts	160 000	Aggression/friendliness judgement on a 5-point scale. (very aggressive, neutral, very friendly)	Wulczyn, E., Thain, N. and Dixon, L., 2017. Ex Machina: Personal Attacks Seen at Scale. ArXiv	A, B
OffensEval 2019, OffensEval 2020	Twitter, except for Danish: Facebook, Reddit, and comments in a local newspaper, Ekstra Bladet	over nine million, 10 000, 3 600, 10 287, 35 000	Three-level hierarchy: • Level A - Offensive Language Detection – NOT: content that is neither offensive, nor profane; – OFF: content containing inappropriate language, insults, or threats. • Level B - Categorization of Offensive Language – TIN: targeted insult or threat towards a group or an individual; – UNT: text containing untargeted profanity or swearing. • Level C - Offensive Language Target Identification – IND: the target is an individual explicitly or implicitly mentioned in the conversation; – GRP: hate speech, targeting group of people based on ethnicity, gender, sexual orientation, religious belief, or other common characteristic; – OTH: targets that do not fall into any of the previous categories, e.g., organizations, events, and issues.	Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ... & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). arXiv preprint arXiv:2006.07235.	A, C, D
Illegal is not a Noun: Linguistic Form for Detection of Pejorative Nominalizations	Twitter, Reddit, news articles and interviews, political debates, and video and written blogs	56 237	Four target adjectives: Illegal, Female, Gay and Poor, two categories: linguistic form and pejorative meaning	Palmer, A., Robinson, M., & Phillips, K. K. (2017, August). Illegal is not a noun: Linguistic form for detection of pejorative nominalizations. In Proceedings of the First Workshop on Abusive Language Online (pp. 91-100).	D
Detecting and Monitoring Hate Speech in Twitter	Twitter	6 000	Binary	Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019).	A

				Detecting and monitoring hate speech in Twitter. Sensors, 19(21), 4654.	
HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection	Twitter and Gap	9 055, 11 093	3-class classification (i.e., hate, offensive or normal), the target community (i.e., the community that has been the victim of hate speech/offensive speech in the post), and the rationales, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal) is based.	Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289.*	A, B, C
Automatic detection of cyberbullying in social media text	social networking site ASKfm	113 69 8, 78 387	four roles are distinguished in the annotation scheme, including victim, bully, and two types of bystanders, a number of textual categories that are often inherent to a cyberbullying event, such as threats, insults, defensive statements from a victim, encouragements to the harasser, etc.	Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, *13*(10), e0203794.	A, B, C

References

Adams, Maurianne, Lee Anne Bell & Pat Griffin. 2007. *Teaching for diversity and social justice*. London: Routledge/Taylor & Francis Group.

Alhujailli, Ashraf, Waldemar Karwowski, Thomas Wan & Peter Hancock. 2020. Affective and stress consequences of cyberbullying. *Symmetry* 12.9. 1536.

Allan, Keith. 2015. When is a slur not a slur? the use of nigger in 'pulp fiction'. *Language Sciences* 52. 187–199.

Allan, Keith & Kate Burridge. 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press.

Anderson, Luvell & Ernie Lepore. 2013. A brief essay on slurs. Alessandro Capone, Franco Lo Piparo & Marco Carapezza (eds.), *Perspectives on Pragmatics and Philosophy*, 507–514. Cham: Springer.

Andersson, Lars-Gunnar & Peter Trudgill. 1990. *Bad Language*. London: Penguin Books Ltd.

Austin, James. 1962. *How to do things with words.* Oxford: Oxford University Press.

Bach, Kent. 1994. Conversational implicature. *Mind and Language* 9. 124–162.

Bach, Kent & Robert Harnish. 1979. *Linguistic Communication and Speech Acts.* Cambridge, MA: MIT Press.

Baider, Fabienne & Monika Kopytowska. 2018. Narrating hostility, challenging hostile narratives. *Lodz Papers in Pragmatics* 14.1–24.

Bączkowska, Anna. 2022. Explicit and implicit offensiveness in dialogical film discourse in *Bridgit Jones* films. *International Review of Pragmatics* 14. 198–225.

Bączkowska, Anna, Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedre Oleskeviciene Valunaite & Marcin Trojszczak. 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.

Bhattacharya, Shiladitya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri & Atul Kr. Ojha, 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 158–168, Marseille, France. European Language Resources Association (ELRA).

Bretschneider, Uwe & Ralf Peters. 2016. Detecting cyberbullying in online communities. *Research Papers*, Paper 61.

Bretschneider, Uwe & Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA.

Brown, Penelope & Stephen, C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Cachola, Isabel, Eric Holgate, Daniel Preoţiuc-Pietro & Junyi Jessy Li. 2018. Expressively vulgar: The socio- dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2927–2938.

Carston, Robyn. 2009. The explicit/implicit distinction in pragmatics and the limits of explicit communication. *International Review of Pragmatics* 1(1). 35–62.

Caselli, Tommaso, Valerio Basile, Jelena Mitrović & Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. Association for Computational Linguistics.

Cepollaro, Bianca. 2015. In defense of a presuppositional account of slurs. *Language Sciences* 52. 36–45.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, & Ion Androutsopoulos. 2020. Legalbert: "preparing the muppets for court"'. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2898–2904.

Chandrasekharan, Eshwar, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Clif Lampe, Jacob Eisenstein & Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 32.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, & Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829, Florence, Italy. Association for Computational Linguistics.

Cousens, Chris. 2020. Are ableist insults secretly slurs? *Language Sciences* 77.101252.

Croom, Adam. 2011. Slurs. *Language Sciences* 33(3). 343–358.

Croom, Adam. 2014. The semantics of slurs: A refutation of pure expressivism. *Language Sciences* 41. 227–242.

Culpeper, Jonathan. 2005. Impoliteness and entertainment in the television quiz show: The weakest link. *Journal of Politeness Research* 1(1). 35–72.

Culpeper, Jonathan. 2011. *Impoliteness: Using language to cause offence*. Cambridge: Cambridge University Press.

Culpeper, Jonathan & Michael Haugh. 2014. *Pragmatics and the English Language*. London: Red Globe Press.

Culpeper, Jonathan & Michael Haugh. 2021. The metalinguistics of offence in (British) English: a corpus-based metapragmatic approach. *Journal of Language Aggression and Conflict* 9(2). 185–214.

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*, 512–515.

de Gibert, Ona, Naiara Perez, Aitor García-Pablos & Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 11–20, Brussels, Belgium. Association for Computational Linguistics.

Eelen, Gino. 2014. *A Critique of Politeness Theory: Volume 1*. London: Routledge.

Erjavec, Karmen & Melita Poler Kovacic. 2012. "You don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society* 15. 899–920.

Founta, Antigoni-Maria, Djouvas, Constantinos, Chatzakou, Despoina, Leontiadis, Ilias, Blackburn, Jeremy, Stringhini, Gianluca, Vakali, Athena, Sirivianos, Michael & Nicolas Kourtellis. 2018a. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

Founta, Antigoni-Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos & Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 491–500.

Frege, Gottlob 1956. I.—the thought: A logical inquiry. *Mind* 65. 289–311.

Goossens, Louis. 1990. Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic actions. *Cognitive Linguistics* 1–3. 323–340.

Gao, Lei & Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 260–266, Varna, Bulgaria. INCOMA Ltd.

Goffman, Erving. 1955. On face-work; an analysis of ritual elements in social interaction. *Psychiatry MMC* 18. 213–231.

Golbeck, Jennifer, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja R. Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali M. Naik, Heather L. Nixon, Riyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna S. Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan & Derek Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci 17. 229–233.

Gomez, Raul, Jaume Gibert, Lluis Gomez & Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1459–1467.

Graumann, Carl-Fridrich & Marget Wintermantel. 2015. *Diskriminierende Sprechakte. Ein funktionaler Ansatz*. 147–178. transcript Verlag. In Steffen Herrmann, Sybille Krämer & Hannes Kuch (eds.), *Verletzende Worte: Die Grammatik sprachlicher Missachtung*, 147–178. Bielefeld: transcript Verlag.

Grice, Paul H. 1968, 'Utterer's Meaning, Sentence Meaning, and Word-Meaning', *Foundations of Language* 4. 225–42.

Grice, Paul H. *1989*. S*tudies in the way of words*. Cambridge, MA: Harvard University Press.

Haugh, Michael & Jonathan Culpeper. 2018. Integrative pragmatics and (im)politeness theory. *Pragmatics and its interfaces*, 213–239. Amsterdam: John Benjamins.

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.

Hess, Leopold. 2020. Slurs and expressive commitments. *Acta Analytica* 36. 263–290.

Hom, Christopher. 2008. The semantics of racial epithets. *The Journal of Philosophy* 105. 416–440.

Hornsby, Jennifer. 2001. *Meaning and uselessness: How to think about derogatory words. Midwest Studies in Philosophy* 25(1). 128–141.

Hudson, David L. 2012. *The first amendment: freedom of speech*. West, a Thomson Reuters business.

Jeshion, Robin. 2013. Expressivism and the offensiveness of slurs. *Philosophical Perspectives* 27. 231–259.

Jha, Akshita & Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 7–16.

Jigsaw & Google. 2018. Toxic Comment Classification Challenge.

Jucker, Andreas H. 2000. Slanders, slurs and insults on the road to canterbury: Forms of verbal aggression in Chaucer's canterbury tales. In Irma Taavitsainen, Terttu Nevalainen, Paivi Pahta & Matti Rissanen (eds.), *Placing Middle English in Context*, 369–389. Berlin and New York: Muton de Gruyter.

Kampf, Zohar. 2015. The politics of being insulted: The uses of hurt feelings in Israeli public discourse. *Journal of Language Aggression and Conflict* 3(1). 107–127.

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.

Kennedy, Randall 2002. *Nigger: The strange career of a troublesome word*. New York: Knopf Doubleday Publishing.

Koller, Pavel & Petr Darida. 2020. Emotional behavior with verbal violence: Problems and solutions. *Interdisciplinary Journal Papier Human Review* 1(2). 1–6.

Kunupundi, Deepti, Shamtanu Godbole, Pankaj Kumar & Suhas Pai. 2020. Toxic language using robust filters. *SNU Data Science Review* 3(2). Available at: https://scholar.smu.edu/datasciencereview/vol3/iss2/12 (accessed 20 July 2022).

Lakoff, George. 1987a. *Women, Fire, and Dangerous Things*. Chicago: Chicago University Press.

Lakoff, George. 1987b. 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.

Lakoff, George & Mark Johnson. 1980. *Metaphors We Live By*. Chicago: Chicago University Press.

Langacker, Ronald. 1987. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*. Stanford: Stanford University Press.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So & Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4). 1234–1240.

Leets, Laura. 2001. Explaining perceptions of racist speech. *Communication Research* 28. 676–706.

Lepore, Ernie & Mathew Stone. 2018. Explicit indirection. In Daniel Fogal, Daniel W. Harris & Matt Moss (eds.), *New Work on Speech Acts*, 165–184. Oxford: Oxford University Press.

Lewandowska-Tomaszczyk, Barbara. 2017. Incivility and confrontation in online conflict discourses. *Lodz Papers in Pragmatics* 13. 347–367.

Lewandowska-Tomaszczyk, Barbara. 2020. Culture-driven emotional profiles and online discourse extremism. *Pragmatics and Society* 11. 262–291.

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. Zaragossa, Spain.

Liu, Zhuang, Degen Huang, Kaiyu Huang, Zhuang Li & Jun Zhao. 2020. FinBERT: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Special Track on AI in FinTech*, 4513–4519.

Ljung, Magnus. 2010. *Swearing: A Cross-Cultural Linguistic Study.* London: Palgrave Macmillan.

Modha, Sandip, Thomas Mandl, Prasenjit Majumder & Daksh 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 Forum for Information Retrieval Evaluation*, *FIRE '19: Forum for Information Retrieval Evaluation*, 14–17.

Martínez, José M. & Francisco Yus. 2013. Towards a cross-cultural pragmatic taxonomy of insults. *Journal of Language Aggression and Conflict* 1(1). 87–114.

Mathew, Binny, Punyajoy Saha, Seid M. Yimam, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* 35(17). 14867–14875.

McEnery, Tony. 2004. *Swearing in English: Bad language, purity and power from 1586 to the present* (Routledge Advances in Corpus Linguistics). London: Routledge.

Mills, Sara. 2003. *Gender and politeness*. Cambridge: Cambridge University Press.

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*, WWW 16, 145–153.

Nunberg, Geoffrey. 2018. The social life of slurs. In Daniel Fogal, Daniel Harris & Matt Moss (eds.), *New Work on Speech Acts*, 237–295. Oxford: Oxford University Press.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song & Dit-Yan Yeung. 2019. Multilingual and multiaspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Pamungkas, Endang. W. & Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, 363–370.

Partington, Alan. 2006. *The Linguistics of Laughter: A Corpus-assisted Study of Laughter talk*. London: Routledge.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, & Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 1–47.

Pruksachatkun, Yada, Jason Phang, Haokun Liu, Phu M. Htut, Xiaoyi Zhang, Richard Y. Pang, Clara Vania, Katharina Kann & Samuel Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5231–5247.

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding & William Y. Wang. 2019a. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding & William Y. Wang. 2019b. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764, Hong Kong, China: Association for Computational Linguistics.

Ramirez, Artemio, Palazzolo, Kellie E. & Matthew W. Savage. 2010. New directions in understanding cyberbullying. In Rotimi Taiwo (ed.), *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*, 729–744. Herskey, Pennsylvania: IGI Global.

Razavi, Amir H., Diana Inkpen, Sasha Uritsky & Stan Matwin. 2010. Offensive language detection using multi-level classification. In Atefeh Farzindar & Vlado Kešelj (eds.), *Advances in Artificial Intelligence. Canadian AI 2010. Lecture Notes in Computer Science* 6085, 16–27. Berlin: Springer.

Reynolds, Kelly, April Kontostathis & Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops* 2, 41–244.

Sai, Siva & Yashvardhan Sharma. 2020. Siva@hasoc-dravidian-codemix-fire-2020: Multilingual offensive speech detecxxtion in code-mixed and romanized text. In *FIRE*: *Forum for Information Retrieval Evaluation*, 16-20 December, Hyderabad, India, 336–343.

Tenchini, Maria P. & Aldo Frigerio. The impoliteness of slurs and other pejoratives in reported speech. *Corpus Pragmatics* 4(1). 1–19.

Tirrell, Lynne. 2018. Toxic speech: Inoculations and antidotes. *Southern Journal of Philosophy* 56. 116–144.

Vidgen, Bertie, Tristan Thrush, Zeerak Waseem & Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1667–1682, Online. Association for Computational Linguistics.

Waseem, Zeerak. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142, Austin, Texas. Association for Computational Linguistics.

Waseem, Zeerak & Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics.

Wellsby, Michele, Paul D. Siakaluk, Penny Pexman & William J. Owen. 2010. Some insults are easier to detect: The embodied insult detection effect. *Frontiers in Psychology* 1.

Zadeh, Lotfi. 1965. 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 NAACL*.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019b. Semeval2019 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, *Proceedings of the Thirteenth Workshop on Semantic Evaluation*. Minneapolis, Minnesota: Association for Computational Linguistics.

Zampieri, Marcos, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis & Cagri Coltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May & Ekaterina Shutova (eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics.

Žitnik, Slavko, Chaya Liebeskind, & Jelena Mitrović. 2021. Offensive language organization. Available at: https://github.com/UL-FRI-Zitnik/offensive-language-organization (accessed 10 April 2023).

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 Bączkowska, Dr Habil. Prof. UG, 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 lectured 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

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 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 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 Universit*é* 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