3

Improving NLP Techniques by Integrating Linguistic Input to Detect Hate Speech in CMC Corpora

Idalete Dias and Filipa Pereira

1 Introduction

Hate Speech detection research relies heavily on automatic detection models that make use of machine learning (ML), opinion mining, sentiment analysis and polarity detection (Njagi et al. 2015; Rodríguez et al. 2019; Vidgen et al. 2020; Wanjala and Kahonge 2016). Although automatic models are key to current and future hate speech research, their optimisation requires the integration of more fine-grained lexical, syntactic, semantic and discourse analysis input. This study aims to demonstrate that this can be achieved by applying a mixed methods research

I. Dias (✉)
Department of German Studies, School of Arts and Humanities,
University of Minho, Braga, Portugal
e-mail: idalete@elach.uminho.pt

F. Pereira
Department of Informatics Engineering, School of Engineering,
University of Minho, Braga, Portugal
e-mail: a81712@alunos.uminho.pt

I. Ermida (ed.), *Hate Speech in Social Media*,
https://doi.org/10.1007/978-3-031-38248-2_3

approach that integrates both qualitative and quantitative methods. Although there has been previous work on CMC (Computer Mediated Communication) corpus compilation and annotation on various linguistic levels to aid automatic hate speech detection (Bick, 2020; Davidson et al., 2017; Yang et al., 2011), the task of applying a mixed methods approach to identify and analyse fixed discursive patterns for hate speech detection has, to the best of our knowledge, not yet been addressed and evaluated. Our proposal focuses on opinion markers that exhibit a relatively high degree of fixedness and can act as pointers to hateful content. The highly informal and speech-like nature of CMC poses many challenges for electronic processing and automatic hate speech detection methods (Beiβwenger et al., 2016; Fišer et al., 2020; Nobata et al., 2016; Pereira, 2022; Vidgen et al., 2019), but also for linguistic studies on all levels of analysis, in particular the detection, annotation and analysis of discursive-pragmatic features and strategies used to express prejudice and hate.

The chapter is structured as follows: Section 2 reviews previous related work on CMC hate speech corpora, opinion mining methods, hate speech classification approaches and challenges in automatically detecting CMC hate speech. Section 3 outlines the mixed methods research approach pursued. Section 4 provides an overview of the Portuguese-English CMC NETLANG corpus on which the study is based: the text selection and extraction processes and text classification according to prejudice types and corresponding social variables. Section 5 focuses on specific CMC text features that constitute obstacles to the three NLP text pre-processing phases, namely tokenisation, lemmatisation and part-of-speech tagging. Section 6 presents a detailed description of the fixed opinion markers that can act as pointers to expressions of prejudice and hate extracted from the comment threads categorised under the prejudice types ‘Sexism’ and ‘Racism’.

For the purposes of this study, our definition of hate speech is based on the five-factor model proposed by Ermida (Chap. 2, this volume), who defines hate speech as a publicly transmitted message verbalising prejudiced and discriminatory content towards a disadvantaged social group or representative of such a group based on a particular identity trait with the intention of disseminating hate.

2 Related Work

This section reviews previous related work, taking special interest on the CMC hate speech detection and classification approaches currently available for evaluation purposes.

2.1 CMC Hate Speech Corpora

The main goal of CMC hate speech corpora projects is to define and classify hate speech. The simplest classification model distinguishes between hate speech ('Abusive') and non-hate speech ('Clean') (Nobata et al., 2016). Other models are more complex, such as the two-level schema developed to annotate the bilingual Slovene-English Frenk Corpus containing hate speech toward migrants and LGBT. The first annotation level distinguishes between 'Socially Acceptable Discourse' and 'Socially Unacceptable Discourse' (SUD). Comments that fall under the latter category are further classified according to the type of SUD and the targeted individual or group: Background—violence; Background—offensive speech; Other—threat; Other—offensive speech; Inappropriate speech in the case of an unspecified target (Fišer et al., 2020, this volume; Ljubešić et al., 2019). Davidson et al. (2017), on the other hand, defined a three-category model to classify tweets: hate speech, offensive language, or neither of the two. Despite the specific goals underlying each project, the existence of a multiplicity of classification models highlights the fact that there is still no clear definition of what constitutes hate speech.

The quest to understand the difference between hate speech and other types and degrees of offensive language necessarily entails investigating the respective linguistic distinctive features. This leads us to the importance of annotating CMC hate speech corpora with different levels of linguistic information. The Frenk Corpus was annotated according to the following five categories: Orthography, Lexis, Morphology, Syntax and Word Order (Fišer et al., 2020; Ljubešić et al., 2019). As demonstrated by the results of comparing the linguistic features of SUD comments with non-SUD comments, SUD comments reveal: (i) higher vocabulary richness; (ii) more non-standard properties and (iii) a higher frequency of informal syntactic structures. Another noteworthy project is the XPEROHS Danish and German Twitter corpus compiled for hate speech research that has been enriched with morphological, syntactic and the much less common but valuable semantic annotation (Bick, 2020, this volume). Semantic annotation performed within the XPEROHS Corpus will enable the extraction of animal and/or disease metaphors that are often used to express hate toward minority groups. These two projects highlight that high-level linguistic annotation, namely syntactic and semantic annotation, facilitates the automatic extraction of implicit expressions of hate speech. Our own study is motivated by the importance of annotating CMC hate speech corpora on an even higher level of linguistic input, namely the pragmatic-discursive level.

2.2 Opinion Mining

Opinion Mining, also referred to as Sentiment Analysis, is an NLP text-analysis technique that identifies the sentiment (positive, negative, neutral) expressed by someone in a written text regarding a specific topic, product, etc. (Nasukawa & Yi, 2003; Yi et al., 2003). Opinion Mining is significantly used in marketing research and business intelligence. It is mainly employed to gather customer's reactions and feelings towards specific products and services by analysing text reviews, with the goal of improving the services provided. Nevertheless, this technique has also shown potential in extracting people's opinions, attitudes and feelings about issues of public interest, such as politics and health care services, but also towards a particular individual, community, social phenomenon or idea.

There are two fundamental approaches to perform sentiment analysis and opinion mining with the aim of automatically generating (domain-specific) sentiment lexicons (Darwich et al., 2019): the lexicon-based approach and the machine learning (ML) approach.

The lexicon-based approach categorises the sentiments in a specific text according to a polarity scale (positive, negative or neutral) by employing semantic orientation lexicons. These lexicons can be created manually or generated automatically with dictionary-based or corpus-based methods. Dictionary-based methods take a small set of words in an initial lexicon and, with the help of online lexical resources, such as dictionaries or WordNet, expand the lexicon by adding the synonyms and antonyms of those words along with the general sentiment they express. An example of a dictionary-based method to automatically generate a sentiment lexicon is the *qwn-ppv* (Q-WordNet as Personalised Page Ranking Vector) method that relies on the lexical and semantic relations between words in the WordNet database with an emphasis on *antonymy*, *similarity*, *derived from*, *pertains-to* and *also-see* relations (San Vicente et al., 2014).[1] As demonstrated by the authors, this method can be applied to other WordNet languages and outperforms other automatic methods used to generate lexicons. In turn, corpus-based methods are more context-specific and rely on the co-occurrence of word patterns to assess the word's sentiment, using the semantic distance between a word and a specific set of positive and negative words to determine sentiment polarity (Darwich et al., 2019).

Machine learning approaches are frequently applied in sentiment analysis endeavours, by employing classification techniques, such as Naive Bayes, Support Vector Machines (SVM) and Decision Trees. Most projects perform sentiment analysis by applying one of these two approaches. However, ML approaches can also be combined with lexicon-based approaches. Yang et al. (2011) combined ML techniques with semantic-oriented approaches in order to identify radical opinions in hate group web forums. Messages from two extremist/hate forums were collected and pre-processed. Four different types of text features were defined and extracted as classification predictors and three classification techniques were performed on the datasets: Naive Bayes, SVM and Adaboost. The four feature categories defined were syntactic, stylistic, content-specific and lexicon features. Syntactic and stylistic features, such as POS n-grams, function words and vocabulary richness, are more generic and can be used in different analyses of social media text. To represent domain-specific knowledge, content-specific features (named entities, noun phrases) were chosen. These three types of features arise from ML

[1] San Vicente et al. (2014: 89–90) provide an overview of relevant dictionary-based methods used to both manually and automatically build polarity lexicons.

approaches while lexicon features come from a semantic-oriented approach. These lexicon features (subjective/objective term lists, hate terms, etc.) are used to capture more terms pertaining to the expression of negative emotions, such as hate and violence.

Yang et al. (2011) then compared the performance of the three classifiers with different feature sets. F1-scores[2] improved with the addition of more feature types and SVMs outperformed the two other classifiers regardless of the feature set chosen. The lexicon features also significantly improved the performance of the classifiers.

Wanjala and Kahonge (2016) aimed to go further than only identifying radical opinions, seeking to identify and investigate hate mongers and cyber criminals. Their goal was to create a platform that extracts social media comments, classifies them as positive or negative and provides tools for cyber forensics. After a URL is crawled and the respective comments are extracted and stored in a database, the text is pre-processed (tokenisation, text normalisation and POS tagging) and a Naive Bayes classifier is used to obtain the classification of each comment. The project aims to contribute to hate speech opinion mining research via the application of cyber forensics tools.

2.3 Hate Speech Detection

2.3.1 Hate Speech Classification Approaches

The detection of hate speech has become imperative in our online-immersed society. In recent years, considerable classification efforts have been undertaken to mitigate the propagation of hate speech in online mediums, such as social media. The lack of sufficiently accurate hate speech detection and the proliferation of disinformation in online platforms has led to real-life attacks in discriminated communities, bringing attention and outrage from the media and public to the lack of effective and timely moderation employed by social media platforms. Consequently,

[2] F1-score is an ML metric to measure the accuracy of classification models that combines both measures of precision and recall.

many projects have attempted to solve this complex problem with varied approaches.

Most approaches employ machine learning and deep learning methods, creating models to classify messages as hateful, offensive or clean. Other efforts use linguistic rule-based approaches with the help of lexicons. To better moderate its platform, Facebook designed an AI dubbed XLM-R (Conneau et al., 2019), which combines two distinct models: XLM (Cross-lingual Language Model) and RoBERTa (Robustly Optimised BERT Pretraining Approach). RoBERTa (Liu et al., 2019) is an adaptation of Google's algorithm BERT (Bidirectional Encoder Representations from Transformers). BERT (Devlin et al., 2018) is pre-trained with bidirectional representations of unlabelled text where a masked language model is used in order for the BERT model to learn the specific masked tokens based on its left and right contexts. Facebook improved on this model by pre-training it for a longer time and with more data, and by adopting a dynamic masking pattern for the masked language model. After pre-training, the model is fine-tuned to complete the task of classifying posts with possible hate speech with the help of previous posts already identified on the platform. By incorporating RoBERTa with the cross-lingual language model, which was trained on 100 different languages, Facebook obtains a multilingual model with state-of-the-art performance in the understanding of texts in multiple languages.

In another project that applies deep learning in order to classify prejudice, Vidgen et al. (2020) focused specifically on prejudice against East Asians amidst the coronavirus pandemic, collecting and classifying a 20,000-tweet dataset in four different categories: Hostility against East Asia, Criticism of East Asia, Discussion of East Asian prejudice and Non-related. Different contextual embedding models were tested, and RoBERTa proved to be the one that achieved better results.

Other approaches use lexicons to collect potentially offensive and hate-related speech, and then apply sentiment analysis to filter these candidate cases (Njagi et al., 2015; Rodríguez et al., 2019). Sentiment analysis can also be combined with ML methods (Njagi et al., 2015), providing more information that will help the model categorise the possible hateful expressions. Other types of opinion mining techniques such as emotion analysis (Rodríguez et al., 2019) and objectivity analysis (Njagi et al., 2015), are also used to complement sentiment analysis. Sentiment lexicons, such as SentiWordNet (Baccianella et al., 2010), are usually employed to obtain the polarity of the sentences.

Njagi et al. (2015) apply the Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto & Gilbert, 2014) which goes beyond just attributing polarities ranging from −1 to 1 to a list of words as is the case with SentiWordNet. Hutto and Gilbert (2014) create a human-validated dictionary which classifies the sentiment intensity of the words on a scale ranging from slightly to extremely negative or positive, taking into account punctuation, word capitalisation, degree modifiers and other semantic features. These specific semantic rules allow for a more dynamic and context-based identification of sentiment in phrases, especially for microblog texts where these features are more commonly used to convey the tone of the opinions expressed by the users.

The employment of hate lexicons, such as HateBase, can also be combined with the creation of certain lexical and/or linguistic patterns in order to identify possible phrases habitually used in the expression of prejudiced and hateful views (Pereira, 2022; Silva et al., 2016). Even if this approach doesn't detect all occurrences of hate speech present in a certain dataset, it helps to filter large amounts of potentially hateful comments and improved performance can be achieved via the application of further hate speech detection measures on the filtered comment set (Pereira, 2022).

Nonetheless, the research in automatic hate speech detection is still relatively new and faces many challenges. Fortuna and Nunes (2018) found that research projects in this field have been increasing in the last few years but that the field is still in an early stage, with research scopes often being broader than just hate speech, focusing on cyberbullying or general abuse.

2.3.2 Challenges for Automatic Detection of Hate Speech in CMC

Automatic detection and annotation tools have mostly been optimised for well-formed written production. The highly informal and speech-like nature of CMC text productions have significantly contributed to major shortcomings in automatic hate speech detection methods and classification models (Beißwenger et al., 2016; Vidgen et al., 2019). Main challenges for automatic detection and electronic processing tools include word and sentence boundary detection due to non-standard use of punctuation, not clearly identifiable word boundaries and contracted forms. Analysing and describing linguistic irregularities, lexico-semantic features and discourse strategies that are characteristic of social media content generated by users is of paramount importance to improve the performance of NLP and ML tools. A phenomenon worth highlighting is the large amount of orthographic variation found in CMC texts that may be categorised as typographic errors, spelling errors due to ignorance or intentional deviations from the norm to obscure the true meaning of the expression or for emphatic purposes: the replacement of characters by asterisks or other typographical symbols, as in 'YOU NPC'S seriously need to get your heads out of your @$$e$!'[3]; the repetition of characters in a word or phrase used as a means of emphasising the intensity of a specific emotion, as is the case of the interjection 'argh' in Example (1). Another important aspect to mention is the rapid evolution of language in social media that involves the creation of new words by employing different word-formation processes to convey a specific standpoint, opinion or attitude, as shown by the ad hoc formed words 'Liebour' (*lie* + *Labour*) and 'Con-swerve-atives' (a word play with the noun 'Conservatives' and the verb 'to swerve') in Example (1). Both these ad hoc derived lexical units are context-dependent and intricately linked to the user's opinion of the Labour and Conservative Parties, respectively. As a result, prolific lexical creativity on social media networks poses problems for automatic

[3] All examples in the chapter are taken from our NetLang corpus and are reproduced as they appeared, including typographical errors, misspellings, and non-standard grammar.

models that were trained on older datasets which don't account for these new language characteristics (Vidgen et al., 2019).

(1) Lack of representation and being governed by incompetent aspholes do my head in. Failure to report real news, removal of freedom of speech, **aaaaarrrrrgggggghhhhhh**! What choice do we have at election time? Do you want **Liebour** to hit you with a piece of 4 × 2 or the **Conswerve-atives** to belt you with a lump of 2 × 4? If you are an OAP commit to hate speach each and everyday and save on nursing home fees.

In Example (2), the commenter, who is clearly anti-feminist, uses the word 'wimmins' that was adopted by feminists to avoid the use of the word 'men' at the end of the commonly used word 'women'. Since the word 'wimmin' is not lexicalised and is not found in the dictionary resources, it is not recognised by NLP pre-processing tools.

(2) Women are STRONG and capable....but they lack confidence?!?!?! So we've got to engineer society to satisfy every whim of the **wimmins**. Pathetic!

The use of irony and sarcasm is an additional challenge for automatic methods as sarcastic comments and humorous remarks introduce ambiguity in the automatic models. It is often necessary to be aware of the context about the community and the user in order to understand the irony present in written text (Nobata et al., 2016). In some cases, even humans have difficulties in determining if a specific comment in written discourse is meant sarcastically or not.

3 Research Approach

Our mixed methods research approach is focused on combining linguistic knowledge and qualitative analysis with the quantitative results obtained from NLP techniques (see our joint paper, Ermida et al., 2023). We started by carrying out a qualitative fine-grained linguistic and pragmatic analysis of comment threads categorised under the prejudice types 'Sexism' and 'Racism' in the NETLANG English subcorpus in order to identify specific pragmatic patterns that seem to act as pointers to hateful content. Having identified a number of fixed pragmatic cues that exhibit possible anaphoric and cataphoric relations to expressions of prejudiced opinion within the dialogic nature of user-generated CMC content, we had a look at the frequency of occurrence of each pattern in the selected subcorpora and in the entire English corpus. We then proceeded to the qualitative analysis of the behaviour of the most frequent patterns in their dialogic context, including anaphoric and cataphoric phenomena that may contribute to optimising automatic methods of hate or prejudiced speech detection.

4 Corpus Compilation

The NETLANG Corpus is a collection of English and Portuguese CMC texts, more specifically of comments present in the comment boards of online newspaper websites and YouTube. The English subcorpus is composed of articles and comment threads originating from the newspapers *Metro*, *Daily Express* and *Daily Mail*, while the Portuguese subcorpus includes articles and comment threads from the newspapers *Sol*, *O Público* and *Observador*. NETLANG's corpus contains 50.5 million words of which around 43 million pertain to the English subcorpus. YouTube is the biggest contributor to our corpus with close to 48 million words extracted.

The selection of texts was performed by targeting news articles and videos that might incite hateful responses, increasing, in that way, the probability of capturing potential occurrences of hate speech. Prior to the text extraction process, the NETLANG project compiled a table of keywords and expressions, matching them to specific social variables and the respective prejudice type associated with the word or expression.

The web scraping process resulted in a dataset in JSON output format composed of (i) the full newspaper article texts and the YouTube post texts that originated the comments and (ii) the multiple posts that constituted the entire threads available on the date of extraction. In addition, an array of metadata related to the newspaper article, YouTube post text and individual comments was also retrieved, including article or video

Table 3.1 Total number of occurrences of opinion markers in the racism and sexism subcorpora

Patterns	Occurrences (English Corpus)	Occurrences (racism only subcorpus)	Occurrences (sexism only subcorpus)	Occurrences (sexism and racism subcorpus)
(a | you) bunch of ((ADJ) + NN)	4537	1059	740	1465
if you think	2562	490	526	919
(x) people like you	2313	524	427	972
((if | whether) you) like it or not | like it or not	552	145	107	208

title, date of publication of the article, post text and comments, extraction date, number of likes, article URL and the most frequently occurring keywords in the comment thread (Henriques et al., 2019). The extracted comment threads were then automatically classified according to the social variables defined using the keyword analysis tool NetAC (Elias et al., 2021), a statistical framework which applies the keyword table to assess the most frequently occurring keywords in the text and consequently assign the corresponding social variables. It is important to point out that comment threads may be categorised as belonging to more than one prejudice type, as is often the case with racism and sexism (Table 3.1).

5 Pre-Processing

The corpus compilation phase was followed by a fundamental set of text pre-processing operations that facilitate annotation and analysis of linguistic phenomena in electronic texts, namely tokenisation, lemmatisation and part-of-speech tagging.

5.1 Tokenisation

The pre-processing pipeline starts with recognising sentence boundaries and performing tokenisation that refers to the division of the input string, in this case, the full comment text, into tokens based on whitespaces, punctuation at the beginning or end of words or specified delimiters to be further analysed individually. A token can represent a word, a punctuation mark or an emoji. Example (3) displays language-specific and user-generated content features that pose a problem for tokenisation:

(3) Trump hasn't caused a nuclear war yet so **id** say **hes** doing okay. Aoc wants america to be socialist so i **dont** like her. **Whos** gonna foot the bill for free healthcare. The united state has a ton of debt, if the governed **cant** pay off the debt how are they **gonna** pay for free healthcare. The Somalian politician said 911 was some people doing stuff. She hates isreal. Screw her. **Dont** know much about the others other then the African american one is using her race as a shield. Obama never did those things when he was president. He used his talents which had nothing to do with race.

In example (3), standard contracted forms 'I'd' (*I would*), 'he's' (*he is*), 'don't' (*do not*), 'who's' (*who is*) and 'can't' (*can not*) lack the apostrophe and will be interpreted by the tokeniser as one token when in fact these forms represent two tokens. As a result, common non-apostrophe variants in user-generated CMC content constitute an obstacle for higher-level NLP tools, such as lemmatisers and POS taggers that are not always able to effectively deal with ambiguous tokens, as is the case with 'id' that can be recognised as the contracted form 'I'd' or the noun 'ID' referring to a form of personal identification. Additional problems arise with informal contracted forms such as 'gonna' that is a widely used variant for 'going to'. For specific linguistic analyses this token is required to be treated as two tokens.

Another important aspect to mention is that defining punctuation marks as sentence and word delimiters may lead to tokenisation errors due to the multiple functions they perform. For example, periods also appear in abbreviations and ellipses that should both be interpreted as one token. Ideally the tokeniser should identify the periods in the abbreviated form 'P.C.' (*politically correct*) in example (4a) as part of the abbreviation. Example (4b) illustrates that in CMC communication ellipses may also be represented by commas and be composed of more than the three standard punctuation marks.

(4) a. well stated brother. Well said! These **P.C. warriors** are outta control and gonna cost some lives. Glad I'm at the tail end of an over twenty year career. Transgendered is the latest change coming. And coming quickly.

b. ISRAEL,,,,,,,,,,,,,,,,,,,,,,,,,why? bec GOD almighty is on there side no one can argue period

Multiple adjacent punctuation symbols as represented by the shrug emoticon in example (5a) should be rendered as one token so as to preserve the meaning behind the emotion icon. Emojis are also a common non-verbal communication feature of CMC that may take various forms to convey affective states, intent, make opinion statements, as in example (5b), etc. Since the repetition of the same emoji in a sequence reinforces the opinion expressed, the entire sequence should be considered one token, instead of the seven tokens rendered by the tokeniser.

(5) a. 2:09 Funny how the girl who prefers women in EVERY possible way, cuts her hair like a man \\_(ツ)_/¯

b. shes an ugly peice of 💩💩💩💩💩💩💩

At this point, it is important to emphasise that the quality of the output of the tokenisation process will have a direct influence on the results of the subsequent processing steps.

5.2 Lemmatisation

The second phase of the pre-processing pipeline is lemmatisation. After dividing each sentence in its individual parts, each token is then lemmatised. This dictionary-based process that takes into account the morphological features of a word consists in removing the inflectional endings of the input word, returning it to its *lemma* or base form. A lemmatised corpus is of considerable use in linguistic studies as it allows users to: (i) search for the base form of a word and obtain as a result all its inflected and derived forms and (ii) carry out frequency and distribution counts.

Performing a search query in our corpus to search for occurrences of lemma 'kill' used as a verb retrieves instances containing inflected forms of the verb as shown in Example (6).

(6) a. Of corse the muslim women don't have "hate"....they would be beaten up or **killed** if they had an opinion or spoke up, like most oppressed people do.
b. I would pay a hit man £1000 for every feminist he **kills**
c. It's pretty obvious why we prefer older women. This third wave feminism has **killed the chances** of getting with any younger girls.
d. Always **kills me** when gay dudes talk about how hetero sexual men treat women. If we are so bad at the job why don't you all step in?

In examples (6a) and (6b) the verb 'to kill' is used in its literal sense, as opposed to its metaphorical sense in Examples (6c) and (6d). The verb-noun collocation in (6c) 'kill the chances' followed by the phrasal verb 'to get with' in the gerund, meaning 'to eliminate the possibility of something happening', exhibits a fixed lexical and syntactic pattern that can be used to extract all the instances of this verb-noun construction. This is also the case of the hyperbolic expression in (6d), '(It) always kills me', normally followed by the conjunction 'when', meaning that the user finds the situation described preposterous. The lexical and syntactic fixedness of expressions can be exploited by NLP resources to perform higher-level linguistic annotation, including metaphor annotation.

Given the prevalence of misspellings in CMC texts, it is important to point out that this feature results in the incorrect lemmatisation of the word in question, as is the case in Examples (5b) and (6a), with the incorrect spelling of the noun 'piece' ('peice') and the second element of the fixed expression 'of course' ('corse'). When it comes to intentional misspellings, as is the case of the word 'Liebour' in example (1), in which the first part of the word is strategically replaced by another to convey the user's opinion, running a spellchecker that compares 'Liebour' with correctly spelled words in a large dictionary and uses algorithms, such as approximate string matching algorithms, to correct its spelling to 'Labour' will distort the intended meaning the user wishes to express.

5.3 Part-of-Speech Tagging

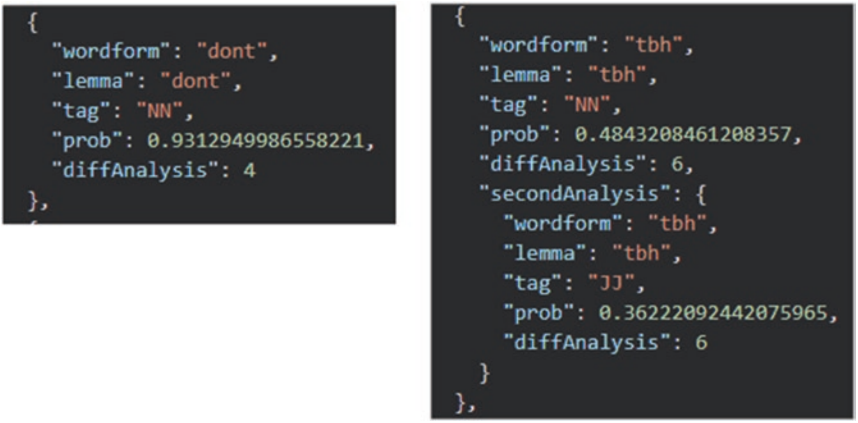
The third phase entails automatically assigning each token with a part-of-speech tag according to its morpho-syntactic properties: noun, verb, adjective, adverb, etc. Example (7) shows the result of POS tagging the tokens in the comment '@[USERNAME] i dont think you can compare pregnancy to lung cancer tbh' using the FreeLing POS tagger and respective tagset.[4]

(7) **@** (*punctuation*) **[USERNAME]** (*common noun*) **i** (*personal pronoun*) **dont** (*common noun*) **think** (*verb*, *present tense*) **you** (*personal pronoun*) **can** (*modal verb*) **compare** (*verb*) **pregnancy** (*common noun*) **to** (*particle*) **lung** (*common noun*) **cancer** (*common noun*) **tbh** (*common noun*). (*punctuation*)

The tagger was not able to classify the non-standard contracted form of 'dont' (*do not*) and the informal abbreviation 'tbh' (*to be honest*) correctly. As clearly demonstrated, non-canonical spelling and non-lexicalised abbreviations complicate the POS tagging task. The POS tags are assigned based on the probability of a particular tag occurring. In the case of 'tbh' (Fig. 3.1), the first POS analysis carried out indicates that there is a low probability of the token 'tbh' being a noun (NN). In a second iteration, the tagger classified the token as an adjective (JJ) with an even lower likelihood of occurrence.

The results of these automatic processes add a layer of linguistic input to the text of each comment which allows the search of specific combinations of words/lemmas and POS tags. In fact, the annotated comment

[4] http://nlp.lsi.upc.edu/freeling/



```
{
  "wordform": "dont",
  "lemma": "dont",
  "tag": "NN",
  "prob": 0.9312949986558221,
  "diffAnalysis": 4
},
```

```
{
  "wordform": "tbh",
  "lemma": "tbh",
  "tag": "NN",
  "prob": 0.4843208461208357,
  "diffAnalysis": 6,
  "secondAnalysis": {
    "wordform": "tbh",
    "lemma": "tbh",
    "tag": "JJ",
    "prob": 0.36222092442075965,
    "diffAnalysis": 6
  }
},
```

Fig. 3.1 Pre-processing result for tokens ‘dont’ and ‘tbh’

threads that resulted from this pre-processing enabled the creation of a query engine, SAQL (Pereira, 2022), in which one can search for lexical and morpho-syntactic patterns in the corpus.[5]

6 Using Linguistic-Pragmatic Patterns to Detect HS

Although there are many studies on CMC corpus compilation and annotation at various linguistic levels to automatically detect hate speech (Bick, 2020; Davidson et al., 2017; Yang et al., 2011, to mention just a few), little mixed methods research has focused on the identification of pragmatic patterns to detect hate speech in user-generated content. We address this gap by qualitatively identifying and analysing fixed pragmatic patterns that constitute potential pointers to hate speech. In this process, we took a subset of comment threads categorised under the prejudice types ‘Sexism’ and ‘Racism’ and manually identified a number of opinion markers that seemed to point to prejudiced and discriminatory content.

[5] The query engine is freely available for research purposes at http://netlang-corpus.ilch.uminho.pt:10400/

We then carried out an absolute frequency count of these markers in these specific subcorpora and extracted the candidate comment texts for further analysis. We focused on linguistic and contextual aspects of a subset of those instances that express hateful and discriminatory opinions. The results of this qualitative analysis were used to aid the sentiment polarity analysis that was run on the candidate comment texts containing instances matching the regular expression *(a | you) bunch of ((ADJ) + NN)*.

6.1 Linguistic Pattern Identification

Whilst analysing the subset of comment threads categorised under the prejudice types 'Sexism' and 'Racism', we picked up on the following opinion markers that exhibit a certain degree of fixedness and seem to behave as potential pointers to hate speech:

- Expressions containing the noun 'bunch' in the pattern *(a | you) bunch of ((ADJ) + NN)*, as in '**a bunch of primates**'.
- The expression 'If you think', as in '**If you think** millennials are bad,wait till they raise their kids 😂😂 kill it before it breeds.'
- The expression '(x) people like you', as in 'Black **people like you** are racist. Stop blaming white people because you are racist'.
- Fixed phrases '((if | whether) you) like it or not', as in 'Amongst our specie,Men are meant to lead ,**like it or not**'.

Table 3.1 shows the total number of occurrences of the above-listed patterns in the 'Racism Only', 'Sexism Only' and 'Sexism and Racism' categorised subcorpora[6] and in the entire English corpus. The pattern matching the regular expression *(a | you) bunch of ((ADJ) + NN)* is the most frequently occurring pattern with a total of 4537 instances in the corpus. It should be pointed out that 71.9% of the pattern occurrences are found in the racism and sexism categories. The patterns 'if you think' and '(x) people like you' are relatively comparable in terms of their

[6] The 'Sexism and Racism' category refers to files that have been identified as containing both prejudice types. It is important to point out that the files categorised as 'Sexism Only' and 'Racism Only' in our study may also contain other prejudice types.

distribution in the categories 'Racism' and 'Sexism' representing 75.5% and 83% of the total occurrences, respectively. The pattern '((if | whether) you) like it or not' appears a total of 552 times in the corpus with 83.3% of instances in the subcorpora.

In the following section, we provide a detailed analysis of instances matching the patterns in Table 3.1 based on the mixed methods approach.

6.2 Analysis of Results

6.2.1 Pattern 1: *(a | you) bunch of ((ADJ) + NN)*

A qualitative analysis of a subset of the total instances containing the word 'bunch' highlights the importance of taking into account the surrounding context when evaluating expressions of hate speech. Consider the following examples taken from the subset:

(8) a. By allowing the Muslims to march, pray wherever they want and slaughter animals in our streets like **the bunch of Neanderthals** they are the police are failing OUR community!

b. Even a century after the UK first gave female citizens the vote, feminists are still holding each other back by jealousy and attacking each other like **a bunch of alley cats**.

c. Well, you know, if your video is just **a bunch of stats** bundled up and not especially interesting, at least make sure to not assume that correlation implies causality...

d. You just said **a bunch of nothing**. Don't preach your ancient fairy tales to me.

Having identified instances matching the regular expression *(a | you) bunch of ((ADJ) + NN)* that target a specific group in a prejudiced and hateful form, as in Examples (8a) and (8b), and other instances that are not considered hate-related speech, such as (8c) and (8d), we conducted an automatic sentiment polarity analysis using SentiWordNet on the total occurrences of Pattern 1 in the entire corpus with the two-fold aim of (i) obtaining an overview of the number of these instances that are

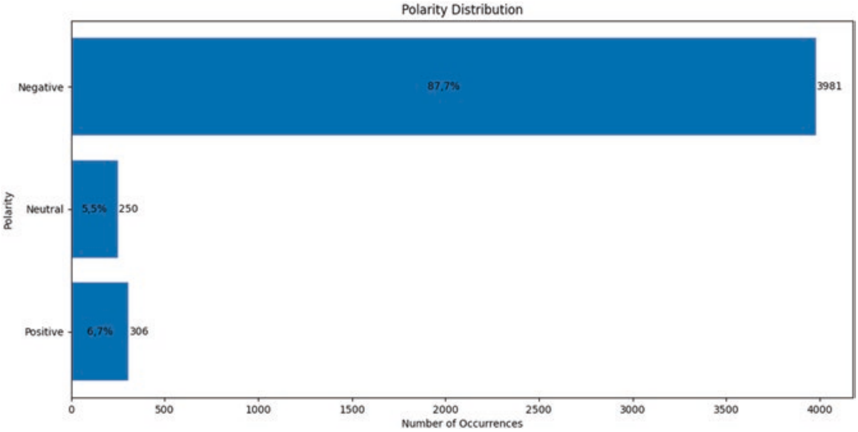


Fig. 3.2 Polarity D=distribution for instances of Pattern 1

assigned a negative connotation and (ii) understanding the behaviour of sentiment polarity analysis tools. As will be demonstrated by the sentiment polarity analysis carried out with Pattern 1, the automatic assignment of polarity based on polarity lexicons presents some shortcomings.

The results in Fig. 3.2 show that 87.7% of the total instances of Pattern 1 are assigned a negative polarity, in contrast to the number of instances assigned a neutral polarity (5.5%) and a positive polarity (6.7%). The very large difference between the negative and the other polarity values seems to indicate that Pattern 1 can function as a pointer to prejudiced and discriminatory content. A closer qualitative analysis of the results reveal that there are instances that were assigned a negative polarity that should be classified as neutral or positive. There are also cases of neutral comments that bear negative connotations and positive comments that are in effect negative.

SentiWordNet operates on the basis of annotated WordNet synonym sets (synsets) that are classified according to the three degrees of sentiment polarity (positive, negative and objective) of the terms that form each synset (Baccianella et al., 2010). Words with different senses, as is the case with the noun ‘bunch’, appear in more than one synset, each with its own polarity scores. The assignment of a polarity score for an instance matching the regular expression *(a | you) bunch of ((ADJ) + NN)* results from the sum of the polarity scores of the words that compose the pattern. This process is exemplified in Example (9):

(9) Anyone else think that the parents in this show don't give a shit about their kids and are using them to make money like **a bunch of degenerate pieces** of fucking garbage?

The noun 'bunch' has been assigned a negative polarity score of 0.125. Since the word 'degenerate', which was wrongly classified as a noun by the POS tagger, has a positive polarity of 0.125, these two values cancel each other out, resulting in a neutral polarity assignment. As can be seen, errors in the pre-processing phases have implications for higher-level operations. A qualitative analysis of Example (9) clearly identifies the instance as bearing a negative sentiment. This analysis also demonstrates the importance of sentiment analysis systems taking into account larger units of meaning and consequently larger chunks of discourse in order to improve their performance.

Example (10) illustrates the difficulty sentiment analysis tools have when dealing with words that are not lexicalised or words that are not contained in the dictionaries used to train them, as is the case of the word 'conspiritards', a blend composed of the words 'conspirator' and 'retard'.

(10) I'm looking in and I see a **bunch of schizophreniec conspiritards** thinking the cops are making black people look like criminals.

Given the above, SentiWordNet is not able to assign the words 'schizophreniec' and 'conspiritards' a polarity score. Due to the misspelling of the first word and the absence of the second in the dictionaries, SentiWordNet does not take them into account when calculating the sentiment polarity of the sentence(s). This is a shortcoming of sentiment analysis, since these unrecognised words combined with 'bunch' convey the main opinion of the user. Once again, a qualitative analysis makes clear that this comment expresses a negative prejudiced opinion.

Moving on to analyse the other three patterns in Table 3.1, it is important to highlight that the pattern *(a | you) bunch of ((ADJ) + NN)*: (i) contains the collective noun 'bunch' that functions as an anchor in a WordNet structure that facilitates automatic hate speech detection and

(ii) has a specific syntactic construction that can be easily queried in a POS-tagged corpus. As will be demonstrated in Sect. 6.2.2, automatic hate speech detection becomes a very difficult task when the above conditions are not satisfied.

6.2.2 Remaining Patterns: *if you think*, *(x) people like you*, *((if| whether) you) like it or not*

In contrast to Pattern 1, the fixed expressions *if you think*, *(x) people like you*, *((if| whether) you) like it or not* do not contain a content word that can aid in automatic hate speech detection. As demonstrated by Examples (11) to (13), separating each of the words in the expressions into tokens destroys the potential of these opinion markers operating as pragmatic cues to hateful content. For this reason, these fixed expressions should be treated as a unit and labelled accordingly. Automatic hate speech detection models, which normally make use of sentiment analysis and polarity detection, should consider incorporating pragmatic-based features, such as fixed opinion markers in their approach.

Another key factor that poses problems for automatic hate speech detection, specifically when handling opinion cues, is the very fragmented nature of CMC interaction due to sequential incoherence (Herring, 1999). The opinion cues in examples (11) to (13) serve as anaphoric and/or cataphoric anchors to stereotypical, toxic and racist opinions towards black people.

(11) I hate how blacks think we owe them something because we enslaved there ancestors over a 100 years ago. I see this all of the time!! I think of it as where would blacks be if we hadn't enslaved them and brought them to America. They would be in Africa swating flies off of there face. **If you think** we are so racist here you can go back to Africa to be with your own race!!!!

(12) i love being white. i'm proud of being white and i hate anyone who hates white people... including **white people like you** who hate white people. anyone who's against whites is my enemy. that means YOU are my enemy. you're too stupid to realise that negroes hate white people. that stupidity may end up getting you badly hurt one day. don't be looking for any sympathy from whites when it happens.

(13) **Whether you like it or not**, BLM is just a sorry excuse to get black supremacy and sadly you're buying into it. Seriously if the word "hypocrisy" could take a physical form it would be BLM.

The expressions '*if you think*', '*(x) people like you*', '*((if | whether) you) like it or not*' in the examples refer back and forth to the comment(s) of (a) previous user(s) and refer back and forth to the commenter's own opinion. NLP models are not able to effectively deal with complex behaviour of anaphoric and cataphoric referencing in CMC texts to capture hateful content.

A linguistic analysis of the content and formal syntactic properties of the utterances initiated by the opinion markers '*if you think*' (11) and '*Whether you like it or not*' (13) reveals a certain degree of fixedness that can be formalised: '*If you think (that) x, (then) y*' and '*(Whether you) like it or not, x*'. Notwithstanding the drawbacks associated with non-standard features of CMC content (e.g., lack of punctuation), it would certainly be interesting to explore how automatic detection models handle fixed opinion cues whose content structure can be formalised and their contribution to hate speech detection in conjunction with other resources.

Detecting hateful content automatically in comments characterised by an extensive argumentative mode, as is the case of Example (12), is an extremely difficult task. A more fine-grained qualitative analysis of the lexical, syntactic, contextual and pragmatic features of the comment is required.

7 Conclusion

In our attempt to demonstrate how linguistic input can improve NLP techniques applied to CMC Corpora in order to detect Hate Speech, we started by conducting a qualitative analysis of a subset of the texts present in the NETLANG Corpus categorised according to the prejudice types 'Racism' and 'Sexism' with the aim of identifying pragmatic patterns that can serve as cues to prejudiced, discriminatory and hateful content. In this process, we observed the occurrence of fixed opinion markers, namely *(a | you) bunch of ((ADJ) + NN)*, *if you think*, *(x) people like you*, *((if | whether) you) like it or not*, that seemed to operate as anaphoric and cataphoric cues to hateful speech. Subsequently, we extracted all instances of the identified opinion markers in the above-mentioned subcorpora and analysed a small set of the instances qualitatively focusing on their potential to effectively point to hateful, prejudiced and discriminatory opinions in the comment. As demonstrated by the results of our study, the selected opinion markers do serve as pointers to prejudiced or toxic opinions. The lexical and/or syntactic fixedness of the marker facilitates automatic hate speech detection. Identifying hate speech using opinion markers found in long argumentative comments has proven to be a difficult task.

As shown in various examples from the corpus, certain characteristics of CMC texts pose challenges to automatic methods and result in errors that have consequences for the entire NLP pre-processing pipeline. Identifying and understanding these characteristics is crucial to fine-tune pre-processing tools and improve the results for higher-level processing tasks. Enriching the NLP tools with linguistic input, such as new terms, new variants and neologisms, is necessary for a more accurate attribution of POS tags, and consequently lead to a more precise extraction of the fixed patterns that will later be analysed and submitted to other automatic NLP methods.

Our study highlights the importance of applying a mixed methods research approach that integrates both qualitative and quantitative methods. The fixed patterns analysed in this study would have been difficult to identify in the corpus if we had not started exploring the data from a qualitative perspective. The identification of fixed patterns that function as opinion markers and a pragmatic analysis of their anaphoric and cataphoric behaviour in CMC content constituted a valid approach to understand aggressive, hateful and prejudiced opinion in CMC, thus demonstrating the importance of integrating pragmatic-discursive knowledge in hate speech detection and extraction. ML techniques and NLP methods, such as content-specific lexicons and semantic corpus-based approaches can greatly benefit from this knowledge.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA).

Beißwenger, M., Bartsch, S., Evert, S., & Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In P. Cook et al. (Eds.), *Proceedings of the 10th web as corpus workshop (WAC-X) and the EmpiriST shared task* (pp. 44–56). Association for Computational Linguistics.

Bick, E. (2020). An annotated social media corpus for German. In N. Calzolari et al. (Eds.), *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)* (pp. 6127–6135). European Language Resources Association (ELRA).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised cross-lingual representation learning at scale*. CoRR.

Darwich, M., Mohd Noah, S. A., Omar, N., & Osman, N. (2019). Corpus-based techniques for sentiment lexicon generation: A review. *Journal of Digital Information Management, 17*, 296–305.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512–515.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. CoRR.

Elias, C., Gonçalves, J., Araújo, M., Pinheiro, P., Araújo, C., & Henriques, P. R. (2021). NetAC, an automatic classifier of online hate speech comments. In *Trends and applications in information systems and technologies* (pp. 494–505). Springer.

Ermida, I., Dias, I. & Pereira, F. (2023). *Social media mining for hate speech detection: Adversative constructions as markers of opinion and emotion conflict*.

Fišer, D., Smith, P., & Ljubešic, N. (2020). Nonstandard linguistic features of Slovene socially unacceptable discourse on Facebook. In *The dark side of digital platforms: Linguistic investigations of socially unacceptable online discourse practices*. Ljubljana University Press.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Survey, 51*(4), 85.

Henriques, P. R., Araújo, C., Ermida, I., & Dias, I. (2019). Scraping news sites and social networks for prejudice term analysis. In H. Weghorn (Ed.), *Proceedings of the IADIS international conference applied computing 2019* (pp. 179–189). IADIS Press.

Herring, S. (1999). Interactional coherence in Cmc. *Journal of Computer-Mediated Communication, 4*(4), JCMC444.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. CoRR.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In K. Ekštein (Ed.), *Text, speech, and dialogue* (pp. 103–114). Springer.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on knowledge capture (K-CAP'03)* (pp. 70–77). Association for Computing Machinery.

Njagi, D., Zuping, Z., Hanyurwimfura, D., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*, 215–230.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153). International World Wide Web Conferences Steering Committee.

Pereira, A. (2022). *SAQL: Query language for corpora with morpho-syntactic annotation*. Master's thesis, University of Minho.

Rodríguez, A., Argueta, C., & Chen, Y. (2019). Automatic detection of hate speech on Facebook using sentiment and emotion analysis. In *2019 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 169–174). IEEE.

San Vicente, I., Agerri, R., & Rigau, G. (2014). Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 88–97). Association for Computational Linguistics.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016) Analyzing the targets of hate in online social media. In *Proceedings of the tenth international AAAI conference on web and social media* (pp. 687–690). The AAAI Press.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online* (pp. 80–93). Association for Computational Linguistics.

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., & Tromble, R. (2020). Detecting east Asian prejudice on social media. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 162–172). Association for Computational Linguistics.

Wanjala, G. W., & Kahonge, A. M. (2016). Social media forensics for hate speech opinion mining. *International Journal of Computer Applications, 155*(1), 39–47.

Yang, M., Kiang, M., Ku, Y., Chiu, C., & Li, Y. (2011). Social media analytics for radical opinion mining in hate group web Forums. *Journal of Homeland and Emergency Management, 8*(1). https://doi.org/10.2202/1547-7355.1801

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE international conference on data mining* (pp. 427–434). IEEE.