Computational Linguistics in the Netherlands Journal 12 (2022) 33-48 Submitted 12/2022; Published 12/2022

Linguistic Analysis of Toxic Language on Social Media

Ine Gevers* Ine.Gevers@uantwerpen.be
Ilia Markov** i.markov@vu.nl
Walter Daelemans* Walter.Daelemans@uantwerpen.be

* *University of Antwerp, CLiPS, Lange Winkelstraat 40, Belgium*

** *Vrije Universiteit Amsterdam, CLTL, De Boelelaan 1105, The Netherlands*

Abstract

The increasing popularity of online communication platforms entails a profound interest in the automatic detection of toxic language, since the effects of user anonymity or issues with content moderation can result in hostile environments. Linguistic analysis can be an important tool for discovering language patterns discriminating between toxic and non-toxic language, leading to the development of more robust detection systems. In this paper, we investigate several linguistic features of online Dutch toxic comments compared to non-toxic comments. We focus on three main research questions investigating the differences between the two types of comments: average length, lexical diversity, and linguistic standardness of comments. More specifically, we compared the average number of tokens per comment, the type-token ratio, (variants of) the content-to-function-word ratio, the propositional idea density, the use of emoji and emoticons, the punctuation to non-punctuation ratio, and measured the level of linguistic standardness combining features such as word choice, character flooding, and unconventional capitalization. The analysis was performed on the LiLaH dataset, which contains over 36,000 Dutch Facebook comments related to the LGBT community and migrants. We conclude that toxic comments are different from their non-toxic counterpart regarding all the investigated linguistic features. Additionally, we compared our results to Slovene and English. Our analysis suggests that there are commonalities but also remarkable differences in the linguistic landscape of toxic language across the three languages that may lead to further research.

1. Introduction

Communication of the last years has been marked by the popularity of online platforms. It has never been this easy to connect with someone from anywhere on the planet. Unfortunately, this easy access and broader scale of communication in addition to medium-related aspects such as anonymity seems to have contributed to antisocial communication behaviour (Chui 2014). The growing presence of toxic language has attracted the attention of researchers in multiple fields. Often, this issue is mentioned in connection with the term "hate speech". While the expressions of hate speech imply legal consequences, toxic language also includes utterances that cannot be prosecuted, but can still be harmful to the target. For this reason, we adopt the term "toxic language" in this paper.

Research about toxic language, and by consequence also the detection of toxic language, is marked by its interdisciplinarity. Numerous disciplines offer their interpretation and analyses of – according to them – the most important aspects of the phenomenon. In March 2019, the top ten fields that published about this topic included among others science information systems, psychology, and communication science (Waqas et al. 2019). It is noteworthy that while toxicity is inherently expressed by means of language, an in-depth analysis from this point of view has lagged behind until recently. However, this perspective could provide important insights into the internal structure of toxic language.

©2022 Ine Gevers, Ilia Markov, Walter Daelemans.

In this paper, we investigate the differences between Dutch toxic and non-toxic comments from a linguistic point of view. We focus on three main research questions, that can be divided into subcategories. These research questions are based on existing research that focused on surface linguistic features of both the comparison between positive and negative emotions, and the comparison between acceptable and non-acceptable comments in Slovene using the FRENK dataset (Vitez and Fišer 2016, de Maiti et al. 2020). The findings indicate that in Slovene, toxic and non-toxic comments have similar lengths, and toxic comments turn out to be lexically more diverse but linguistically less standard (de Maiti et al. 2020).

Given that this paper by de Maiti et al. (2020) researched toxicity in Facebook comments on a similar dataset, this study provided useful insights and a good framework for the research questions addressed in this paper. We decided to continue with a comparable subset of research questions, because those features (i.e., average comment length, lexical diversity, and linguistic standardness) can provide insights into the possible superficial structural differences between toxic and non-toxic language (de Maiti et al. 2020). While our primary focus is on linguistic features of Dutch toxic language, this setup allows us to investigate whether the previously obtained results hold in a multilingual context.

First, we compare the average length of toxic and non-toxic comments. Second, we explore the lexical diversity, which consists of the vocabulary diversity (measured by the type-token ratio (TTR), the content-to-function-word ratio (CTFW), and the propositional idea density (PID)), and the analysis of emoji and emoticons. Third, we zoom in on the linguistic standardness of comments (i.e., the (non-)adherence to the linguistic norms of that language), which is constructed from multiple features such as punctuation use, word choice, and unconventional capitalization.

The rest of the paper is structured as follows. In Section 2, we start with an overview of related work providing more details on our interpretation of toxic language, and a framework of the automatic detection of this discourse type. Next, in Section 3, we discuss our research questions and hypotheses as well as our methodology. Further on, in Section 5, we present the results of our analysis and the comparison between results obtained for Dutch, Slovene, and English. The final Section 6 concludes our research, giving an overview of the findings.

2. Related Work

2.1 Toxic language

Given the rise in research focusing on hate speech on social media platforms, a general definition of the phenomenon is needed. The most widely accepted definition of hate speech is the one by (Nockleby 2000): hate speech is "any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic". While this definition covers most aspects, MacAvaney et al. (2019) highlight that it could be optimized. The authors argue that individual attacks, over-generalisations about certain social groups, and the expressions of agreement with hateful comments should also be included.

Additionally, the term "hate speech" is used in legal contexts. As mentioned before, there is no unified global definition of hate speech, resulting in governmental bodies adopting different interpretations, which makes it hard to generalize penalties for the offence over country boundaries. However, countries or unifying institutions such as the U.S., U.N., Council of Europe, or the E.U. do delineate form(s) of hate speech and assign consequences[1] (de Maiti et al. 2020).

While comments that belong to this category evidently have to be detected, other comments that fall out of this classification do deserve attention as well. To ensure a broad scope in this paper, we do not use the term "hate speech", taking example from previous studies (de Maiti et al. 2020, Ljubešić

1. An overview of legislations that apply in different countries can be found here: `https://futurefreespeech.com/global-handbook-on-hate-speech-laws/\#post-1391-footnote-2`

et al. 2019). Instead, we will apply the term "toxic language". We consider toxic language to be an inclusive term, stretching over subfields such as abusive and offensive language, hate speech, and cyberbullying. Our interpretation of the term is compatible with the definition of "socially unacceptable discourse", which is used in the FRENK datasets for Slovene, English, and Croatian (Ljubešić et al. 2021), a related corpus to LiLaH, as used in (Markov et al. 2021). In their definition, socially unacceptable discourse covers a wide range of offensiveness, including "prosecutable hate speech, threats, abuse and defamations, but also not prosecutable but still indecent and immoral insults and obscenities" (Fišer et al. 2017). This is also what we include in our definition of "toxic language".

Besides definitions presented by researchers, Big Tech companies that host the social media platforms also provide their interpretations. Those companies have been criticised in the past for inactivity and lack of effort to prevent or address toxicity on their platforms, but in time they have joined the battle against this issue by formulating clear policies and penalties (Davidson et al. 2017). In its community standards, Facebook defines hate speech as "a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease"[2]. Twitter's definition does not differ greatly from Facebook's: according to Twitter's policy, it is forbidden to "promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease"[3].

This multitude of definitions and interpretations is a witness to the inherent difficulty and subjectivity surrounding the topic. Therefore, some researchers suggest to stop the search for the ideal and universal definition, but instead to take advantage of the subjectivity this phenomenon entails and adapt it to the researchers' need (Khurana et al. 2022).

2.2 Computer-mediated communication (CMC)

As mentioned before, social media platforms and computer-mediated communication in general provide environments that are suitable for the expression of toxic language, arguably partly because of the anonymity of the users and the creation of environments for extreme ideologies (Chui 2014).

While analysing the linguistic characteristics of toxic comments on social media, it is important to bear in mind that there are factors other than discourse type that can affect the linguistic structure of expressions. As discussed in (de Maiti et al. 2020), the medium of communication itself can influence the standardness of the messages. Most studies concerning linguistic norms online focus on the language use of young people such as teenagers and adolescents, because these are the individuals that are growing up with social media, and are most likely to be influenced by these platforms.

Research investigating CMC in Dutch revealed that language use on this medium is distinct from "analogue" contexts (Hilte et al. 2017). Possible arguments why this is the case include the limited message size online on certain platforms such as Twitter, the importance of efficiency and speed over correctness, and the creative use of orthography to make up for the absence of other strategies present in speech such as body language, volume, etc. In addition, deviations of the norms can be an expression of belonging to a certain social group (Verheijen 2015).

The linguistic features that have gained the attention of researchers are among others the use of emoticons and emoji, the omission of function words and the use of borrowings, as well as punctuation, capitalization of words and repetition of letters (Verheijen 2015, Hilte et al. 2017). However, it is important to note that the use of non-standard features also depends on social variables such as individual preferences, age, gender, familiarity with CMC customs or discourse topic (Verheijen 2015). In summary, deviations from the linguistic norms in online toxic language could be explained not only by the discourse type, but by its medium as well.

2. https://transparency.fb.com/policies/community-standards/hate-speech/
3. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

2.3 Toxic language detection

Given the increasing presence of toxic comments on online platforms, attempts have been made to counter this phenomenon. Besides public sensitization, moderators of platforms can delete or proactively censor hateful comments, but this might not be the most effective approach (Vandenbosch n.d.). The accurate detection of and adequate responses to toxicity remain problematic, yet the deployment of computational techniques helps tackling the issue. The urge to combat online toxic comments with automatic means has been around since 1997 (Spertus 1997), and lately the research community witnessed a renewed interest in the field. This can be corroborated by the number of publications on this topic: there is a significant rise of publications since 2014 (Fortuna and Nunes 2018, Poletto et al. 2021). This rise can be linked to the growing computing power and data accessibility and availability: those innovations allowed the entrance of deep learning models, which have taken the lead in recent years. Especially ensemble learning seems a fruitful direction (Markov et al. 2022, Zampieri et al. 2019). Before deep learning, researchers frequently adopted classical machine learning approaches such as Support Vector Machines (SVM), Random Forests, or Decision Trees (Fortuna and Nunes 2018).

The majority of the datasets used to train and evaluate the models are in English. Less global languages have remained underrepresented for a while, but this is starting to change with more publications featuring languages other than English (Fortuna and Nunes 2018, Poletto et al. 2021). In that respect, our research deals both with the current lingua franca (i.e., English), but mainly with two "minority" languages (i.e., Dutch and Slovene). Besides, it is noteworthy that researchers do not restrict their attention to the binary classification of toxic language. The task has also been redefined as a multiclass classification problem, including more fine-grained categories of toxicity. For example, the LiLaH dataset consists of six more fine-grained categories, which will be detailed in Section 4. Moreover, related classification tasks focus on target or participant identification, exposing the complex social links between offenders, targets, and bystanders (Van Hee et al. 2015, Van Aken et al. 2018). This expansion of the task is proof of the complexity that comes with not only the definition, but also the accurate detection of toxic language.

Most researchers apply their methods on newly made datasets. While the acquisition of more data in itself is desirable, most annotated datasets are not shared publicly, which complicates the comparison of research done on toxic language detection (Vidgen and Derczynski 2020, Fortuna and Nunes 2018). Often, however, this is legitimately so because of restrictions imposed regarding privacy (e.g., the GDPR law in Europe). With reference to these datasets, Poletto et al. (2021) rightfully argue that more researchers should include annotator guidelines and annotator agreements of these datasets, since this information is still not regularly shared. The absence of these clarifications complicate the evaluation of the datasets regarding possible annotator biases and leaves the readers speculating what the researchers consider to be toxic language. The disadvantages related to the high number of new datasets that are used only by a small number of researchers have not gone unnoticed by the research community. Initiatives such as shared tasks address this issue (Poletto et al. 2021). There has been an increase in the number of organised shared tasks in the last years, and given the elevated number of participants in these tasks we can observe a growing interest in the field (Zampieri et al. 2019). For instance, the OffensEval 2020 task counted 145 teams who submitted their runs on the test dataset, which broke the all-time record of the SemEval shared tasks (Zampieri et al. 2020). A major advantage of the setup of shared tasks is that models and techniques can be fully and objectively compared to one another, since all have been trained on the same dataset, and all teams follow the same procedure provided by the organizers.

While the state-of-the-art detection models are promising, challenges remain. For instance, the use of hateful words in non-hateful contexts can lead to false positives, or inversely the absence of hateful words in hateful contexts can lead to false negatives (Markov and Daelemans 2021, Van Aken et al. 2018). Related to that, correctly interpreting the context in which hateful words are used is still challenging (Markov and Daelemans 2022). Additionally, long-range dependencies or misspelled

and idiosyncratic words can confuse the classifiers (Van Aken et al. 2018, Vidgen et al. 2019). Currently, one of the main challenges resides with implicit language use such as sarcasm, irony, or humour (Van Aken et al. 2018, Vidgen et al. 2019). Recently, Lemmens et al. (2021) showed that type and target classification of hate speech can be improved by providing the models with hateful metaphors as a feature. On a more basic level, researchers have pointed out that biases in the annotation procedure could lead to biased models. For instance, researchers found that words based on the African American English lexicon tend to be perceived as offensive more often, while this is not necessarily the case (Sap et al. 2019). These challenges show that the task of automatically detecting toxic language use is far from being solved. While this study does not deal with open challenges head-on, we hope that our linguistic analysis sheds light on different facets of the phenomenon, and encourages new approaches.

3. Methodology

We investigate the linguistic differences between toxic and non-toxic Facebook comments. In light of this, we provide three research questions and hypotheses:

1. Research question 1: Average length
   (a) Hypothesis 1.1: Toxic and non-toxic comments have a similar length.
2. Research question 2: Lexical diversity
   (a) Hypothesis 2.1: Vocabulary diversity is larger in toxic comments compared to non-toxic comments.
   (b) Hypothesis 2.2: Non-toxic comments contain more emoticons and emoji than toxic comments.
3. Research question 3: Linguistic standardness
   (a) Hypothesis 3.1: Punctuation to non-punctuation ratio is lower in toxic comments.
   (b) Hypothesis 3.2: toxic comments are linguistically less standard than non-toxic comments.

The research questions we discuss have also been addressed for the Slovene language (de Maiti et al. 2020). We hypothesized that the results would be similar because both papers investigate a global interlingual issue. Since the LiLaH dataset is constructed in the same way as the FRENK dataset, on which the previous study by de Maiti et al. (2020) is based, we assume that the observations regarding comment length (i.e., no significant difference) generalizes to Dutch. We hypothesize that the vocabulary diversity is larger in toxic comments, because "people tend to use more colourful and creative language for emotionally-charged content" (de Maiti et al. 2020). In contrast, the less frequent use of emoji in toxic comments is argued to be caused by the relative difficultly of accessing more specific emoji through the emoji keyboard, "which can be perceived as too time-consuming during the creation of an emotionally-charged comment" (Bočková 2019, de Maiti et al. 2020). Next, also based on conclusions of previous research we expect toxic comments to be linguistically less standard (de Maiti et al. 2019). Additionally, we performed the analysis on the English part of the FRENK dataset for further interlingual comparison. We will discuss the comparison between the results for the different languages in Section 5.2.

In what comes next, we provide more detail about the specific methodologies to address our research questions.

We performed tokenization and part of speech (POS) tagging by applying the StanfordNLP library. For the POS tags, we use the 17 universal tags (uPOS)[4]. As mentioned before, the LiLaH

4. https://universaldependencies.org/u/pos/

and FRENK datasets are constructed to be as similar as possible, which is why the StanfordNLP library was used to guarantee that the same tool was used across the different languages (i.e., Dutch, Slovene, and English).

Given the imbalance in comment lengths between toxic and non-toxic comments, we calculated the type-token ratio (TTR), the content-to-function-word ratio (CTFW), and the propositional idea density (PID) by making 100 samples of 1,000 tokens from the entire subset, instead of comparing individual comments[5].

3.1 Lexical diversity

Type-token ratio We calculated the TTR by dividing the number of types (unique words) by the number of tokens (all words) within the sample, after removing punctuation marks.

Content-to-function-word ratio We divided the count of content words by the count of all words. For this, we consider the following POS categories as function words: adposition, auxiliary, coordinating conjunction, determiner, numeral, particle, pronouns, and subordinating conjunction. These are therefore excluded from the content words.

Propositional idea density PID is used to calculate the number of propositions, or new ideas, in a text. To the best of our knowledge, PID has not been researched in the context of toxic language detection, but since this could be seen as a variant of the content-to-function-word ratio, we include this measure as well. To calculate the PID in a comment, we divided the number of words that are related to such ideas by all words. We use a baseline technique as suggested in Marckx et al. (2018), which is to divide the total count of verbs, adjectives, adverbs, adpositions, nouns, and proper nouns by the total number of tokens in that comment. For this purpose, we used the universal part of speech (uPOS) tagset. Usually, the PID is measured on longer pieces of texts, such as novels or blogs. However, as we applied it to typically shorter texts (social media posts), this could skew the results.

Emoji To account for the use of emoji in comments, we apply the libraries emojis and emoji (emoji version 1.6.3) from PyPI.

Correlation coefficients Additionally, we compute the correlation coefficients for each feature with the toxicity of the comments. To do this, we add the values of each feature for each comment in a separate vector, and compared these with the binary annotated label. Because of the binarity of the toxicity labels, we use the biserial correlation coefficient, employing the appointed method from SciPy[6]. While we calculate the significance for the TTR, CTFW, and PID by making samples from the entire dataset, here we proceed at the comment level.

3.2 Linguistic standardness

The standardness, or non-standardness, of a message is hard to measure. Some researchers prefer to annotate these messages manually on several subcategories of standardness such as orthography, lexis, morphology, syntax, or word order (de Maiti et al. 2020). However, this approach is time-consuming and labour-intense, and becomes harder to uphold for larger datasets. Therefore, encoding the level of standardness automatically is advantageous. We were very fortunate to be able to build on a research about the linguistic non-standard features of chat conversations between Dutch-speaking youth (Hilte 2019). While not all aspects described in this research are applicable to toxic language detection, the work provided an essential baseline for this section. In what follows, we will provide a verbal description of how the features were encoded.

5. We also calculated the PID on the comment level, to account for the contextual influence within one comment on the PID. However, the scores do not differ greatly from the sampling technique. Hence, we will proceed with the sampling method.
6. scipy.stats.pointbiserialr

First, we merged all toxic comments together, and all non-toxic comments. In doing this, comments belonging to the two topics (LGBT and migrants) are distributed over the two categories (toxic and non-toxic). Next, we made token lists of all comments for each category. Then, we filtered out all references to hyperlinks, e-mail addresses, filenames, and emojis.

Standard or non-standard To separate all standard tokens from the non-standard tokens, word lists were made with words belonging to substandard Dutch, or other languages (English and Arabic). Each token was attributed to one of the word lists, else it was interpreted as standard Dutch.

Flooding Flooding is the excessive overuse of characters to gain expressiveness (Hilte 2019). Since in standard Dutch the double repetition of a letter is rather frequent, the cut-off point for flooding was the repetition of three or more times the same character. The flooding was investigated both for letters and punctuation marks.

Emoticons and emoji By the use of regular expressions different types of emoticons (described as 'western' such as ":)", 'asian' such as " ^_^", or 'hearts' such as "<3") were encoded. Besides emoticons, unicode emoji were also taken into account.

Unconventional capitalization This feature includes words with (a) all caps, (b) inverse caps, and (c) alternating caps, but excludes standard abbreviations or emoticons (such as "XD").

Combination of question and exclamation marks Regular expressions were used to encode the combination of question and exclamation marks.

Laughter Also by the use of regular expressions variants of "haha" and "hihi" were counted.

4. Dataset

In this study we used the LiLaH dataset, which contains 5,094 comments related to the LGBT community, and 31,571 comments related to migrants (Markov et al. 2021). The data was collected from three prominent Flemish news providers: VRT[7] (biggest public provider), Het Nieuwsblad[8] and Het Laatste Nieuws[9] (biggest private provider).

We would like to point out that research focusing on classification tasks such as toxic language detection often rely on annotated datasets, the LiLaH corpus being no exception. Unfortunately, the annotating guidelines and annotator trainings are not universal, which complicates the comparison between multiple studies. In this light, previous studies have argued for the use of common labels and annotation guidelines in hate speech detection (Schmidt and Wiegand 2019).

To address this problem and to ensure reproducibility, we note that the annotation guidelines for the LiLaH corpus are the same as those used for the compilation of the FRENK corpus, which enables us to compare the results obtained on the two datasets (de Maiti et al. 2020). The complete annotation guidelines can be found in (Ljubešić et al. 2019).

Two trained annotators and one expert annotator decided on type and target of toxic language. For the binary classification (toxic - non-toxic), the annotators obtained an inter-annotator agreement of 70.5%, which resulted in a Cohen's Kappa score of 0.43 (a moderate agreement). Besides this binary label, they classified comments according to more fine-grained categories. These data properties, as will be described further, can be found in Table 1. First, there are six options concerning type. The annotators decided between violent speech/ threats and offensive speech, and whether these are based on the target's background (such as religion, gender, sexual orientation etc.), or on individual traits. If there is no clear target, annotators should choose inappropriate speech. The last label is appropriate speech, when none of the above is true. Second, for both the LGBT and Migrants corpora possible targets are (a) the individuals that belong to these social

7. https://www.vrt.be/nl/
8. https://www.nieuwsblad.be/
9. https://www.hln.be/

groups themselves, (b) people related to one or both of the communities, (c) the journalist or media platform that produced the article, (d) another commenter that posted a message as reaction to the article, (e) or other target(s).

Type	Toxic	Non-toxic
Acceptable speech	0	17,720
Other offensive	12,562	
Background offensive	5,727	
Other violence	241	
Background violence	233	
Inappropriate	182	
TOTAL	18,945	17,720

(a) Distribution of hate speech types

Target	Toxic	Non-Toxic
No target	176	17,719
Other	8,066	1
Migrants	5,684	
Commenter	3,520	
Related to migrants	639	
Journalist or medium	634	
LGBT	223	
Related to LGBT	3	
TOTAL	18,945	17,720

(b) Distribution of hate speech targets

Table 1: Data properties.

5. Results and discussion

The sections below will discuss the results obtained according to the research questions.

5.1 Quantitative analysis

5.1.1 Average length

We compared the median length of toxic comments and non-toxic comments by looking at the number of tokens in each comment. We measured a median of 10 tokens for non-toxic comments, while for toxic comments the median was 21. It is safe to conclude that toxic comments are generally longer than non-toxic comments, rejecting our first hypothesis.

5.1.2 Lexical diversity

First, we calculated the TTR. This is slightly lower for toxic comments (0.55) compared to non-toxic comments (0.57).

Next, we compared the content-to-function-word ratio (CTFW) for the two categories. We conclude that non-toxic comments have a higher ratio (1.63 vs 1.47). Additionally, as a variant of the CTFW, we calculated the propositional idea density (PID). While the CTFW is higher for non-toxic messages, the PID is slightly higher for toxic comments (toxic 0.56, non-toxic 0.55).

To verify if the length of comments influences the result, we also compared longer toxic and non-toxic comments. For this, we only selected the Facebook posts that had more than 10 tokens. This restriction did not change the general trend of the results.

While de Maiti et al. (2020) suggest that toxic comments might have a more extensive lexical diversity because "people tend to use more colourful and creative language for emotionally-charged content", the results from our study suggest the opposite. The vocabulary diversity is larger for non-toxic comments, rejecting our hypothesis.

However, we note that there are generally more individual ideas proposed in a toxic comment. This could be explained by the fact that toxic comments are on average twice as long as non-toxic comments, allowing for more propositions than in short comments.

Besides the TTR and (variants of) the CTFW, we also looked at the relative frequency of emoji. On token level, there are fewer emoji in toxic comments (0.005) than in non-toxic comments (0.013).

	Toxic	Non-toxic
Average length	22 tokens	11 tokens
Type-token ratio	0.53	0.57
Content-to-function-word ratio	1.47	1.63
Propositional idea density	0.56	0.55
Relative frequency emoji	0.005	0.013
Unique emoji	251	363

Table 2: Summary of the findings relating to the lexical diversity of toxic and non-toxic comments.

Feature	Correlation
Average length	0.19
Type-token ratio	-0.09
Content-to-function-word ratio	-0.08
Propositional idea density	-0.01
Relative frequency emoji	-0.10
Unique emoji	-0.10
Punctuation to non punctuation ratio	-0.06

Table 3: Correlation coefficients of each feature with the toxicity of the comments. All correlations are significant at p <0.01.

Considering the number of unique emoji and emoticons being used, there are again fewer of these in the toxic subcorpus (251) than in the non-toxic one (363). If we consider the emoji (excluding emoticons), there is an overlap between the two categories of 32%. This indicates that we can accept our hypothesis that there are more emoji in non-toxic comments. A possible explanation for this imbalance is provided in (de Maiti et al. 2020): they argue that toxic comments might include fewer emoji because the search for the right emoji is more time-consuming.

All results described above are significant with a p-value of <0.05. We summarize the findings in Table 2.

Lastly, we include the correlation coefficient for each feature with the toxicity of the comments. We provide the results in Table 3. The correlation coefficients show that longer comments are positively correlated with the appearance of toxic comments, while higher TTR, CTFW, PID, relative and unique frequencies of emoji and punctuation to non-punctuation ratios are associated with non-toxic comments. We note that the average length has the largest coefficient, indicating its importance. While all correlations are significant at p <0.01, we point out that most correlation values are rather low.

5.1.3 Linguistic non-standardness

In this subsection, we describe several features that together measure linguistic non-standardness, and compare these properties for both toxic and non-toxic comments. We summarize the results in Table 4.

The first aspect of the linguistic non-standardness that we considered was the punctuation to non-punctuation ratio. We note that non-toxic comments have a higher ratio than toxic comments (0.11 and 0.10, respectively), but this difference is rather small.

The rest of the features related to linguistic non-standardness we investigated are based on the sociolinguistic research we described earlier in Section 3.

In both toxic and non-toxic contexts, the majority of the tokens are standard Dutch (95%). Less than 1 % are English or influenced by English. However, when we investigate the features of non-standard language, we can observe interesting differences between toxic and non-toxic comments.

	Toxic	Non-toxic
Punctuation to non-punctuation	0.10	0.11
Word choice	95% Dutch	95% Dutch
Flooding (letter)	0.09	0.01
Flooding (punctuation)	8%	6%
Combination '!' and '?'	21 times	3 times
Capitalization	0.8%	0.5%
Laughter	0.05%	1%

Table 4: Summary of the findings relating to the linguistic standardness of toxic and non-toxic comments.

Comparing character flooding (the deliberate overuse of a character within word boundaries), we distinguish between letter flooding, and punctuation flooding. Regarding letter flooding, 0.09% of the tokens in the toxic corpus are part of this. Mainly the flooding of the letters 'f', 'e', and 'a' are present (e.g. 'pfff', 'veeeel' ('muuuch'), 'aaah'). 0.01% of the tokens in non-toxic comments include letter flooding. The most frequently flooded letters are 'e', 'f', and 'o' (such as 'zeeeer' ('veeeery'), 'pffff', 'loooool'). Considering the punctuation flooding, it is noteworthy that in toxic comments, 8% of all exclamation marks or question marks appear in a flooding context, while this is only 6% in non-toxic comments. This is not surprising, since the overuse of characters can contribute to the expressiveness of the comment, which is an inherent characteristic of emotionally-charged language use such as toxic language. In both toxic and non-toxic contexts, the punctuation mark that is most often flooded is the exclamation mark. Often, this is in combination with a question mark and/ or a full stop.

Given this observation, we also zoomed in on the specific combination of the exclamation mark and the question mark. We noticed that this combination is notably more frequent in toxic comments (21 times) than in non-toxic comments (3 times). While the difference is relatively substantial, the absolute values are rather low. Therefore, we would like to draw attention to the possibility that these results might be influenced by a minority group of individuals with a distinct preference for this combination, and might not be generalizable to the whole corpus of toxic versus non-toxic. However, due to privacy regulations and the structure of the dataset used, this is impossible to verify.

Next, we considered the unconventional capitalization of letters. We observe that letters are more often capitalized in toxic comments (0.8%) than in non-toxic comments (0.5%). As suggested before, this might be because this non-standard feature marks the expressiveness.

Lastly, we focused on instances of laughter (variations of "haha" or "hihi"). This phenomenon is more frequent in non-toxic comments (1%) than in toxic comments (0.05%).

Taking into account all observations, we can conclude that Dutch toxic comments are linguistically less standard than Dutch non-toxic comments. Most of the distinguishing factors can be related to the higher expressiveness of comments, which is often desired in negative contexts.

5.2 Comparison to Slovene and English

Existing literature already discussed these research questions for the Slovene subset of the FRENK dataset (de Maiti et al. 2020). Additionally, we performed part of this analysis on the English data, which is also part of the FRENK datasets. Given that LiLaH and FRENK are created following the same annotator guidelines, we can compare the results obtained on these languages. However, we cannot compare the linguistic standardness of these languages; the methodology that was used on the Slovene dataset differs from ours (Dutch), and due to time constraints we have no specific software available for English.

In Figure 1, the properties relating to average length and lexical diversity are visualised for each of the three languages.

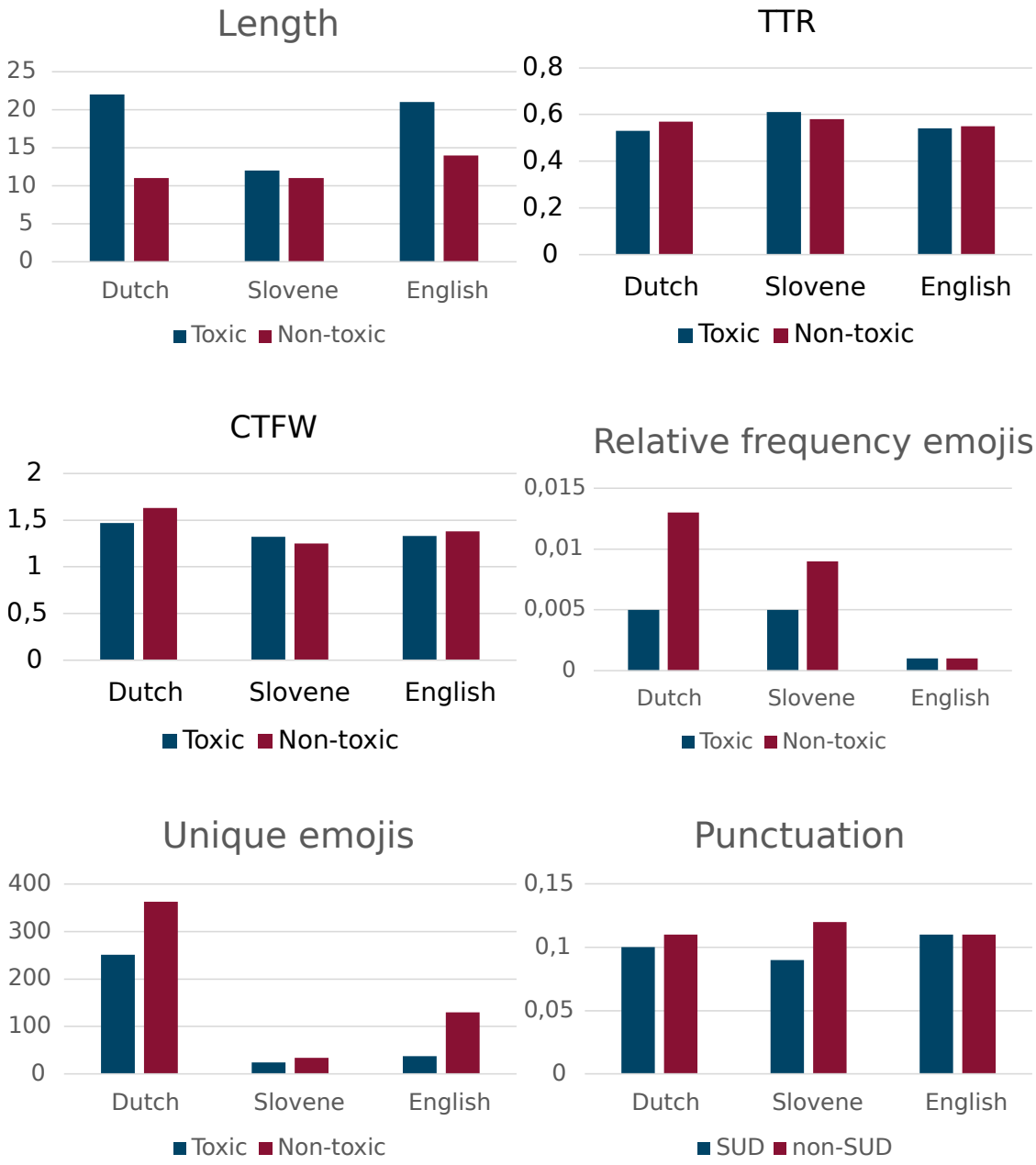

Figure 1: The comparison of the different properties concerning comment length and lexical diversity (type-token ratio (TTR), content-to-function-word ratio (CTFW), relative frequency of emoji, number of unique emoji, and punctuation-to-non-punctuation ratio) between Dutch, Slovene, and English.

5.2.1 Average length

Both in Dutch and in English, toxic comments are on average longer than non-toxic comments, but this is more outspoken in Dutch (22 vs 11 for Dutch, 21 vs 14 for English). In Slovene, the researchers did not find a significant difference between the discourse types.

5.2.2 Lexical diversity

The results concerning the TTR and CTFW are opposed in Dutch and Slovene. While in Dutch the TTR is higher for non-toxic comments, a higher TTR for toxic comments was noted in Slovene (0.61 vs 0.58). In like manner, in Dutch the CTFW is higher for non-toxic, which is the opposite of the Slovene findings (1.32 for toxic comments, 1.25 for non-toxic ones).

The results for English are comparable to Dutch; the TTR and CTFW are higher for non-toxic contexts, but in this language, the differences are small (TTR: 0.54 for toxic, 0.55 for non-toxic; CTFW: 1.33 for toxic, 1.38 for non-toxic). We also calculated the PID for English. These results reflect the conclusions made for Dutch: toxic comments generally utter more propositional ideas than non-toxic comments (English toxic comments have a PID rate of 0.53, non-toxic comments of 0.52). In all languages, the differences between the discourse types are significant with a p-value < 0.05.

Considering the relative frequency of emoji and emoticons, both in Slovene and in Dutch the relative frequency of emoji is higher for non-toxic messages (0.009 for Slovene, 0.01 for Dutch) than toxic messages (0.005 for Slovene, 0.005 for Dutch). There is no difference to be observed in English (both discourse types 0.001). In all three languages, there are more unique emoji in the toxic corpus than in the non-toxic corpus. Slovene toxic comments "contained 24 different emoticons and emojis while non-SUD (sic. non-toxic) contained 34, 35% of which overlap with those found in SUD (sic. toxic) comments" (de Maiti et al. 2020). In English, there are 37 unique emoji in the toxic corpus, 130 in the non-toxic corpus, and an overlap of 27%. However, the Dutch corpus has a significantly higher count of emoji in general.

We suggest this might be the result of the release of new emoji in 2019 (Broni 2019). The FRENK corpus was compiled approximately one year before the LiLaH corpus, so this new release might have been reflected in the Dutch-speaking users' rising quantitative use of emoji.

The punctuation to non-punctuation ratio is slightly higher for non-toxic comments in both Dutch (0.11 vs 0.10) and Slovene (0.12 vs 0.09). There is no difference for English (0.11).

To summarize, we observe from the TTR, CTFW, and punctuation to non-punctuation ratio that while in Dutch and English the lexical diversity is larger for non-toxic comments, this is not the case for Slovene. However, in Dutch and Slovene, toxic comments have fewer unique emoji and a smaller relative frequency compared to non-toxic comments, while there is no difference in English. This suggests that English and Dutch are more similar regarding the distribution of lexical diversity over toxic and non-toxic comments, but Dutch and Slovene are more similar when it comes to emoji use.

6. Conclusion

Toxic language use seems omnipresent in social media content. This topic has been examined thoroughly from many points of view. The aims of this study were to provide insights into the structural differences between Dutch toxic and non-toxic online comments, and to compare these results to English and Slovene.

For this purpose we used the LiLaH corpus, which contains over 36,000 Dutch Facebook comments related to two social minority groups: LGBT community and migrants. Based on the study by de Maiti et al. (2020), we compared several linguistic features between toxic and non-toxic comments. We formulated three main points of interest: average length of comments, lexical diversity, and linguistic non-standardness. The last two research questions were divided into multiple subquestions: lexical diversity consisted of the type-token ratio (TTR), the content-to-function-word

ratio (CTFW), the propositional idea density (PID), the number of unique emoji, and the relative frequency of emoji; linguistic non-standardness of punctuation to non-punctuation ratio, character flooding, combination of punctuation marks, unconventional capitalization, and instances of laughter ("haha" and "hihi").

We observed that Dutch toxic comments are generally longer, lexically less diverse (indicated by a lower TTR and CTFW), and linguistically less standard (indicated by more frequent character flooding, combinations of exclamation marks and question marks, and more unconventional capitalization). We note that the non-toxic counterpart on average uses more (unique) emoji, and we find more instances of laughter. We would like to point out that some of the correlations between features and the toxicity of a comment are, although all are significant, rather low. This could limit the usability of these features in the detection of toxic language. Our results suggest that average length would be the most promising feature to include in a machine learning setup. Besides, we also suggest to reproduce this analysis including other social media platforms, to investigate the generalizability of these results. While this study focused on the analysis of linguistic features, further research could use these results to investigate the importance of these features in automated methods for toxic language detection.

Next, we compared our results to the ones obtained in a previous study on Slovene, and performed the analysis on the English FRENK dataset (Ljubešić et al. 2019). We note similar patterns regarding average length and relative frequency of emoji in all three languages, but we reach opposite conclusions about lexical diversity: Slovene toxic comments are lexically more diverse than Slovene non-toxic comments, which is opposite in Dutch and English. The significant difference in the use of emoji in the FRENK (English and Slovene) and LiLaH (Dutch) datasets can be explained by the addition of new emoji in the standard library in the time between the creation of the two datasets: we hypothesize that there are more emoji to be found in the Dutch corpus not necessarily because of cultural or linguistic differences with Slovene or English, but simply because of the grown availability of emoji when the Dutch corpus was assembled. In conclusion, our analysis confirmed a number of expected commonalities in the linguistic landscape of toxic language in the studied languages, but we also found discrepancies that can lead to further research.

Acknowledgements This research has been supported by the Flemish Research Foundation through the bilateral research project FWO G070619N "The linguistic landscape of hate speech on social media". The research also received funding from the Flemish Government (AI Research Program). We would like to thank the reviewers for their valuable contributions.

References

Bočková, Renata (2019), The use of punctuation, emoji and emoticons in youtube abusive comments.

Broni, Keith (2019), Emoji updates in 2019, *Emoji updates in 2019*. https://blog.emojipedia.org/emoji-updates-in-2019/.

Chui, Rebecca (2014), A multi-faceted approach to anonymity online: Examining the relations between anonymity and antisocial behaviour, *Journal For Virtual Worlds Research* **7** (2), pp. 1–13.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber (2017), Automated hate speech detection and the problem of offensive language, *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11, pp. 512–515.

de Maiti, Kristina Pahor, Darja Fišer, and Nikola Ljubešić (2019), How haters write: analysis of nonstandard language in online hate speech, Cergy-Pontoise, France, pp. 37–42.

de Maiti, Kristina Pahor, Darja Fišer, and Nikola Ljubešić (2020), Nonstandard linguistic features of Slovene socially unacceptable discourse on facebook, *Fišer, D., & Smith, P. The Dark Side of Digital Platforms: Linguistic Investigations of Socially Unacceptable Online Discourse Practices* pp. 12–35.

Fišer, Darja, Tomaž Erjavec, and Nikola Ljubešić (2017), Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene, *Proceedings of the first workshop on abusive language online*, Association for Computational Linguistics, Vancouver, BC, Canada, pp. 46–51.

Fortuna, Paula and Sérgio Nunes (2018), A survey on automatic detection of hate speech in text, *ACM Computing Surveys* **51** (4), pp. 1–30.

Hilte, Lisa (2019), *The social in social media writing: the impact of age, gender and social class indicators on adolescents' informal online writing practices*, PhD thesis, University of Antwerp.

Hilte, Lisa, Reinhild Vandekerckhove, and Walter Daelemans (2017), Modeling non-standard language use in adolescents' cmc: the impact and interaction of age, gender and education, *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora17), 3-4 October 2017, Egon W.[edit.]; et al.*, Italy/Stemle, pp. 11–15.

Khurana, Urja, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens (2022), Hate speech criteria: A modular approach to task-specific hate speech definitions, *Proceedings of the Sixth Workshop on Online Abuse and Harms*, Association for Computational Linguistics, Seattle, Washington (Hybrid, pp. 176–191.

Lemmens, Jens, Ilia Markov, and Walter Daelemans (2021), Improving hate speech type and target detection with hateful metaphor features, *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Association for Computational Linguistics, pp. 7–16.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec (2019), The FRENK datasets of socially unacceptable discourse in Slovene and English, *Proceedings of the international conference on text, speech, and dialogue*, Springer, Ljubljana, Slovenia, pp. 103–114.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Ajda Šulc (2021), Offensive language dataset of Croatian, English and Slovenian comments frenk 1.1. Slovenian language resource repository CLARIN.SI.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder (2019), Hate speech detection: Challenges and solutions, *PloS one* **14** (8), pp. e0221152.

Marckx, Silke, Ben Verhoeven, and Walter Daelemans (2018), The claus case: Exploring the use of propositional idea density for Alzheimer detection, *Computational Linguistics in the Netherlands Journal* (8), pp. 66–82.

Markov, Ilia and Walter Daelemans (2021), Improving cross-domain hate speech detection by reducing the false positive rate, *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Association for Computational Linguistics, Online, pp. 17–22.

Markov, Ilia and Walter Daelemans (2022), The role of context in detecting the target of hate speech, *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying*, Association for Computational Linguistics, Gyeongju, Republic of Korea, pp. 37–42.

Markov, Ilia, Ine Gevers, and Walter Daelemans (2022), An ensemble approach for Dutch cross-domain hate speech detection, *Proceedings of the International Conference on Applications of Natural Language to Information Systems*, Springer, Valencia, Spain, pp. 3–15.

Markov, Ilia, Nikola Ljubešić, Darja Fišer, and Walter Daelemans (2021), Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Online, pp. 149–159.

Nockleby, John T (2000), Hate speech, *Encyclopedia of the American constitution* **3** (2), pp. 1277–1279.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti (2021), Resources and benchmark corpora for hate speech detection: a systematic review, *Language Resources and Evaluation* **55** (2), pp. 477–523.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and A Noah Smith (2019), The risk of racial bias in hate speech detection, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Assocation for Computational Linguistics, Florence, Italy, pp. 1668–1678.

Schmidt, Anna and Michael Wiegand (2019), A survey on hate speech detection using natural language processing, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017,*, Association for Computational Linguistics, Valencia, Spain, pp. 1–10.

Spertus, Ellen (1997), Smokey: Automatic recognition of hostile messages, Aaai/iaai, pp. 1058–1065.

Van Aken, Betty, Julian Risch, Ralf Krestel, and Alexander Löser (2018), Challenges for toxic comment classification: An in-depth error analysis, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Association for Computational Linguistics, Brussels, Belgium, pp. 33–42.

Van Hee, Cynthia, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste (2015), Automatic detection and prevention of cyberbullying, *Proceedings of the International Conference on Human and Social Analytics*, International Academy, Research, and Industry Association, Saint Julians, Malta, pp. 13–18.

Vandenbosch, Sam (n.d.), Omgaan met publieksreacties op online nieuwsartikels.

Verheijen, Lieke (2015), Out-of-the-ordinary orthography: The use of textisms in Dutch youngsters' written computer-mediated communication, *Proceedings of the York Papers in Linguistics* **Series 2 (Online)**, pp. 127–142.

Vidgen, Bertie, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts (2019), Challenges and frontiers in abusive content detection, *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, pp. 80–93.

Vidgen, Bertie and Leon Derczynski (2020), Directions in abusive language training data, a systematic review: Garbage in, garbage out, *Plos one* **15** (12), pp. e0243300.

Vitez, Ana Zwitter and Darja Fišer (2016), Linguistic analysis of emotions in online news comments-an example of the eurovision song contest.

Waqas, Ahmed, Joni Salminen, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen (2019), Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate, *PloS one* **14** (9), pp. e0222194.

Zampieri, Marcos, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin (2020), SemEval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), pp. 1425–1447.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (2019), SemEval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA.