ISSN 2499-4553

IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 10, Number 1
june 2024

aAccademia
university
press

editorial board

editorial office

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

AILC Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000983

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_10_1

aAccademia
university
press

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 10, Number 1
june 2024

CONTENTS

Adapting BLOOM to a new language: A case study for the Italian
Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, Marco Polignano, Giovanni Semeraro 7

U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models
Claudiu Daniel Hromei, Danilo Croce, Roberto Basili 21

Investigating Text Difficulty and Prerequisite Relation Identification
Chiara Alzetta 39

Italian Linguistic Features for Toxic Language Detection in Social Media
Leonardo Grotti 65

Publishing the Dictionary of Medieval Latin in the Czech Lands as Linked Data in the LiLa Knowledge Base
Federica Gamba, Marco Carlo Passarotti, Paolo Ruffolo 95

Italian Linguistic Features for Toxic Language Detection in Social Media

Leonardo Grotti[*]
University of Antwerp

This study addresses the urgent issue of toxic language, prevalent on social media platforms, focusing on the detection of toxic comments on popular Italian Facebook pages. We build upon the framework suggested by the LiLaH project: a standardized framework for analyzing hateful content in multiple languages, including Dutch, English, French, Slovene, and Croatian. We start by examining the linguistic features of Italian toxic language on social media. Our analysis reveals that toxic comments in Italian tend to be longer and have fewer unique emojis compared to non-toxic comments, while both exhibit similar lexical diversity. To evaluate the impact of linguistic features on state-of-the-art models' performance, we fine-tune three pre-trained language models (PoliBERT, UmBERTo, and bert-base-italian-xxl-uncased). Despite their significant correlation with comments' toxicity, the inclusion of linguistic features worsens the best model's performance.

1. Introduction

Warning: *This paper contains comments that may be offensive or upsetting.*

Throughout the last 10 years, the rapid growth in social media usage has exacerbated the issue of toxic language (Aljero and Dimililer 2021). Because platforms such as Facebook and Twitter make interactions between individuals faster, easier, and oftentimes anonymous, they are ideal environments for the propagation of harmful content (Del Vigna et al. 2017). Such content may be targeted at an individual or at groups of individuals (De Maiti, Fišer, and Ljubešić 2020) and carried out by individuals or groups of individuals (Del Vigna et al. 2017). Also, online toxic language has been shown to incite and drive violent acts in the offline world (Siege 2020).

In the context of Natural Language Processing (NLP), this phenomenon is often referred to as hate speech (HS). However, the use of the term HS is problematic since it is associated with the legal context. In other words, in legal systems, HS is used to indicate a type of rhetoric that can be prosecuted (Siege 2020). Here and throughout, we use the term "toxic language" instead. Toxic language is "an inclusive term, stretching over subfields such as abusive and offensive language, hate speech, and cyberbullying" and includes “prosecutable hate speech [...] but also not prosecutable but still indecent and immoral insults and obscenities" (Fišer, Erjavec, and Ljubešić 2017). By doing so, we extend the scope of our analysis to online phenomena that are not legally prosecutable and insert the present contribution in a coherent and comparable framework for the

[*] CLiPS Research Center, Faculty of Arts, Prinsstraat 13, B-2000, Antwerp, Belgium.
E-mail: `leonardo.grotti@uantwerpen.be`

analysis of online harmful content (Fišer, Erjavec, and Ljubešić 2017; De Maiti, Fišer, and Ljubešić 2020; Gevers, Markov, and Daelemans 2022).

Regardless of its definition, legal authorities, social media platforms, and companies have shown an increasing interest in countering this phenomenon (Grotti and Quick 2023; Markov and Daelemans 2021). Facebook, Twitter, and YouTube, as well as other websites, often ban toxic language. However, research has highlighted how companies often do not dispose of well-organized control systems and often rely too much on users to signal comments/posts (Sanguinetti et al. 2018). Moreover, manually filtering messages containing toxic language has proven to be not only highly time-consuming but also damaging for human annotators (Zampieri et al. 2019). Furthermore, human-labeled data has been shown to reflect annotators' individual biases (Markov and Daelemans 2021).

Recent surveys (Fortuna and Nunes 2019; Poletto et al. 2021; Yin and Zubiaga 2021), however, have highlighted a number of limitations related to the automated detection of toxic language: first, even though languages other than English have received increasing attention (Poletto et al. 2021), researchers have not addressed the task systematically (Nozza, Bianchi, and Attanasio 2022). Also, scholars often do not agree on what constitutes toxic language and how it differs from, e.g., offensive or aggressive language (Caselli et al. 2018). Finally, many datasets compiled and annotated for research in this field remain unavailable (Fortuna and Nunes 2019; Plaza-del arco, Nozza, and Hovy 2023) and their labels are strongly biased and are often not comparable (Yin and Zubiaga 2021). Because of the above-mentioned issues, the generalisability of toxic language detection models remains so far the biggest challenge (Poletto et al. 2021).

The Linguistic Landscape of Hate Speech in Social Media LiLaH project[1] tries to address these issues. The objective of LiLaH is to create a common framework for the analysis of toxic language in multiple languages (for now, Dutch, English, French, Slovene, and Croatian). These efforts include: creating comparable benchmark datasets (Ljubešić, Fišer, and Erjavec 2019), such as the FRENK dataset, analyzing the grammatical (De Maiti, Fišer, and Ljubešić 2020) and lexical (Ljubešić et al. 2020) features, and giving a working definition of HS (Socially Unacceptable Discourse, SUD) (De Maiti, Fišer, and Ljubešić 2020) and toxic language (Gevers, Markov, and Daelemans 2022).

This contribution aims to create a similar framework for the Italian language. As noted by (Nozza, Bianchi, and Attanasio 2022), scholars interested in the detection of HS in Italian have put a great effort into improving the models (see, e.g., the three latest EVALITA shared tasks, in (Bosco et al. 2018; Sanguinetti et al. 2020; Lai et al. 2023) for an overview); however, their efforts have lacked a coherent approach to the task. In other words, researchers have approached the task without a common framework of analysis (e.g., lacking a common definition for the investigated phenomenon, comparable results, systematic experiments, etc.).

As such, the paper presents a combination of quantitative analysis and Natural Language Processing (NLP) techniques. First, following (De Maiti, Fišer, and Ljubešić 2020; Gevers, Markov, and Daelemans 2022), it assesses the difference between toxic v. non-toxic comments' average length, vocabulary diversity, and linguistic standardness. Then, similarly to (Markov, Gevers, and Daelemans 2022), we verify whether the inclusion of linguistic features improves toxic language-detection models.

The analysis is carried out on the HaSpeeDe Facebook dataset (Bosco et al. 2018) containing Facebook comments extracted from a set of Italian groups. This particular

[1] https://lilah.eu/

dataset was chosen because, among the existing Italian toxic language corpora[2], it aligns most closely with those discussed within the LiLaH project. The similarities include Facebook page selection mode and inter-annotator agreement

In the remainder of this paper, Section 2 discusses the related work. Section 3 outlines our methodology. In Section 4, we first report on the findings about linguistic features of toxic language. Then, we test different language models for the task of toxic language detection. Finally, we explore how the inclusion of the investigated linguistic features through a GradientBoostClassifier[3] algorithm affects the best model's performance. We further explore and analyze our results with an in-depth error analysis. Finally, Section 5 presents the conclusions.

2. Related work

2.1 Hate Speech, Toxic Language, or Socially Unacceptable Discourse?

Hate speech is a complex concept: scholars, legal advisors, policymakers, and ethical commissions have all long debated its meaning (Siege 2020). However, some points of agreement exist in the literature. For instance, it is commonly agreed (Sanguinetti et al. 2018) that what is considered HS is prohibited and does not—or rather, cannot—fall under the right to freedom of expression. In other words., HS is a term that has legal implications depending on the country. Additionally, HS is often understood "to be bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived characteristics"[4] (Siege 2020).

Different countries and social media use the term hate speech and give different definitions. For the sake of completeness, we here report a selection of these. The European Union (EUR-Lex 2008) defines hate speech as "the public incitement to violence or hatred based on certain characteristics, including race, color, religion, descent, and national or ethnic origin"[5]. However, the decision on whether to extend it to gender identity, sexual orientation, and disability depends on the country. Similarly, social media companies have different definitions. For example, Facebook defines HS as “a direct attack against people based on what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease” (Meta 2022). Other social media, such as YouTube, extend this definition to people with a veteran status and victims of a major violent event and their kin and distinguish it from harassment and cyberbullying (YouTube 2019).

In automatic detection studies, many have adopted the term HS (Poletto et al. 2021). This preference dates back to earlier publications in the field (Warner and Hirschberg 2012; Djuric et al. 2015; Gitari et al. 2015). However, the implications of using such a term are too constraining: as highlighted above, it is too dependent on its legal context. Furthermore, other phenomena that do not always fall under its scope (e.g., personal

[2] for an overview, (Poletto et al. 2021).

[3] See the https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.htmlScikit-Learn documentation on GradientBoostingClassifier.

[4] Note that the term 'bias' is here used to indicate the offender's personal bias. I.e., insults are targeted at certain (groups of) people because the offender identifies the victim as a member of some group

[5] Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia through criminal law: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l33178

attacks and cyberbullying) also deserve attention (Gevers, Markov, and Daelemans 2022). To avoid such issues, scholars have adopted two strategies. The first one is to use umbrella terms to indicate harmful content: these include offensive language (Davidson et al. 2017), Socially Unacceptable Discourse (SUD, (De Maiti, Fišer, and Ljubešić 2020), and abusive language (Vidgen et al. 2019)—to mention a few. The second is to study subcategories of HS, such as misogyny (Bosco et al. 2018; Sanguinetti et al. 2020) and homophobia (Akhtar, Basile, and Patti 2019).

The result is a non-coherent framework in which the same label is used to describe a variety of phenomena and the same phenomenon is labeled differently. As such, we follow (De Maiti, Fišer, and Ljubešić 2020) and (Gevers, Markov, and Daelemans 2022) recommendation: rather than HS, we use the term toxic language, which covers a variety of harmful speech (ranging from cyberbullying and obscenity to harassment and threats), both prosecutable and non-prosecutable (Gevers et al, 2022).

In adopting such a definition we keep in mind that the HaSpeeDe Facebook dataset (used for the present paper) was not developed as part of the LiLaH project. As such, while the definitions of toxic language and SUD—as well as the FRENK and LiLaH corpus—were developed with a cohesive framework in mind, the HaSpeeDe Facebook dataset was not.

Although the comparison is not exact, the above definition of toxic language is similar to the description of HS used for the annotation of the HaSpeeDe Facebook dataset (Del Vigna et al. 2017). Even if (Del Vigna et al. 2017) does not provide an exact definition of HS, they use it to talk about online harmful content in general, including subcategories of HS (e.g., threats, cyberbullying, incitement to violence, profanity, and trolling) but also broader trends such as offensive and abusive language. Furthermore, (Del Vigna et al. 2017) do not specify being legally prosecutable as a necessary condition for the identification of HS.

By adopting a common definition to those in other studies (De Maiti, Fišer, and Ljubešić 2020; Gevers, Markov, and Daelemans 2022) and comparing our results to theirs, the aim is to create a common framework for the investigation of toxic language in more languages.

2.2 Computer Mediated Communication (CMC) and Social media language

The previous section discussed what toxic language is and how it relates to other online phenomena. However, toxic language is present not only online and is shaped differently according to the *medium* of propagation. For instance, (Baumgarten, Bick, and Geyer 2019) have demonstrated how speakers manipulate prosody to change a listener's perception of toxic language. In other words, verbal cues can be used to render toxic language sarcastic or ironic in face-to-face communication. Moreover, (Gevers, Markov, and Daelemans 2022) suggest that various social media platforms have different character limits, which influences toxic language's linguistic characteristics. Thus, before proceeding any further in our analysis, we here offer a brief summary of how language varies in social media and CMC in general.

Scholars have investigated how CMC impacts various languages. (Yin and Zubiaga 2021) report that non-standard English is fairly common on social media platforms: omission of punctuation, alternative spellings, the use of code words to disguise toxic language, and unconventional capitalization to achieve emotional emphasis are the most common characteristics. (Hilte, Vandekerckhove, and Daelemans 2017) have carried out a similar analysis of adolescents' chat conversations in Dutch, highlighting seven expressive markers that characterize linguistic non-standardness, such as character/letter flooding, renderings of kisses and laughs, the use of emoticons/emojis, non-standard capitalization, and unconventional punctuation markers combinations. Finally, (De Maiti, Fišer, and Ljubešić 2020) report that shortenings are a common feature of Slovene Twitter messages and extend beyond orthography to lexicon and syntax. Furthermore, (De Maiti, Fišer, and Ljubešić 2020) analyzed the strandedness of toxic language in the FRENK dataset: the use of emojis/emoticons to mitigate a message's illocutionary force and unconventional punctuation is typical of social media comments.

When it comes to Italian, (Frenguelli 2020) has shown how CMC has changed the linguistic norm: the rise of social media has made the acceptance and spread of neologisms and otherwise ungrammatical linguistic forms easier. (Sedda and Demuru 2019) have focused on the markers of online populism in social networks. Their analysis highlights that CMC Italian has similar characteristics to the ones found in other languages (Sedda and Demuru 2019). These include unconventional capitalization, punctuation, and syntax, which are used to catch the attention of the reader and incite similar messages.

From the above summary, we can observe that CMC affects linguistic forms across different languages in similar manners. Unconventional linguistic forms seem to be common in the mentioned languages and are present at different levels of a language: punctuation (e.g., unconventional capitalization), grammar, semiotics (e.g., emojis/emoticons), lexicon (e.g., unconventional spellings), and syntax. As such, one must take these factors into account when analyzing toxic language online.

2.3 Toxic language detection

Early efforts in the field of automatic toxic-language detection date as far back as 1997. Although early contributions focused on specific subcategories of toxic language, their impact laid the foundation for the development of more comprehensive approaches to detecting and mitigating toxic language in online communication. (Spertus 1997) used a decision tree classifier to classify hostile messages from online feedback forms based on semantic and syntactic features that were extracted from the texts. (Greevy and Smeaton 2004) were amongst the first to suggest a Support Vector Machine approach: they apply SVM to classify racist documents. For cyberbullying, (Karthik, Reichart, and Lieberman 2011) explored the effectiveness of topic-specific classifiers.

Such interest is reflected in the numerous tasks that are concerned with toxic language and some of its subcategories: Aggression Identification (Kumar, Lahiri, and Ojha 2021), Offensive Language Identification (Zampieri et al. 2019), Misogyny Identification (Fersini, Rosso, and Anzovino 2018; Gajo et al. 2023), and HS detection in Italian Facebook and Twitter messages (Bosco et al. 2018; Sanguinetti et al. 2020; Lai et al. 2023), to mention a few. The available models for toxic language detection have not only increased in number but also quality: as noted by (Markov, Gevers, and Daelemans 2022), the "large amount of user-generated content available on social media and the arrival of transformer-based pre-trained language models" has significantly improved accuracy.

Despite the advancements in the field, (Yin and Zubiaga 2021) have noted how state-of-the-art models have often been overestimated. Generalisability, or a model's performance on a dataset different from the training one (Wiegand, Ruppenhofer, and Kleinbauer 2019), is one of the primary goals of automatic detection systems (Yin and Zubiaga 2021). A model's ability to perform in cross-domain (i.e., coming from different domains, such as Facebook and Twitter) and cross-genre (i.e., text belonging to different genres, e.g., social media posts and journal articles) setups is highly desirable. Thus, scholars, e.g., see (Markov and Daelemans 2021), have put great effort into improving cross-domain performance.

For Dutch, (Markov, Gevers, and Daelemans 2022) proposed an ensemble method combining BERTje and RoBERT pre-trained models with an SVM algorithm, achieving improved performance with features like comment length and personal pronouns. In Italian, efforts center around EVALITA shared tasks (Bosco et al. 2018; Sanguinetti et al. 2020; Lai et al. 2023), with participants using pre-trained models like ALBERTo and UmBERTo (Polignano et al. 2019; Parisi, Francia, and Magnani 2020; Nozza, Bianchi, and Attanasio 2022; Grotti and Quick 2023) addressed data scarcity for Italian by leveraging multi-language pre-trained models (XLM-base, XLM-large, and mBERT)[6]. Furthermore, (Fersini, Nozza, and Boifava 2020) have explored how lexical components, punctuation, and pragmatic particles can be leveraged to improve the performance of machine learning algorithms (such as SVM, Naïve Bayes, and Multi-Layer Perceptron) for misogyny detection. While their findings highlighted an improvement in detecting misogynistic comments, the authors (Fersini, Nozza, and Boifava 2020) did not explore how these features could be incorporated in transformers-based models and suggested that the inclusion of more stylometric features could further improve the models' performance.

However, several issues prevent models from generalizing well: (Fortuna and Nunes 2019) report low agreement between human annotators and the constant evolution of language. (Poletto et al. 2021) also note that the frequent use of toxic lexicon often introduces bias in datasets. In simpler terms, using biased or offensive language in datasets can make models focus too much on those specific patterns, causing them to become overly specialized and potentially unfair. (Yin and Zubiaga 2021) further mention that misspellings and code words are often problematic as models do not recognize them; this issue is amplified in languages other than English due to the limited availability of data (Arango, Pérez, and Poblete 2021). Thus, even models like Long Short-Term Memory (LSTM), and Convolutional Neural Network-Gate Recurrent Unit (CNN-GRU), drop as much as 30 points in macro-averaged F1 when tested on cross-domain datasets (Yin and Zubiaga 2021).

3. Data and Method

3.1 Hypotheses

As seen in Section 2, CMC affects language at different levels (e.g., syntax, punctuation, lexicon, etc.) in Italian. To our knowledge, no existing research has explored in detail how toxic and non-toxic comments differ in social media communications for Italian. Previous literature on Italian toxic language detection has not provided a systematic study of the differences between toxic and non-toxic comments. As such, we formulate our hypotheses to investigate the linguistic differences between toxic and non-toxic comments on Facebook. Because previous research (De Maiti, Fišer, and Ljubešić 2020; Gevers, Markov, and Daelemans 2022) highlighted a significant difference in length between toxic and non-toxic comments, our first hypothesis explores this aspect for Italian. Then, we move on to lexical diversity and linguistic standardness. We do so because past research (Yin and Zubiaga 2021) has shown how the inclusion of such features

[6] Unfortunately, the model resulting from the latter paper is not available on Hugging Face as the authors have decided to not make the model publicly available.

can improve a model's performance. Furthermore, to make the results of our research comparable to those found in (De Maiti, Fišer, and Ljubešić 2020) and (Gevers, Markov, and Daelemans 2022), we formulate the same hypotheses. Thus, in the qualitative part of the study, we answer the following research questions:

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

Our hypotheses are formulated based on the results obtained in previous research on Slovene (De Maiti, Fišer, and Ljubešić 2020), English, and Dutch (Gevers, Markov, and Daelemans 2022). As mentioned before, the HaSpeeDe dataset was not developed using the same annotation guidelines as the LiLah and FRENK datasets. Thus, we do not expect previous findings to perfectly generalize to our research. For average length, we expect toxic comments to be longer than non-toxic ones: although research on Slovene has highlighted that toxic and non-toxic comments have a similar length (De Maiti, Fišer, and Ljubešić 2020), (Gevers, Markov, and Daelemans 2022) have found that toxic comments are longer in both English and Dutch. The latter study also suggested that toxic comments have a greater vocabulary diversity in both English and Dutch, while non-toxic comments seem to contain a higher relative number of emojis and a total number of unique emojis. We expect these results to partially generalize to our research.

With regards to linguistic standardness in Dutch comments, (Gevers, Markov, and Daelemans 2022) has noted that toxic comments are overall less standard since comments tend to be more expressive and such expensiveness is transmitted through non-standard forms (Gevers, Markov, and Daelemans 2022). However, when it comes to the Punctuation-to-Non-punctuation ratio (PNR), there was no significant difference between the two categories. Since we analyzed a similar dataset, we expect our findings to be similar to the ones described above.

3.2 Data

The data was selected based on a similarity criterion. I.e., the dataset should be as similar as possible to those used in (De Maiti, Fišer, and Ljubešić 2020) and (Gevers, Markov, and Daelemans 2022). The selection of a similar dataset to (De Maiti, Fišer, and Ljubešić 2020) and (Gevers, Markov, and Daelemans 2022) facilitates a meaningful comparison, enabling a more accurate evaluation of our results. Additionally, employing a comparable dataset allows us to assess the generalizability of the models deployed. Of the seven publicly available Italian datasets described in (Poletto et al. 2021), one has been annotated only for homophobia (Akhtar, Basile, and Patti 2019), two are multilingual and include short texts (hate-speech v. counter-narrative pairs) (Chung et al. 2019) or comments from news websites (Steinberger et al. 2017), three contain comments only from social media other than Facebook (i.e., Twitter) (Poletto et al. 2017; Sanguinetti et al. 2018; Corazza et al. 2019). Thus, the choice fell on the HaSpeeDe 2018 Facebook dataset. What follows is a brief description of the FRENK (Ljubešić, Fišer, and Erjavec 2019), LiLaH (Markov et al. 2021), and HaSpeeDe (Del Vigna et al. 2017) datasets. A detailed description of each dataset can be found in Appendix A.

The FRENK dataset includes Facebook comments from three mainstream Slovene newspaper pages, with a focus on discussions about migrants and LGBTQ+. In total, 6,545 comments about migrants and 4,517 about LGBTQ+ were analyzed by (De Maiti, Fišer, and Ljubešić 2020). Thirty-two master students annotated the data based on the presence of SUD and provided more detailed annotations for SUD comments, including background violence, targeted content, and the type of violence. The inter-annotator agreement exceeded the acceptable threshold of 0.66. (Gevers, Markov, and Daelemans 2022) also analyzed the English part of the FRENK dataset, following the same guidelines. The LiLaH dataset, focusing on Dutch content, contains 10,732 comments from Flemish newspaper Facebook pages, annotated using similar guidelines. The inter-annotator agreement for LiLaH was between fair and good (0.56). The HaSpeeDe dataset, originating from Italian Facebook pages, involved crawling 17,567 comments. The annotation process, conducted by five bachelor students, categorized comments into three levels and various types of hate speech. The inter-annotator agreement was measured using Fleiss' kappa κ and reached $\kappa = 0.26$ when merging two classes. To offer a better overview of the data, Table 1 below summarizes the data distribution across the three sub-corpora.

Table 1
Toxic language distribution across FRENK, LiLaH, and HaSpeeDe.

	Train		*Test*		*Total*
Dataset	**Toxic**	**Non-toxic**	**Toxic**	**Non-toxic**	**Toxic\Non-toxic**
HaSpeeDe$_{\textbf{Italian}}$	1,382	1,618	677	323	4,000
FRENK$_{\textbf{Slovene}}$	3,506	3,238	882	821	8,847
FRENK$_{\textbf{English}}$	2,848	5,091	744	1,351	10,034
LiLaH$_{\textbf{Dutch}}$	3,753	4,821	949	1,209	10,732

3.3 Lexical Diversity

We consider lexical diversity to be expressed by four different features: type-to-token ratio (TTR), Content-to-function-words ratio (CFR), number of unique emojis, and relative number of emojis. Because of the difference in length between the comments, we report the mean and median TTR and CFR for 100 random samples of 1000 tokens.

Type-token ratio To calculate the TTR, we first removed all punctuation. Then, all occurrences of user tags, URLs, and emojis were substituted with the token TAG, URL, and EMOJI respectively. To tokenize the comments, we used the nltk Italian tokenizer, which was further modified to handle cases where an apostrophe was present to ensure correct tokenization. Finally, the TTR for each comment is obtained by dividing the number of unique words (types) by the total number of words (tokens).

Content-to-function-words ratio CFR was calculated using the NLTK library (NLTK 3.7): words that were not included in the NLTK Italian stopwords list were considered content words. Thus, the CFR is calculated by dividing the number of content words by the number of tokens.

Emojis The number of unique emojis and the relative number of emojis were computed using the PiPy package emoji (emoji 2.2.0).

Correlation coefficient Following (Gevers, Markov, and Daelemans 2022), we also compute the correlation coefficient between each feature and the comment's toxicity. Because the toxicity annotation is binary (i.e., toxic vs. non-toxic comments), we use SciPy biserial correlation coefficient[7]. Each feature was added to the original dataset in a separate vector and compared to the binary toxicity annotation.

3.4 Linguistic standardness

Tagging comments' linguistic standardness is a complex task. Authors have adopted different approaches: (De Maiti, Fišer, and Ljubešić 2020) manually annotated comments based on codified spelling and grammar standards; in other words, (De Maiti, Fišer, and Ljubešić 2020) used the Solvene Normative Orthography Guide to determine which forms were to be considered standard. Their annotation schema consisted of four categories[8], each having a range of subcategories. (Hilte 2019) and (Gevers, Markov, and Daelemans 2022) have instead used a computational approach: i.e., linguistic standardness is encoded in a set of linguistic features that can be calculated automatically. (Hilte 2019) investigated chat conversations between adolescent Dutch speakers. As such, not all the identified features are relevant in toxic language analysis. We use the same code used in (Hilte 2019); however, the code was adapted for the analysis of Italian. Before the analysis was carried out, the dataset was grouped based on the toxicity label to obtain values for toxic vs. non-toxic comments.

Flooding Flooding is defined as the deliberate repetition of a character (Hilte, Vandekerckhove, and Daelemans 2017). Like (Gevers, Markov, and Daelemans 2022), we consider flooding any repetition greater than or equal to three characters.

Emoticons and emoji Unicode emojis, 'western emoticons' (e.g., ":P"), 'hearts' ("<3"), and 'Asian emoticons' were encoded and detected with the use of regular expressions and the emoji package.

Unconventional capitalization (Hilte 2019) highlights three different forms of unconventional capitalization: all capitalized (e.g., HELLO), alternate capitalization (e.g., HeLlO), and inverse capitalization (e.g., hELLO).

Combination of question and exclamation marks Combinations of questions and exclamation marks were encoded using regular expressions.

Laughter We encode all occurrences of 'hahaha', 'hihihi', and 'ahahah' (the latter has been added specifically for Italian as it is commonly used) using regular expressions.

[7] Note that SciPy biserial correlation coefficients package uses a shortcut formula but the final results is the same that would be obtained with a *Pearson's ρ*. See `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pointbiserialr.html`SciPy stats.

[8] Orthography, morphology, syntax, and word order

3.5 Model

To further explore the relationship between lexical diversity/linguistic standardness and toxic language detection, a set of experiments was run to determine how the inclusion of linguistic features impacts the performance of toxic detection models.

We do so by fine-tuning pre-trained language models. Fine-tuning is a form of transfer learning and an often-used technique in NLP that allows researchers to train a pre-trained model on new data to fit a downstream task (Durrani, Sajjad, and Dalvi 2021). In (Pan and Yang 2010) words:

> Given a source domain D_S and a learning task T_S, a target domain D_T and a learning task T_T, inductive transfer learning aims to help improve the learning of the target predictive function $f_T(.)$ in D_T using the knowledge in D_S and T_S, where $T_S \neq T_T$.

(Howard and Ruder 2018) have highlighted how fine-tuning has been successfully used to transfer between similar tasks. Furthermore, this approach has been widely used in toxic language detection in Italian[9].

In this paper, we fine-tune three popular Italian pre-trained models:

- bert-base-italian-xxl-uncased[10] is a BERT-based model which was trained on over 80GB of data (13 billion tokens). The model was pre-trained on a combination of data which includes the OPUS and OSCAR (Open Super-large Crawled ALMAnaCH coRpus) corpora as well as a Wikipedia dump. Note that for ease of readability, we refer to this model as BERT-ita from now on.
- PoliBERT[11] is a BERT-based (bert-base) model which was fine-tuned on Italian political tweets for sentiment analysis.
- UmBERTo[12] is a RoBERTa-based model which was trained on the OSCAR Italian large corpus[13]. The model is used for both Named Entity Recognition (NER) and Part-of-speech (POS) tagging and reached excellent performance on different datasets.

All three models were fine-tuned using PyTorch Trainer[14] on a random sample (75%) of the training data across three epochs and evaluated on the remaining 25%. We select this specific split to mirror the task's original train-test split (3000-1000). The models were then used to make predictions on the test data. The text was lowercase and emojis were removed.

Next, we select the best-performing model and conduct an ablation study on the effect of the investigated linguistic features on the model's performance. The additional features were computed for both the training and the test set and implemented one at a time into the model through GradientBoosterClassifier, an ensemble algorithm that se-

[9] E.g., (Lavergne et al. 2020; Tamburini 2020; Nozza, Bianchi, and Attanasio 2022).

[10] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased?text=Roma+%C3%A8+la+%5BMASK%5D+d%27Italia.

[11] `https://huggingface.co/unideeplearning/polibert_sa`

[12] https://github.com/musixmatchresearch/umberto

[13] 70gb of plain text

[14] `https://huggingface.co/docs/transformers/training#train-with-pytorch-trainer`

quentially trains weak models, ultimately producing a strong model that is a weighted sum of the weak models. Differently from other algorithms, GradientBoosterClassifier uses decision trees as the weak learners and it is optimized via gradient descent.

To incorporate the linguistic features into our model, we initially generate output labels from the best-performing transformer-based model. These labels are subsequently included as input features for training the GradientBoostingClassifier, alongside the linguistic features. The linguistic features are implemented one at a time to assess their respective contributions to improving or worsening the model's overall performance. The predictions were then computed on the test set and compared to the performance of the best model with no linguistic features. The aim is to verify whether the inclusion of linguistic features improves or worsens toxic language classifiers.

4. Results and discussion

4.1 Quantitative analysis

In this subsection, we report the results obtained from the quantitative analysis of the data. Each section answers one of the three research questions presented in Section 4.1. The quantitative analysis was carried out on the full dataset (i.e., the train and test set were merged for the analysis, for a total of 4000 comments).

4.1.1 Average length

We computed the mean and median length of toxic and non-toxic comments in tokens. Our analysis shows that toxic comments are, on average, 13.22 tokens long, with a median of 10 tokens. In contrast, non-toxic comments are 8.87 tokens long on average and their median was measured at 5 tokens. Thus, we can accept Hypothesis 1.1: toxic comments are, on average, longer than non-toxic ones.

4.1.2 Lexical diversity

To calculate lexical diversity, we computed four different measurements: TTR, CFR, normalized emoji frequency, and the number of unique emojis. Because of the limited size of the dataset and the difference in comment length, TTR and CFR were calculated on 100 samples of 1000 tokens. Additionally, we also report the median and mean values.

For TTR, we first observe that toxic and non-toxic comments have, on average, the same type-token ratio: 0.95. This is considered a high TTR because a value close to 1 suggests that the text consists mostly of unique words, indicating a rich and varied vocabulary rather than repetitive or redundant word usage. We hypothesize that such high TTR may be caused by the large presence of alternative spellings or an effect of text length since the TTR was not normalized for comment length. However, when computed over samples of 1000 tokens across different comments, the TTR decreases significantly to 0.53 for toxic comments and 0.52 for non-toxic ones. Next, we look at CFR. Toxic comments have a slightly lower CFR (0.61) than non-toxic ones (0.63). Similarly to TTR, the CFR decreases to 0.44 and 0.43 for toxic and non-toxic comments respectively when computed over 100 random samples of 1000 tokens. This pattern indicates that, overall, both toxic and non-toxic comments have a low vocabulary diversity within our dataset. Given our observations, we reject hypothesis 2.1. Although toxic comments have slightly higher TTR and CFR, such difference is minimal. As such, we can conclude that toxic and non-toxic comments have similar vocabulary diversity.

This is in contrast with (De Maiti, Fišer, and Ljubešić 2020)'s findings, which suggested that toxic comments tend to have higher lexical diversity due to the more creative and colorful nature of the message.

Finally, we look at the number of unique emojis and the relative number of emojis and emoticons. Our analysis highlights that non-toxic comments have both a higher number of unique emojis (54) in total and the relative number of emojis and emoticons (0.011) per comment compared to toxic comments (39 unique emojis, 0.004). This is consistent with the literature: looking for emojis or typing emoticons is often considered too time-consuming when it comes to writing emotionally intense comments (Bočková 2019). We thus accept hypothesis 2.2: non-toxic comments contain more emoticons and emojis than toxic comments. The findings are summarized in Table 2.

Table 2
Lexical diversity features for toxic and non-toxic comments.

	Toxic	Non-toxic
Average length	13.22 tokens	8.87 tokens
Type-token ratio	0.53	0.52
Content-to-function-words ratio	0.44	0.43
Relative frequency emoji	0.004	0.011
Unique emoji	39	54

All the investigated characteristics correlated significantly (threshold $p < 0.01$) to the toxicity of the comments except for TTR. The correlation coefficients show that comment length is positively correlated to the comment's toxicity. I.e., toxic comments are associated with a higher number of tokens. In contrast, CFR, the number of unique emojis, the relative number of emojis and emoticons, as well as the punctuation-to-non-punctuation ratio (PNR) are all associated with non-toxic comments. Table 3 shows the correlation coefficients for all the investigated characteristics. It is worth noting that all correlation coefficients are rather low. As such, we expect that they will not have any significant impact on the model's performance.

Table 3
Correlation coefficients between lexical features and type of comments (toxic and non-toxic). Positive coefficients denote a positive correlation with the toxic category.

	Coefficient	P-value	Significance
Average length	$+0.18$	$1.01^{\text{e-30}}$	**
Type-token ratio	$+0.006$	0.66	None
Content-to-function-words ratio	-0.05	0.001	**
Relative frequency emoji	-0.06	$5.76^{\text{e-40}}$	**
Unique emoji	-0.05	0.0006	**
Punctuation-to-non punctuation ration	-0.06	$1.7^{\text{e-05}}$	**

4.1.3 Linguistic non-standardness

We now discuss the findings related to the comments' linguistic non-standardness. The investigated features are based on (Hilte 2019)'s and (Gevers, Markov, and Daelemans

2022)'s research on linguistic standardness. The results described below are summarized in Table 4. First, we computed flooding for both letters and punctuation. As a reminder, we investigated all combinations of letters or punctuation longer than three characters (e.g., 'nooo', '!!!').

Non-toxic comments have a higher relative frequency when it comes to letter flooding. Of the 17,223 tokens that form the non-toxic comments, 141 present letter flooding (0.008). In the toxic comments, 136 tokens are 'flooded'; however, because the toxic corpus is larger (27,214 tokens) than the non-toxic one, the relative frequency of letter flooding is lower (0.004). A more in-depth analysis of the comments revealed that in non-toxic comments the most flooded letters (i.e., 'i', 'e', and 'o') are often positive adverbs, such as 'vaiii'[15], 'certooo'[16], or adjectives, e.g., 'grandeee'[17]. In toxic comments, on the other hand, the most frequent letter flooding ('a' and 'i') is used to emphasize offensive nouns (e.g., 'ruspaaaa'[18] or 'gheiiiiii'[19]).

For punctuation flooding, we calculated how many times the '!' and '?' marks are flooded. Additionally, we determined the relative frequency of punctuation flooding. I.e., how many of the total number of '!' and '?' occurrences are flooded. We find that a high percentage of punctuation tokens presents flooding for both toxic (54%) and non-toxic (44%) comments. It is worth noting that the top occurring !- and ?-variants are longer in toxic comments ('??????' v. '???', '!!!!!!' v. '!!!!').

Given the high percentage of punctuation flooding, we also analyzed the combinations of exclamation and question marks (i.e., '!?'). The results show no significant difference between the two types of comments (12 occurrences in non-toxic comments, 10 in toxic). As noted by (Parkins 2018, 2018; Hilte 2019), flooding is a creative tool used to communicate a writer's emotional states (both positive and negative) or a message's intensity. This is reflected in our data: toxic comments convey a stronger intensity through longer punctuation flooding variants and emphasize negative words by flooding their endings. For the '!?', (Gevers, Markov, and Daelemans 2022) pointed out that individual or restricted groups of authors may influence the results, which seems to be the case in our data.

Following our analysis of specific punctuation marks, we compute the average PNR for toxic and non-toxic comments. Despite a high presence of punctuation flooding, toxic comments have a lower PNR (0.11) in comparison with non-toxic ones (0.17). In turn, this pattern implies that on average toxic comments use less punctuation.

Finally, we looked at unconventional capitalization laughter. Unconventional capitalization occurs more in non-toxic comments (1124). Although such a difference may appear significant, it is not when we look at the relative frequency: 0.03 for non-toxic comments compared to toxic comments' 0.02. Moreover, it is worth noting that in both cases 99% of unconventional capitalizations are entire word capitalization, considered the most common and expressive type of capitalization (Hilte 2019). For laughter, we found 49 instances of laughter in non-toxic comments (0.002 of all tokens) and 22 in toxic ones (0.0007). We further report that in both cases 50% of the occurrences are ahahah-variants, a type of laughter that was specifically added for the analysis of Italian. No instances of hihihi-variants were retrieved.

[15] 'Let's go'
[16] 'Sure'
[17] 'Great'
[18] 'Bulldozer', a motto often used to reference a proposal to remove Romani's camps using a bulldozer
[19] 'Gay'

Table 4
Linguistic standardness features for toxic and non-toxic comments.

	Toxic	Non-toxic
Letter flooding	0.004	0.008
Punctuation flooding	0.54	0.44
Combination '!?'	10 occurrences	12 occurrences
Capitalization	0.02	0.03
Laughter	0.0007	0.002
Average PNR	0.11	0.17

Given the above-described results, we conclude that toxic are only in some aspects (i.e., punctuation flooding, average PNR) linguistically less standard than non-toxic comments. We thus reject hypothesis 3. Indeed, it is worth noting that non-standard features are used differently in toxic and non-toxic comments. Punctuation and letter flooding is used to convey a stronger intensity and to emphasize negative words. Likewise, unconventionally capitalized words are often offensive (e.g., 'STRONZE') in toxic comments. On the other hand, non-toxic comments often use flooding to emphasize encouraging words (e.g., 'vaiiii', 'dajeeee', and 'grandeee') and present shorter punctuation flooding. Furthermore, toxic comments use more unconventional punctuation: 54% of all '!' and '?' are flooded and the average punctuation-to-non-punctuation ratio is lower (11%).

4.1.4 Comparison with Dutch, English, and Slovene

Before proceeding with the toxic language detection system, we here offer a brief comparison of our results with those outlined in the existing literature for Slovene (De Maiti, Fišer, and Ljubešić 2020), English, and Dutch (Gevers, Markov, and Daelemans 2022).

Starting from the median average length, our results mirror those found in Dutch, Slovene, and English data: toxic comments are longer in all four languages (Italian, 10_{tox} v. $5_{\text{non-tox}}$, Dutch, 22_{tox} v. $11_{\text{non-tox}}$, English, 21_{tox} v. $14_{\text{non-tox}}$). However, (De Maiti, Fišer, and Ljubešić 2020) found that in Slovene toxic and non-toxic comments have a similar length (12 tokens for the former, 11 for the latter).

Next, we looked at lexical diversity. Unlike Dutch and English, Italian toxic comments have both a higher TTR and CFR (for Dutch, 0.53_{tox} v. $0.57_{\text{non-tox}}$ and 1.47_{tox} v. $1.63_{\text{non-tox}}$ respectively, for English, 0.54_{tox} v. $0.55_{\text{non-tox}}$, and 1.33_{tox} v. $1.38_{\text{non-tox}}$). The results obtained in Italian are closer to those described in (De Maiti, Fišer, and Ljubešić 2020): Slovene non-toxic comments TTR and CFR are lower compared to toxic ones (TTR, 0.61_{tox} v. $0.58_{\text{non-tox}}$, CFR, 1.32_{tox} v. $1.25_{\text{non-tox}}$). It is worth mentioning that like Slovene, differences in TTR (0.53_{tox} v. $0.52_{\text{non-tox}}$) and CFR (0.44_{tox} v. $0.43_{\text{non-tox}}$) are minimal in Italian. Furthermore, Italian comments CFR is significantly lower compared to all other languages.

The unique number of emojis and the relative number of emoticons and emojis is consistently higher for non-toxic comments across all four languages. However, in the literature, the difference between toxic and non-toxic comments is rather low for both Slovene (unique emojis, 25_{tox} v. $35_{\text{non-tox}}$, relative number, 0.005_{tox} v. $0.009_{\text{non-tox}}$) and English (unique emojis, 37_{tox} v. $130_{\text{non-tox}}$, relative number, 0.001_{tox} v. $0.001_{\text{non-tox}}$). One exception is Dutch. (Gevers, Markov, and Daelemans 2022) found a significant difference in the Dutch corpus: 251 emojis were found in the toxic comments (relative frequency 0.005) while 363 were retrieved from non-toxic ones (0.013). Our findings align with those in Slovene and English for the number of unique emojis. In Italian, the number of unique emojis in toxic comments is slightly lower than the number for non-toxic ones (39_{tox} v. $54_{\text{non-tox}}$). Nonetheless, the relative frequency is significantly lower: 0.004_{tox} v. $0.011_{\text{non-tox}}$). There are two possible explanations for this pattern. First, the overall number of tokens in the toxic comments is higher than the non-toxic ones. Second, non-toxic comments may also contain a higher number of emoticons, further emphasizing the difference in the relative number of emojis and emoticons.

Finally, the Italian average PNR for toxic and non-toxic comments (0.11_{tox} v. $0.17_{\text{non-tox}}$) reflects the PNR[20] found in Dutch (0.10_{tox} vs $0.11_{\text{non-tox}}$) and Slovene (0.09_{tox} vs $0.12_{\text{non-tox}}$).

Table 5
Lexical diversity feature comparison for Italian, Dutch, Slovene, and English

	Toxic				Non-toxic			
	Italian	Dutch	Slovene	English	Italian	Dutch	Slovene	English
Average length	10	22	12	21	5	11	11	14
Type-token ratio	0.53	0.53	0.61	0.54	0.52	0.57	0.58	0.55
Content-to-function-words ratio	0.44	1.47	1.32	1.33	0.43	1.67	1.25	1.38
Relative Frequency emoji	0.004	0.005	0.005	0.001	0.013	0.013	0.009	0.001
Unique emojis	39	251	25	37	54	363	35	130
Punctuation-to-non-punctuation ratio	0.11_{avg}	0.10	0.09	0.11	0.17_{avg}	0.11	0.12	0.11

Looking at Table 5, it is possible to conclude that the average length and the unique number of emojis are consistent in all languages (except for Slovene, where the length is similar across the two categories). I.e., toxic comments are longer on average than non-toxic comments but contain fewer unique emojis. It is worth mentioning that in Dutch and Italian, there is a more accentuated difference in the relative frequency of emojis, compared to Slovene and English.

Coming to lexical diversity, Italian does not reflect any of the findings in the literature. Both toxic and non-toxic comments have a similar TTR and CFR (although both values are marginally higher for toxic comments). Furthermore, while TTR values are similar to those found in the literature, CFR for Italian comments is overall significantly lower compared to other languages. Thus, while there is no significant difference in lexical diversity between Italian toxic and non-toxic comments, Italian comments have a lower lexical density (i.e., CFR) than English, Slovene, and Dutch. In contrast, the average PNR values are similar to English and Dutch: non-toxic comments have a higher punctuation-to-non-punctuation ratio.

Thus, it can be inferred that lexical diversity in Italian toxic and non-toxic comments is similar, but the lexical density of Italian comments is lower compared to other languages. This pattern suggests that Italian comments are less complex and use fewer unique words and content words compared to comments in other languages. The similar PNR values of non-toxic comments in Italian, English, and Dutch may indicate

[20] Both (De Maiti, Fišer, and Ljubešić 2020) and (Gevers, Markov, and Daelemans 2022) report the overall PNR for toxic and non-toxic comments, not the average. I.e., the PNR was calculated over the entire corpus of toxic and non-toxic comments rather than per comment.

that these languages have similar patterns of punctuation usage in non-toxic comments. Additionally, it may imply that toxic comments use overall less standard punctuation.

4.2 Classification models

In this section, we report the results obtained from fine-tuning BERT-ita, PoliBERT, and UmBERTo. Additionally, we evaluate how the addition of linguistic features affects the model's performance.

4.2.1 Transfer Learning

BERT-ita, PoliBERT, and UmBERTo were fine-tuned using PyTorch Trainer. All three models were trained for 3 epochs[21]. Two training arguments were tuned: learning rate (1^{e-3}, 2^{e-5}, and 5^{e-05})[22] and batch size 8, 16, and 32. In line with the literature on the correlation between batch size and generalization (He, Liu, and Tao 2019), we found that increasing the batch size improved the performance on the evaluation set but worsened generalizability. As such, we found the best combination of parameters to be 8 batch size length with a 5^{e-05} learning rate. To better assess the performance of our models, we also trained a simple SVM classifier using a Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer. Table 6 reports the results of the fine-tuning experiments.

Table 6
UmBERTo and PoliBERT performance on the test set after fine-tuning on (75% of) the training data

		precision	recall	F1 score	accuracy	macro avg
Baseline SVM	non-toxic	0.51	0.69	0.59	0.69	0.67
	toxic	0.82	0.68	0.74		
UmBERTo	non-toxic	0.74	0.67	0.70	0.82	0.79
	toxic	0.85	0.89	0.87		
PoliBERT	non-toxic	0.86	0.71	0.78	0.87	0.84
	toxic	0.87	0.94	0.91		
BERT-ita	non-toxic	0.88	0.71	0.79	**0.88**	**0.85**
	toxic	0.87	0.95	0.91		

To begin with, all models outperform the baseline SVM classifier by a large margin: although the latter performs close to the other models when it comes to precision in predicting the toxic comments class, this is to be expected since toxic comments are the majority class in the test set. As can be seen in Table 6, PoliBERT outperforms UmBERTo across precision, recall, accuracy, and macro avg. The overall better performance can be attributed to the similarity between the dataset original annotation of the HaSpeeDe dataset and the source task's output. (Pan and Yang 2010) have highlighted how the similarity between target and source tasks improves a model's performance. As mentioned before, the HaSpeeDe data (Del Vigna et al. 2017) was annotated on three different levels: not hateful, weak hate, and strong hate. In (Bosco et al. 2018),

[21] The number of three epochs was determined using EarlyStoppingCallback with patience of 3

[22] These learning rates are found in (Nozza, Bianchi, and Attanasio 2022), HuggingFace's fine-tuning guide, and in the standard parameters, respectively

however, the weak and strong hate labels were grouped into one. Likewise, PoliBERT source task (sentiment analysis) assigns three labels probability to a text: 0 for Negative, 1 for Neutral, and 2 for Positive. During the fine-tuning process, the labels were adapted to the two in the downstream classification task. Indeed, the effectiveness of transfer learning may have been improved since the source task, as represented by the original model, exhibits a high degree of alignment with the granularity of the annotations within the target dataset.

BERT-ita outperforms PoliBERT by marginally improving precision for non-toxic comments (0.88 v. 0.86) and recall for toxic ones (0.95 v. 0.94). The better performance of the minority class is likely related to the larger amount of data on which it was pre-trained. As such, the model has been exposed to a larger and more diverse set of examples, which allowed it to learn more robust and generalizable features[23].

PoliBERT and BERT-ita also outperform the best model in the EVALITA 2018 task[24]. While the latter achieved a macro-f1 score of 0.82, PoliBERT reached 0.84. However, the recall for non-toxic comments remains relatively low (0.71) when compared to the overall performance. Figure 1 shows the confusion matrix for BERT-ita.

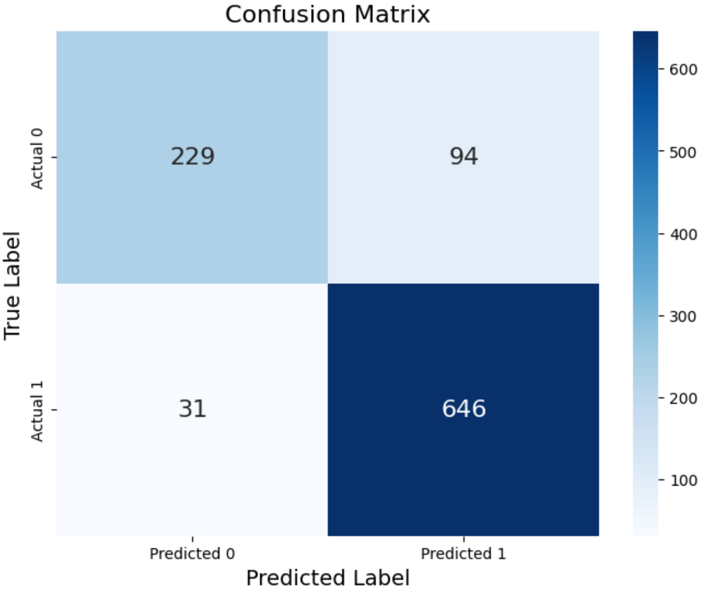

Figure 1
BERT-ita (no additional features) confusion matrix on the test set

In Figure 1, we can observe a high number of false positives (FP, 94, top-right quadrant), in contrast with the low number of false negatives (FN, 31, bottom-left). In turn, this pattern relates to the data imbalance of the test set. In contrast to the training set, the test data contained more toxic comments (677) than non-toxic ones (323). As noted by (Johnson and Khoshgoftaar 2019), the problem can be mitigated through a two-phase learning strategy: i.e., pre-training the model on a thresholded dataset and fine-

[23] For an overview of how a larger amount of pertaining data in the source task can improve models see Pan and Yang (2010)

[24] A Bi-LSTM improved with the use of external lexical resources, see (Cimino, Mattei, and Dell'Orletta 2018).

tuning it on the full version of the same dataset. This method improves the performance of the minority class while maintaining the performance of the majority class. Applying this technique was not possible in our case as a substantial amount of data is needed to implement this solution.

4.2.2 Ablation

For the ablation study, we implement the lexical features discussed in Section 4.1.2 in our best-performing model through GradientBoostClassifier. To do so, we first generate labels for each comment in the training data using BERT-ita. Following, GradientBoostClassifier was trained on these labels and the relevant linguistic features. In other words, we used the output of BERT-ita as ground truth to train the GradientBoostClassifier. Additionally, we also input the desired lexical features together with the label. By doing so, we allowed the model to make predictions based on the additional lexical feature(s). Finally, the resulting model(s) was used to make predictions on the test set.

We first add all linguistic features and see how they affect the model performance. Then, we add them individually and verify whether single features have a specific impact on accuracy and f1-macro. Figure 2 below shows the confusion matrix for BERT-ita and GradientBoostClassifier with the inclusion of all linguistic features.

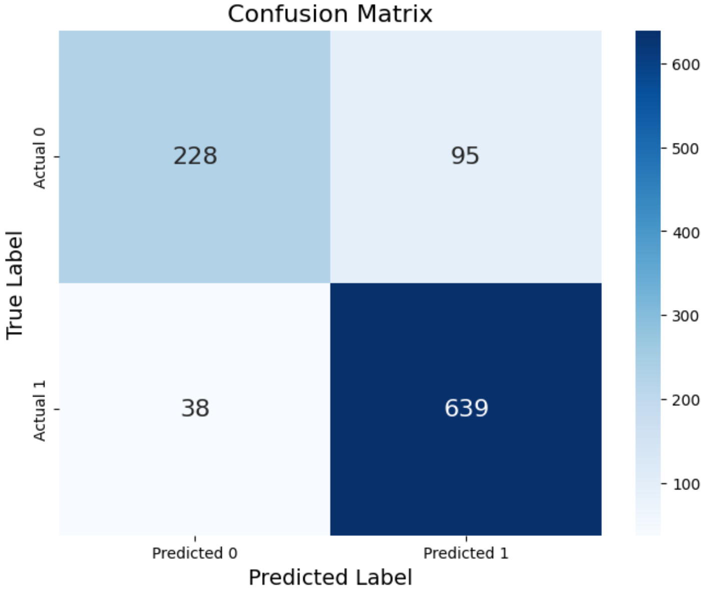

Figure 2
BERT-ita (with additional features) confusion matrix on the test set

Here, we can observe how the inclusion of linguistic features does not improve but rather marginally worsens the model. Compared to Figure 1, the number of true positives and true negatives has marginally decreased (228 v. 229 and 639 v. 649). Thus, we can conclude that including all the investigated linguistic features slightly worsens the model by increasing the FP and FN rates.

Next, we look at each linguistic feature individually. To better capture the difference in performance we also include accuracy and f1-macro in Table 6 below. We additionally report the results for the previously discussed model with all linguistic features to make a comparison.

Table 7
Accuracy and f1-macro of BERT-ita and GradientBoostClassifier (GBC) models with the inclusion of the different linguistic features

	accuracy	f1-macro
BERT-ita	0.875	0.849
$\text{GBC}_{\text{all features}}$	**0.867**	**0.840**
$\text{GBC}_{\text{comment length}}$	0.874	0.848
GBC_{TTR}	0.874	0.848
GBC_{CFR}	0.874	0.847
$\text{GBC}_{\text{unique emojis}}$	*0.875*	*0.849*
$\text{GBC}_{\text{relative n° emoji}}$	0.874	0.848
GBC_{PNR}	0.871	0.844

As shown in Table 6, none of the linguistic features improves the model. Overall, the addition of individual features has only a limited impact on performance. All individual features marginally worsen the model with one exception: the number of unique emojis (in italics). In this case, the performance is the same as BERT-ita without any features. In contrast, the addition of all the features worsens the model more than any other individual feature. This pattern is in accordance with the findings in Section 4.1.2 on individual features' correlation with the comments' toxicity. Although all but one feature (TTR, see previous sections) significantly correlated with comments' toxicity, the coefficients were low. As such, the correlation coefficients are likely too low to improve accuracy and f1-macro. In other words, the models' performance is not improved by the addition of linguistic features since their correlation with the comment's toxicity is too weak.

Our findings are in disagreement with those of (Fersini, Nozza, and Boifava 2020), in which linguistic features improved the classification performance of SVM, Naïve Bayes, and Multi-Layer Perceptron. However, it is worth mentioning that the algorithms and systems used by the authors (Fersini, Nozza, and Boifava 2020) to integrate the linguistic features into the models differ significantly from those used in the present study.

Although there are linguistic similarities across various social media platforms, stylistic variations between Twitter and Facebook are expected due to differences in the genres of content analyzed (i.e., $\text{Tweets}_{\text{Twitter}}$ vs. $\text{Comments}_{\text{Facebook}}$) and the thematic focus. Misogynistic tweets typically center on gender-related issues and prevalent stereotypes (Hewitt, Tiropanis, and Bokhove 2016), whereas comments in our data reflected broader forms of discrimination and intolerance[25].

Therefore, while linguistic features proved beneficial in (Fersini, Nozza, and Boifava 2020) for Twitter-based misogyny profiling, our findings highlight the need for more fine-grained approaches when applying these techniques to different social media platforms and classifying distinct types of hate speech and discriminatory behavior.

[25] For a more comprehensive description of the dataset, see Appendix A

4.3 Error Analysis

We here analyze the output of BERT-ita to check what are the more common errors. In total, we analyzed 125 comments, of which 31 are FN and 94 FP (see Figure 1 for reference).

We first filtered out all the occurrences of true positives and true negatives. Then, we go through the data frame manually to identify recurrent patterns in both FP and FN. The results for FN are shown in Figure 3 while those for FP are in Figure 4.

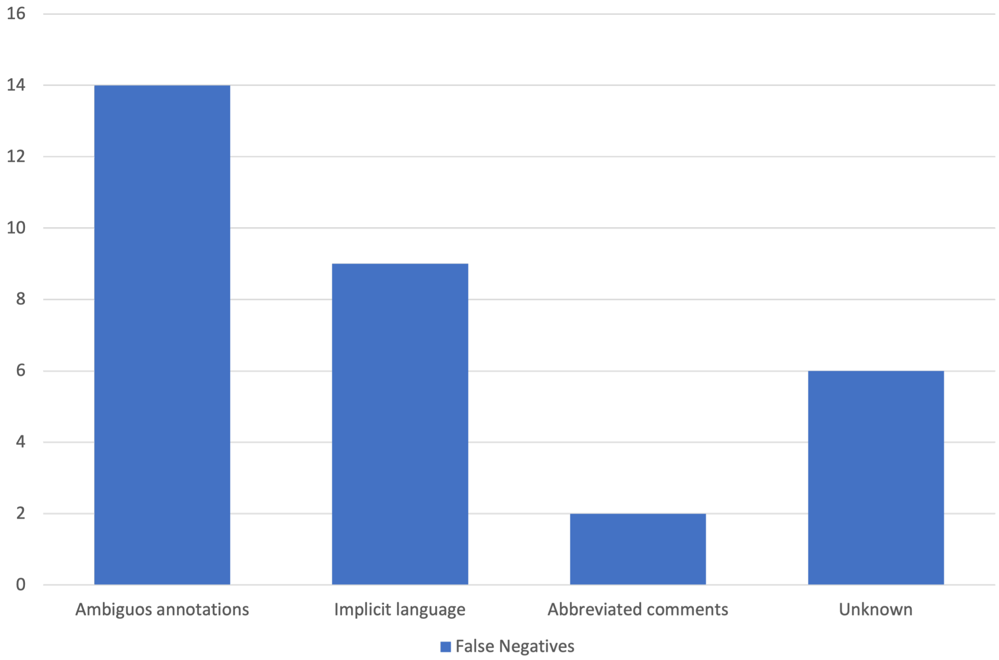

Figure 3
False negatives across four different categories

Starting from Figure 3, we identify four different types of FNs: ambiguous annotation[26], implicit toxicity[27], abbreviated comments[28], unknown[29]. In the left figure, we can observe that most FNs are related to ambiguous annotations (14) and the use of implicit language (9). Look, for instance, at examples 1 and 2 below.

E1: "*Tolleranza zero! Bastaaaaa*"
(Zero tolerance! Stop it)

E2: "*rivoluzione*"
(revolution)

[26] Here we use the term ambiguous because it better reflects the nature of the label. We were not present when the comments were annotated and cannot know why an annotator labeled a comment as toxic or non-toxic. Ambiguous comments contain text which could have more interpretations depending on its context or that may have been incorrectly labeled

[27] I.e., all the comments that contain toxic language but express it implicitly

[28] Comments that contain offensive language but in an unconventional spelling

[29] all those comments that do not fit in the previously described patterns

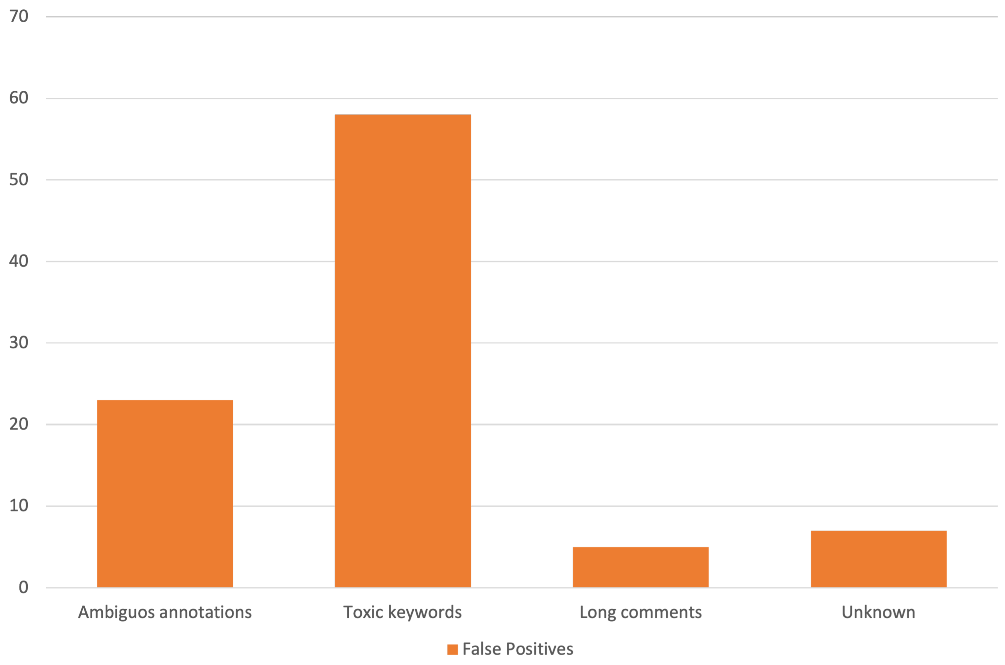

Figure 4
False positives across four different categories

Although E1 contains letter flooding ("Bastaaaaaa"), its content is implicit: no toxic language is used. The combination of the context, the flooding, and the phrase "zero tolerance" (often referred to immigrants) is what makes the comment toxic. However, our model was not able to pick it up. E2 is an example of ambiguous annotation: the word "rivoluzione" was annotated as a toxic comment. Without its context, however, it is impossible to assess the quality of the label. It is not surprising that implicit comments and ambiguous annotations constitute the larger part of FNs. (Fortuna and Nunes 2019) have highlighted how annotators' bias often affects the model's performance. Considering the rather low inter-annotator agreement reported in (Del Vigna et al. 2017), this is likely the case for BERT-ita. Likewise, (Lemmens, Markov, and Daelemans 2021)have highlighted how implicit toxic language remains one of toxic language detection's biggest challenges.

During the analysis, only two FNs were identified as belonging to the category of abbreviations or alternative spellings. The model has likely learned many alternative spellings of offensive words during the fine-tuning phase, as these occur frequently. 6 FNs did not belong to any of the identified patterns. These FNs were relatively short and may not have provided enough context for the model to classify them accurately.

In the analysis of FPs, we identified four common patterns: ambiguous annotations, unknown patterns, and two additional categories specific to FPs. The first two patterns, ambiguous annotations, and unknown patterns are also observed in FPs. The remaining two categories are unique to FPs: long comments[30] and comments containing keywords

[30] length in Figure 4

associated with toxic language[31]. Looking at Figure 4, we can observe that the largest part of FPs are comments containing offensive words (58) and ambiguous annotations (23).

E3: "*Date un premio a quello che ha tirato la macchina giocattolo a quello che era nel campo rom abusivo*"

(Give a prize to the guy who threw the toy car to the guy in the abusive Romani camp)

E4: "*Ma fatele tacere!!!!*"
(Make them shut up!)

Figure E3 presents an example of ambiguous annotation. Upon initial inspection, the comment appears to be offensive as it appears to suggest that a person who threw a toy into a Romani camp should be rewarded and uses the term "rom," which is considered a slur. However, the phrase "dare un premio" is often used ironically in Italian. As such, the annotator is likely to have interpreted the comment as non-toxic in its context. E4, on the other hand, is an example of a comment containing a keyword ("tacere") which is often associated with toxic comments. Thus, the comment was labeled as toxic by the model although it is not. Our findings reflect what has been described in the literature. A number of surveys (Yin and Zubiaga 2021; Poletto et al. 2021) have noted that FPs are often related to the presence of keywords associated with toxic comments and to subjective annotations.

Finally, a part of the FPs (5) was found to be significantly longer than average. As noted in section 4.1.3, toxic comments are on average longer than non-toxic ones. Also, although well-articulated and syntactically complex, these comments often contain a series of words that may be associated with toxicity. For the unknown (7) category, we were not able to identify any overarching pattern.

5. Conclusion

This study aimed to understand the linguistic differences between toxic and non-toxic comments in Italian on Facebook and how these differences impact the performance of toxic language detection models. To do this, we analyzed comments from the HaSpeeDe dataset, which consists of 4000 comments from eight Facebook pages that are suspected of containing toxic language. The study was divided into two parts: a quantitative analysis and a natural language processing study.

In the quantitative analysis, we examined three linguistic differences between toxic and non-toxic comments: comment length, lexical diversity (as measured by TTR, CFR, relative emoji frequency, and the number of unique emojis), and linguistic standardness (as measured by letter flooding, punctuation flooding, combinations of '!?', unconventional capitalization patterns, presence of laughter, and PNR). To better understand the significance of our results, we computed the correlation coefficients between average length/lexical diversity and the comments' toxicity. The differences were then compared to the results obtained by previous research in other languages (i.e., Dutch, English, and Slovene).

Overall, we found that all investigated features (except TTR) correlated significantly ($p < 0.01$) with the comments' toxicity. When it comes to specific linguistic differences,

[31] Such keywords may be swear words or slang. However, their sole presence does not make the comment toxic

Italian toxic comments are on average longer than non-toxic ones but contain fewer emojis (both unique and relative numbers). This result aligns with our hypotheses (1.a and 2.b) and indicates that the difference in average length and number of unique emojis/relative number of emojis between toxic and non-toxic comments is consistent across Italian, Dutch, English, and Slovene. Furthermore, the results echo Hilte's (Hilte 2019) observation on how users typing toxic comments often do not want to invest additional time to look for the right emoji to add to the text. However, it is worth noting that the difference in length and emojis found for Slovene in (De Maiti, Fišer, and Ljubešić 2020) is marginal compared to the other three languages.

In terms of lexical diversity, Italian differs from all the other three languages. In contrast with our expectations, both TTR and CFR values are only marginally higher for toxic comments. While the Italian comments' TTR value resembles those found in Dutch, English, and Slovene, the CFR is significantly lower. I.e., in the other three languages, the CFR is always above one and this is not the case for Italian. These results suggest that Italian social media users use not only fewer unique words but also a greater number of function words. Thus, Italian comments are less lexically diverse compared to the other languages.

Looking at linguistic standardness, we find that Italian toxic comments are less standard than non-toxic ones only in some aspects (i.e., punctuation flooding, PNR). Looking at the literature (Hilte 2019; Gevers, Markov, and Daelemans 2022), we expected toxic comments to be less standard since users tend to express emotionally charged content through inventive and expressive language. Italian toxic comments tend to have less punctuation than non-toxic ones and punctuation marks are more often flooded. In contrast, the values found for combinations of '!?' and unconventional capitalization patterns are similar in both types of comments. A more in-depth look into the comments revealed that non-standard features are used differently in toxic and non-toxic comments: in toxic comments, letter flooding and unconventional capitalization are often used to convey a stronger intensity and to emphasize negative words. Non-toxic comments, on the other hand, tend to use flooding to emphasize encouraging words and have shorter punctuation flooding.

Conclusively, Italian toxic comments are in general longer and use less punctuation, but do not significantly differ from non-toxic ones in lexical diversity and are only less standard in some of the investigated features (PNR and punctuation flooding). On the other hand, non-toxic comments use more emojis and contain more instances of laughter and letter flooding.

In the second part of our study, we fine-tuned three large language models (BERT-ita, PoliBERT, and UmBERT) across three different epochs on 75% of the training data and tuned two different training arguments: batch size and learning rate. We found that all three language models outperform our baseline classifier, with BERT-ita reaching an accuracy of 0.88 and a macro avg of 0.85. Next, we encoded average length, TTR, CFR, the unique and relative number of emojis, and PNR as input features to evaluate how their inclusion affected our best model's performance. The additional features were implemented through a GradientBoostClassifier. Given that all features (except for TTR) significantly correlated with comments' toxicity, we would have expected the addition of linguistic features to improve the model, but our result highlights a different pattern. We find that the inclusion of linguistic features does not improve the model's performance. In fact, most individual features slightly decrease the model's performance. The only exception is the use of the number of unique emojis, which does not affect the model's performance. On the other hand, using all of the linguistic features together actually decreases the model's performance more than any individual feature.

We encourage future research on toxic language to look not only at more languages but to further explore the issue of linguistic features, analyzing additional characteristics of toxic comments that were not investigated in this study and testing different implementation methods (such as fine-tuning the language models with the additional linguistic features). Moreover, we hope that the present study can prompt a more systematic and cohesive analysis of toxic language to facilitate the comparability and replicability of the obtained results.

6. Acknowledgements

I would like to express my deep gratitude to Professor Walter Daelemans, my research supervisor, and Dr. Lisa Hilte, the assessor of this work, for their patient guidance and feedback. A special thanks also to Professor Luna De Bruyne and Jens Lemmens for their insights and useful suggestions.

References

Akhtar, Sohail, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In Mario Alviano, Gianluigi Greco, and Francesco Scarcello, editors, *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603. Springer International Publishing.

Aljero, Mona K. and Nazife Dimililer. 2021. A novel stacked ensemble for hate speech recognition. *Applied Sciences*, 11(24):1–15, December.

Arango, Aymé, Jorge Pérez, and Barbara Poblete. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *CoRR*, abs/2104.14728.

Baumgarten, Nicole, Eckhard Bick, and Klaus Geyer. 2019. Towards balance and boundaries in public discourse: Expressing and perceiving online hate speech (xperohs). *RASK: International Journal of Language and Communication*, 50(2):87–108.

Bosco, Cristina, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of the Final Workshop of the 6th Evaluation Campaign EVALITA 2018*, pages 67–74, Turin, Italy, December 10-12. Associazione Italiana di Linguistica Computazionale (AILC), Accademia University Press.

Bočková, Renata. 2019. The Use of Punctuation, Emoji and Emoticons in YouTube Abusive Comments.

Caselli, Tommaso, Viviana Patti, Nicole Novielli, and Paolo Rosso. 2018. Evalita 2018: Overview on the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 3–8, Turin, Italy, December 12-13. Accademia University Press.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July. Association for Computational Linguistics.

Cimino, Andrea, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 86–95, Turin, Italy, December 12-13. Accademia University Press.

Corazza, Michele, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Cross-Platform Evaluation for Italian Hate Speech Detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy, November 13-15.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh*

International AAAI Conference on Web and Social Media (ICWSM-17), Montreal, Quebec, Canada, May 15-18. Association for the Advancement of Artificial Intelligence.

De Maiti, K. Pahor, Darja Fišer, and Nikola Ljubešić. 2020. Nonstandard linguistic features of Slovene socially unacceptable discourse on Facebook. *Znanstvena založba Filozofske fakultete*, 3.

Del Vigna, Fabio, Andrea Cimino, Felice Dell'Orletta, Marinelli Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, Venice, Italy, January.

Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, New York, NY, USA, May. ACM.

Durrani, Nadir, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models. *arXiv: Computation and Language*, 5.

EUR-Lex. 2008. Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law. Technical Report 2008/913/JHA, EU, 12.

Fersini, Elisabetta, Debora Nozza, and Giulia Boifava. 2020. Profiling Italian misogynist: An empirical study. In Johanna Monti, Valerio Basile, Maria Pia Di Buono, Raffaele Manna, Antonio Pascucci, and Sara Tonelli, editors, *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France, May. European Language Resources Association (ELRA).

Fersini, Elisabetta, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on Automatic Misogyny identification at IberEval 2018. In Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Sevilla, Spain, September 18.

Fišer, Darja, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online (ALW '17)*, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Fortuna, Paula and Sérgio Nunes. 2019. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30, 7.

Frenguelli, Gianluca. 2020. La norma linguistica nell'epoca dei social network: da petaloso a scendi il cane. *Circula*, 1(11):86–105.

Gajo, Paolo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. 2023. On the Identification and Forecasting of Hate Speech in Inceldom. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 373–384, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Gevers, Ine, Ilia Markov, and Walter Daelemans. 2022. Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands Journal*, 12:33–48.

Gitari, N. Dennis, Zhang Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 4.

Greevy, Edel and Alan F. Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, July. ACM Press.

Grotti, Leonardo and Patrick Quick. 2023. BERTicelli at HaSpeeDe 3: Fine-tuning and Cross-validating Large Language Models for Hate Speech Detection. In Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, and Giulia Venturi, editors, *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, volume 3473, Parma, Italy, September 7-8.

He, Fengxiang, Tongliang Liu, and Dacheng Tao. 2019. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hewitt, Sarah, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA, May 22-25.

Association for Computing Machinery.

Hilte, Lisa. 2019. The social in social media writing: the impact of age, gender and social class indicators on adolescents' informal online writing practices.

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. In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2017)*, Bolzano, Italy, October. Eurac Research.

Howard, Jeremy and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July. Association for Computational Linguistics.

Johnson, Justin M. and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(27):1–54, 3.

Karthik, Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. *International Conference on Weblogs and Social Media*, 5(3):11–17, 7.

Kumar, Ritesh, Bornini Lahiri, and Atul Ojha. 2021. Aggressive and offensive language identification in hindi, bangla, and english: A comparative study. *SN Computer Science*, 2, 02.

Lai, Mirko, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, and Giulia Venturi. 2023. EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, and Giulia Venturi, editors, *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy, September 7-8. CEUR-WS.org.

Lavergne, Eric, Rajkumar Saini, György Kovács, and Killian Murphy. 2020. TheNorth @ HaSpeeDe 2: BERT-based Language Model Fine-tuning for Italian Hate Speech Detection. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 142–147, Online, December 17. Accademia University Press.

Lemmens, Jens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Online, June.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. *Text, Speech, and Dialogue*, pages 103–114.

Ljubešić, Nikola, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The lilah emotion lexicon of croatian, dutch and slovene. In Malvina Nissim, Viviana Patti, Barbara Plank, and Esin Durmus, editors, *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, volume Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 153–157, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Markov, Ilia and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov, editors, *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Online, June. Association for Computational Linguistics.

Markov, Ilia, Ine Gevers, and Walter Daelemans. 2022. An Ensemble Approach for Dutch Cross-Domain Hate Speech Detection. In *Proceedings of the 27th International Conference on Applications of Natural Language to Information Systems*, pages 3–15, Valencia, Spain, June 15–17. Springer International Publishing.

Markov, Ilia, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In Orphee De Clercq, Alexandra Balahur, Joao Sedoc, Valentin Barriere, Shabnam Tafreshi, Sven Buechel, and Veronique Hoste, editors, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online, April. Association for Computational Linguistics.

Meta. 2022. Hate Speech, 7.

Nozza, Debora, Federico Bianchi, and Giuseppe Attanasio. 2022. Hate-ita: Hate speech detection in italian social media text. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, volume Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Seattle, Washington (Hybrid), July. Association for Computational Linguistics.

Pan, Sinno J. and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 10.

Parisi, Loreto, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Parkins, Róisín. 2018. Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Studies in Linguistics and Literature*, 2(3).

Plaza-del arco, Flor Miriam, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Yi-ling Chung, Paul Röttger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada, July. Association for Computational Linguistics.

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):477–523, 9.

Poletto, Fabio, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In Roberto Basili, Malvina Nissim, and Giorgio Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December 11-13. Italian Association for Computational Linguistics. Fourth Italian Conference on Computational Linguistics.

Polignano, Marco, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In Rachele Sprugnoli, Franco Cutugno, Sara Tonelli, Giulia Venturi, Simonetta Montemagni, Francesca Frontini, and Monica Monachini, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy, November 13-15.

Sanguinetti, Manuela, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop*, pages 93–101. Accademia University Press, December 17. Event conducted online.

Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. *Language Resources and Evaluation*, pages 1–8, 5.

Sedda, Franciscu and Paolo Demuru. 2019. La rivoluzione del linguaggio social-ista: umori, rumori, sparate e provocazioni / The revolution of social-ist language: moods, noises, shots, provocations. *Rivista Italiana di Filosofia del Linguaggio*, 13(2), 1.

Siege, Alexandra A. 2020. *Online Hate Speech*. Cambridge University Press.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence (AAAI'97/IAAI'97)*, Providence, Rhode Island, July. AAAI Press.

Steinberger, Josef, Tomáš Brychcín, Tomáš Hercig, and Peter Krejzl. 2017. Cross-lingual Flames Detection in News Discussions. *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, 11.

Tamburini, Fabio. 2020. How "bertology" changed the state-of-the-art also for italian nlp. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 415–421, Bologna, Italy, March 1-3. Accademia University Press.

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

Warner, William and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada, June.

Wiegand, Michael, Joseph Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2-7. Association for Computational Linguistics.

Yin, Wenjie and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 6.

YouTube. 2019. Hate speech policy, 6.

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). In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Appendix A: Datasets Description

What follows is an in-detail description of the datasets described in Section 3.

FRENK$_{\text{Slovene/English}}$ The Slovene part of the FRENK dataset contains Facebook comments from posts on three mainstream Slovene newspaper pages. In total, (De Maiti, Fišer, and Ljubešić 2020) analyzed 6,545 comments about migrants and 4,517 about LGBTQ+. In the original dataset, thirty-two master students received a brief training and annotated the data based on the presence of SUD (SUD vs non-SUD) and added more fine-grained annotation on whether SUD comments contained (i) background violence (background violence vs offensive speech) (ii) a target (targeted vs non-targeted) and (iii) violence (threat vs offensive speech). The inter-annotator agreement was calculated using Krippendorff's α and reached a score above 0.66, i.e., the minimum acceptable threshold, across all annotations. In their paper, (Gevers, Markov, and Daelemans 2022) also analyzed the English part of the FRENK dataset[32] and compared their results to those obtained on the LiLaH Dutch dataset and those on Slovene in (De Maiti, Fišer, and Ljubešić 2020).

LiLaH$_{\text{Dutch}}$ The Dutch part of the LiLaH dataset contains 10,732 comments from prominent Flemish newspaper Facebook pages. The LiLaH dataset was annotated by one expert and two trained annotators using the same annotation guidelines used for the FRENK dataset and its comments also regard the LGBTQ+ community and migrants. The inter-annotator agreement was calculated using Cohen's Kappa, which resulted in an agreement between fair and good (0.56) (Markov et al. 2021).

HaSpeeDe$_{\text{Italian}}$ For the original HaSpeeDe dataset (Del Vigna et al. 2017), the authors crawled 17,567 (of which at least 6,502 received at least one and at max five annotation) comments from eight Italian Facebook pages/groups[33]. The annotation process involved 5 bachelor students who were asked to distinguish between three levels (*No hate*, *Weak hate*, *Strong hate*) and types (*Religion*, *Physical and/or mental handicap*, *Socio-economical status*, *Politics*, *Race*, *Sex* and *Gender issues*, and *Other*). The inter-annotator agreement was calculated using Fleiss' kappa κ and reached a $\kappa = 0.26$ when *Weak*

[32] It is worth noting that the English part of the FRENK dataset was annotated following the above-described Slovene part following the same guidelines and procedures.

[33] *salviniofficial, matteorenziufficiale, lazanzarar24, jenusdinazareth, sinistracazzateliberta2, ilfattoquotidiano, emosocazzi, noiconsalviniufficiale.*

hate and *Strong hate* were merged into one class. Unfortunately, the full dataset is not available: a sub-corpus, consisting of 4000 comments, has been made available for EVALITA 2018 (Bosco et al. 2018) shared task. The sub-corpus maintains the different annotation levels: however, the three classes of toxic language have been grouped into two for the task: *Toxic v. non-toxic*.