# Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection

**Ilia Markov**
CLIPS Research Center
University of Antwerp, Belgium
ilia.markov@uantwerpen.be

**Nikola Ljubešić**
Dept. of Language Technologies
Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

**Darja Fišer**
Faculty of Arts
University of Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

**Walter Daelemans**
CLIPS Research Center
University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

## Abstract

In this paper, we describe experiments designed to evaluate the impact of stylometric and emotion-based features on hate speech detection: the task of classifying textual content into hate or non-hate speech classes. Our experiments are conducted for three languages – English, Slovene, and Dutch – both in in-domain and cross-domain setups, and aim to investigate hate speech using features that model two linguistic phenomena: the writing style of hateful social media content operationalized as function word usage on the one hand, and emotion expression in hateful messages on the other hand. The results of experiments with features that model different combinations of these phenomena support our hypothesis that stylometric and emotion-based features are robust indicators of hate speech. Their contribution remains persistent with respect to domain and language variation. We show that the combination of features that model the targeted phenomena outperforms words and character n-gram features under cross-domain conditions, and provides a significant boost to deep learning models, which currently obtain the best results, when combined with them in an ensemble.

## 1 Introduction

Hate speech is commonly defined as communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). The exact definition of hate speech, however, remains a disputed topic, as it is a subjective and multi-interpretable concept (Waseem et al., 2017; Poletto et al., 2020).

The lack of a consensus on its definition poses a challenge to hate speech annotation. Annotating hateful content remains prone to personal bias and is culture-dependent, which often results in low inter-annotator agreement and therefore scarcity of high quality training data for developing supervised hate speech detection systems (Ross et al., 2016; Waseem, 2016; Sap et al., 2019).

Hate speech online presents additional challenges for natural language processing (NLP): offensive vocabulary and keywords evolve fast due to their relatedness with the hate speech triggering events (Florio et al., 2020), moreover, users may adapt their lexical choices as a countermeasure against identification or introduce minor misspellings to bypass filtering systems (Berger and Perez, 2006; Vidgen et al., 2019). Therefore, we intend to investigate more abstract features, less susceptible to specific vocabulary, topic or corpus bias, which we examine in in-domain and cross-domain settings: training and testing on social media datasets belonging to same/different domains, for three languages: English, Slovene, and Dutch.

Our hypothesis is that the style and emotional dimension of hateful textual content may provide useful cues for its detection. We investigate this through a binary hate speech classification task using features that model such information, i.e., function words and emotion-based features. The latter are operationalized in terms of the types of emotions expressed and the frequency of emotion-conveying words in the data.

Function word usage is one of the most important and revealing aspects of style in written language, as shown by numerous studies in stylometric analysis for authorship attribution (Grieve, 2007; Kestemont, 2014; Markov et al., 2018). While stylometric characteristics have been implicitly included in some hate speech detection studies (e.g., in bag-of-words or character-level models),

*Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159
April 19, 2021. ©2021 Association for Computational Linguistics

their impact on the task has not been studied. We propose the hypothesis that stylometric characteristics of hateful writing are distinctive enough to contribute to the hate speech detection task. In other words, hate speech acts as a specific text type with an associated writing style.

On the other hand, we are motivated by psychological and sociological studies, which correlate toxic behaviour online with the emotional profile of the user (Kokkinos and Kipritsi, 2012). However, unlike previous research that used sentiment information for detecting unacceptable content (Davidson et al., 2017; Dani et al., 2017; Van Hee et al., 2018; Brassard-Gourdeau and Khoury, 2019), we test whether we are able to capture some of these phenomena by going beyond the sentiment level (positive / negative / neutral) to a more fine-grained emotion level.

We compare the performance of stylometric and emotion-based features with commonly used features for hate speech detection: words, character n-grams, and their combination, and with more recent deep learning models that currently provide the state-of-the-art results for the hate speech detection task (Mandl et al., 2019; Basile et al., 2019): convolutional neural networks (CNN), long short-term memory networks (LSTM), and bidirectional encoder representations from transformers (BERT). The results of these experiments indicate that the combination of stylometric and emotion-based features performs better than words and character n-grams under cross-domain conditions, and allows to further improve the results of deep learning models when combined with them in an ensemble.

In summary, the contributions of the research presented here are the following: (i) evaluating the contribution of stylometric and emotion-based features to hate speech detection, (ii) examining how robust and persistent their contribution is with respect to domain and language variation, (iii) comparing their performance with commonly used features for the hate speech detection task: words and character n-grams, and with the state-of-the-art deep learning models.

## 2 Methodology

### 2.1 Datasets

To investigate the role of stylometric and emotion information in the hate speech detection task, we conducted experiments on several recent social media datasets in hate speech detection research.

#### 2.1.1 In-domain datasets

**FRENK** (Ljubešić et al., 2019) The FRENK dataset consists of Facebook comments in English and Slovene covering LGBT and migrant topics. The dataset was manually annotated for fine-grained types of socially unacceptable discourse (e.g., violence, offensiveness, threat). We used the coarse-grained (binary) hate speech classes: hate speech or non-hate speech messages, selecting the messages for which more than four out of eight annotators agreed upon the class. The detailed description of the dataset collection and annotation procedures can be found in (Ljubešić et al., 2019).

**LiLaH** The LiLaH dataset consists of Facebook comments on LGBT and migrant topics in Dutch. The dataset was collected using the same procedure and annotated following the same annotation guidelines as the FRENK dataset by two trained annotators and one expert annotator. For the binary classes used in this paper, the Percent Agreement for two annotators equals 78.7% and Cohen's Kappa 0.56, which corresponds to an inter-annotator agreement halfway between “fair” and “good” (Fleiss, 1981).

The statistics of the datasets used are shown in Table 1. We used training and test partitions splitting the datasets by post boundaries in order to avoid comments from the same discussion thread to appear in both training and test sets, that is, to avoid within-post bias. The partitions were performed in such a way that the distribution of hate (‘1’) and non-hate speech (‘0’) classes is as balanced as possible, while the proportion of 80% training and 20% test messages for the addressed languages is preserved.

The balanced subsets of the datasets, in terms of the number of messages for each of the languages, were used for 10-fold cross-validation experiments in order to provide a fair comparison across the targeted languages (the maximum number of available hate and non-hate speech examples across the three languages was selected; marked with ‘*’ in Table 1), while the merged training and test partitions were used as training sets for the cross-domain experiments in order to provide more examples for training the supervised models described further in the paper.

#### 2.1.2 Cross-domain datasets

For cross-domain experiments, we merged the training and test splits of the FRENK and LiLaH

|  |  | Training |  | Test |  | Training & test |  | Balanced |  |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Label** | **# mes** | **%** | **# mes** | **%** | **# mes** | **%** | **# mes** | **%** |
| **English (FRENK)** | ‘1’ | 2,848 | 35.9 | 744 | 35.5 | 3,592* | 35.8 | 3,500 | 46.7 |
|  | ‘0’ | 5,091 | 64.1 | 1,351 | 64.5 | 6,442 | 64.2 | 4,000 | 53.3 |
|  | Total | **7,939** | (79.1) | **2,095** | (20.9) | **10,034** | (100) | **7,500** |  |
| **Slovene (FRENK)** | ‘1’ | 3,506 | 52.0 | 882 | 51.8 | 4,388 | 51.9 | 3,500 | 46.7 |
|  | ‘0’ | 3,238 | 48.0 | 821 | 48.2 | 4,059* | 48.1 | 4,000 | 53.3 |
|  | Total | **6,744** | (79.8) | **1,703** | (20.2) | **8,447** | (100) | **7,500** |  |
| **Dutch (LiLaH)** | ‘1’ | 3,753 | 43.8 | 949 | 44.0 | 4,702 | 43.8 | 3,500 | 46.7 |
|  | ‘0’ | 4,821 | 56.2 | 1,209 | 56.0 | 6,030 | 56.2 | 4,000 | 53.3 |
|  | Total | **8,574** | (79.9) | **2,158** | (20.1) | **10,732** | (100) | **7,500** |  |

Table 1: Statistics of the datasets used for in-domain experiments.

datasets (see Table 1) and used it as training data, while the following social media datasets belonging to other domains were used as test sets.

**HASOC** (Mandl et al., 2019) We used the training subset in English of the HASOC-2019 (Hate Speech and Offensive Content Identification in Indo-European Languages) shared task dataset. It contains Twitter and Facebook messages covering various topics (e.g., Brexit, cricket).

**Ask.fm** (Van Hee et al., 2015). We used the Dutch cyberbullying dataset, which contains 85,485 posts from the social networking website Ask.fm annotated with fine-grained cyberbullying categories (e.g., general insult, sexual harassment, sexism, racism). We selected the same number of positive and negative messages as for English in order to provide a fair comparison between the two languages.

The statistics of the datasets used as test sets for the cross-domain experiments is shown in Table 2.[1]

| **Dataset** | **Label** | **# mes** | **%** |
|---|---|---|---|
| **English (HASOC)** | ‘1’ | 2,261 | 38.6 |
|  | ‘0’ | 3,591 | 61.4 |
|  | Total | **5,852** |  |
| **Dutch (Ask.fm)** | ‘1’ | 2,261 | 38.6 |
|  | ‘0’ | 3,591 | 61.4 |
|  | Total | **5,852** |  |

Table 2: Statistics of the datasets used for the cross-domain experiments as test sets.

## 2.2 Experiment setup

We performed tokenization, lemmatization, and POS tagging using the StanfordNLP library (Qi et al., 2018) for all the languages addressed in this work, removing metadata and URL mentions in preprocessing. We used the sets of features described below, a term frequency (tf) weighting scheme and the liblinear scikit-learn (Pedregosa et al., 2011) implementation of Support Vector Machines (SVM) with optimized parameters for classification (we selected the optimal liblinear classifier parameters: penalty parameter (C), loss function (loss), and tolerance for stopping criteria (tol) based on grid search). The effectiveness of SVM has been shown by numerous experiments on hate speech detection (Fortuna and Nunes, 2018; Basile et al., 2019).

We used a CNN model (Kim, 2014) to learn discriminative word-level features with the following architecture: first, an embedding layer transforms sparse vectors into dense vector representations. To process the word embeddings, we used a convolutional layer (kernel size: 3) followed by a global max pooling layer. Then, a dense layer with a ReLU activation is applied, followed by a dropout of 0.6, and finally, a dense layer with a sigmoid activation to make the prediction for the binary classification. We used an LSTM model (Hochreiter and Schmidhuber, 1997), which takes a sequence of words as input and aims at capturing long-term dependencies. We processed the sequence of word embeddings with a unidirectional LSTM layer with 300 units, followed by a dropout of 0.4, and a dense layer with a sigmoid activation for predictions. The multilingual BERT model (BERT-base, multilingual cased (Devlin et al., 2019)) was used for all the languages addressed in this work. The implementation was done in PyTorch (Paszke et al., 2019) using the simple transformers library.[2] Deep learning models currently achieve the state-of-the-art results for the hate speech detection task, which are in 80%–90% F1-score range in in-domain set-

[1] For Slovene, we did not find an annotated dataset belonging to a different hate speech domain.

[2] https://simpletransformers.ai/

tings, depending on the languages being considered, amount of data, etc. (Mandl et al., 2019; Basile et al., 2019; Zampieri et al., 2020).

We report the results in terms of precision, recall, and F1-score (macro-averaged). Note that we used similar settings, tools, and models, i.e., size of the training and test data, StanfordNLP for tokenization, lemmatization, and POS tagging, multilingual BERT models, in order to provide a fair comparison across the different languages covered in the paper.

### 2.3 Features

The experiments we report were designed to investigate and quantify the impact of stylometric information, modeled through function word usage, and emotion-based features on hate speech detection. While representations of documents through word/character n-grams provide good results for detecting abusive language (Nobata et al., 2016; Van Hee et al., 2018), these features cover – and at the same time obscure – a wide range of phenomena, and therefore, it is not clear what the impact is of subsets of these features representing specific linguistic information. Moreover, these features include content words and are susceptible to topic, genre, and domain bias, which often results in overfitting the data. Because of this we abstract away from domain-dependent content word patterns and use more abstract POS n-gram features, to which we add stylometric features (function words) and emotion-based features, to evaluate their impact on the hate speech detection task.

**Part-of-speech (POS)** POS features capture the morpho-syntactic patterns in a text, and are indicative of hate speech, especially when used in combination with other types of features (Warner and Hirschberg, 2012; Robinson et al., 2018). POS tags were obtained with the Stanford POS Tagger (Toutanova et al., 2003). We used the same 17 universal POS tags for the three languages and built n-grams from this representation with n = 1–3.

**Stylometric features** Function words (FW) are considered one of the most important stylometric feature types (Kestemont, 2014). They clarify the relationships between the content-carrying elements of a sentence, and introduce syntactic structures like verbal complements, relative clauses, and questions (Smith and Witten, 1993). With respect to emotion features, FW can appear as quantifiers, intensifiers (e.g., very good) or modify the emotion expressed in other ways. We used linguistically-defined FW, that is, words belonging to the closed syntactic classes.[3] FW are incorporated into the POS representation, as shown in Table 3.

**Emotion-based features** To encode emotion information in our data, we used the 14,182 emotion words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) from the NRC emotion lexicon (Mohammad and Turney, 2013). We used the LiLaH emotion lexicon for Slovene and Dutch (Ljubešić et al., 2020; Daelemans et al., 2020), which contains manual translations of the NRC emotion lexicon entries. The emotion information was modeled through (i) incorporating emotion-conveying words into the POS & FW representation, as shown in Table 3, (ii) counting the number of such words in a message (count), and (iii) using the emotion associations of the emotion words in a message. These features were used to encode the types of emotions in a message and to capture how high-emotional or low-emotional a message is.

Consider the following English comment from our data belonging to the hate speech class: *Mental illness on parade*. Table 3 shows an example of the representation of this message through the features described above. From the POS & FW & emotion word representations, n-grams (n = 1–3) are built.[4] The count of emotionally-charged words and the emotion associations were added as additional feature vectors.

| phrase | *Mental illness on parade* |
|---|---|
| POS | ADJ NOUN ADP NOUN |
| POS & FW | ADJ NOUN on NOUN |
| POS & FW & emo. words | ADJ illness on parade |
| POS & FW & emo. words & count | ADJ illness on parade & 2 |
| POS & FW & emo. features | ADJ illness on parade & 2 & fear sadness surprise ... |

Table 3: Example of features used.

## 3 Results and Discussion

The main goal of this paper is to identify and analyze specific linguistic phenomena with respect to their role in hate speech detection. In particular,

[3] https://universaldependencies.org/u/pos/

[4] Representing the messages in the following way provided higher 10-fold cross-validation results than combining separate feature vectors (0.8%–1.2% depending on the language).

we focused on function word usage (as an expression of style of the hateful content) and emotion (as personality and psychological state indicators with respect to the usage of emotion terms in written messages).

First, we perform a separate analysis of the contribution of the phenomena we target, and then compare the performance of combining features that encode these phenomena with the commonly used features for the hate speech detection task: words, character n-grams, and their combination (tf weighting scheme and the liblinear SVM classifier with optimized parameters), and with more recent deep learning models that achieve state-of-the-art results for hate speech detection (Mandl et al., 2019; Basile et al., 2019): CNN, LSTM, and BERT (see Section 2.2).

Sections 3.1 and 3.2 show the results of these experiments in in-domain and cross-domain settings, respectively.

### 3.1 In-domain experiments

**10-fold cross-validation** First, we analyze the features that capture the phenomena we target in isolation and in combination on the balanced subsets of the datasets (see Table 1) for 10-fold cross-validation results. The results of these experiments, presented in Table 4, are compared with words (BoW), character 1–3-grams (char), and their combination. We also present the results when all the feature sets are combined. As a reference point we provide the random baseline (stratified strategy).

The results of the experiments presented in Table 4 indicate that stylometric features indeed contribute to the hate speech detection task, as evidenced by their positive impact to the POS representation for all the considered languages (POS & FW representation). Likewise, emotion-conveying words (POS & FW & emotion words), the count of such words in a message (POS & FW & emo words & count) and ten features that correspond to the type of emotions being conveyed (POS & FW & emo words & emo feats) further contribute to the results. While their performance in isolation is moderate, higher results are achieved when features representing each of these phenomena are combined, indicating that they are complementary for the hate speech detection task (this representation is marked in bold and in the remainder of this paper referred to as 'our' approach). Feature importance analysis revealed that negative emotion words, such as 'disgusting', 'sick', 'invasion', are the most indicative features in our model.

We also note that words (BoW) and character n-gram features perform well in in-domain conditions and achieve higher results than our approach. While words are the best unique features for English, the combination of words and character n-grams shows the highest results for Slovene and Dutch. This may be related to the fact that Slovene and Dutch are morphologically richer languages than English, as character n-grams are able to capture morphological affixes.

When stylometric and emotion features are combined with words and character n-grams, the best results are obtained for English and Slovene. For the Dutch language, the combination of words and character n-grams performs very well in in-domain settings, as confirmed by the additional experiments we present further in the paper. It is also interesting to note that for the English language adding the combination of BoW and character n-grams to our approach provides a higher boost than adding BoW features only, the best unique feature type for this language.

**Train-test partitions** Next, we present the results when splitting the training and test sets by post boundaries in order to avoid within-post bias, as described in Section 2.1. In this scenario, we additionally compare the performance of stylometric, emotion, and BoW and character n-gram features with more recent deep learning models: CNN, LSTM, and BERT. The results for these experiments are shown in Table 5.

The results presented in Table 5 indicate that splitting the data by post boundaries is a more challenging scenario, as evidenced by the drop in performance for the BoW approach for all the languages. Nonetheless, the trends observed in the 10-fold cross-validation experiments remain consistent: stylometric and emotion-based features provide substantial improvements when added to the POS representation, while the gap in performance when compared to BoW and character n-grams is smaller for all the languages but Dutch. For the Dutch language, character n-grams provide higher results than in the 10-fold cross-validation experiments, and higher boost when combined with BoW. The combination of character n-grams and BoW for this language shows even higher results than BERT: the best deep learning model across the three languages. For all the targeted languages, adding BoW

|  | English |  |  |  | Slovene |  |  |  | Dutch |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Features** | Prec. | Rec. | F1-score | # feats | Prec. | Rec. | F1-score | # feats | Prec. | Rec. | F1-score | # feats |
| Random baseline | 52.0 | 52.0 | 52.0 | – | 52.0 | 52.0 | 52.0 | – | 52.0 | 52.0 | 52.0 | – |
| BoW | 80.2 | 79.2 | **79.3** (± 1.0) | 15,091 | 72.5 | 71.3 | 71.3 (± 1.6) | 29,898 | 74.4 | 73.2 | 73.3 (± 1.6) | 18,043 |
| Character 1–3-grams (char) | 78.3 | 77.5 | 77.6 (± 0.7) | 19,030 | 74.1 | 73.5 | 73.5 (± 1.7) | 18,648 | 74.1 | 73.2 | 73.3 (± 1.1) | 17,165 |
| BoW & char | 79.1 | 78.3 | 78.5 (± 0.6) | 34,121 | 74.6 | 74.0 | **74.1** (± 1.6) | 48,546 | 74.8 | 74.0 | **74.1** (± 1.2) | 35,208 |
| Function words (FW) | 65.0 | 62.2 | 61.2 (± 1.6) | 745 | 63.1 | 61.4 | 60.6 (± 2.2) | 1,207 | 65.4 | 62.7 | 61.8 (± 1.7) | 988 |
| Emotion words | 74.0 | 72.1 | 72.1 (± 1.6) | 2,376 | 66.8 | 65.1 | 64.7 (± 1.9) | 2,451 | 69.2 | 67.3 | 67.1 (± 1.8) | 2038 |
| POS 1–3-grams (POS) | 64.6 | 63.5 | 63.2 (± 1.7) | 3,423 | 61.6 | 60.5 | 60.0 (± 1.9) | 3,417 | 65.5 | 64.3 | 64.0 (± 1.2) | 3,055 |
| POS & FW (1–3-grams) | 69.1 | 67.6 | 67.4 (± 1.7) | 36,463 | 65.0 | 63.8 | 63.5 (± 1.8) | 47,027 | 68.1 | 66.9 | 66.7 (± 1.4) | 45,561 |
| POS & FW & emotion words | 74.9 | 73.9 | 74.0 (± 1.4) | 105,400 | 67.7 | 66.9 | 66.9 (± 1.3) | 117,265 | 70.4 | 69.8 | 69.8 (± 1.2) | 109,169 |
| POS & FW & emo words & count | 75.2 | 74.4 | 74.5 (± 1.4) | 105,401 | 68.4 | 67.7 | 67.7 (± 1.0) | 117,266 | 70.3 | 69.8 | 69.9 (± 1.4) | 109,170 |
| **POS & FW & emo features (emo)** | 75.3 | 74.5 | 74.6 (± 1.4) | 105,411 | 68.9 | 68.2 | 68.3 (± 1.4) | 117,276 | 71.1 | 70.6 | 70.6 (± 1.2) | 109,180 |
| POS & FW & emo & BoW | 79.2 | 78.7 | 78.8 (± 1.0) | 120,502 | 72.7 | 72.1 | 72.2 (± 1.7) | 147,174 | 73.4 | 72.9 | 73.0 (± 2.2) | 127,223 |
| POS & FW & emo & char | 79.6 | 78.9 | 79.0 (± 1.0) | 124,441 | 75.3 | 74.8 | 74.8 (± 1.1) | 135,924 | 73.2 | 73.0 | 73.0 (± 1.5) | 126,345 |
| POS & FW & emo & BoW & char | 80.1 | 79.4 | **79.6** (± 0.9) | 139,532 | 75.7 | 75.1 | **75.2** (± 1.4) | 165,822 | 73.8 | 73.5 | **73.6** (± 1.7) | 144,388 |

Table 4: Performance of the features explored in isolation and in combination on the balanced subsets of the datasets under 10-fold cross-validation. The results for bag-of-words (BoW), character 1–3-grams (char), and their combination with each other and with the stylometric and emotion-based features (emo) as well as the number of features for each experiment (# feats) are also provided.

|  | English |  |  | Slovene |  |  | Dutch |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | Precision | Recall | F1-score | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Random baseline | 50.7 | 50.7 | 50.7 | 50.9 | 50.9 | 50.9 | 48.3 | 48.3 | 48.3 |
| (1) BoW | 71.0 | 70.8 | 70.9 | 68.5 | 68.5 | 68.5 | 72.0 | 70.9 | 71.1 |
| (2) Char 1–3-grams | 69.0 | 69.2 | 69.1 | 72.1 | 72.1 | 72.1 | 74.5 | 73.4 | 73.7 |
| (3) BoW & char | 70.6 | 70.6 | 70.6 | 72.4 | 72.4 | 72.4 | 75.0 | 74.4 | **74.6** |
| (4) CNN | 73.4 | 73.6 | 73.5 | 67.7 | 67.7 | 67.7 | 72.6 | 72.9 | 72.5 |
| (5) LSTM | 71.0 | 69.9 | 70.4 | 68.5 | 67.3 | 67.1 | 70.5 | 70.5 | 70.5 |
| (6) BERT | 74.9 | 74.6 | **74.8** | 73.0 | 72.9 | **72.9** | 74.3 | 74.1 | 74.2 |
| (7) POS | 57.3 | 57.0 | 57.1 | 63.2 | 63.1 | 62.8 | 63.9 | 62.9 | 62.9 |
| (8) POS & FW | 64.3 | 63.6 | 63.8 | 63.5 | 63.4 | 63.1 | 70.2 | 67.7 | 67.8 |
| (9) **POS & FW & emo** | 70.9 | 69.9 | 70.3 | 68.0 | 68.0 | 67.8 | 73.1 | 70.6 | 70.8 |
| (10) POS & FW & emo & BoW & char | 74.4 | 73.7 | **74.0** | 74.3 | 74.3 | **74.3** | 75.1 | 74.5 | **74.7** |

Table 5: In-domain results (training-test splits by post boundaries).

and character n-grams to our approach further improves the results, outperforming BoW, character n-grams, and their combination, and achieving competitive results with the deep learning models.

Having confirmed that due to stylometric choices and emotion expression we can distinguish the hateful messages in in-domain settings, we proceed with a cross-domain analysis of the targeted phenomena.

### 3.2 Cross-domain experiments

In this section, we evaluate the robustness of stylometric and emotion-based features under cross-domain conditions: training and testing on out-of-domain social media datasets described in Section 2.1. Cross-domain scalability is essential to identify features of online hate speech that generalize well across domains. Table 6 presents the results for the cross-domain experiments.

The results in Table 6 show that using out-of-domain data for testing leads to a drop in performance for all the models. The drop for the English language is much higher than for Dutch, despite that for English we used a dataset annotated for the same task (hate speech detection) and for a different task, cyberbullying detection, for Dutch.[5] The descriptive analysis showed that the Jaccard similarity coefficient (Jaccard, 1901) for the cross-domain training and test sets is 20.6% for English and 12.4% for Dutch, which implies that a large part of the training and test vocabularies do not overlap.

[5] We also tested the robustness of cross-domain settings by experimenting with other subsets for Dutch, e.g., balanced class distribution, more examples, as well as with other English datasets, achieving similar results with the same trends.

| | **English** | | | | **Dutch** | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | Precision | Recall | F1-score | F1 drop | Precision | Recall | F1-score | F1 drop |
| Random baseline | 49.2 | 49.3 | 49.2 | – | 50.7 | 50.7 | 50.6 | – |
| (1) BoW | 60.5 | 57.4 | 56.6 | 14.3 | 71.6 | 65.9 | 66.3 | 4.8 |
| (2) Char 1–3-grams | 55.8 | 56.1 | 55.1 | 14.0 | 72.3 | 66.0 | 66.3 | 7.4 |
| (3) BoW & char | 56.5 | 56.8 | 55.6 | 14.9 | 73.7 | 67.4 | 67.8 | 6.8 |
| (4) CNN | 58.7 | 58.2 | 58.3 | 15.2 | 72.3 | 70.0 | **70.6** | 1.9 |
| (5) LSTM | 57.5 | 57.5 | 57.5 | 12.9 | 71.7 | 66.6 | 67.1 | 3.4 |
| (6) BERT | 59.3 | 59.8 | **59.1** | 15.7 | 74.0 | 69.5 | 70.2 | 4.0 |
| (7) POS | 52.9 | 52.5 | 52.0 | 5.1 | 65.9 | 60.6 | 60.0 | 2.9 |
| (8) POS & FW | 55.2 | 54.5 | 54.2 | 9.6 | 69.7 | 63.6 | 63.5 | 4.3 |
| **(9) POS & FW & emo** | 59.1 | 57.8 | 57.7 | 12.6 | 73.1 | 68.8 | **69.5** | 1.3 |
| (10) POS & FW & emo & BoW & char | 58.1 | 58.5 | **57.9** | 16.1 | 73.8 | 68.6 | 69.3 | 5.4 |
| Ensemble (4 & 6 & 9) | 60.7 | 60.1 | **60.2*** | 16.5 | 77.1 | 71.6 | **72.5*** | 2.9 |

Table 6: Cross-domain results (testing on out-of-domain datasets). The F1 drop column reports the drop in F1 points for each model when compared to the in-domain experiments. Statistically significant gains of the ensemble model over the best deep learning models (BERT or CNN) according to McNemar's statistical significance test (McNemar, 1947) with $\alpha < 0.05$ are marked with '*'.

Therefore, the asymmetric drop in performance across the two languages cannot be explained by lexical overlap. The lower drop for Dutch may be related to the relative non-complexity of the cyberbullying content. An analogous effect was observed in (Emmery et al., 2020), where a similar behaviour is reported when training on toxic messages and using Ask.fm as out-of-domain test set, and which is also evidenced by the high precision scores obtained for the models used in this paper when testing on the Dutch cyberbullying data.

We note that, similarly to the in-domain experiments, stylometric and emotion-based features provide substantial improvements when combined with the POS representation. Being more abstract features, they cope well with domain variation and show a lower drop in cross-domain conditions when compared to the baseline models. This indicates that the features that capture the targeted phenomena are robust and portable across social media domains.

Words and character n-gram features, on the contrary, show a high drop in cross-domain settings and provide marginal improvement for English and no improvement for Dutch when combined with our approach. For the Dutch language, stylometric and emotion-based features partly compensate for the loss in performance brought by words and character n-grams under cross-corpus conditions. We can also observe that our approach outperforms commonly used features for the hate speech detection task: words, character n-grams, and their combination both for English and Dutch.

The deep learning models provide high results in cross-domain settings. Combining our approach with the best performing deep learning models: CNN and BERT, using a hard majority-voting ensemble, significantly improves the results when compared to the performance of CNN and BERT in isolation (according to McNemar's statistical significance test (McNemar, 1947) with $\alpha < 0.05$) and achieves the highest cross-domain results for the both languages. We conclude that this significant improvement is brought by the ability of our approach to capture the style of hateful content and the emotional peculiarities present in hateful messages, which are hard to encode by deep learning models on relatively small datasets. A detailed analysis presented below provides deeper insights into the nature of these improvements.

### 3.3 Error analysis

In this section, we report on a manual error analysis performed on the difference in the output of the BERT model (the best-performing deep learning model in our experiments) and the stylometry-emotion-based approach (POS & FW & emo, 'our' model) by inspecting hate speech instances that were correctly identified by one model but not by the other. We perform the analysis on the Slovene in-domain dataset and the English cross-domain dataset. We inspect 50 random misclassified hate speech instances per dataset (in-domain vs. cross-domain) and the model correctly identifying hate

|  |  | **In-domain** |  | **Cross-domain** |  |
|---|---|---|---|---|---|
| Type | Element | Style-emo (%) | BERT (%) | Style-emo (%) | BERT (%) |
| explicit | violence | 4 | 22 | 0 | 4 |
|  | insult | 26 | 40 | 40 | 34 |
|  | swearword | 2 | 2 | 4 | 50 |
| implicit | violence | 4 | 4 | 4 | 0 |
|  | argument | 10 | 6 | 0 | 0 |
|  | accusation | 0 | 0 | 24 | 2 |
|  | othering | 20 | 10 | 0 | 0 |
| other | quotation | 2 | 2 | 0 | 0 |
|  | multilingual | 0 | 0 | 4 | 0 |
|  | unclear | 32 | 16 | 24 | 10 |

Table 7: Distribution of correctly classified hate speech instances by one model and not by the other, with respect to the type and element of hate speech, in in-domain and cross-domain settings.

speech (BERT vs. our). Our error analysis is based on annotating each instance with the type of hate speech (explicit vs. implicit), the element of hate speech (call to violence, insult, swearword, argument, accusation, and othering) and, where relevant, the reason why it was undetected (informal expression, unconventional spelling, foreign language, creative language, metaphorical language, unconventional tokenization). In Table 7, we present the quantitative results of this manual analysis, reporting for each model and dataset the distribution of correctly classified hate speech instances, the other model failing on these instances, given the type and element of hate speech.

On the first level of the type of hate speech, where we discriminate between explicit and implicit hate speech, both in the in-domain and the cross-domain settings, we observed that BERT is better at identifying explicit cases of hate speech (overall 72% of instances correctly identified by BERT but not by the stylometry-emotion-based approach vs. 32% of instances correctly identified by the stylometry-emotion-based approach but not by BERT on the in-domain dataset, on the cross-domain dataset this relation being 84% vs. 44%), while the stylometry-emotion-based approach deals better with implicit instances of hate speech (34% vs. 20% in-domain; 28% vs. 2% cross-domain).

The results for the hate speech elements reflect the difference between the in-domain and the cross-domain datasets. For the explicit cases of hate speech, in both datasets insults tend to be the dominant elements caught by one model but not by the other, while the in-domain dataset contains much more calls to violence, and swearing prevails in the cross-domain dataset. For the explicit cases of hate speech, arguments and othering strategies were most frequently misclassified in the in-domain dataset, while accusations were the most frequent issues in the cross-domain dataset. These differences can be followed back to the differences in the medium mostly containing Facebook discussions in the in-domain case, which are more discursive and implicit and Twitter messages in the cross-domain case, which are much shorter and direct.

We also performed a closer inspection of insults and swearwords, which are lexical categories and should in principle be simple to identify via means of supervised machine learning, but were missed because they were highly informal, non-canonically spelled or tokenized, taken from foreign languages, incomplete or idiosyncratic. Another type of undetected insults were words from the general vocabulary from topics such as animals, hygiene, intelligence, etc. that were used metaphorically or with a distinctly negative connotation.

For the category of misclassifications by one model but not by the other, where the reason for not detecting the element of hate speech was unclear, we observed a trend that our approach is prevailing in that category both for the in-domain and the cross-domain settings. Furthermore, we observed that these instances were rather long, which brought us to question whether there is a consistent difference in the length of instance given which model correctly classified the hate speech instance. The analysis of the median length in characters for hate speech instances correctly classified by one model but not by the other for all the languages both in the in-domain and cross-domain settings revealed that there is a drastic tendency for the longer instances to be correctly identified by our approach, while BERT performs better on shorter instances. The only deviation from this trend is the cross-domain Dutch dataset where the instances are overall very short.

We conclude that the stylometry-emotion-based approach performs better on less explicit and longer instances of hate speech, while it lags behind BERT on capturing the more explicit cases.

## 4 Conclusions

The goal of this work was to evaluate and quantify the role of stylometric and emotion-based features in the hate speech detection task. We showed that stylometric and emotional dimensions of hateful content provide useful cues for its detection, as evidenced by the positive impact of stylometric and emotion-based features in various in-domain experiments for all the considered languages. Their contribution remains persistent with respect to domain variations. Under cross-domain conditions, our approach that combines features that capture the targeted phenomena performs better than commonly used features for hate speech detection such as words, character n-grams, and their combination. Finally, we showed that in cross-domain settings our approach that incorporates stylometric and emotion-based features significantly contributes to the recent deep learning models when combined through a majority-voting ensemble, which allows to achieve the highest results for the languages addressed in this work. A manual error analysis showed that this improvement is brought by the ability of stylometric and emotion-based features to capture implicit and longer instances of hate speech. The consistent and substantial improvement in hate speech detection brought by including stylometric and emotion-based features in the different setups and for different languages explored indicates that their usage is a robust indicator of the hateful content.

The importance of stylometric features points in the direction of the existence of a linguistic register for hate speech messages with specific stylistic properties and a negative emotional load. Focusing on these features in text representation leads to more cross-domain robustness.

## Acknowledgments

This work has been supported by the Slovenian Research Agency and the Flemish Research Foundation through the bilateral research project ARRS N6-0099 and FWO G070619N "The linguistic landscape of hate speech on social media", the Slovenian Research Agency research core funding No. P6-0411 "Language resources and technologies for Slovene language", and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (grant no. 875263).

## References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. ACL.

JM Berger and Heather Perez. 2006. The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. Technical report, George Washington University.

Eloi Brassard-Gourdeau and Richard Khoury. 2019. Subversive toxicity detection using sentiment information. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 1–10, Florence, Italy. ACL.

Walter Daelemans, Darja Fišer, Jasmin Franza, Denis Kranjčić, Jens Lemmens, Nikola Ljubešić, Ilia Markov, and Damjan Popič. 2020. The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. Slovenian language resource repository CLARIN.SI.

Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–67, Skopje, Macedonia. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. 2020. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*.

Joseph Fleiss. 1981. *Statistical methods for rates and proportions*, 2nd edition. New York: John Wiley, Heidelberg, Germany.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

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

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Digital Scholarship in the Humanities*, 22(3):251–270.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–579.

Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, pages 59–66, Gothenburg, Sweden. ACL.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar. ACL.

Constantinos Kokkinos and Eirini Kipritsi. 2012. The relationship between bullying, victimization, trait emotional intelligence, self-efficacy and empathy among preadolescents. *Social Psychology of Education*, 15(1):41–58.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Proceedings of the 22nd International Conference on Text, Speech, and Dialogue*, pages 103–114, Ljubljana, Slovenia. Springer.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The LiLaH emotion lexicon of Croatian, Dutch and Slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 153–157, Barcelona, Spain (Online). ACL.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, page 14–17, New York, NY, USA. ACM.

Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2018. Improving cross-topic authorship attribution: The role of pre-processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 10762, pages 289–302, Budapest, Hungary. Springer.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29:436–465.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Switzerland. IW3C2.

John Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution*, pages 1277–1279.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. ACL.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on Twitter: Feature engineering v.s. feature selection. In *Proceedings of the Satellite Events of the 15th Extended Semantic Web Conference*, pages 46–49, Heraklion, Greece. Springer.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, Bochum, Germany. Ruhr-Universität Bochum.

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

Tony Smith and Ian Witten. 1993. Language inference from function words. Technical Report 93/3, Department of Computer Science, University of Waikato. Computer Science Working Papers.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, Edmonton, Canada. ACL.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste1. 2018. Automatic detection of cyberbullying in social media text. *PLoS ONE*, 13(10).

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA.

Bertie Vidgen, 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. ACL.

William Warner 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. ACL.

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

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, Canada. ACL.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *CoRR*, abs/2006.07235.