# The Effects of User Features on Twitter Hate Speech Detection

**Elise Fehn Unsvåg** and **Björn Gambäck**
Department of Computer Science
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
elisefol@stud.ntnu.no, gamback@ntnu.no

## Abstract

The paper investigates the potential effects user features have on hate speech classification. A quantitative analysis of Twitter data was conducted to better understand user characteristics, but no correlations were found between hateful text and the characteristics of the users who had posted it. However, experiments with a hate speech classifier based on datasets from three different languages showed that combining certain user features with textual features gave slight improvements of classification performance. While the incorporation of user features resulted in varying impact on performance for the different datasets used, user network-related features provided the most consistent improvements.

## 1 Introduction

Detecting hate speech has become an increasingly important task for online communities, but automatic hate speech detection is a challenging task, which the majority of the research in the field is targeting through textual features. However, as shown by, e.g., Gröndahl et al. (2018), there is a need for further efforts to improve the quality and efficiency of detection methods, motivating for studies on how non-textual features can be utilised to enhance detection performance.

The goal of this research is to investigate information related to users in the Twitter community that can be helpful in identifying online hate speech, and use this as features in hate speech classification. Information about the users could be either known factors, such as age and gender, or factors derived from behaviour. There exists research that investigates the impact of different features, and research about the personality and behaviour of users expressing hate speech. However, there is little research that combines the two topics.

Most early studies on automatic recognition of online hate speech focused on lexicon-based approaches for detecting "bad" words, with Kwok and Wang (2013) finding that 83% of their data was annotated racist due to the presence of offensive words. However, these approaches tend to give low precision by mistakenly classifying all messages containing specific terms as hate speech, which is particularly problematic on social media sites that have a relatively high prevalence of offensive words (Wang et al., 2014). After all, hate speech can be much more sophisticated than that.

Finding the features that best represent the underlying phenomenon of hate speech is challenging. Later studies have mainly focused on content-based text classification using features such as the appearance or frequency of words, spelling mistakes or semantic meaning, but while these methods perform relatively well, there is still need for improvements to increase the quality of detection.

The rest of the paper is structured as follows: Section 2 discusses previous studies related to the authors of hate speech and Section 3 presents the datasets used together with an analysis of user characteristics. Section 4 describes the classifier developed, while Section 5 details the experiments conducted to measure the impact of user features. Section 6 sums up the research contributions along with suggestions for potential future work.

## 2 Related Work

Including user information in methods for detecting hate speech is an under-researched area. However, related to hate speech detection are studies of the people that post hateful content online, including characteristics and behavioural traits that are typical of the authors behind aggressive behaviour, hate speech or trolling. Chen et al. (2012)



*Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 75–85
Brussels, Belgium, October 31, 2018. 

proposed a Lexical Syntactic Feature architecture to bridge the gap between detecting offensive content and potential offensive users in social media, arguing that although existing methods treat messages as independent instances, the focus should be on the source of the content. Waseem and Hovy (2016) stated that among various extra-linguistic features, only gender brought improvements to hate speech detection. Papegnies et al. (2017) mention a plan to use context-based features for abuse detection, and especially those based on the networks of user interactions. Several authors share this intention, but face the challenge that user information often is limited or unavailable.

Wulczyn et al. (2017) qualitatively analyzed personal attacks in Wikipedia comments, showing that anonymity increases the likelihood of a comment being an attack, although anonymous comments only contributed to less than half of the total attacks. The study also suggested that personal attacks cluster in time, which may be because one attack triggers another. In another qualitative analysis, Cheng et al. (2015) characterized forms of antisocial behaviour in online discussion communities, comparing the activity of users that have been permanently banned from a community to those that are not banned. The study found the banned users to use less positive words and more profanity, and to concentrate their efforts in a small amount of threads. They also receive more replies and responses than other users.

Hardaker (2010) defined a *troller* as a user who appears to sincerely wish to be part of a group, including professing, or conveying pseudo-sincere intentions, but whose real intentions are to cause disruption or to trigger conflict for the purposes of their own amusement. Buckels et al. (2014) studied the characteristic traits of Internet trolls by looking at commenting styles and personality inventories, and found strong positive relations among commenting frequency, trolling enjoyment and trolling behaviour and identity. Cheng et al. (2017) proposed that an individual's mood and seeing troll posts by others trigger troll behaviour.

Most similar to the objectives of the present work, Chatzakou et al. (2017) investigated user features that can be utilized to enhance the detection and classification of bullying and aggressive behaviour of Twitter users. They found that network-based features (such as the number of friends and followers, reciprocity and the position in the network) were particularly useful and effective in classifying aggressive user behaviour.

## 3 Data Analysis

Creating datasets of hate speech is time consuming, as the number of hateful instances in online communities is relatively low. The datasets available are also often created for different tasks, and from different types of media and languages, and therefore vary in characteristics and types of hate speech. Sources include Twitter (Waseem and Hovy, 2016; Fortuna, 2017; Ross et al., 2016), Wikipedia (Wulczyn et al., 2017), and Fox News (Gao and Huang, 2017). Furthermore, many datasets (from Yahoo, SpaceOrigin and Twitter) are not publicly available (Djuric et al., 2015; Nobata et al., 2016; Papegnies et al., 2017; Chatzakou et al., 2017), while others are available only under some restrictions (Davidson et al., 2017; Golbeck et al., 2017). This may be due to privacy issues or considering the content of the datasets: Pavlopoulos et al. (2017) made their Greek Gazzetta dataset available by using an encryptor to avoid directly publishing hate speech content.

Here, three datasets were used to investigate the characteristics of users for increased insight and to allow comparisons of the findings. All datasets have Twitter as their source, ensuring that the same information could be retrieved. However, the datasets differ in terms of annotations, size and characteristics, and come from three different languages: English (Waseem and Hovy, 2016), Portuguese (Fortuna, 2017), and German (Ross et al., 2016). The datasets contain tweet IDs that can be used to retrieve the actual text, information about the tweet or information about the user who has posted it. As user information is something that should be handled with care, it is important to mention that no attempt was made to directly identify the actual users.

Tweet IDs may become unavailable, either by the tweet having been deleted, or if the user who posted the tweet has become suspended or has deleted their account. Therefore, a review of the availability of the tweets in all datasets was conducted prior to the investigation of characteristics, and will be described first below, before going into details of the analysis of the user characteristics in the three datasets. The statistics of the actually available tweets and posting users in the datasets as included in this work are shown in Table 1.

| Label | ENG | | POR | | GER | |
|---|---|---|---|---|---|---|
| | Tweets | Users | Tweets | Users | Tweets | Users |
| Hate | 4,968 | 539 | 649 | 376 | 98 | 47 |
| None | 10,759 | 1,569 | 2,410 | 634 | 243 | 123 |
| Total | 15,727 | 2,108 | 3,059 | 1,010 | 341 | 170 |

Table 1: Available tweets and users in the datasets

### 3.1 Datasets

The English dataset by Waseem and Hovy (2016) is publicly available on GitHub.[1] The Twitter search API was used to collect the corpus, and in total 16,907 tweets (from 2,399 users) were annotated either as racist, sexist or neither. The dataset contains more instances of neutral than racist or sexist tweets. This unbalance was intended by the developers, to make the corpus more representative of the real world, where hate speech is a limited phenomenon. Since the dataset was developed in 2016, the Python library Tweepy was used here to filter out any unavailable tweets and users. Furthermore, the original "Sexism" and "Racism" classes were merged into one "Hate speech" class. 1,180 of the original tweets were no longer available, which also impacted the number of users in the dataset. The remaining tweets and users are presented in Table 1, in the 'ENG' column.

Fortuna (2017) developed a dataset consisting of 5,668 Portuguese tweets and made it available through the INESC TEC research data repository.[2] Tweets were collected through the Twitter API with searches based on keywords related to hate speech and Twitter profiles known for posting hate messages. Fortuna aimed to have a higher proportion of hate speech messages than other related datasets, and 22% of the tweets were annotated as hate speech. She annotated nine direct hate speech sub-classes, but in the present work those will be merged into one hate speech class. In total there are 5,668 annotated tweets by 1,156 distinct users; however, the distribution of users within the target classes was not specified. Today, close to half of the tweets in both classes are unavailable; however, as shown in the 'POR' column of Table 1, there are still 1,010 users available, meaning that the unavailability of tweets did not heavily affect the number of users. While the original dataset had a binary value for the presence of hate speech and subcategories as labels, the target classes were here changed to "Hate speech" and "None".

[1] github.com/ZeerakW/hatespeech

[2] rdm.inesctec.pt/dataset/cs-2017-008

To investigate the issue of reliability concerning hate speech annotation, Ross et al. (2016) compiled a German hate speech corpus with tweets linked to the refugee crisis in Europe. By using known insulting or offensive hashtags, a total of 13,766 tweets were collected, 469 of which were annotated by two annotators for presence or absence of hate speech. In Table 1 the column 'GER' shows the availability of the tweets in the dataset and the number of users in each target class. It was beneficial to transform the labels of the dataset into binary classes, to equal the labelling of the other datasets. Therefore, a tweet that was labelled "Yes" by one or both of the annotators was assigned to the "Hate speech" class. Hence, the "Hate speech" class consists of 65 available tweets labelled as hate speech by one annotator, and 33 labelled hate speech by both annotators.

### 3.2 Characteristics

A quantitative analysis was conducted to better understand the characteristics of the users in the datasets, based on the proposed features in Section 2 and other information about the user available through the Twitter API. All datasets included several tweets from the same users; tweets that then can be present in both target classes. However, to better distinguish between users and avoid redundancy in the analysis, users who are present in both target classes are here only included as users within the "Hate speech" class.

**Gender:** Twitter does not require users to register their gender, so no explicit gender field is retrievable through the Twitter API. Finding the gender distribution for users in the dataset is therefore challenging. Waseem and Hovy (2016) investigated the distributions of gender in their original dataset through extracting gender information by looking up usernames and names in the user profiles, and then comparing these to known male or female given names. A similar approach was used here, by incorporating lists of common international, Portuguese, German, and English names. In addition, the user descriptions were also considered, as users often give a more detailed description there of who they are, e.g., "I am a mom of three boys". A risk with this approach is that names or descriptions may mistakenly be classified as the wrong gender, and therefore the gender findings may not be entirely accurate. Names that can be both female and male have been avoided.

(a) English (Waseem and Hovy, 2016) (b) Portuguese (Fortuna, 2017) (c) German (Ross et al., 2016)

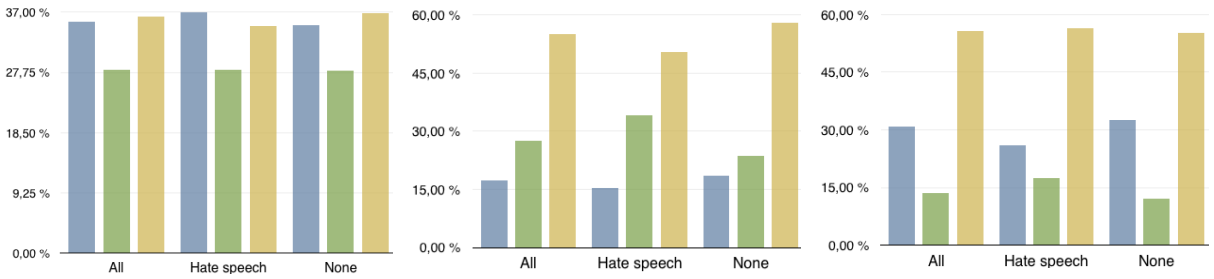



Figure 1: Gender distribution of users in the datasets: blue = male, green = female, beige = unidentified

Waseem and Hovy (2016) expressed that the gender of about half the users could not be identified by their approach, and that the male gender was over-represented in all categories. Figure 1a presents the gender distribution derived here, with significant differences to their findings. In particular, the female users are identified to a much larger degree, with the distribution of male, female and unidentified users being more equal; the fraction of unidentified users has decreased from 50% to 36%. Still, a higher amount of male users are identified than female, which is also the case for the dataset by Ross et al. (2016). In contrast, the gender distribution derived from the dataset by Fortuna (2017) shows a majority of the identified user genders to be female. In that dataset there is also a large number of unidentified genders (55%), which is equal to the number of unidentified users in the dataset by Ross et al. (2016).

**User Network:** The user networks are defined here as their social networks on Twitter, i.e., who a user follows (called 'following' or 'friends' on Twitter) and who follows that user ('followers'). Chatzakou et al. (2017) found network-based features to be very useful in classifying aggressive user behaviour. They investigated features such as the ratio of followers to friends, the extent to which users reciprocate the follow connections they receive from others, and the users' tendency to cluster with others.

In Figure 2a, the relationship between a user's friends and followers in the dataset by Waseem and Hovy (2016) is illustrated. The majority of users form a cluster in the area below 10,000 friends and 50,000 following. Beyond this cluster, it appears as users of the "None" class are most common, with the exception of one outlier of the "Hate speech" class with about 228,00 followers and no friends. It is difficult to say whether this trend can be generalized, or is caused by the uneven number of users in the two target classes.

Figure 2b shows the distribution of friends and followers for the users in the dataset by Fortuna (2017). A general observation is that the users of this dataset often tend to have more followers than friends. Furthermore, there is little that distinguishes the users of the two classes regarding the number of friends and followers. The number of users in the dataset by Ross et al. (2016) is considerably lower than the other datasets, and may explain the lower number of friends and followers for the users, as shown in Figure 2c. There is

(a) English (Waseem and Hovy, 2016) (b) Portuguese (Fortuna, 2017) (c) German (Ross et al., 2016)

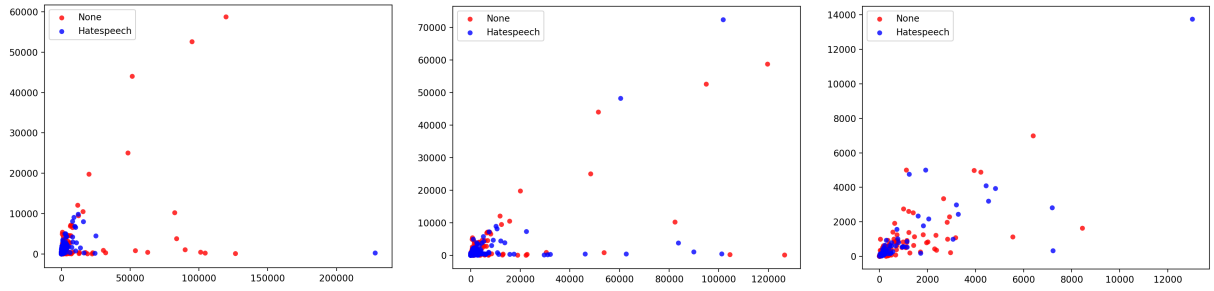



Figure 2: Distribution of users based on their network (number of friends vs number of followers)

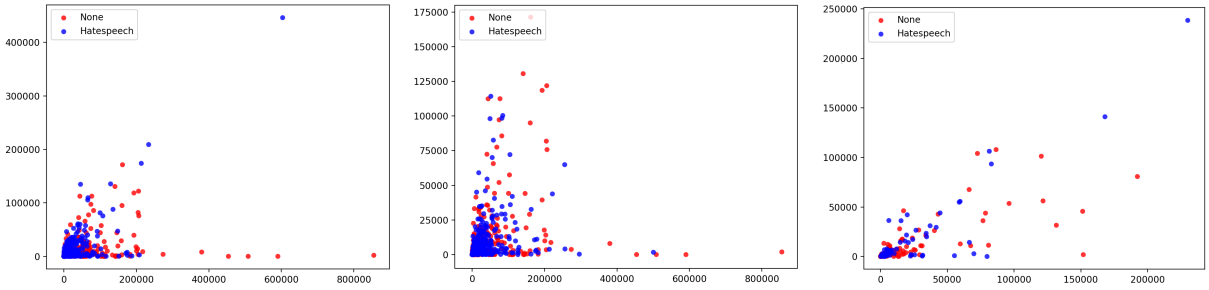



Figure 3: Distribution of users based on activity (number of favourites given vs number of status updates)

an outlier in the "Hate speech" class with about 13,000 followers and 14,000 friends, but the rest of the users are somewhat evenly distributed.

**Activity:** Previous research suggests that both high and low activity levels can be related to posting hate speech content. Buckels et al. (2014) found commenting frequency to be positively associated with trolling enjoyment and Cheng et al. (2015) suggested that frequently active users are often associated with anti-social behaviour online. In contrast, Wulczyn et al. (2017) found users that launched personal attacks on Wikipedia regardless of activity level. Here, activity is defined by the information that can be extracted through the Twitter API. Tweepy enables the retrieval of the number of tweets a user has posted (also known as 'status updates' on Twitter) and the number of 'favorites' they have given to tweets by others (corresponding to 'likes' in other online media).

In Figure 3 the relationship between a user's number of favourites given and total number of statuses is illustrated, showing that there is a general tendency to have a larger number of status updates than favourites. With the exception of one outlier in the "Hate speech" class with over 400,000 favourites and over 600,000 statuses, the majority of users in both classes of the English dataset form a cluster below 50,000 favourites and 200,000 statuses. In the Portuguese dataset, the users of both target classes are somewhat evenly distributed, and in general the users of this dataset have posted below 200,000 tweets and given below 25,000 favourites. The number of status updates and the number of favourites for the users in the German dataset are much lower than in the other datasets, and similarly to the findings investigating the users' network, there is no clear distinction between the activity characteristics of the users in the target classes.

| | ENG | | POR | | GER | |
|---|---|---|---|---|---|---|
| Feature | Hate | None | Hate | None | Hate | None |
| Geotag | 51.7 | 48.6 | 58.8 | 58.2 | 16.1 | 26.6 |
| Profile | 60.1 | 72.4 | 75.1 | 67.4 | 50.0 | 45.9 |
| Image | 98.1 | 96.2 | 99.6 | 98.2 | 98.4 | 98.2 |

Table 2: User profile characteristics (%)

**User Profile:** Twitter enables users to customize their profile pages, e.g., by changing theme colour, or by adding a profile or header picture. In addition, users can add a bio description, a geographical location or a web page link. Wulczyn et al. (2017) found personal attacks to be more prevalent among anonymous users than registered users. Therefore the elements of a user's profile that can be personalized were examined with the underlying assumption that personalizing the profile is contradicting to remaining anonymity. The elements retrieved were the number of public lists a user has joined, geotagging of tweets, the profile image, and whether or not the user has altered a default theme or background of the profile.

The users in the English data are somewhat equally divided between enabling and disabling of geotagging for both target classes, as seen in Table 2. The distribution is similar for the geotagging characteristic of users in the Portuguese data. However, the majority in both target classes in the German data have disabled geotagging, with a slightly higher percentage for the users in the "Hate speech" class. There is a tendency of the users in the English and Portuguese datasets to rather have a customized profile page than a standard, while the German data is more evenly distributed. Nearly all the users in the three datasets have changed their profile image. For all the datasets, the percentage of changed profile images is also marginally higher for the users in the "Hate speech" class than the users in the "None" class.

## 4 Classification Setup

The analysis of the datasets presented in the previous section indicated that none of the investigated user characteristics could be clearly used to differentiate textual tweets annotated "Hate speech" and "None". However, the impact of user features in detection may become more visible when tested through a classifier. To investigate the possible effects user features have on the performance of hate speech classification, a baseline hate speech classifier was implemented and trained only on the textual tweets from the datasets, and then compared to a classifier that also incorporated user features. Along with observing the overall effects of user feature inclusion, the impacts of the individual features and feature subsets were investigated.

A basic hate speech classifier needs to include preprocessing of the textual input, feature extraction, and a choice of actual classification model. These will be addressed in turn below, while classification results will be given in the next section.

**Preprocessing:** Text processing is a difficult task due to the noise contained in language and should be done with care, to avoid losing any important features. This is particularly proliferent in social media such as Twitter, which also introduces domain-specific challenges: the character limit in a tweet increases the use of abbreviations, while including non-textual content (e.g., URLs, images, user mentions and retweets) is common.

The Natural Language Toolkit (Bird et al., 2009) was used for preprocessing of the data, through: (i) removal of Twitter specific information (user mentions, emoticons, retweets, URLs, and hashtag symbols; only retaining textual content), (ii) tokenization, (iii) lowercasing, and (iv) stop word removal (with different stop word lists for the datasets, due to the different languages).

**Feature Extraction and Representation:** Having found many tweets to be annotated racist due to the appearance of offensive words, Kwok and Wang (2013) constructed a vocabulary using unigram features only. However, this fails to capture relationships between words, so Nobata et al. (2016) added syntactic features, while also employing n-grams and distributional semantic derived features. They found combining all features to yield the best performance, but character n-grams made the largest individual feature contribution. Mehdad and Tetreault (2016) specifically investigated character-based approaches and showed them to be superior to token-based approaches and other state-of-the-art methods.

Since n-grams thus have been shown to be very useful in hate speech classification, both character n-grams and word n-grams were tested here to represent the textual content of the tweets. A TF-IDF approach was used to represent the n-gram features, and ranges up to $n$=6 tested. Higher values of $n$ were not considered due to the computational effort required. The most suitable type of n-gram and n-gram range were explored through a grid search, and finding different alternatives for representing the tweets suiting the different datasets.

**Classification Model:** Supervised machine learning classifiers have been the most frequently used approaches to hate speech detection, in particular Support Vector Machines (SVM) and Logistic Regression (LR). Davidson et al. (2017) found LR and linear SVM to perform better than other models, such as Naïve Bayes, Decision Trees, and Random Forests. A comparative study performed by Burnap and Williams (2015) concluded that an ensemble method seemed most promising. Deep learning methods have also been investigated, both Recurrent Neural Networks (Pavlopoulos et al., 2017; Mehdad and Tetreault, 2016), Convolutional Neural Networks (Gambäck and Sikdar, 2017), and combinations (Zhang et al., 2018). Badjatiya et al. (2017) used various deep learning architectures to learn semantic words embeddings and showed these to outperform character and word n-grams.

Here a Logistic Regression model was chosen due to its simplicity and its common usage in language classification. This is also in line with the note by Gröndahl et al. (2018) that a simple LR model performed on par with more complex models in their comparison of hate speech detection classifiers. As the aim here was not to implement the best performing classifier or to compare methods, but to investigate the effects of user features, no other classification models were tested.

## 5 Experiments and Results

The datasets were initially split into training data and test data to ensure that the model performance was evaluated on unseen data. A grid search with 10-fold cross-validation over the training data was used for selecting model parameters. The classification model with the chosen hyperparameters

| n-gram | ENG | | POR | | GER | |
|---|---|---|---|---|---|---|
| | Word | Char | Word | Char | Word | Char |
| [1, 1] | .8166 | .7399 | **.7769** | .6927 | .7227 | .7185 |
| [1, 2] | .8168 | .8020 | .7718 | .7383 | .7185 | **.7269** |
| [1, 3] | .8147 | .8201 | .7688 | .7525 | .7227 | .7101 |
| [1, 4] | .8119 | .8226 | .7667 | .7657 | .7227 | .7101 |
| [1, 5] | .8117 | **.8248** | .7637 | .7698 | .7227 | .7143 |
| [1, 6] | .8110 | .8237 | .7612 | .7759 | .7227 | .7143 |

Table 3: Grid search of n-gram parameters

| Class | ENG | | | POR | | | GER | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| None | .83 | .94 | .88 | .79 | .96 | .87 | .68 | .99 | .81 |
| Hate | .82 | .58 | .68 | .82 | .40 | .54 | .50 | .03 | .06 |
| Avg. | .83 | .83 | .83 | .80 | .79 | .79 | .62 | .68 | .65 |

Table 4: Baseline model performance on test data

was then evaluated on the test set. This section first presents results from baseline classification with only n-gram features, and then discusses the effects of incorporating user features.

### 5.1 Classifier with Text Features

The dataset provided by Waseem and Hovy (2016) contained 15,727 available English tweets, that were split into a training set of 11,008 tweets and a test set containing 4,719 tweets, of which 3,275 were classified as non-hate speech. A grid search found that character n-grams in range [1,5] provided the best performance, as shown in the column 'ENG' in Table 3. Table 4 shows the performance metrics of the model, where 0.83 was the macro average $F_1$-score. Both the precision and recall values are higher for the "None" class. However, the recall value for the "Hate speech" class obtained for this dataset is higher than for the other datasets, most probably due to the larger amount of available training data.

3,059 tweets from the Portuguese dataset by Fortuna (2017) were used, with the training set containing 2,636 tweets and the test set 423, of which 126 were annotated as hate speech. Word unigrams yielded the best performance (Table 3), and the macro average $F_1$-score obtained for the test data was 0.79 (Table 4). The precision obtained for "Hate speech" is slightly higher than for the "None" class, while the recall is much lower.

The German dataset by Ross et al. (2016) is considerably smaller than the other datasets, containing only 341 tweets, that were split into a training set of 238 and a test set of 103 (33 hateful). A grid search of the n-gram parameters ('GER' in Table 3) showed a character n-gram with the range [1,2] produced the best 10-fold cross validation results on the training data. On the unseen test data, this model received a macro average $F_1$-score of 0.65 (Table 4), with the score severely hampered by the classifier only being able to identify 3% of the instances of the "Hate speech" class. This is most likely due to small the size of the dataset, resulting in an insufficient amount of training instances. Notably, Ross et al. (2016) did not develop this dataset primarily for classification, but for investigating hate speech annotation reliability. Their study concluded that the presence of hate speech perhaps should not be considered a binary yer-or-no decision; however, this is how the current classification model is operating.

| Features | ENG | | | POR | | | GER | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| n-grams | .83 | .83 | .83 | .80 | .79 | .79 | .62 | .68 | .65 |
| + gender | .83 | .83 | .83 | .80 | .79 | .79 | .69 | .70 | .69 |
| + network | .84 | .85 | .84 | .81 | .81 | .81 | .63 | .68 | .65 |
| + activity | .83 | .84 | .83 | .79 | .79 | .79 | .68 | .69 | .68 |
| + profile | .83 | .83 | .83 | .80 | .80 | .80 | .71 | .71 | .71 |
| + all | .86 | .86 | .86 | .79 | .79 | .79 | .63 | .68 | .65 |

Table 5: Impact of different user feature sets

### 5.2 Classifier with Text and User Features

In the second part of the experiments, the classifier was expanded to incorporate various user features and subsets. Four types of in total ten features were experimented with:

**Gender:** male and female,
**Network:** number of followers and friends,
**Activity:** number of statuses and favourites,
**Profile:** geo enabled, default profile, default image, and number of public lists,

where the "number of" features are integer valued, while all the other features are binary (boolean).

Table 5 repeats the performance of the baseline model (n-grams only, in row 1) and then shows n-grams along with various subsets of user features. Including all user features yielded the largest improvement over the baseline on the Waseem and Hovy (2016) dataset, with the 'Network' feature subset making the largest difference. 'Gender' did not improve performance at all, while 'Activity' and 'Profile' provided very slight improvements. Each individual feature was also tested along with

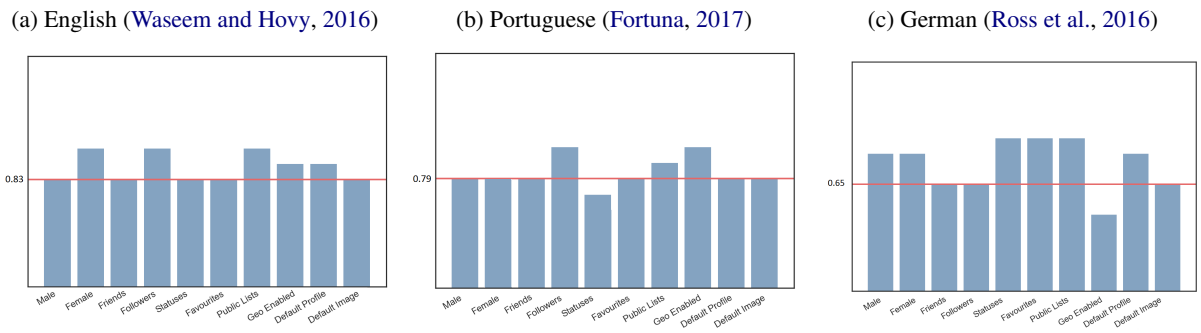


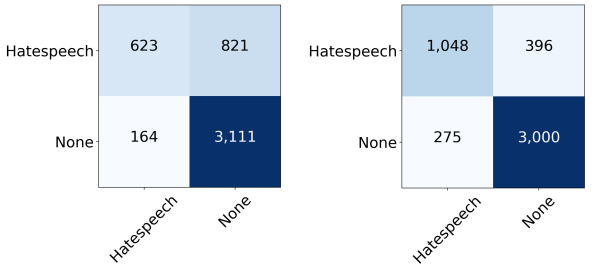

Figure 4: $F_1$ of individual features along with n-grams. (Red lines = average $F_1$ using n-grams only.)
[Features: Male, Female, Friends, Followers, Statuses, Favourites, Public Lists, Geo Enabled, Default Profile, Default Image.]

n-grams, as shown in Figure 4a. Half of the features had no impact on performance; 'Default profile' and 'Geo enabled' increased $F_1$ by 0.1, while 'Female', 'Followers' and 'Public lists' had the most impact, increasing $F_1$ by 0.2.

The incorporation of all user features on the Fortuna (2017) dataset resulted in a slightly worsened performance. This was also the case for inclusion of the 'Activity' subset, while including 'Network' improved performance. 'Gender' and 'Profile' made no major impact on the scores. Of the individual features, 'Followers' and 'Geo enabled' resulted in the largest $F_1$-score increase when used in combination with n-gram features, as shown in Figure 4b. In addition, the inclusion of 'Public lists' also slightly improved the $F_1$-score. Interestingly, the inclusion of 'Statuses' actually worsened model performance.

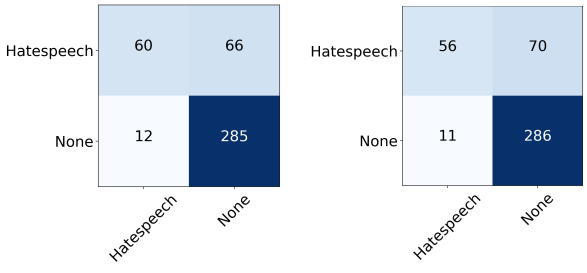
By only using word unigrams, the baseline classifier only received a recall value of 0.03 for the hate speech class of the dataset by Ross et al. (2016), as shown in Table 4. Looking at Table 5, we see that the 'Gender', 'Activity' and 'Profile' feature subsets resulted in improvements of the average $F_1$-score. The inclusion of all the features and the 'Network' subset had no effect on the average $F_1$ score. The inclusion of 'Activity' increased the $F_1$ by 0.02, 'Gender' increased it by 0.04, and 'Profile' had the largest impact by increasing $F_1$ by 0.06. These results are consistent with the testing of the individual features shown in Figure 4c, where 'Male', 'Female' and 'Profile' have a large impact on performance. However, the 'Statuses', 'Favourites', and 'Public lists' had the largest improvement by 0.4. Of the individual features included, only 'Geo Enabled' lead to a decreased $F_1$ score over the baseline.

The results are notably affected by the uneven distribution of instances in the target classes, as shown by significantly lower $F_1$-scores for "Hate speech" than "None" for all datasets. This was reflected clearly by a closer comparison of classifier output for the English data with n-grams and with all user features (i.e., the setup which yielded the best classifier performance on this dataset): the number of correctly labelled "Hate speech" instances increased from 623 to 1,048 (of 1,444) while the correctly labelled "None" instances decreased slightly, from 3,111 to 3,000 (of 3,275), as illustrated by the confusion matrices in Figure 5.




(a) Only n-gram features (b) Adding all user features

Figure 5: English dataset confusion matrices




(a) Only n-gram features (b) Adding 'Network' subset

Figure 6: Portuguese dataset confusion matrices

A similar pattern was observed also for the German data, for which the optimal classifier setup was to include the 'Profile' feature subset. However, for Portuguese where the best classifier utilised only the 'Network' user features, Figure 6 shows that adding those features produced a decrease of correctly labelled "Hate speech" (from 60 to 56 of 116) with marginally increased correctly labelled "None" instances (from 285 to 286 of 297), possibly since data sparsity made the model interpret the added features as noise.

## 6 Conclusion and Future Work

There are several challenges linked to the detection of harmful online behaviour, such as detection beyond simply recognising offensive words. Aiming to address this gap, the paper investigated the potential and effects of including user features in hate speech classification, focusing on Twitter. A quantitative analysis of three datasets in three different languages indicated that there were no particular characteristics distinguishing the users who have had tweets annotated as hate speech and those who have seemingly not.

However, systematically incorporating the user features into a Logistic Regression-based hate speech classifier in conjunction with word and character n-gram features allowed observations of the effects of individual features and feature subsets. Experimental results showed that the inclusion of specific user features, in addition to n-grams, caused a slight improvement of the baseline classifier performance.

Of all individually tested feature subsets, 'Network' (i.e., the number of followers and friends) caused the largest improvement of the classifier performance on the English and Portuguese datasets, corroborating the findings of Chatzakou et al. (2017) that network-based features are powerful for detecting aggressive behaviour. This subset improvement may have been affected by the individual feature (number of) 'Followers', which also increased the $F_1$-score on the two datasets. The other features had inconsistent effects on the different datasets, suggesting that the impact is highly dependent on the data or the subtask the data was created for. The experiments also found the inclusion of some user features to be detrimental to model performance, while some user features were ineffective alone, but improved model performance when combined with others.

Interestingly, the 'Gender' feature subset mainly failed to give any $F_1$-score improvements, in contrast to the result by Waseem and Hovy (2016). While other user features are easily retrievable through the Twitter API, user gender was derived from a comparative method, classifying more users by gender than in the work by Waseem and Hovy. However, also the method used here is still unable to identify the gender of a large amount of users in all datasets, so combinations with other gender identification methods would be needed to properly investigate the impact gender has in hate speech detection. As of now, it can be argued that gender is not a useful feature to include, at least where it cannot be directly extracted.

One limitation of using several datasets is that they were developed for different subtasks and languages, with different geographical areas of the users in the datasets, and in particular with different interpretations and annotations of hate speech. However, the main difference of the datasets is the size and hence number of instances available for model training, which probably is the main reason for the different results. Still, the results combine to show a potential for incorporating user features to improve hate speech detection performance.

There is a great amount of information related to the users of Twitter that was not used in the experiments, but that could be retrieved or derived from user behaviour. Examples include considering the time of tweeting, investigations of relationships with other users, communication with other users, and what content users are exposed to. It is in general important to not only consider who the users are or what they have written, but also their context and how they are affected by surrounding factors in their online communities, as well as combinations of those issues, since what can be considered as hate speech by one user in a specific context may be considered as non-hate speech if written by another user or in another context.

## Acknowledgements

Thanks to Johannes Skjeggestad Meyer for fruitful discussions and to the providers of the datasets used in this work: Benjamin Cabrera, Guillermo Carbonell, Paula Fortuna, Dirk Hovy, Nils Kurowsky, Michael Rist, Björn Ross, Zeerak Waseem, and Michael Wojatzki.

## References


Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, Perth, Australia. International World Wide Web Conferences Steering Committee.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, California.

Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus. 2014. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102.

Peter Burnap and Matthew Leighton Williams. 2015. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. *Policy and Internet*, 7(2):223–242.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 13–22, Troy, New York, USA. ACM.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 71–80, Amsterdam, Netherlands. IEEE.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 1217–1230, Portland, Oregon, USA. ACM.

Justin Cheng, Christian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the 9th International Conference on Web and Social Media*, pages 61–70, University of Oxford, Oxford, UK. AAAI Press.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 512–516, Montréal, Québec, Canada. AAAI Press.

Nemanja Djuric, 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*, pages 29–30, Florence, Italy. ACM.

Paula Fortuna. 2017. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master's thesis, Faculdade De Engenharia Da Universidade Do Porto, Porto, Portugal, June.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, Canada. ACL.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *CoRR*, abs/1710.07395.

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

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. *CoRR*, abs/1808.09115.

Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 6(2):215–242.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1621–1622, Bellevue, Washington. AAAI Press.

Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, USA. ACL/SIGDIAL.

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, Montréal, Québec, Canada. International World Wide Web Conferences Steering Committee.

Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linares. 2017. Impact of content features for automatic online abuse detection. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 10762 of *Lecture Notes in Computer Science*, Budapest, Hungary. Springer.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. ACL.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, Germany.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in English on Twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 415–425, Baltimore, Maryland, USA. ACM.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) Student Research Workshop*, pages 88–93, San Diego, California, USA. ACL.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Perth, Australia. International World Wide Web Conferences Steering Committee.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *15th European Semantic Web Conference*, pages 745–760, Heraklion, Greece. Springer.