International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00437-1

REGULAR PAPER

Enhancing hate speech detection with user characteristics

Rohan Raut[1] · Francesca Spezzano[1]

Received: 15 April 2023 / Accepted: 26 July 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract
Social media provide users with a powerful platform to share their ideas. Using one's right to expression to incite hate toward a particular group of people is inappropriate. However, hate speech is pervasive in our society. Spreading hate through online social networks like Facebook, Twitter, Tiktok, and Instagram is commonplace in today's milieu. One such case is the unprecedented COVID-19 pandemic, which engendered anti-Asian hate. In the current literature, there is limited study on using user features in conjunction with textual features to detect hate speech. In this paper, we propose to combine tweet textual features with a variety of user features to improve the state-of-the-art hate speech detection techniques. The user feature we propose consists of demographic, behavioral-based, network-based, emotions, personality, readability, and writing style. To test our approach, we used four different English datasets gathered from Twitter and available in the public domain. Our results show that combining tweet textual features with the proposed user features improves hate speech detection up to +0.32 in F1 score and beats previously proposed approaches that use a limited number of user features. The analysis of the most important features confirms that hateful tweets or their authors express more negative emotions and use more swear words.

Keywords Hate speech · Tweet text classification · User information and behavior

1 Introduction

Online social networks (OSN) have become a powerful means for people to express their views and opinions. The ability to express oneself is a fundamental human right. However, one should not leverage that right to direct hate toward a certain group of people. According to the American Bar Association people can express hate toward a topic without it being considered hate speech. However, speech that demeans based on race, ethnicity, gender, religion, age, disability, or any other similar ground is understood as hate speech. The unprecedented COVID-19 pandemic sparked a plethora of Anti-Asian hate. As per a Forbes article, Anti-Asian hate speech skyrocketed with a rise of 1662 percent in anti-Asian hate speech in 2020 compared with 2019. Furthermore, the same article states that a new analysis of 263 million online conversations in the UK and USA on social media sites, blogs, and forums has found that hate speech increased dramatically between 2019 and mid-2021.

Various organizations have been working in order to regulate social media platforms in order to reduce hate speech. For instance, the European Union has agreed on a Digital Services[1] [2]Act[3] to combat hate speech and misinformation aggressively. Social media websites have policies on speech, which are allowed on their platform. As per Twitter Policy,[4] users may not promote violence against or directly attack or threaten other people based on race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. As per a report compiled by UNESCO and the Oxford Internet Institute,[5] between July and December 2020, 1,628,281 pieces of content were found to violate Twitter's hate speech policy. Due to

✉ Francesca Spezzano
francescaspezzano@boisestate.edu

Rohan Raut
rohanraut@u.boisestate.edu

[1] Boise State University, Boise, ID, USA

[1] https://abalegalfactcheck.com/articles/hate-speech.html.

[2] https://www.forbes.com/sites/emmawoollacott/2021/11/15/anti-asian-hate-speech-rocketed-1662-last-year/.

[3] https://www.consilium.europa.eu/en/press/press-releases/2022/04/23/digital-services-act-council-and-european-parliament-reach-deal-on-a-safer-online-space/.

[4] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[5] https://unesdoc.unesco.org/ark:/48223/pf0000379177.

Published online: 16 August 2023

the accessibility to spread hate and the ability of social media websites to be used for hate, accurate automatic detection of hate speech is imperative these days. Twitter receives over six hundred tweets per second and five hundred million per day [1]. Using human resources to manually filter hate speech with such gigantic traffic is next to impossible.

Hate speech detection is an exacting task. Depending on the reader, the content may or may not be considered hate. Niceties in the content add to the challenge of hate speech detection. Due to this fact, dataset quality highly depends on the inter-annotator agreement. As per the Cambridge Dictionary,[6] hate speech is public speech that expresses hate or encourages violence toward a person or group based on something such as race, religion, sex, or sexual orientation.

There has been active research going on in the hate speech detection field. The current state of the art has used linguistic (semantic and psycho-linguistic) and hashtag features [2] or neural networks composed of convolutional and bidirectional gated recurrent unit (BiGRU) [3] to classify hate speech in the text of the tweets.

Adding user information can help improve hate speech detection. However, to the best of our knowledge, there is limited literature on combining text features with user features to enhance hate speech detection. Just a few features have been investigated so far, e.g., gender, location, number of followers and friends, or user profile features, and not across multiple datasets [4–6].

In this paper, we study a variety of user features, including demographics, emotions, personality, readability, and writing style, as well as network and behavioral characteristics to complement tweet textual features to improve hate speech detection. Moreover, we compare our proposed approach across multiple publicly available datasets. We have also collected additional user information for these datasets that we plan to make publicly available to the research community upon publication of the paper.

We posit that user features that embody details like average emotions of their tweets, their number of followers, and other user details will improve hate speech detection compared to the current state-of-the-art. We tested this by training a classification model combining BERT, psycho-linguistic features, and user features. Our experimental evaluation conducted on four publicly available datasets shows that enhancing tweet textual features with features characterizing the author of the tweet improves hate speech detection up to +0.32 in F1 score and beats previously proposed approaches that use a limited number of user features. As expected, our feature importance analysis revealed that hate speech tweets or their authors express more negative emotions like anger, fear, sadness, and annoyance and use more swear words than normal tweets and their authors.

The paper is organized as follows. Section 2 summarizes related work, Sect. 3 describes the datasets we used in this paper, Sect. 4 presents our proposed approach to hate speech detection and presents the user features we used, Sect. 5 reports on our experimental evaluation and, finally, conclusions are drawn in Sect. 6.

2 Related work

Hate speech detection has been performed in various contexts, including but not limited to hate against women, immigrants, etc. Recent studies have shown a rise in anti-Asian hate speech on social media due to the COVID-19 pandemic [2, 7, 8]. He et al. [2] explored a large dataset of 206 million tweets they collected and showed that nodes exposed to hate are more likely to spread hate. Also, nodes exhibited homophily in both hate spreaders and counter-speech users. Counter-speech reduced the probability of neighbors becoming hateful. To mitigate racial bias, Xia et al. [9] used adversarial training to introduce a hate speech classifier that detects toxic sentences. The study found that there is a high correlation between annotators' perceptions of toxicity and signals of African American English (AAE). The method can reduce the false positive rate for AAE text while only minimally affecting the performance of hate speech classification. Davidson et al. [10] examined five different datasets containing hate and abusive tweets. The study trained classifiers and compared the prediction of their classifier written in AAE with the ones written in Standard American English. The results showed systematic racial bias in all the datasets the study analyzed. Overall, hate speech detection is a challenging task as it is dependent on time and context. The generalizability of hate speech detection is difficult due to the nature of online hate speech, limits of existing NLP methods, and dataset building [11].

Recently, several studies have proposed deep learning models for hate speech detection. Ding et al. [12] used a stacked BiGRUs model with a capsule network system, while Khan et al. [3] used a deep neural network with convolutional and BiGRU layers to outperform previous models such as the ones by Ding et al. [12] and Roy et al. [1]. Alatawi et al. [13] explored two approaches, one with a bidirectional LSTM model and another with BERT, with BiLSTM detecting intentional misspellings and common slang better. Roy et al. [1] proposed a deep CNN using GloVe embedding vectors with multiple convolution layers. Gambäck and Sikdar [14] used word2vec embeddings with CNN models. Lastly, Park and Fung [15] used a two-step approach with Hybrid-CNN, taking character and word features as input to classify abusive language into specific labels such as homophobic, sexist, profanity, and racist.

[6] https://dictionary.cambridge.org/us/dictionary/english/hate-speech.

Some recent studies have explored the problem of identifying users who may spread hate speech. One study by Irani et al. [16] used the PAN 2021 dataset to perform social media author profiling specifically for hate speech directed toward immigrants and women. They found that user-level representations improved accuracy more than document-level representations and User2Vec user embeddings induced with contextualized word embeddings performed better than static word embeddings. Another study by Rangel et al. [17] presented the Author Profiling shared task at PAN 2021, which aimed to determine if an author is likely to spread hate speech or not. The best result for the English language task was achieved by Dukic and Kržic [18], who used a combination of fine-tuned BERT embeddings, indicator binary variables, and logistic regression classifier. Previous work has considered the user social graph for detecting hateful users [19].

Nevertheless, fewer studies have investigated the use of user data to improve hate speech detection. Qian et al. [20] proposed a novel model that uses both intra-user and inter-user representation learning to improve hate speech detection on Twitter. This involves analyzing a user's historical posts to model intra-user Tweet representations and using reinforced inter-user representation learning techniques to model similar Tweets posted by all other users. Other work has considered social and conversational interactions modeled through their corresponding graph [21]. Mosca et al. [6] investigated the integration of context features represented as the follower–followee graph and its effect on the hate speech detection model. They found that user features have an impact on the model's decision and affect the feature space learned by the model. Waseem and Hovy [4] and Fehn Unsvåg and Gambäck [5] considered adding user gender, location, activity, and profile features to tweet text features to improve hate speech detection. Recently, Nagar et al. [22] hypothesized that an individual's hateful content is influenced by their social circle and creates a framework that merges text content with social context to detect hate speech.

A deep dive into the literature shows a cornucopia of studies on hate speech detection. Nevertheless, there is little research done on leveraging user characteristics along with textual features for hate speech detection. Our study aims to explore the combination of textual features with several features characterizing user behavior, emotions, personality and writing style in combination with demographics and networks features to improve the classification task. To the best of our knowledge, all these user features have never been considered before.

3 Datasets

In this paper, we consider four datasets, as detailed below. These datasets contain various labels, but for our study, we only considered normal and hate speech labels and data points with available user information, which means the user account is active, not suspended, and not protected. The label distribution is unbalanced in all the datasets considered.

Dataset 1 (DS1): This is a dataset of 80,000 tweets, annotated for abusive behavior, developed by Founta et al. [23] via a crowdsourced annotation process. The tweets are annotated according to four labels, namely normal, spam, abusive, and hateful, and the dataset is publicly available on GitHub.[7] In this dataset, 70% of the tweets reach overwhelming annotators' agreement (i.e., three out of five annotators agree), and there is high disagreement in a few tweets. As shown in Table 1, there are 5385 normal and 2064 hateful tweets with the user information available.

Dataset 2 (DS2): This dataset has been developed by Waseem and Hovy [4]. The authors proposed a list of criteria based on critical race theory to identify if a tweet is offensive.

The dataset contains 136,000 tweets collected over two months. The authors randomly sampled these tweets to get 16,914 tweets and got them labeled by expert annotators. The inter-annotator agreement was k=0.84, with 85% of disagreements in the sexism class. The dataset is made available on GitHub.[8] As shown in Table 1, there are 5129 normal and 1322 hateful tweets with the user information available.

Dataset 3 (DS3): This dataset is publicly available on Kaggle.[9] The dataset contains 31,962 tweets with their corresponding label (2242 hateful and 29,720 normal). As the dataset does not contain tweet IDs, we were not able to retrieve user information. This dataset does not provide any information about the inter-annotator agreement, data collection, data pre-processing, or the criteria used for annotation. We consider this dataset in our paper as it has been used by Khan et al. [3] in their experiments.

Dataset 4 (DS4): The spread of the COVID-19 pandemic gave rise to hate directed toward Asian communities on social media. He et al. [2] created an anti-Asian hate speech and counter-speech dataset spanning 14 months, containing over 206 million tweets. Among these tweets, 3355 tweets have been manually labeled by two trained undergraduate annotators as hate, normal, or counter-speech tweets. We have not considered tweets labeled as counter-speech, as it is not pertinent to our study. As shown in Table 1, there are 720 normal and 415 hateful tweets with the user information available.

[7] https://github.com/ENCASEH2020/hatespeech-twitter.

[8] https://github.com/zeeraktalat/hatespeech.

[9] https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech.

Table 1 Distribution of labels for datasets DS1, DS2, and DS4 when the tweet author information is available

Dataset	None/normal	Hateful	Total
DS1	5385	2064	7449
DS2	5129	1322	6451
DS4	720	415	1135

3.1 User data collection

We further gathered user-related information from tweets using Tweepy[10] and Snscrape.[11] Tweepy is a Python library to access Twitter APIs. We created a developer account and applied to get the API keys for research purposes. Given the tweet ID, we were able to retrieve information related to retweet count, tweet favorite count, and timestamp. We also gathered user information related to the tweet author such as username, user ID, description, protected status, followers count, favorites count, statuses count, verified status, statuses count, language, and URL of the profile image.

We used Snscrape to get a hundred tweets per user posted before the tweet timestamp in our dataset. We used a hundred tweets per user to extract user features as described in Sect. 4.2.

4 Hate speech detection

We propose combining state-of-the-art textual features extracted from the tweets with user features to improve hate speech detection. In this section, we first describe the tweet features we used and then the user features we propose.

4.1 Encoding hate speech tweets

The most recent works for classifying hate speech tweets have been proposed by He et al. [2] and Khan et al. [3]. He et al. [2] propose three approaches based on fine-tuned BERT (to incorporate sentence-level semantics), Linguistic Inquiry and Word Count (LIWC) features [24] (to incorporate stylistic and psycho-linguistic patterns—see Appendix for a description of this set of features) and hashtag features. Specifically, the first approach relies on computing the tweet embedding from the BERT base uncased text embedding model [25] and providing it in input to a neural network classifier with one feed-forward layer. (This model is also fine-tuned for improved performance.) The other two proposed models consist of machine learning classifiers with input LIWC features or a vectorial representation of the number of occurrences of hashtags in the tweet. In this paper, we only use BERT and LIWC features and did not consider hashtags as particular hashtags are often used to collect the data regarding particular events [2, 26]. So, they may bias the classifier toward the class to detect or make the classifier specific only toward particular events.

Khan et al. [3] propose a convolutional, bidirectional gated recurrent (BiGRU) and capsule network-based deep learning model, named HCovBi-Caps, for hate speech detection. The deep neural network model extracts hate speech-related contextual information from the text, accounting for the order of words and various orientations. This model outperforms classical deep learning architectures such as CNN, LSTM, GRU, BiLSTM, BiGRU, and DNN, as well as previous work by Ding et al. [12] and Roy et al. [27].

As the methods proposed by He et al. [2] and Khan et al. [3] do not directly compare each other in their respective papers, we compare them in Sect. 5.2 and use the resulting best model for encoding hate speech tweets.

4.2 User features

Nagar et al. [28] showed that homophily is exhibited by the users generating hateful content, where to compute homophily among the users they considered features including the writing style and the readability of the users. Thus, we expand upon this set of features and propose the following groups of user features to improve hate speech detection from tweet text: demographics, Twitter behavior, personality, readability level, emotions, and writing style. These features are computed from the set of additional 100 tweets we collected for each user as described in Sect. 3.

4.2.1 Demographics

As demographic information is not typically readily available on social networks, we applied machine learning-based methods to infer such features for all datasets except DS3 since the required metadata is unavailable. Specifically, we used the m3inference[12] tool provided by Wang et al. [29], which is a deep-learning-based system trained on Twitter data to infer user age, gender, and organization information. The m3inference tool takes in input user id, name, screen name, description, language, and the profile picture path and predicts: (i) the gender of the user as male or female; (ii) the age of the user into four categories (≤ 18, 19–29, 30–39, and ≥ 40); and (iii) whether the provided account is administered by an organization or not. The m3inference tool achieves an F1 score of 0.918 for gender prediction, 0.522 for the four-class age prediction problem, and 0.898 for predicting the organization status.

[10] https://www.tweepy.org/.

[11] https://github.com/JustAnotherArchivist/snscrape.

[12] https://github.com/euagendas/m3inference.

4.2.2 Network-based features

We consider the *number of friends* and the *number of followers* of the Twitter account as network-based features.

4.2.3 Behavioral-based features

Users' tweeting/sharing behavior and engagement is measured by the following features:

Weekend index: We computed the normalized difference in the number of tweets on weekdays and weekends.

Insomnia index: We analyzed the user's daily tweeting behavior (24 h). Based on the user's local time, we divided the time into day and night. We considered the '6:01 AM-8:59 PM' window as day and the '9 PM-6 AM' window as night. Then, we computed the normalized difference in the number of posts made during the night window and the day window.

We also used the *count of day, night, weekend, and weekday posts* as additional features.

4.2.4 Emotion-based features

Hate speech is purposely sparked by emotionally charged statements to sway public opinion and damage feelings of a particular group by inciting their anger, fear, and mistrust in the direction of the event, individual, and agency. Here, we investigate whether users' emotional traits extracted from the tweets they wrote have a connection with hate speech sharing. We compute emotion features including anger, joy, sadness, fear, disgust, anticipation, surprise, and trust by using the emotion intensity lexicon (NRC-EIL) given in Mohammad [30] and happy, unhappy, angry, do not care, inspired, afraid, amused, and aggravated using Emolex.[13] We have also calculated the objectiveness of the text.

Firstly, we cleaned up the tweets by expanding contraction phrases, using LanguageTool 6 to fix spelling and grammar errors, swapping out negated terms for their WordNet antonyms, removing forestall phrases, and lemmatizing the words. Following that, we calculated emotion vectors using the methods suggested by Milton et al. [31] and Milton and Pera [32]. We specifically looked up each word in the two emotion dictionaries and mapped the corresponding affect values of the words that matched. To create an emotion vector, we then normalized the ranks of each emotion class using the full range of emotions that were retrieved from a tweet. If two lexicons had the same emotion, such as sad in NRC-EIL and unhappiness in Emolex, we took the average of the two computed values into consideration.

4.2.5 Personality

The Big Five is a broadly used taxonomy to explain human beings' personality traits under five main traits, namely Openness to experience, Consciousness, Extroversion, Agreeableness, and Neuroticism (also known as OCEAN or Five-factor Model) [33].

We compute the user personality traits from the text the user wrote in their tweets by using the code available on GitHub[14] which implements a personality prediction model inspired by the one proposed by Majumder et al. [34]. Specifically, words are encoded via GloVe embeddings, which are then aggregated at the sentence and then document level via a convolutional neural network. In our context, single tweets are sentences, and the concatenation of all the used tweets constitutes the document. Then, the embedding of the whole document is classified by using a random forest classifier. We used the provided pre-trained model, which is trained on combining several textual datasets with ground truth on the Big Five or the Myers–Briggs type indicator (MBTI) model. The latter is an introspective self-report questionnaire indicating differing psychological preferences in how people perceive the world and make decisions.[15] The datasets are combined on the correlating traits as shown in Table 2.

4.2.6 Readability

Readability measures the complexity of the textual content, and when computed from tweet content written by the user, it represents which level of textual content complexity a person is able to understand. Hence we used popular readability measures in our analysis, including Flesh Reading Ease, Flesh Kincaid Grade Level, Coleman Liau Index, Gunning Fog Index, Simple Measure of Gobbledygook Index (SMOG), Automatic Readability Index (ARI), Lycee International Xavier Index (LIX), and Dale-chall Score.

The Flesch scale ranges from 0 to 100. Higher Flesch reading-ease ratings suggest that the text is easier to read, while lower scores indicate that it is more difficult to read. The Coleman Liau Index gauges the text's readability based on the word's characters. The Gunning Fog Index, the Flesh Kincaid Grade Level, the SMOG Index, the Automatic Readability Index, and the Gunning Fog Grade Level are algorithmic heuristics used to determine readability based on the number of educational years needed to comprehend the text. Lastly, the Dale–Chall reading test gauges the text's difficulty using a vocabulary list that fourth-graders are familiar with.

[13] https://sites.google.com/site/emolexdata/.

[14] https://github.com/jkwieser/personality-detection-text.

[15] https://en.wikipedia.org/wiki/Myers-Briggs$_$Type$_$Indicator.

Table 2 Correlation between the Big Five and MBTI models shown by Furnham [35]

MBTI	Big five
Intuition/sensing	Openness to experience (correlates with N)
Feeling/thinking	Agreeableness (correlates with F)
Perception/judging	Conscientiousness (correlates with J)
Introversion/extraversion	Extraversion (correlates with E)
Not available in MBTI	Neuroticism

The syllable count and sentence count are the number of syllables and sentences, respectively, present in the given tweet text. We also have lexicon count, which is the number of words present in the text after removing punctuation.

4.2.7 Writing style

This set of features captures the writing style of the tweets authored by the same user. Specifically, we computed the average number of words, the average number of uppercased words, the average number of characters per user tweet, the percentage of stop-words, and the use of part of speech such as the number of nouns, proper nouns, personal nouns, possessive nouns, pronouns, determinants, adverbs, interjections, verbs, and adjectives.

5 Experimental evaluation

5.1 Experimental setting

We addressed the problem of identifying hate speech as a binary classification task. After computing the features, we chose the best classifier by comparing the performances of various traditional machine learning algorithms, including support vector machine, logistic regression, random forest, CatBoost, and XGBoost. When combining BERT with other features, i.e., LIWC or LIWC + User Features, we concatenated the BERT features of the last hidden layer with the other features and put them in input to a classical machine learning algorithm. For a better readability, we report, for each classification task, the performance score obtained by the best machine learning algorithm. Detailed results are reported in Appendix. We used class weighting to deal with the class imbalance, tuned hyper-parameters by using a 10% validation set, performed a fivefold cross-validation, and used the F1 score as the performance metric.

5.2 Determining the best tweet text-based approach for hate speech detection

As a first experiment, we want to compare the most recent approaches for hate speech detection, namely [2] and Khan et al. [3], that only consider the text of the tweet (no user features) in order to determine the best method to be combined with user features.

We used the results reported in Table 3 by Khan et al. [3] which show that HCovBi-Caps is the best method as compared to several other baselines on datasets DS1 and DS3,[16] including an LSTM model that achieved an F1 score of 0.35 on DS1 and 0.55 on DS3. We reproduced the same experimental setup as Khan et al. [3] and reported on the performances achieved on the same datasets by the approaches proposed by He et al. [2].

Results are shown in Table 3. As we can see, the combination of BERT and LIWC features gave the highest F1 score (0.84 vs. a score of 0.76 achieved by HCovBi-Caps) on dataset DS1.

In the case of DS3, HCovBi-Caps results in a better F1 score as compared to the combination of BERT and LIWC (0.84 vs. 0.79). However, there are certain shortcomings in DS3 that might question the dataset's quality. The dataset description was missing on the Kaggle webpage, and we have no information regarding the quality of data collection, pre-processing, or annotation criteria, in contrast to other datasets used in our study. For instance, Waseem and Hovy [4] had a list of criteria to identify hate as stated in their paper, while He et al. [2] trained undergraduate annotators to identify hate toward Asian people. Also, we could not find whether the data collection was based on a particular event using a hashtag or collected randomly. As per our knowledge, various studies have used DS1 and DS2 as reported in Sect. 2, but only a few have used DS3 [1, 3]. Therefore, we trust the results obtained for DS1 and we are using BERT + LIWC in the rest of our paper to encode the tweet text. Also, these features can be easily combined with other sets of features such as user features.

5.3 Adding user features

We propose to combine user features and tweet text-based approaches to improve hate speech detection. Hence, we compared the performances of the best tweet text-based

[16] For these datasets, Khan et al. [3] sampled 2615 hate and 5385 non-hate tweets from DS1 and 1421 hate and 9579 non-hate tweets from DS3 for their experiments, hence we did the same.

Table 3 F1 score comparison between the methods proposed by He et al. [2] and Khan et al. [3] for hate speech detection (tweet text only) on datasets DS1 and DS3

Method	DS1	DS3
LIWC [2]	0.67 (with CatBoost)	0.54 (with XGBoost)
BERT [2]	0.75	0.77
LIWC + BERT	**0.82** (with XGBoost)	0.79 (with XGBoost)
HCovBi-Caps [3]	0.76	**0.84**

The best F1 scores are bolded

method (LIWC + BERT) as determined in the previous experiment to the case where we also add the user features proposed in this paper (cf. Section 4.2), and other methods existing in the literature that also consider both tweet text and user features. Specifically, we compare with:

- ***Approach by*** Waseem and Hovy [4]: They merged word n-grams of length one to four with other extra-linguistic factors including length features (i.e., length of the tweet, length of the user description, and average length of the n-grams of the considered size) and user gender and location.
- ***Approach by*** Unsvåg and Gambäck [5]: Unsvåg and Gambäck [5] combined a TF-IDF approach to represent the n-gram (words and characters) features up to size six with several user features that include gender, network features (number of followers and number of friends), user activity (number of statuses and and number of favorites), and user profile (geo enabled, default profile, default image, and number of public lists).

Results are reported in Table 4. As we can see, our proposed approach, which combines both tweet text (LIWC + BERT) and user features (demographics, network, behavioral, emotions, personality, readability, and writing style), always achieves the best F1 score in all the three considered datasets as compared to previous work also using user features [4, 5] and the best approach only using tweet text features. Specifically, adding user features improves from 1% on DS1 to 3% on DS2 and 32% on DS4 as compared to just using BERT + LIWC features. (Improvements are statistically significant.)

5.4 Feature importance

We start by comparing the main three groups of features of our proposed approach, namely LIWC, BERT, and User Features, across the three datasets considered. F1 score results are reported in Table 5. As we can see, tweet text-based features (BERT for DS1 and DS2, and LIWC for DS4) are always the best group of features, while the user-based features are the second most important group of features in two out of the three considered datasets (DS2 and DS4).

Next, we computed feature importance with a forest of trees to study the top-20 most important LIWC, and the top-20 most important user features across all the considered datasets. In the case of tweet LIWC features, there are 11 common features among the top-20 most important features across all the four datasets considered, as reported in Table 6. These features are: ‘AllPunc’ (use of punctuation symbols such as.,:;!?-“”’(’, ‘OtherP’ (use of other punctuation symbols), ‘Dic’ (presence of LIWC dictionary words), ‘Sixltr’ (use of words with more than six letters), ‘Tone’ (emotional tone), ‘WPS’ (words per sentence), ‘affect’ (use of words related to affective processes, i.e., happy, cried), ‘anger’ (use of anger-related words), ‘function’ (use of function words, e.g., it, to, no, very), ‘negemo’ (negative emotions), and ‘swear’ (use of swear words).

Regarding user features, there are ten common features among the top-20 most important features across datasets DS1, DS2, and DS4, as reported in Table 7. (We do not have user features in the case of DS3.) These features are ‘Fear,’ ‘Objective,’ ‘angry,’ ‘annoyed,’ ‘count_day_posts,’ ‘dont_care,’ ‘followers_count,’ ‘happy,’ ‘inspired,’ and ‘syllable_count.’

For a given dataset D and feature f, the presence of ‘H’ in Tables 6 and 7 means that the average value of f in D is higher among examples of hate speech vs. normal tweets. Vice versa, an ‘L’ means that the average value of f in D is lower among examples of hate speech vs. normal tweets. The value in parentheses report the rank of that feature according to feature importance.

As shown in Table 6, we found out that in the majority of the datasets, hateful tweets use, on average, fewer punctuation symbols (‘AllPunc’ and ‘OtherP’) than normal tweets, have more negative tone (the higher the score of the feature ‘Tone,’ the more positive the tone [24]), express more ‘Anger’ and negative emotions in general (‘Negemo’), have more ‘Swear’ words, and use more ‘Function’ words.

As Table 7 shows, the emotions expressed by users in their tweets play an important role in enhancing hate speech detection. We found out that users who write hateful tweets express, on average, more fear, anger, annoyance, objectiveness, and inspired emotions, and fewer ‘Don’t Care’ and happiness emotions as compared to users who write normal tweets in the majority of the datasets. Moreover, users who write hateful tweets use, on average, simpler words (less syl-

Table 4 F1 score comparison between our proposed approach (LIWC + BERT + User Features) and related work on datasets DS1, DS2, and DS4

Approach	DS1	DS2	DS4
LIWC + BERT (tweet text only)	(0.77 with CatBoost)	0.89 (with CatBoost)	0.52 (with XGBoost)
Unsvåg and Gambäck [5]	0.62	0.87	0.63
Waseem and Hovy [4]	0.67	0.89	0.67
LIWC + BERT + User Features	**0.78** (with XGBoost)	**0.92** (with XGBoost)	**0.84** (with XGBoost)

Best scores are bolded

Table 5 F1 score comparison among group of features (LIWC, BERT, and proposed User Features) on datasets DS1, DS2, and DS4

	DS1	DS2	DS4
LIWC	0.65 (with CatBoost)	0.65 (with XGBoost)	0.65 (with XGBoost)
BERT	0.75	0.77	0.52
User Features	0.58 (with XGBoost)	0.73 (with CatBoost)	0.59 (with XGBoost)

lable count on average), have more followers, and post more during the day (‘Count Day Posts’) in the majority of the datasets.

6 Conclusion

This paper addressed the problem of automatically identifying hate speech on online social networks such as Twitter. Our proposed approach relies on enhancing tweet textual features such as BERT and LIWC with several user features, including demographics, emotions, personality, readability, writing style, and behavioral and network features. Although there is substantial literature on hate speech detection in online social networks, there is limited literature that leverages user information on hate speech detection.

We conducted an extensive experimental evaluation on four publicly available datasets. We showed that combining user features and tweet textual features improves hate speech detection up to +0.32 in F1 score compared to just using tweet textual features and improved over previous work using a limited number of user characteristics. Furthermore, our feature importance analysis showed that hate speech tweets or their authors express more negative emotions and swear words.

There is a lack of hate speech datasets in the current literature containing user information. In accordance with Twitter policies for releasing collected data, we plan to make available to the research community the additional user data we gathered upon publication of this paper.

We have used automated tools to infer user demographics, emotions and personality which could be a potential limitation of our study. Thus, inferred characteristics of some of the users might not be entirely accurate. However, it is impossible to test the tools’ efficiency in the considered datasets as such metadata are not explicitly available to be used as ground truth.

Table 6 Common LIWC features among the top-20 most important features across all the considered four datasets. For a given dataset D and feature f, ‘H’ (resp. ‘L’) means that the average value of f in D is higher (resp. lower) among examples of hate speech vs. normal tweets. The value in parentheses report the rank of that feature according to feature importance

LIWC feature	DS1	DS2	DS3	DS4
AllPunc	L (8)	L (6)	L (3)	H (3)
Dic	H (6)	H (3)	L (2)	L (4)
OtherP	L (9)	L (4)	L (6)	H (2)
Sixltr	L (10)	L (8)	H (4)	H (1)
Tone	L (4)	L (9)	L (1)	L (8)
WPS	L (11)	H (7)	H (8)	L (6)
Affect	H (5)	H (10)	L (7)	L (7)
Anger	H (3)	H (1)	H (9)	L (9)
Function	H (7)	H (5)	H (5)	L (5)
Negemo	H (2)	H (2)	H (10)	L (10)
Swear	H (1)	H (11)	H (11)	L (11)

Our proposed approaches can be further improved and extended in the future. For example, most of the user features we have used are on a personal level. It would be interesting to further study network features such as user centrality and network density or explore the concept of homophily (the tendency of individuals to associate and bond with similar others). Another interesting aspect to explore could be analyzing the conversation (tweet replies in our case) that goes on in the tweet threads. Finally, we have used the pre-pandemic datasets and a dataset containing anti-Asian tweets due to the COVID-19 pandemic in this paper. It would be interesting to extend further our study of hate speech in the post-pandemic era.

Table 7 Common User features among the top-20 most important features across datasets DS1, DS2, and DS4. For a given dataset D and feature f, 'H' (resp. 'L') means that the average value of f in D is higher (resp. lower) among examples of hate speech vs. normal tweets. The value in parentheses reports the rank of that feature according to feature importance

User Feature	DS1	DS2	DS4
Fear	H (9)	H (3)	L (10)
Objectiveness	L (5)	H (4)	H (5)
Anger	H (1)	H (2)	H (2)
Annoyance	H (4)	H (1)	L (8)
Don't care	H (6)	L (7)	L (7)
Happiness	L (2)	L (5)	L (6)
Inspired	L (10)	H (10)	H (4)
Followers count	L (7)	H (9)	H (1)
Syllable count	L (3)	L (8)	L (9)
Count day posts	H (8)	H (6)	L (3)

Author Contributions FS designed and directed the research and wrote the paper; RR collected additional data, performed the experiments, and wrote the paper.

Funding The authors did not receive external funding to perform this research.

Data Availability The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from any Contributing Author) by another publisher. All of the materials are owned by the authors, and/or no permissions are required.

Declarations

Conflict of interest The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Appendix

Linguistic inquiry and word count (LIWC)

The transparent text analysis tool LIWC counts words in groups that have psychologically significant meanings. We examine the text's cognitive, emotional, and linguistic processes using the LIWC 97 measures. We categorize the LIWC characteristics into four groups [24] to compare the differences between the hate and normal content writing styles.

Linguistics features (28 features) refer to features that represent the functionality of text, such as the average number of words per sentence and the rate of misspelling. This feature category also includes negations and part of speech (Adjective, Noun, Verb, Conjunction) frequencies.

Punctuation features (11 features) quantifies dramatization or sensationalization of content. Various punctuation types used in the text, such as Periods, Commas, Question, Exclamation, and Quotation marks, dramatize or sensationalize the content.

Similarly, *psychological features* (51 features) target emotional, social, and cognitive processes. The affective processes (positive and negative emotions), social processes, cognitive processes, perceptual processes, biological processes, time orientations, relativity, personal concerns, and informal language (swear words, nonfluencies) scrutinize the emotional part of the content.

Summary features (seven features) define the frequency of words that reflect the writer's thoughts, perspective, and honesty. It consists of analytical thinking, clout, authenticity, emotional tone, words per sentence, words with more than six letters, and dictionary words under this category.

Table 8 Detailed experimental results in support of Table 3. F1 score comparison among different machine learning algorithms with LIWC features in input

	DS1	DS3
CatBoost	**0.67**	0.34
XGBoost	0.64	**0.54**
Support vector machine	0.63	0.05
Logistic regression	0.66	0.19
Random forest	0.63	0.37

Best scores are in bold

Table 9 Detailed experimental results in support of Table 3. F1 score comparison among different machine learning algorithms with LIWC + BERT features in input

	DS1	DS3
CatBoost	0.4	0.37
XGBoost	**0.82**	**0.79**
Support vector machine	0.22	0.32
Logistic regression	0.28	0.2
Random forest	0.39	0.38

Best scores are in bold

Detailed experimental results

This section shows the comparison in terms of F1 score among different machine learning algorithms, namely logistic regression, support vector machine, random forest, CatBoost, and XGBoost with different set of features in input to support the results reported in Tables 3, 4 and 5.

Table 3 shows the F1 score comparison between the methods proposed by He et al. [2] and Khan et al. [3] for hate speech detection (tweet text only) on datasets DS1 and DS3. Table 8 shows that considering the LIWC features in input, CatBoost achieves the best F1 score on DS1, while XGBoost achieves the best F1 score with LIWC features on DS3; Table 9 shows that considering the LIWC + BERT features in input, XGBoost achieves the best F1 score on both DS1 and DS3.

Table 10 Detailed experimental results in support of Table 4. F1 score comparison among different machine learning algorithms with LIWC + BERT features in input

	DS1	DS2	DS4
CatBoost	**0.77**	**0.89**	0.5
XGBoost	0.60	0.92	**0.52**
Support vector machine	0.42	0.33	0.24
Logistic regression	0.45	0.42	0.38
Random forest	0.68	0.89	0.52

Best scores are in bold

Table 11 Detailed experimental results in support of Table 4. F1 score comparison among different machine learning algorithms with LIWC + BERT + User features in input

	DS1	DS2	DS4
CatBoost	0.75	0.90	0.81
XGBoost	**0.78**	**0.92**	**0.84**
Support vector machine	0.49	0.42	0.40
Logistic regression	0.33	0.25	0.38
Random forest	0.55	0.52	0.46

Best scores are in bold

Table 12 Detailed experimental results in support of Table 5. F1 score comparison among different machine learning algorithms with LIWC features in input

	DS1	DS2	DS4
CatBoost	**0.65**	0.64	0.63
XGBoost	0.64	**0.65**	**0.65**
Support vector machine	0.33	0.52	0.2
Logistic regression	0.42	0.53	0.44
Random forest	0.63	0.60	0.54

Best scores are in bold

Table 4 shows the F1 score comparison between our proposed approach (LIWC + BERT + User Features) and related work. Table 10 shows that considering LIWC + BERT features (tweet text only) in input, CatBoost achieves the best F1 score on DS1 and DS2, while XGBoost achieves the best F1 score on DS3[17]; Table 11 shows that considering LIWC + BERT + User features in input, XGBoost achieves the best F1 score on all the three datasets.

Table 5 shows the F1 score comparison among group of features (LIWC, BERT, and proposed User Features) on datasets DS1, DS2, and DS4. Table 12 shows that when considering only the LIWC features in input, CatBoost achieves the best F1 score on DS1, while XGBoost achieves the best F1 score on DS2 and DS4. Table 13 shows that when considering only the User features in input, CatBoost achieves the best F1 score on DS2, while XGBoost achieves the best F1 score on DS1 and DS4.

Table 13 Detailed experimental results in support of Table 5. F1 score comparison among different machine learning algorithms with User features in input

	DS1	DS2	DS4
CatBoost	0.58	**0.73**	0.56
XGBoost	**0.58**	0.72	**0.59**
Support vector machine	0.35	0.33	0.19
Logistic regression	0.30	0.25	0.2
Random forest	0.45	0.62	0.52

Best scores are in bold

[17] Results for DS1 in Table 10 are different from the results reported in Table 9 as the dataset size is different. Table 9 refers to the setting adopted by Khan et al. [3] as reported in Footnote 16, while Table 10 refers to the entire dataset size which is reported in Table 1.

References

1. Roy, P.K., Tripathy, A.K., Das, T.K., Gao, X.-Z.: A framework for hate speech detection using deep convolutional neural network. IEEE Access **8**, 204951–204962 (2020). https://doi.org/10.1109/ACCESS.2020.3037073
2. He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., Kumar, S.: Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 90–94 (2021)
3. Khan, S., Kamal, A., Fazil, M., Alshara, M.A., Sejwal, V.K., Alotaibi, R.M., Baig, A.R., Alqahtani, S.: Hcovbi-caps: Hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. IEEE Access **10**, 7881–7894 (2022). https://doi.org/10.1109/ACCESS.2022.3143799
4. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93. Association for Computational Linguistics, San Diego, California (2016). http://www.aclweb.org/anthology/N16-2013
5. Fehn Unsvåg, E., Gambäck, B.: The effects of user features on Twitter hate speech detection. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 75–85. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/W18-5110. https://aclanthology.org/W18-5110
6. Mosca, E., Wich, M., Groh, G.: Understanding and interpreting the impact of user context in hate speech detection. In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pp. 91–102. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.socialnlp-1.8. https://aclanthology.org/2021.socialnlp-1.8
7. Kim, J.Y., Kesari, A.: Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. J. Online Trust Saf. **1**(1), 647–667 (2021)
8. Li, J., Ning, Y.: Anti-asian hate speech detection via data augmented semantic relation inference. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 607–617 (2022)
9. Xia, M., Field, A., Tsvetkov, Y.: Demoting racial bias in hate speech detection. CoRR arXiv:2005.12246 (2020)

10. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 25–35. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3504. https://aclanthology.org/W19-3504
11. Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science **7**, 598 (2021)
12. Ding, Y., Zhou, X., Zhang, X.: YNU_DYX at SemEval-2019 task 5: A stacked BiGRU model based on capsule network in detection of hate. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 535–539. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/S19-2096. https://aclanthology.org/S19-2096
13. Alatawi, H.S., Alhothali, A.M., Moria, K.M.: Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. IEEE Access **9**, 106363–106374 (2021). https://doi.org/10.1109/ACCESS.2021.3100435
14. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90. Association for Computational Linguistics, Vancouver, BC, Canada (2017). https://doi.org/10.18653/v1/W17-3013 . https://aclanthology.org/W17-3013
15. Park, J.H., Fung, P.: One-step and Two-step Classification for Abusive Language Detection on Twitter. arXiv (2017). https://doi.org/10.48550/ARXIV.1706.01206 . https://arxiv.org/abs/1706.01206
16. Irani, D., Wrat, A., Amir, S.: Early detection of online hate speech spreaders with learned user representations (2021)
17. Rangel, F., Sarracén, G., Chulvi, B., Fersini, E., Rosso, P.: Profiling hate speech spreaders on twitter task at pan 2021. In: CLEF (2021)
18. Dukic, D., Kržic, A.S.: Detection of hate speech spreaders with bert (2021)
19. Ribeiro, M., Calais, P., Santos, Y., Almeida, V., Meira Jr., W.: Characterizing and detecting hateful users on twitter. Proceedings of the International AAAI Conference on Web and Social Media **12**(1), (2018). https://doi.org/10.1609/icwsm.v12i1.15057
20. Qian, J., ElSherief, M., Belding, E.M., Wang, W.Y.: Leveraging intra-user and inter-user representation learning for automated hate speech detection. arXiv preprint arXiv:1804.03124 (2018)
21. Mishra, P., Tredici, M.D., Yannakoudakis, H., Shutova, E.: Abusive language detection with graph convolutional networks. In: North American Chapter of the Association for Computational Linguistics (2019)
22. Nagar, S., Barbhuiya, F.A., Dey, K.: Towards more robust hate speech detection: using social context and user data. Soc. Netw. Anal. Min. **13**(1), 47 (2023)
23. Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Twelfth International AAAI Conference on Web and Social Media (2018)
24. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015. Technical report (2015)
25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
26. Chen, L., Lyu, H., Yang, T., Wang, Y., Luo, J.: In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19. arXiv preprint arXiv:2004.10225, 1–8 (2020)
27. Roy, P.K., Tripathy, A.K., Das, T.K., Gao, X.-Z.: A framework for hate speech detection using deep convolutional neural network. IEEE Access **8**, 204951–204962 (2020)
28. Nagar, S., Gupta, S., Bahushruth, C., Barbhuiya, F.A., Dey, K.: Empirical assessment and characterization of homophily in classes of hate speeches. In: AffCon@ AAAI, pp. 30–34 (2021)
29. Wang, Z., Hale, S., Adelani, D.I., Grabowicz, P., Hartman, T., Flöck, F., Jurgens, D.: Demographic inference and representative population estimates from multilingual social media data. In: The World Wide Web Conference, pp. 2056–2067 (2019)
30. Mohammad, S.: Word affect intensities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018). https://aclanthology.org/L18-1027
31. Milton, A., Batista, L., Allen, G., Gao, S., Ng, Y.-K.D., Pera, M.S.: "don't judge a book by its cover": Exploring book traits children favor. In: Fourteenth ACM Conference on Recommender Systems, pp. 669–674 (2020)
32. Milton, A., Pera, M.S.: What snippets feel: Depression, search, and snippets (2020)
33. Neuman, Y.: Computational Personality Analysis: Introduction, Practical Applications and Novel Directions. Springer, Cham (2016)
34. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
35. Furnham, A.: The big five versus the big four: the relationship between the Myers–Briggs type indicator (MBTI) and neo-pi five factor model of personality. Pers. Individ. Differ. **21**(2), 303–307 (1996)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.