User-Centric Modeling of Online Hate Through the Lens of Psycholinguistic Patterns and Behaviors in Social Media

Zeinab Noorian, Amira Ghenai, Hadiseh Moradisani, Fattane Zarrinkalam, *Member, IEEE*, and Soroush Zamani Alavijeh

Abstract—**Hate speech in social media is a growing problem that reinforces racial discrimination and mistrust between people, leading to physical crimes, violence, and fragmentation in world communities. Although previous studies showed the potential of user profiling in hate speech detection in social media, there has not been a thorough analysis of users' characteristics and dispositions to understand the development of hate attitudes among users. To bridge this gap, we investigate the role of a wide range of psycholinguistic and behavioral traits in characterizing and distinguishing users prone to post hate speech on social media. Considering anti-Asian hate during the COVID-19 pandemic as a case study, we curate a dataset of 5 417 041 tweets from 3001 Twitter users prone to publish hate content (aka hateful-to-be users) and a corresponding matched set of 3001 control users. Our findings reveal significant statistical differences in most dimensions of psycholinguistic attributes and online activities of hateful-to-be users compared to control users. We further develop a classifier and demonstrate that features derived from user timelines are strong indicators for automatically predicting the onset of hateful behavior.**

Index Terms—**Hate speech prediction, observational studies, social media analytics, user-centric analysis.**

I. Introduction

WHILE social media provides an easy and efficient communication venue, the increased usage of these platforms created the perfect medium for users to generate content disparaging or judging specific groups of people based on race, religion, nationality, etc. In literature, such content is related to several concepts such as hate speech, abusive language, harmful content, and prejudiced speech. While social media policy makers impose different regulations to combat hate speech, eliminating such harmful content is still challenging.

Importantly, recent research indicates that hate speech generated on online platforms does not stay online and has negative offline consequences for individuals including damaging the victim's mental health state [1] as well as groups [2]. The COVID-19 pandemic is a recent example for which a huge amount of hateful content discussing its origin was propagated which caused the spread of East Asian prejudice in social media [3]. Several news articles also reported offline abuse and physical attacks against the East Asian community during the first period of the pandemic [4], [5].

To overcome these challenges, there is a good amount of work in literature that computationally formalizes hate speech detection in the context of social media. The mainstream research focus is to demonstrate the effectiveness of *content analysis* by developing methods ranging from classical machine learning to deep learning to detect hateful content by leveraging language used in social media. Such state-of-the-art approaches are oblivious to users and only focus on identifying whether a given text contains abusive language or hateful content [6], [7], [8].

In a complementary perspective to online hate research, some studies adopt a *user-centric* approach and focus on the user instead of the content to fight hateful content spread. The intuition behind this approach is that perpetrators of hate speech may carry certain characteristics and online behaviors that are different from other users [9], [10], [11]. Hence, they argue that fighting hate speech from the user-level is an efficient way to early intervention thus better controlling the dissemination of harmful content. In this regard, some researchers utilize social cues, users' level of activity, and their temporality to characterize hateful users [12], [13], [14]. Other studies also leverage language cues such as linguistic patterns, hashtags, uniform resource locators (URLs), and the credibility of information of users' posts and their engagement traces to detect hateful users [11], [15]. The importance of demographic information, political orientation and geolocation features is also shown in characterizing hateful users [10], [16], [17]. Despite the promising insights derived from user-centric approaches in detecting hateful users, none of the previous studies attempted to profile users based on their prospects of initiating hate-motivated behavior in social media (i.e., whether users become hateful in the future).

Manuscript received 10 August 2023; revised 14 December 2023 and 23 January 2024; accepted 24 January 2024. Date of publication 13 February 2024; date of current version 31 May 2024. *(Corresponding author: Zeinab Noorian.)*

Zeinab Noorian and Amira Ghenai are with Ted Rogers School of Information Management, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada (e-mail: znoorian@torontomu.ca; aghenai@torontomu.ca).

Hadiseh Moradisani and Fattane Zarrinkalam are with the School of Engineering, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: hmoradis@uoguelph.ca; fzarrink@uoguelph.ca).

Soroush Zamani Alavijeh is with the Department of Computing and Software, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: szchem20@gmail.com).

Digital Object Identifier 10.1109/TCSS.2024.3359010

The main motivation of this article is to fill the gap of a systematic method to understand the human intrinsic or extrinsic attributes that are most predictive of users prone to create hate content (hereafter *hateful-to-be* users). Considering Twitter as our medium, we speculate that distinct human attributes reflected in users' generated content and their online activities are determinants of the development of hateful attitudes in social media. Thus, in this article, we present the methodology for hate speech prediction that leverages a wide range of psycholinguistic or behavioral characteristics (including linguistic patterns, emotions, attitude polarization, personality traits, topical analysis, social engagement, posting behaviors, readability, and communication style), and investigate their effectiveness in differentiating between regular users (hereafter *control* users) and *hateful-to-be* users *prior to* posting their hate content.

The questions that we seek to answer are as follows.

1) *RQ1:* How different are the psycholinguistic characteristics of *hateful-to-be* users compared to *control* individuals?
2) *RQ2:* How different are the behavioral characteristics of *hateful-to-be* users compared to *control* individuals in terms of social engagement and posting trends, quality of information, and personality traits?
3) *RQ3:* To what extent do the timelines of *hateful-to-be* users are characterized by their attitude polarization?
4) *RQ4:* Are the linguistic and behavioral characteristics of users on social media strong indicators to automatically predict the susceptibility of users to post hateful content?

The contributions of this article are summarized as follows.

1) We develop a methodology to investigate a wide range of psycholinguistic and behavioral features of *hateful-to-be* users by analyzing in depth their past activities and differentiate them from *control* users.
2) We develop a prediction model for identifying future online hate spreaders for any controversial topic. We also perform an ablation study to evaluate the effectiveness of the derived features in the prediction performance.
3) We create a dataset of 3001 *hateful-to-be* users with its matching 3001 *control* users with a total of 5 417 041 tweets and make it publicly available.[1]

The organization of this article is as follows. Section II details the related work. Data collection and user selection are explained in Section III. Section IV elaborates on the design experiments to analyze users' timelines. The insights into online hate language and behavior are shown in Section V. Section VI explains the hate-mongering prediction process. The discussion and implications are elaborated in Section VII. We conclude the article and discuss future work in Section VIII.

II. Related Work

Online hate speech detection is an active research area widely studied in the field of computational linguistics. As evidence of its popularity, a vast number of datasets and resources have been created to train abusive language classifiers. For example, Vidgen and Derczynski [18] presented an analysis of a 63 publicly available datasets from different platforms (i.e., Twitter, Gab, and Facebook) curated for different categories of hate speech and presented critical insight into their annotations and their topical focus. Poletto et al. [6] systematically reviewed available resources and benchmark corpora including their language coverages and development methodology used for hate speech detection.

Given the ground truth data is available, existing hate speech detection mechanisms consider the problem as a *supervised* learning task. A more common approach is to extract surface features such as term frequency-inverse document frequency (TF-IDF), bag-of-words (BOW), linguistic features such as part of speech (POS) dependency relations and named entity recognition (NER) and further combine them with classical machine learning algorithms such as support vector machines (SVM), logistic regression (LR), and random forest to perform hate speech classification [19], [20], [21], [22], [23]. Recent advances in deep learning have empowered hate speech detection approaches. Kshirsagar et al. [24] used word embedding features and adopted a neural-network-based approach to differentiate between different abusive language use (hateful, racist, and sexism). They demonstrated significant improvement in classification performance compared to a number of existing approaches [17]. The work in [25] extracted different features such as word embeddings, word and character n-grams, and psycholinguistic features and investigate a number of data-driven and psycholinguistics-motivated models such as regularized LR, convolutional neural networks (CNN), and transformers [i.e., bidirectional encoder representations from transformers (BERT)] to determine the models that are best suited to detect hate speech in Spanish and English languages. Vidgen et al. [26] curated a specific dataset to detect aggression toward East Asian during COVID-19. They have implemented and fine-tuned several contextual embedding models such as DistilBERT, RoBERTA, and ELECTRA to detect abusive language and classify between hostility, criticism, and prejudice languages. Melton et al. [27] developed an ensemble of tunable deep learning models that incorporates higher-order features from contextual word embedding matrix to differentiate between hate speech and offensive language via multiclass hate speech model.

Although most studies in hate speech detection are content-based, there are some recent studies that focus on detecting hate speech from the user account level perspective. For instance, Mathew et al. [14] investigated the general growth of speech in Gab by building temporal snapshots on hateful users' activity patterns and profile hateful users based on hate intensity using the opinion dynamic model. Ribeiro et al. [12] analyzed the differences between hateful users and not-hateful users considering their activity patterns, linguistic features as well as their social network. Researchers in [28] formalized the pattern of hate speech spread through retweets and examine how localized structural properties of Twitter's information network influences the propagation of the hate speech. An et al. [11] studied the profile, activity level, and linguistic differences between highly hateful and reference users and found that hateful users are more active; use more words discussing controversial topics, and significantly share more diverse social media

[1] https://www.kaggle.com/datasets/hdsmrd/twitter-hate-speech-dataset

URLs. Qian et al. [29] proposed a model that takes into account intrauser (users' historical posts) and interuser (similar tweets posted by other users) representation learning for hate speech detection. Lyu et al. [16] characterized users based on their demographics, political orientation, and geolocation features to detect hateful users who use controversial terms on Twitter. Finally, in the shared tasks organized by Rangel et al. [9], in PAN 2021, different approaches have been proposed to profile Twitter users based on their intentions to spread hateful contents. Considering the historical tweets, most of the proposed solutions were based on building a classification model use deep learning models like CNN, recurrent neural network (RNN), and transformer-based models. Different types of features were also suggested such as n-grams stylistics, emotions, embedding features, users morality, named entities, and communicative behaviour [30].

All of aforementioned user-centric approaches for hate speech detection either explore the prevalence of hate speech in different social media platforms with different moderation policies or seek to profile hateful users and characterize hateful behaviors in such platforms. However, there is still a gap to understand what triggers a hateful attitude and find significant language and behavioral features that have potential to initiate hate-motivated behavior in social media. Thus, unlike previous studies, our work focuses on not only profiling hateful users but also understanding the development of hateful behaviors among social media users.

III. Data Collection

The focus of this study is to predict the onset of hateful behaviors of users in social media. Considering anti-Asian hate as our case study of interest, we adopt COVID-HATE [31], the largest publicly available dataset of anti-Asian hate and counterspeech as our *seed* dataset. COVID-HATE consists of over 206 million tweets gathered between the periods of 15 January 2020 and 26 March 2021. Tweets are classified as: 1) *hate* tweets which are hate speeches against Asian groups; 2) *counter-hate* tweets that are speeches that disapprove the abuse against Asian community; or 3) *neutral* tweets that are speeches which discuss the topic of COVID-19 without a positive or a negative sentiment.

As the purpose of this study is to understand the development of hate attitude in social media (i.e., Twitter), our dataset should consist of tweets belonging to two groups of users: 1) a "hate" group that consists of users promoting hate speech targeted toward the Asian community; and 2) a "control" group that either posts general ideas about COVID-19 and China/Asians or counter-hate speech to support Asians. From the COVID-HATE dataset, we define the hate group considering users posting tweets classified as hateful, and the control group as users posting tweets classified as counter-hate or neutral. We further exclude users with both hate and counter-hate tweets. That accounts for a total of 305 199 users in the hate group, 132 000 users in the counter-hate, and 910 166 neutral users, in which the latter two form our control group.

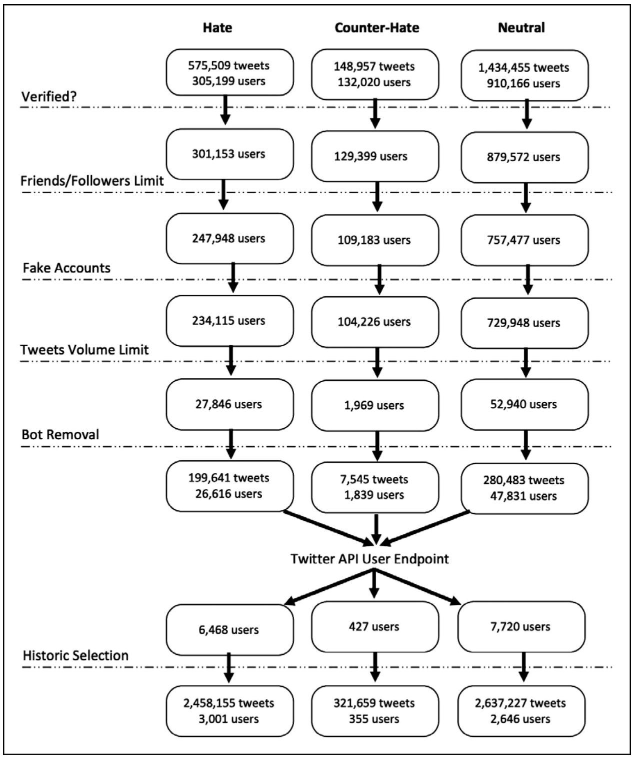

Fig. 1. Data collection and user selection process.

A. User Selection

Following [32], we refine our hate and control users to ensure that only private accounts and nonautomated communications are considered. We 1) remove verified accounts (such as media institutions, political, famous figures, etc.,) to focus on hate speech within individuals; 2) only consider accounts with more than ten friends and more than ten followers (users within a network of connections) and exclude users with more than 5000 friends and more than 10 000 followers (likely to be automated accounts); 3) remove users with significantly more friends than followers (identified as fake accounts in prior studies) where the friends to followers ratio was set to 10:1; 4) filter out users with less than three tweets in each group of users to ensure the existence of some social engagements within the dataset; and 5) use the Botometer Pro API [33] to filter out bot-like accounts (higher scores means a more bot-like account) where the CAP threshold is 80%.

After the data cleaning steps, we ended up with 26 616 hate users; 1831 counter-hate users; and 47 813 neutral users. The detailed steps of our data collection and user selection strategy is illustrated in Fig. 1.

B. Users' Historic Selection and Matching

In order to predict future hate-mongering behavior, we need to analyze the user-generated content before posting hateful tweets. Inspired by previous studies [9], [34], we consider the first hate tweet from the COVID-HATE dataset as our time reference for the prediction part.

The process starts by using Twitter streaming application programming interface (API) to extract the timeline of users in hate and control groups. However, for each user in the hate group, we *only* consider tweets from her timeline posted *before* the first hateful tweet. Similarly, for each user in the control group, we sample such a date with plus or minus a 5-day frame of the first hate tweet and match it with a hate user. Applying this avoids biasing the selection of different time periods which may trivially differentiate users. As a second matching criteria, we ensure that matching users in the hate and control groups have at least 100 tweets in their timelines. This process resulted in a total of 3001 hate users having 2 458 155 tweets matched with 3001 control users having 2 958 886 tweets.

Appendix B illustrates the process for selecting hate users (aka G_H) and matching them to control users (aka G_C) in Fig. 10. This dataset will be adopted for the analysis for the rest of the article.

IV. Experimental Design

In this section, we outline how we design experiments to analyze users' activities on Twitter in order to answer our research questions (*RQ1*, *RQ2*, and *RQ3*). To do so, given our hate and control users, we investigate the linguistic and behavioral differences and similarities between users in those two groups (i.e., G_H and G_C) by applying six different data analytical modules on their corresponding timelines: 1) word analysis; 2) emotion analysis; 3) polarization analysis; 4) topical analysis; 5) personality traits analysis; and 6) social engagement, posting trends and information quality analysis. In the following subsections, we provide more details about how we design each module.

A. Word Analysis

In this module, we analyze the timeline of users in hate and control groups in terms of *vocabulary uniqueness*, *linguistic style*, *readability,* and *communication style* as follows.

1) Vocabulary Uniqueness: To address *RQ1*, we investigate whether hate triggers a specific unique vocabulary that is not being used by others. Considering that each group defines a set, this is achieved by computing the relative size of the intersection by calculating Jaccard's index which quantifies the similarity between finite sample sets [35].

2) Linguistic Style: We also study different stylistic patterns that could potentially distinguish users in the hate group from the control group (*RQ1*). To this aim, given the set of unique words in timelines posted by two group G_H and G_C, i.e., V_{G_H}, V_{G_C}, we count the number of words belonging to different categories of language obtained by using the well-known linguistic inquiry and word count (LIWC) tool [36]. We measure the proportion of tweets posted by each user in group G that has at least one word on certain LIWC categories; and then produce the distribution of V_G with positive scores on LIWC categories for each group G.

3) Readability and Communication Style Assessment: In this module, we aim to estimate the complexity of a text in order to analyze whether users in hate and control groups generate content that is understandable by a reader of a certain level of literacy. Following prior work [37], we employ two well-established readability metrics [38], [39], namely, *Flesch reading ease* and *dale-chall readability score* to investigate whether the readability features have a potential in differentiating hateful users and the control users.

To complement our psycholinguistic analysis for *RQ1*, we assess the communication style of users by studying their intentions behind posting which can be useful for detecting hateful users. We adopt Symanto API[2] which returns a score for the communication style used to classify posts across four dimensions: 1) *action-seeking* (calling for action or attention); 2) *fact-oriented* (discussing about factual information); 3) *information-seeking* (asking questions, seeking advice); and 4) *self-revealing* (sharing one's own opinion and experience). We then calculate the mean of the communication style scores for each group, i.e., G_H and G_C, and report our observation outcomes.

B. Emotion Analysis

In this module, we again focus on *RQ1* by examining how people in the hate group convey their emotions via the social media posts and whether emotional expressions could be a differential factor in distinguishing hateful users. To this end, we apply the DistilBERT pretrained language model [40], on users' content to detect a set of fine-grained emotions. We adopt DistilBERT as our emotion analysis tool since it is fine-tuned as a multilabel classifier on a benchmark emotion dataset presented in [41]. Given the seven-dimensions of emotions derived for each tweet, namely, *anger* (e_1), *fear* (e_2), *joy* (e_3), *surprise* (e_4), *sadness* (e_5), *disgust* (e_6), and *neutral* (e_7), we represent each group of users G (e.g., hate group or control group) by a vector of weights over the seven emotions denoted by $Q(G) = (qG(e_1), \ldots, qG(e_6), qG(e_7))$ where $q_{G(e_i)}$ is a function that counts the number of tweets associated with the users in group G which are labeled by emotion e_i.

C. Topical Analysis

In this module, to address *RQ1*, our focus is on analyzing the topical interests of users that are implied in their posts. We use BERTopic for effective topic modeling of tweets [42], [43]. To extract topics of each dataset M, by assigning each tweet to a single document, BERTopic generates two artifacts: 1) a set of K topics Z where each topic $z \in Z$ is associated with a topic-word distribution; and 2) The topic of each tweet m, i.e., $z^m \in Z$. Thus, for a given user u, Let M_u be a set of N tweets published by user u and Z be a set of K topics, we represent the topic distribution of user u by $(f_u(z_1), \ldots, f_u(z_K))$, where $f_u(z)$ is the number of tweets of user u that are labeled by topic $z \in Z$. The user-topic representation is normalized using the L1–norm.

Finally, let U_G be the users belonging to the group G, we represent each group of users G by $T(G) =$

[2] https://developers.symanto.net/

$(f_G(z_1), \ldots, f_G(z_K))$, where $f_G(z) = \sum_{u \in U_G} f_u(z)/|U_G|$. The topical representation of the group G is normalized so that the sum of all weights in a profile equals 1.

It is noted that, before applying BertTopic, as suggested in [42], we lowercase the tweets and remove their URLs, mentions, punctuation, and special characters. Further, for topical analysis, we only pick one data published in March 2020 in which the highest number searches were recorded for the term "COVID-19" in Google Search. We set K to be 50 after a manual examination of topics.

D. Social Engagement, Posting Trend, and Information Quality Analysis

In this module, we address *RQ2* by profiling the online behavior of users in the hate and control groups. Thus, given each user in group G, we examine how hateful and control users differ by analyzing several activity-related statistics such as the total number of tweets, replies, followers, and friends (normalized by the users' account age). We further compare the two groups based on the number of URLs, mentions, hashtags, retweets, and the ratio of followers to followees (normalized by their number of tweets).

We also assess the information credibility of social media content shared by two groups of users as one of the facets in modeling online behavior of users. We adopt the common practice used in previous research [34] and examine the credibility of the information source by evaluating whether users share links from credible domains. Finally, we calculate the posting interval (seconds) between two consecutive tweets in group G to study the posting trends and users' regularity.

E. Personality Analysis

In this module, we focus on *RQ2* and analyze the differences in personality traits between users in hate and control groups. We employ the five factor model [44] (aka the big five) that summarize human psychological dimension in five aspects, namely, *openness to experience*, *conscientiousness*, *agreeableness*, *extraversion*, and *neuroticism* each of which presents a positive and complementary negative dimension. Neuman et al. [45] further extended the big five model and proposed a vectorial semantic approach which constructs vectors to represent these ten dimensions as well as the personality assessment for nine different disorders including depression, paranoid, schizoid, to name a few.

We adopt this model and represent each group G by a vector over these nineteen dimensions of personality, denoted by $Q(G) = (q_G(p_1), \ldots, q_G(p_{10}), q_G(p_{19}))$, where $q_G(p_i)$ is a function that calculates the mean value of the tweets associated with the users in group G which are labeled by personality p_i.

F. Polarization Analysis

Here, we focus on *RQ3* and investigate how far the tweets of hateful users can be characterized by their attitude polarization. The notion of attitude polarization is built upon the theory of social psychological research called "attitude strength" [46] in which attitude polarization is attributed to extremity and ambivalence. We adopt a SentiStrength algorithm [47] to inspect the tone of the language used in the timelines posted by both hate and control groups. We extract markers of attitude polarization by modeling: 1) the extremity of the attitude through calculating the strength of sentiment in users' timelines; and 2) the ambivalence of the attitude by measuring the simultaneous presence of positive and negative sentiment in users' posts.

To this aim, for a user $u \in G$, given that SentiStrength returns two values (one for positivity, one for negativity) for each tweet t_j in the user's timeline $T_u = t_1, t_2, \ldots, t_N$, we represent user u by a vector of weights over three criteria $C =$ sentiment direction (c_1), attitude extremity (c_2), attitude ambivalence (c_3) as $Y_u(C) = (y_T(c_1), y_T(c_2), y_T(c_3))$, where $y_T(c_i)$ is a function that calculates the mean value for criteria c_i considering the positive and negative scores of each tweet t in her timeline T_u based on the formulations presented in [32], [48]. Finally, we represent each group G by a mean value of its users vector $Y_{u \in G}(C)$ over different criteria c_i and further normalize them to be valued between [0,1].

V. Insights Into Online Hate Language and Behaviour

In this section, we describe the results associated with each analytical module presented in Section IV to answer our research questions *RQ1*, *RQ2*, and *RQ3*.

A. Word Analysis

1) Vocabulary Uniqueness: Our analysis indicates that hate and control users' vocabulary are highly similar with Jaccard index of 62.67%. One possible explanation for such similarity might be related to the fact that both groups of users share similar interests and topics.

2) Linguistic Style: In this section, we study the language of both hateful and control users by comparing a subset of categories on LIWC. From Fig. 2, we observe that the proportion of tweets with words related to positive emotions (named *posemo*) is larger than negative ones (named *negemo*), *even* for the hateful users (p-value < 0.0001).[3] This can be explained by the Pollyanna effect [49] signifying that human languages exhibit a clear positive bias, which also has been observed in other social media studies (e.g., [35]). Our findings also show that the usage of words with negative emotions among hateful users is significantly higher than in control ones. This trend confirms the outcomes of the emotion analysis which will be discussed in Section V-B.

Based on Fig. 2, we also see significant differences between the two groups in terms of personal concerns, drives, and cognitive and social processes. We observe that users in the hate group often use more words related to their concerns about work, home, religion, death, achievement, affiliation, as well as their friendships and social life as compared with the control group except for words related to personal concerns category (p-value < 0.0001). Finally, showcasing the LIWC categories

[3] Throughout the article, we calculate the p-values from two-sample t-tests to compare the averages across different populations.

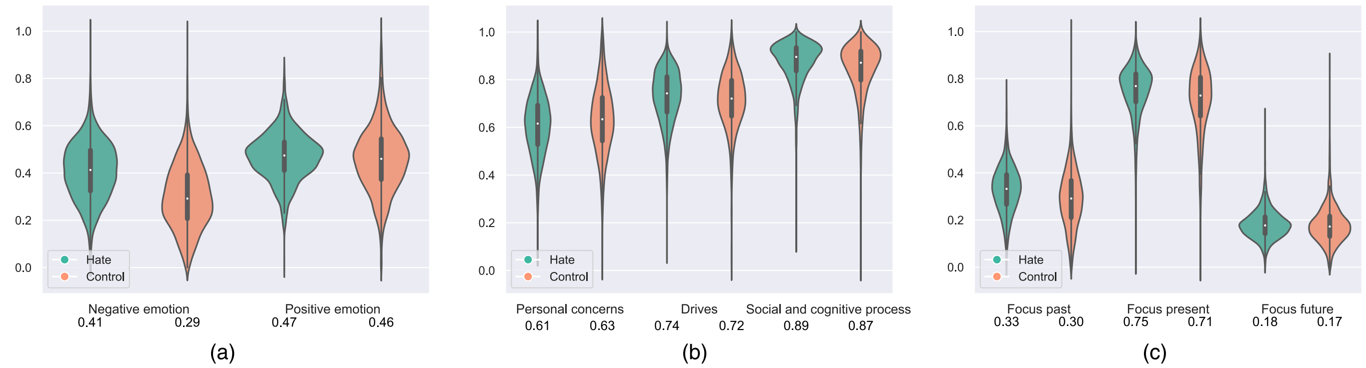

Fig. 2. Distribution of the proportion of tweets that users have on Twitter (y-axis) matching selected LIWC categories. (a) Comparison between positive emotion and negative emotion. (b) Personal concerns, drives and social and cognitive processes categories in LIWC. (c) Time-orientation category of LIWC for hate and control groups.

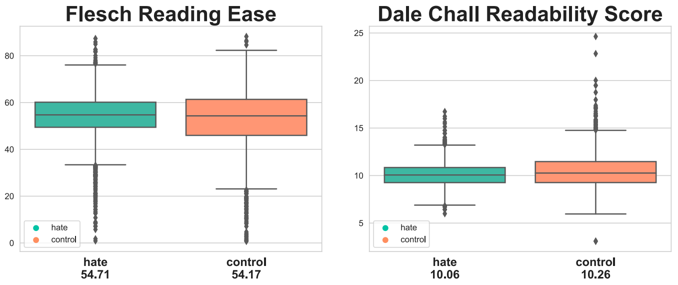

Fig. 3. Comparison of different readability metrics for hate and control groups.

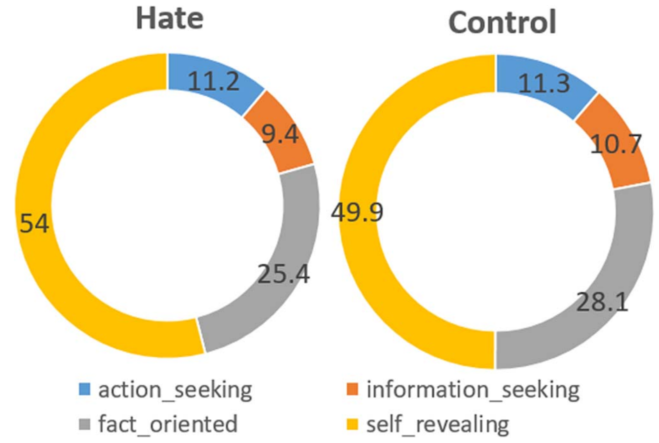

Fig. 4. Comparison of communication style between tweets of hate and control groups.

relative to the time-orientation, we observe that hateful users tend to write their verbs in all tenses more frequently compared to the control users (p-value < 0.0001). However, there is a significant difference in using past and present tenses among users in both groups as compared to writing future-focus content (see the median values of the diagram in Fig. 2(c)).

3) Readability and Communication Style Assessment: Fig. 3 depicts the difference between hateful and control users on the complexity level of their tweets based on different readability metrics. We observe that users in the hate group post tweets that are easier to read and understand. This observation is also confirmed through the *dale-chall* readability score metric, which measures the level of education that is required to understand a written text. We find that users in the control group post content containing more complex words which require a higher level of education to understand compared to the content shared by hateful users.

Furthermore, we attempt to understand the intention of users implied in their posts. As depicted in Fig. 4, our findings show that users in the control group post content that contains more *factual* information than those in the hate group, confirming our observation on information quality analysis in Section V-D, in which we found that control users share more content from credible resources than hateful users. Another interesting observation is on the dominance of *self-revealing* communication style as an underlying intention of the majority of posts in both groups. However, we found that hateful users share more content about their personal opinions compared to control users which is inline with our observation on different categories of LIWC indicated in Section V-A2, specifically for the social and cognitive processes.

B. Emotion Analysis

Fig. 5 provides a pairwise comparison between hate and control users' emotions. These results suggest that, on average, hateful users tend to share more terms related to negative emotions such as *anger*, *disgust*, and *fear* compared to control users. Furthermore, words related to positive sentiment, i.e., *joy* are more prevalent in control users than in hateful users. Looking at the *neutral* emotions, we observe that control users exhibit neutral feelings more frequently than hateful users.

C. Topical Analysis

Applying the topic model explained in Section IV-C, we extracted a total of 50 topics for both hate and control groups. Here we report on the most popular six topics extracted for both user groups and their associated top-20 words in Tables III and IV of Appendix A. We also extracted the discussed themes

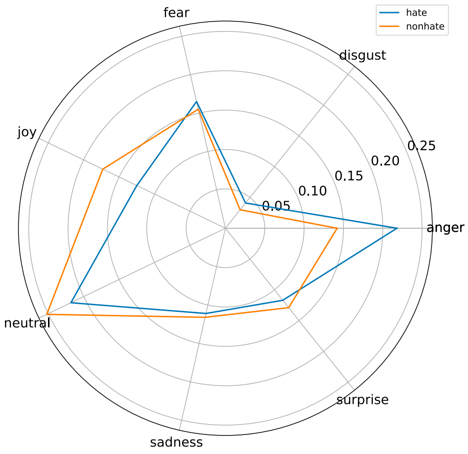

Fig. 5. Radar diagram depicting different dimensions of emotions averaged over users in hate and control groups.

based on the list of words associated with the topics and report them on the tables.

Looking at the topics discussed during the month of March for hate users, the vast majority of topics were around controversial informational (topic #4), political (topics: #1, #2, and #5), or technology-related subjects (topics: #3 and #6). On the other hand, control users discussed mainly news (topics: #1, #2, #5, and #6), political (topic #1), or religious-related themes (topic #4). Taking topic #1 as an example (U.S. presidential elections), while hateful users discussed whether the current U.S. president was qualified or not to be elected as president given his historical medical condition, control users were sharing information about the U.S. presidential elections in general.

Furthermore, even while focusing on the month of March only where COVID-19 topic was trending the most, we noticed that hateful users' content was not necessarily related to the COVID-19 (topics: #2, #3, and #6) compared to control users where most of the topics were COVID-19-related (topics: #2, #3, #5, and #6). In general, while control users had neutral interests in discussing COVID-19, hateful users were sharing more biased controversial topics which is well aligned with the findings of prior work [11].

D. Social Engagement, Posting Trend, and Information Quality Analysis

In this section, we characterize the hate group and a control group based on their activity patterns, tweeting history, and the quality of information they share on social media.

We show different statistics in Fig. 6. Specifically, our results indicate that users in the hate group are generally more "active" in the sense that they retweet more, have more mentions, and favorite more tweets despite having a longer posting interval compared to the control users ($p-$values < 0.0001 for tweets and favorites; and p-values < 0.1 for mentions and intervals). This observation agrees with the findings of a prior study where it was found that hateful users generate more content (i.e., tweet more) than other users [11]. In contrast, we find that hateful users use less hashtags and less URLs per tweet (p-value < 0.0001) than control users. Additionally, we find that users in the control group are more "popular" in that they have more followers and friends and the ratio of followers to friends is significantly larger than the hateful users (p-value < 0.0001). These findings are inline with previous research which suggests that hateful users do not necessarily show spamming behavior and use systematic and automated methods to deliver their content [12]. We also found that the Twitter accounts of hateful users are more recent (i.e., have relatively less age) compared to the control users (p-value < 0.0001)–see the bottom right diagram in Fig. 6. This observation supports previous findings on hate users as reported in various studies (e.g., [12]). Finally, we observe that users in the control group share information (i.e., URLs) from credible sources more often compared to the ones in the hate group as indicated in Fig. 6. Additional experiments on the social engagement behavior of two group of users is provided in Appendix C.

E. Personality Analysis

Following the theoretical framework in [45], we found significant differences between users in the two groups for all big five factors in their positive and negative dimension. As we can see from Fig. 7, users with personality traits in positive senses are more dominant in the control group (p-values < 0.01). In contrast, negative personalities are manifested more in users in the hate group for all characteristics except for Neuroticism($-$). The majority of users in both groups are characterized by agreeableness ($+$) (22.9% in hate group versus 27.1% in control group) and extraversion($-$) (31.7% in hate group versus 30.7% in control group). Our findings align with the previous research in which hateful users are characterized with more agreeableness and extraversion [50]. We also observe that control users show characteristic traits of more empathetic and pleasant personality [aka agreeableness($+$)], more self-disciplined [as reflected in conscientiousness($+$)], and show more emotional stability [according to Neuroticism($-$)] than hateful users. On the contrary, hateful users are shown to be more narrow-minded and nervous and less decisive [as reflected in openness to experience($-$) and extraversion($-$), correspondingly] than control users. We further showcase potential mental disorders as inferred by Neuman and Cohen's personality analytical model [45] in Fig. 8. Interestingly, we observe that different disorders are manifested more in users among the control group than in the hate group with the exception for narcissistic and dependent personality disorders. Furthermore, our findings show that users with avoidant and depression disorders account for more than half of the population in both groups while the number of histrionic and paranoid cases is found to be the lowest among users in both groups.

F. Polarization Analysis

Fig. 9 illustrates the markers of polarization as the distribution of sentiment direction, attitude extremity, and attitude ambivalence using a violin plot among the two groups of hate and control users. We notice that the sentiment direction of the two groups are significantly different (p-value < 0.0001).

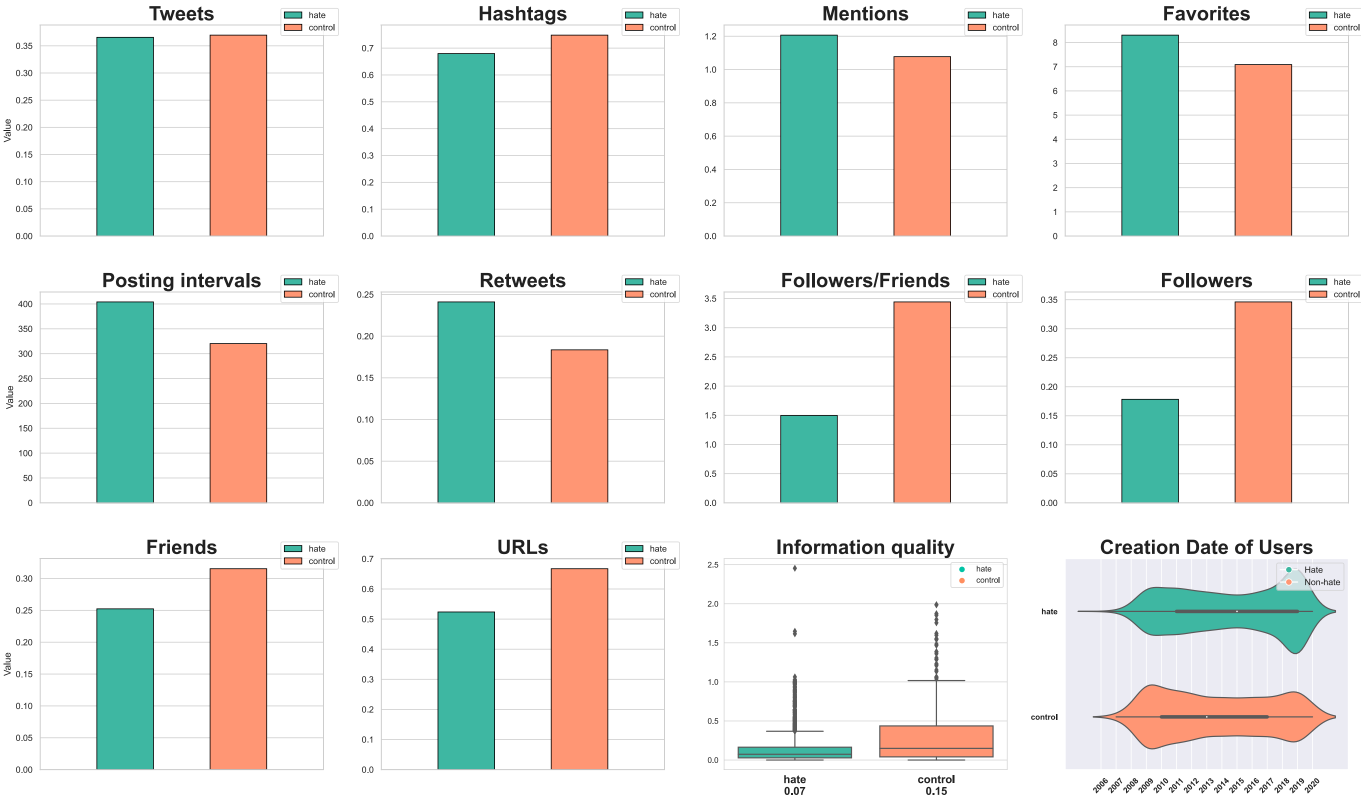

Fig. 6. Summary of characteristics of hate, and control groups. For each feature, a bar chart is shown. “Creation date” shows a violin plot of the account creation dates of each group, where a white dot indicates the median values.

The majority of users in the hate group are shown to use a negative-toned language in their tweets (with the sentiment direction equals to (-1) whereas the violin plot indicates that the sentiment direction of control users are equally distributed in the positive $(+1)$ and negative (-1) range– half of them use negative tone and the other half use the positive-toned language. Additionally, according to Fig. 9, around 50% of the hate crowd are shown to have higher attitude extremity in comparison with the three-quarters of the control folk with a significant difference in their median values (0.44 for hate group versus 0.37 in the control group, p-value < 0.0001). Fig. 9(a) depicts the state of attitude ambivalence that seems similar in both groups (with the median equal to 0.38 and p-value < 0.0001). We found that users in the hate group are more polarized as they exhibit more extremity and express more negativity in their timelines. Our findings are supported by previous research which associates the high extremity and low ambivalence in the language tone of users to high level of the attitude polarization [32].

Note that, we have replicated our experiment in a scenario where the control group consists solely of counter-hate users. The result of our analysis is presented in Appendix D.

VI. Hate-Mongering Prediction

In this section, our goal is to develop a model to predict the susceptibility of Twitter users to post hateful content in the future by looking at their past behavior. We are interested in answering the following research question:

RQ4. Are the linguistic and behavioral characteristics of users on social media indicators to automatically predict the susceptibility of users to post hateful content?

To answer *RQ4*, we train a feature-based binary classifier by utilizing the linguistic and behavioral features based on the results of our observational study in Section V and compare it with a *baseline* model trained on the embedding features of each user’s textual content.

A. Embedding Features

For our baseline, adopted from [51], [52], to embed each user, we first apply sentence BERT (SBERT) to encode each tweet in the user’s timeline and then we average the embedding representations across all her tweets. This results in a 384-dimensional user-level embedding vector for each user which is used as the input features in a binary classifier for identifying hateful-to-be users.

B. Linguistic and Behavioral Features

Based on the results of our observational study, we selected eight categories of features including *linguistic style*, *readability*, *emotion*, *social engagement*, *information quality*, *posting trends*, *personality traits,* and *polarization*. Based on these features, the final user-based representation is a 48-dimensional vector.

We used XGBoost to build a binary classifier for predicting hate speech spreaders. In our experiments, we employed a

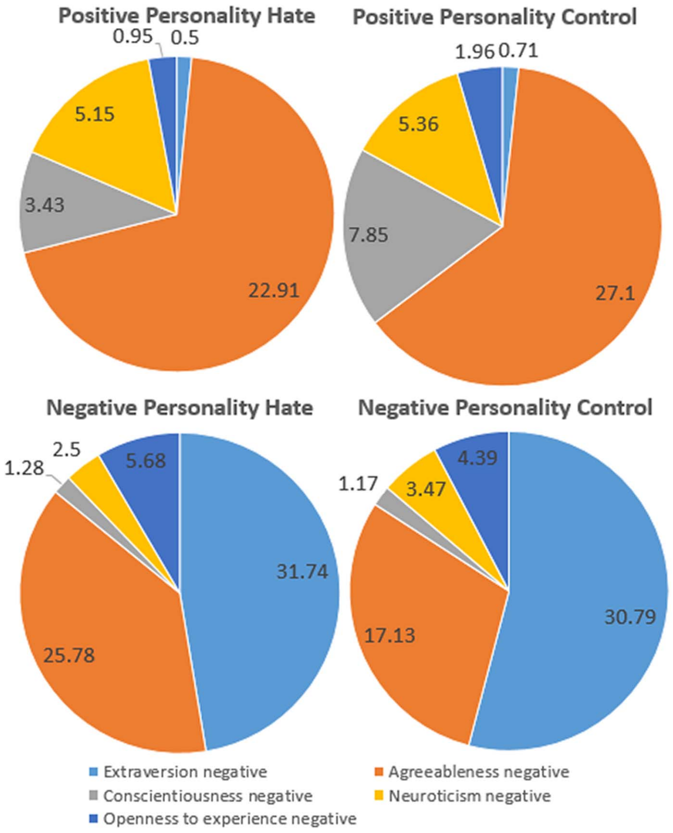

Fig. 7. Pie charts for positive and negative personality traits according to [45] in hate and control group.

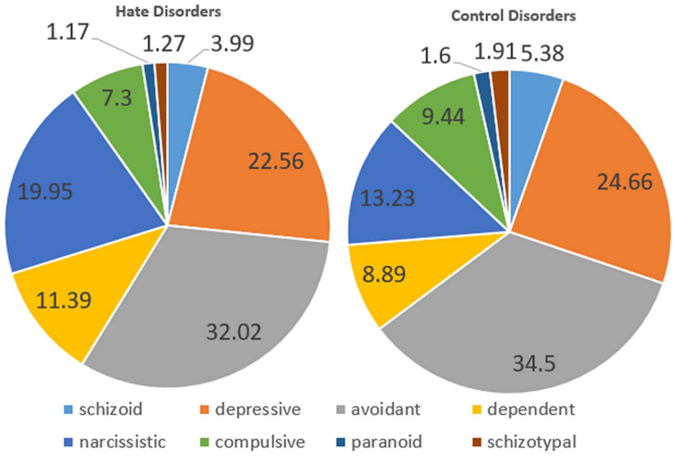

Fig. 8. Pie charts for disorder taxonomies according to [45] in hate and control group.

five-fold cross-validation approach to partition our dataset (3001 *hateful-to-be* users and 3001 control users), ensuring that each fold represented a distinct subset of the data. We applied grid search sklearn's method for hyper-parameter optimization to find the best set of hyperparameters of our model in each fold. The average results are reported in Table I in terms of accuracy (ACC), the area under the curve of the receiver operating characteristic (AUC ROC), precision (P), recall (R), and F1.

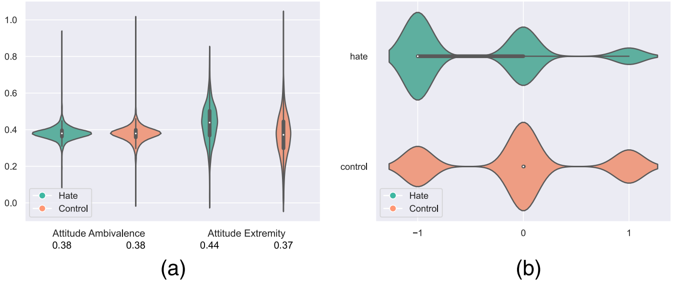

Fig. 9. Comparison of the polarization markers in hate and control groups. (a) Attitude ambivalence versus attitude extremity. (b) Sentiment direction.

TABLE I
RESULTS OF PREDICTING HATEFUL-TO-BE USERS

Feature Categories	ACC	AUC	F1	P	R
Embedding features	0.7517	0.8267	0.7527	0.82893	0.771983
Linguistic and behavioral features	0.8394	0.9268	0.84	0.94233	0.858427
Combination	**0.8544**	**0.9392**	**0.8554**	**0.94049**	**0.862313**

Note: Bold entries represent combinations of our feature categories, outperform other approaches.

TABLE II
ABLATION STUDY RESULTS

Feature List	ACC	AUC	F1	P	R
ALL	0.8544	0.9392	0.8554	0.9405	0.8623
- LIWC	0.8597	0.939	0.8599	0.9399	0.862
- Readability	0.7784	0.8605	0.781	0.8532	0.7906
- Emotion	0.8564	0.9401	0.8571	0.9404	0.8618
- Polarization	0.8602	0.9402	0.8611	0.9416	0.8672
- Social engagement	0.8547	0.9367	0.8548	0.9367	0.8563
- Information quality	0.855	0.9384	0.8558	0.9399	0.8609
- posting trend	0.8545	0.9379	0.8557	0.9391	0.8636
- Personality	0.8476	0.9318	0.8484	0.9339	0.854
- Embedding	0.8394	0.9268	0.84	0.9285	0.8441

Based on the results, the model with linguistic and behavioral features outperforms the baseline in which only embedding features are used, in terms of all the evaluation metrics. We conclude that the linguistic and behavioral features introduced in Section IV are stronger indicators to predict the *hateful-to-be* users compared to embedding features. We also combined the embedding features with our introduced linguistic and behavioral features in one model and reported its results in Table I. In the combination model, to avoid overfitting, we applied principal component analysis (PCA) over the embedding features and reduced the dimension to 50, and then combined it with the rest of the features. Therefore, in the combination model, we represent each user as a 98-feature vector, with the vector being normalized. Based on the results, one can observe that the combination model outperforms the other two models which confirms that adding embedding features helps improving the performance of the predictions.

To explore the relative effectiveness of each of the features in the best model (i.e., the combination model), an ablation study is executed where we remove each feature at a time and retrain our model. The results are reported in Table II. We can see that the *readability* features, *embedding* features, and *personality* features are the top three most effective features for hateful-to-be user prediction.

VII. Discussion

In this work, we propose novel ways on consuming and visualizing the vast amount of social media data (i.e., textual and behavioral data) to derive insights on the development of online hate-mongering behavior. Our methodology complements previous studies by predicting the likelihood of hate speech occurrences hence assisting social media practitioners to develop timely preventive measures to cease the spread of hate speech. The study further enhances research on human behaviors in social media, encompassing areas like mental health detection, fake news identification, and most crucially, the detection of hate speech.

Specifically, results of *RQ1* suggested that the language, communication styles, choice of words, and topics discussed by hateful-to-be users are significantly different from control users. Hateful users' behavior was found to be interestingly similar to fake news spreaders who post self-revealing content with more anger and negative emotions [53]. This observation calls for conducting future research to understand the role of fake news in fueling the spread of hate speech (and vise versa) and online extremism behavior in general. In *RQ2*, findings showed that hateful users shared information from less credible sources which agrees with prior work [11]. Further, similar to prior studies, we also observed that users with hateful attitudes were characterized with more prominent dependent and narcissistic personality disorders [50].

Further, looking at *RQ3*, results showed that hateful users are more polarized and show more extreme behavior. This corresponds with earlier research on individuals engaged in adversarial debates and controversial conversations on social media, where they demonstrate a strong tendency toward polarization [32]. Finally, in *RQ4*, we showed that our model's ACC is comparable to the work presented in [11]. However, our work outperformed in predicting hateful users with a high level of hateful activities.

In the following, we discuss the implications, limitations, and ethical concerns in our study.

1) *Implications.* The results of our work have several implications in different directions. In particular, as online hate speech has been strongly linked to offline behavior, monitoring online communication is important in improving hate speech interventions. For instance in persuasive technologies, one potential intervention would be to attempt to change hateful individuals' views around a topic of interest [54]. Additionally, as people tend to connect with similar-minded individuals in social media (homophily phenomena), it is possible that platforms recommend hateful content to users with a network containing such content neglecting the harmful implications of such recommendations [55]. Adopting the proposed approach in this study, social media platforms can modify their recommender systems to avoid spreading harmful content to users.

2) *Limitations.* Our study on online hate speech in social media has a number of limitations. First, social media adoption and usage heavily depends on different factors (age, gender, education, geographic location, etc.) As a result, the findingsmight under-represent demographic groups which consume less social media platforms [56]. Second, while Twitter is a heavily adopted social media platform [57], it has recently implemented moderation techniques to combat hate speech which heavily limits the research of studying the prevalence of hate speech. Further studies are required to generalize our findings for other moderation-free platforms such as Gab, 4chan, and Bitchute.

   Third, in this study, hateful users are detected based on a list of predefined set of keywords. As a result, we might have missed hate content present in other forms of communication.

3) *Ethical considerations.* Several ethical considerations should be taken into account when studying online hate speech detection. For instance, our work uses a publicly available Twitter dataset that does not contain private or deleted information. Additionally, to mitigate risks of stigmatization, we anonymized users' data and provided the dataset under a data usage agreement (DUA) emphasizing its use solely for research purposes. Next, the use of a Twitter dataset raises concerns about algorithmic bias, a common issue in user-based predictive models [58]. Further, we emphasize the need for ongoing critical examination of the findings to prevent biases and advocate for further studies assessing the real-life impact of interventions on public policies.

VIII. Conclusion and Future Work

This article proposed a methodology for understanding and predicting hate-mongering behaviors on social media. The initial focus is on distinguishing users spreading hateful content from those who do not. Using a curated dataset, we conducted a large-scale study examining online activities of both groups. To understand triggers for hate-motivated behavior, psycholinguistic and behavioral characteristics are extracted from users' posts (linguistic features, readability, communication style, and posting trends). Significant differences in polarity, word usage, emotional expression, personality traits, and social engagement emerge between the two user groups. Additionally, a predictive model was developed to anticipate whether a user is likely to publish hateful content based on past online behavior.

In the future, we will enhance our model's interpretability by employing visualization techniques for high-dimensional feature vectors and conducting a deeper analysis of feature interactions. We will also focus on a detailed examination of evaluation metrics to better understand and improve our model's decision-making process in predicting hate-mongering behavior. We further plan to conduct a time series analysis to investigate how posting hateful content influences users' language and behavior over time. Additionally, we aim to explore potential causal relationships between users' psychological states and online hate attitudes and formalize a prediction task to estimate the intensity of hatred among social media users. Finally, we are interested in analyzing the generalizability potential of our methodology across various topics and other social platforms. Beyond Twitter, it will be also valuable to analyze the expression of hateful speech on other social media platforms.

TABLE III
Topics Extracted From Tweets of Control Users

#	Topic Theme	Topic Words
1	U.S. presidential candidates	“biden,” “bernie,” “joe,” “warren,” “sander,” “tulsi,” “trump,” “elizabeth,” “pete,” “vote,” “supporter,” “dnc,” “tiger,” “win,” “campaign,” “debate,” “progressive”
2	COVID origins	“china,” “chinese,” “communist,” “fake,” “world,” “news,” “virus,” “taiwan,” “medium,” “government,” “american,” “country,” “party,” “ironic,” “blame,” “wuhan,” “account,” “believe”
3	COVID information	“italy,” “spain,” “death,” “italian,” “case,” “coronavirus,” “total,” “andalucia,” “new,” “toll,” “flu,” “lockdown,” “nikon,” “covid-19,” “spanish,” “freelance,” “photographer,” “french,” “china”
4	Religious faith	“god,” “prayer,” “jesus,” “amen,” “pray,” “bless,” “bank,” “eradicatecovid19,” “oneworldonefamily,” “name,” “lord,” “mighty,” “healed,” “christ,” “first,” “shall,” “mercy,” “heal,” “1877theglory,” “preach”
5	COVID statistics	“iceland,” “test,” “day,” “inspired,” “petition,” “sign,” “rate,” “prove,” “spread,” “together,” “fatality,” “testing,” “germany”
6	COVID media reporting	“trump,” “hoax,” “coronavirus,” “fox,” “covidiot45,” “trumppandemic,” “cnn,” “president,” “news,” “donald,” “response,” “donna,” “gob,” “foxnews,” “briefing,” “hell,” “medium,” “speech,” “maga,” “virus”

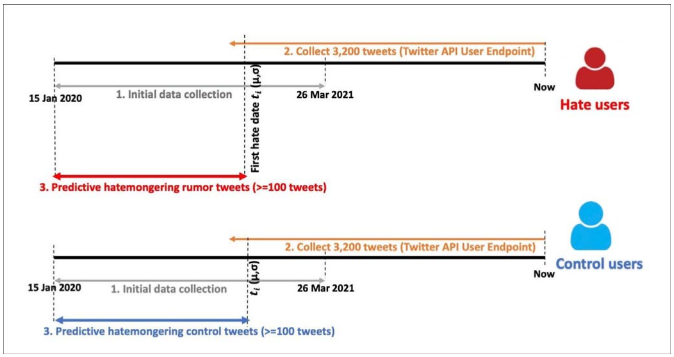
2. Collect 3,200 tweets (Twitter API User Endpoint)
15 Jan 2020
1. Initial data collection
First hate date t_i (μ,σ)
26 Mar 2021
Now
Hate users
3. Predictive hatemongering rumor tweets (>=100 tweets)
2. Collect 3,200 tweets (Twitter API User Endpoint)
15 Jan 2020
1. Initial data collection
t_i (μ,σ)
26 Mar 2021
Now
Control users
3. Predictive hatemongering control tweets (>=100 tweets)

Fig. 10. Users’ historical data section. For both hate and control users: 1) We collect tweets between 15 January 2020 and 26 March 2021. 2) We use Twitter API to collect their timelines, i.e., most recent 3200 tweets. 3) We use the user’s timeline posts up to the first hate tweet for analysis (only users with at least 100 tweets). For control users, we pick such a date from a normal distribution with mean and variance of first hate posts of hate data.

Appendix A
Topic Analysis

See Tables III and IV.

Appendix B
User Selection Process

See Fig. 10.

TABLE IV
Topics Extracted From Tweets of Hateful Users

#	Topic Theme	Topic Words
1	U.S. presidential qualifications	“biden,” “joe,” “bernie,” “sander,” “dementia,” “trump,” “dnc,” “bidens,” “president,” “ukraine,” “he,” “win,” “vote,” “women,” “going,” “vp,” “running,” “loser,” “alzheimers”
2	Hong Kong pro-democracy movement	“hk,” “police,” “boycottmulan,” “hkpolice,” “hkpolicebrutality,” “hongkong,” “hkpoliceterrorists,” “mulan,” “universal,” “kong,” “hong,” “suffrage,” “standwithhongkong,” “hkpolicestate,” “hongkongpolice,” “hongkongprotests,” “actress,” “brutality,” “station,” “support”
3	Digital currency debate	“bitcoin,” “block,” “satoshi,” “learnbitcoin,” “praise,” “hash,” “devotion,” “reward,” “injustice,” “11block,” “satoshinakamoto,” “mod88block,” “fairness,” “darkness,” “unfairness,” “blockchain”
4	COVID statistics	“italy,” “italian,” “death,” “case,” “coronavirus,” “spain,” “covid19,” “china,” “rate,” “lockdown,” “virus,” “corona,” “northern,” “lombardy,” “chinese,” “toll,” “germany”
5	Iran’s response to COVID	“iran,”,“khameneivirus,” “true,” “iranian,” “regime,” “sanction,” “covidsanctionslie,” “ayatollahsspreadcovid19,” “mullah,” “imrankhan,” “islamic,” “coronavirus,” “islamicrepublicvirus,” “occupuied,” “sympathy,” “pilgrimage,” “citizen,” “jammukashmir”
6	Twitter content moderation policy	“twitter,” “tweet,” “blocked,” “conspiracy,” “bot,” “theory,” “retweet,” “tweeting,” “chinese,” “discontent,” “account,” “like,” “warfare,” “delkvfgjfhhtyr65ete,” “nationalism,” “count,” “china,” “patriotism,” “mjkki”

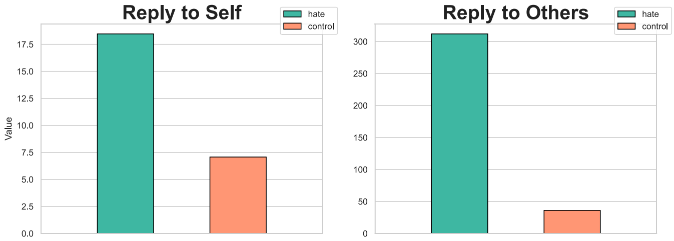

Fig. 11. Comparative analysis of reply tendencies between hate and control groups.

Appendix C
Additional Experiments on Social Engagement

We carried out further experiments to compare the average number of replies by users in the hate and control groups, as well as their tendencies to reply to their own tweets and those of others. Our observations in Fig. 11 indicates that hateful users are more inclined to reply both to their own tweets and to others, compared to those in the control group. Additionally, the average number of replies from hateful users is significantly higher than that of users in the control group.

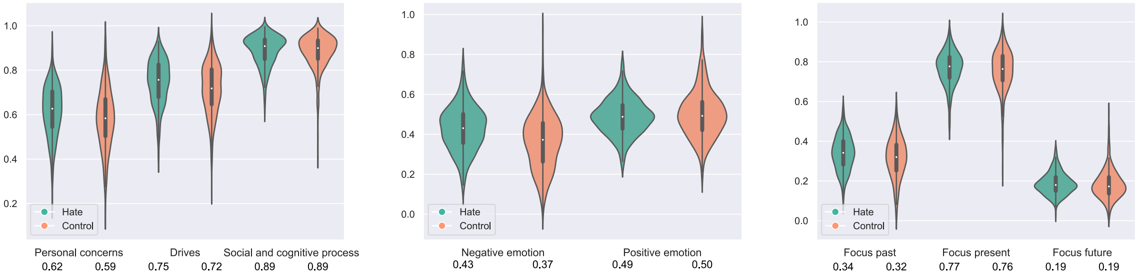

Fig. 12. Distribution of the proportion of tweets among hate and counter-hate users on Twitter based on selected LIWC categories.

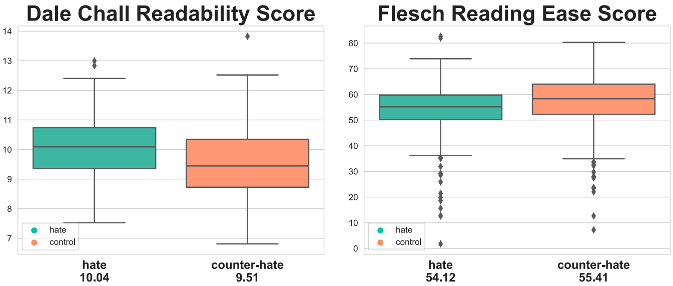

Fig. 13. Comparison of different readability metrics for hate and counter-hate users.

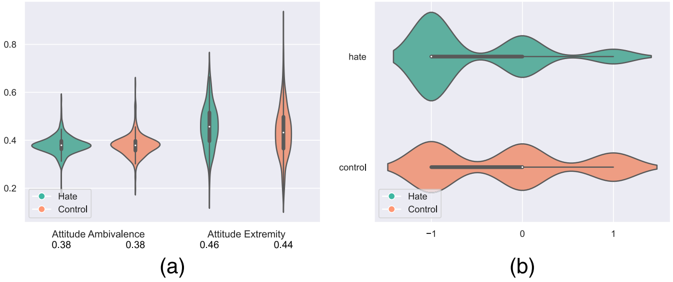

Fig. 14. Comparison of the polarization markers in hate and counter-hate groups. (a) Attitude ambivalence versus attitude extremity. (b) Sentiment direction.

Appendix D
Additional Experimental Setting

We have replicated our experiments in additional experimental settings, where we compare hate users exclusively with counter-hate users. The hate group consist of 355 hateful users with a total of 295 017 tweets and the control group consist of 355 counter-hate users with a total of 321 659 tweets. Fig. 12 displays the results of the LIWC analysis. It reveals that negative emotion is significantly higher among counter-hate users in the control group compared to when neutral users are also present (0.37 versus 0.29). Additionally, counter-hate users exhibit higher levels of positive emotion than hate users in this experimental setup. The analysis also indicates that counter-hate users discuss their personal concerns less frequently than those in the hate group (0.62 versus 0.59), a contrast to the findings in settings that include neutral users in the control group. Similar trends are observed in the drive, social, and cognitive process categories, as well as in the usage of different tenses in the generated content, compared to other experimental conditions.

In evaluating the readability scores of users in two groups, where the control group consists solely of counter-hate users, we note from Fig. 13 that the texts written by counter-hate users are more complex and require a higher education level to understand. This finding is contrary to our observations in settings where neutral users are also part of the control group.

The polarization score analysis, shown in Fig. 14, indicates that control users, comprising only counter-hate users, exhibit more extreme attitudes in their posts compared to settings where the control groups also include neutral users (0.40 versus 0.37). However, hate users consistently demonstrate a higher degree of polarization in their posts across all experimental settings.

References

[1] K. Saha, E. Chandrasekharan, and M. De Choudhury, “Prevalence and psychological effects of hateful speech in online college communities,” in *Proc. ACM Conf. Web Sci.*, 2019, pp. 255–264.

[2] F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on Facebook,” in *Proc. Italian Conf. CyberSecurity*, 2017, pp. 86–95.

[3] F. Tahmasbi et al., “‘Go eat a bat, chang!’: On the emergence of sinophobic behavior on web communities in the face of COVID-19,” in *Proc. Web Conf. (WWW)*, 2021, pp. 1122–1133.

[4] R. Flanagan, “Coronavirus racism: Canada’s top doctor blasts ‘stigmatizing comments’ on social media,” *CTV News*, Jan. 2020. [Online]. Available: https://shorturl.at/fjlDT

[5] T. Wong, “Sinophobia: How a virus reveals the many ways China is feared,” *GBBBC News*, Feb. 2020. [Online]. Available: https://shorturl.at/atFS6

[6] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, “Resources and benchmark corpora for hate speech detection: A systematic review,” *Lang. Resour. Eval.*, vol. 55, pp. 477–523, Sep. 2020.

[7] A. Al-Hassan and H. Al-Dossari, “Detection of hate speech in social networks: A survey on multilingual corpus,” in *Proc. Int. Conf. Comput. Sci. Inf. Technol.*, vol. 10, 2019, pp. 10–5121.

[8] B. AlKhamissi and M. Diab, “Meta AI at Arabic hate speech 2022: MultiTask learning with self-correction for hate speech classification,” in *Proc. Eur. Lang. Resour. Assoc.*, 2022, pp. 186–193.

[9] F. Rangel, G. L. De la Peña Sarracén, B. Chulvi, E. Fersini, and P. Rosso, “Profiling hate speech spreaders on Twitter task at PAN 2021,” in *Proc. Conf. Labs Eval. Forum*, 2021, pp. 1772–1789.

[10] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, “Author profiling for hate speech detection,” 2019, *arXiv:1902.06734*.

[11] J. An, H. Kwak, C. S. Lee, B. Jun, and Y.-Y. Ahn, “Predicting anti-Asian hateful users on Twitter during COVID-19,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4655–4666.

[12] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira Jr, "Characterizing and detecting hateful users on Twitter," in *Proc. AAAI Conf. Web Social Media*, Jun. 2018, pp. 676–679.

[13] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," in *Proc. ACM Conf. Web Sci.*, 2019, pp. 173–182.

[14] B. Mathew, A. Illendula, P. Saha, S. Sarkar, P. Goyal, and A. Mukherjee, "Hate begets hate: A temporal study of hate speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW2, pp. 1–24, 2020.

[15] J. Li and Y. Ning, "Anti-Asian hate speech detection via data augmented semantic relation inference," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16, 2022, pp. 607–617.

[16] H. Lyu, L. Chen, Y. Wang, and J. Luo, "Sense and sensibility: Characterizing social media users regarding the use of controversial terms for COVID-19," *IEEE Trans. Big Data*, vol. 7, no. 6, pp. 952–960, 2021.

[17] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[18] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS One*, vol. 15, no. 12, Dec. 2020, Art. no. e0243300.

[19] J. Salminen et al., "Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 12, no. 1, 2018, pp. 330–339.

[20] S. MacAvaney, H. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS One*, vol. 14, no. 8, 2019, Art. no. e0221152.

[21] F. Vargas, F. R. de Góes, I. Carvalho, F. Benevenuto, and T. Pardo, "Contextual-lexicon approach for abusive language detection," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2021, pp. 1438–1447.

[22] K. Englmeier and J. Mothe, "Application-oriented approach for detecting cyberaggression in social media," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.,* New York, NY, USA: Springer, 2021, pp. 129–136.

[23] G. Kumar, J. P. Singh, and A. K. Singh, "Autoencoder-based feature extraction for identifying hate speech spreaders in social media," *IEEE Trans. Comput. Social Syst.*, early access, 2023.

[24] R. Kshirsagar, T. Cukuvac, K. Mckeown, and S. McGregor, "Predictive embeddings for hate speech detection on Twitter," in *Proc. 2nd Workshop Abusive Lang. Online*, 2018, pp. 26–32.

[25] S. C. d. Silva, T. C. Ferreira, R. M. S. Ramos, and I. Paraboni, "Data-driven and psycholinguistics-motivated approaches to hate speech detection," *Computación y Sistemas*, vol. 24, no. 3, pp. 1179–1188, 2020.

[26] B. Vidgen et al., "Detecting East Asian prejudice on social media," in *Proc. 4th Workshop Online Abuse Harms*, 2020, pp. 162–172.

[27] J. Melton, A. Bagavathi, and S. Krishnan, "DeL-haTE: A deep learning tunable ensemble for hate speech detection," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.,* Piscataway, NJ, USA: IEEE Press, 2020, pp. 1015–1022.

[28] S. Masud et al., "Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 504–515.

[29] J. Qian, M. ElSherief, E. M. Belding, and W. Y. Wang, "Leveraging intra-user and inter-user representation learning for automated hate speech detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., Vol. 2 (Short Papers)*, 2018, pp. 118–123.

[30] M. Lai, M. A. Stranisci, C. Bosco, R. Damiano, and V. Patti, "Analysing moral beliefs for detecting hate speech spreaders on Twitter," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.,* (pp. 149–161), Cham, Switzerland: Springer, 2022, pp. 149–161.

[31] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, "Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2021, pp. 90–94.

[32] J. Buder, L. Rabl, M. Feiks, M. Badermann, and G. Zurstiege, "Does negatively toned language use on social media lead to attitude polarization?" *Comput. Hum. Behav.*, vol. 116, Mar. 2021, Art. no. 106663.

[33] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "Detection of novel social bots by ensembles of specialized classifiers," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, Virtual Event, Ireland: ACM, Oct. 2020, pp. 2725–2732.

[34] A. Ghenai and Y. Mejova, "Fake cures: User-centric modeling of health misinformation in social media," *Proc. ACM Human-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–20, 2018.

[35] E. A. Ríssola, M. Aliannejadi, and F. Crestani, "Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation," *Inf. Process. Manage.*, vol. 59, no. 3, May 2022, Art. no. 102890.

[36] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[37] R. Raut and F. Spezzano, "Enhancing hate speech detection with user characteristics," *Int. J. Data Sci. Analytics*, pp. 1–11, Aug. 2023.

[38] J. R. Davenport and R. DeLine, "The readability of tweets and their geographic correlation with education," 2014, *arXiv:1401.6058*.

[39] P. Jacob and A. L. Uitdenbogerd, "Readability of twitter tweets for second language learners," in *Proc. 17th Annu. Workshop Australas. Lang. Technol. Assoc.*, 2019, pp. 19–27.

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," Feb. 2020, *arXiv: 1910.01108*.

[41] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 3687–3697.

[42] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.

[43] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," *Frontiers Sociol.*, vol. 7, May 2022, Art. no. 886498.

[44] S. Rothmann and E. P. Coetzer, "The big five personality dimensions and job performance," *SA J. Ind. Psychol.*, vol. 29, no. 1, pp. 68–74, Jan. 2003.

[45] Y. Neuman and Y. Cohen, "A vectorial semantics approach to personality assessment," *Sci. Rep.*, vol. 4, no. 1, Apr. 2014, Art. no. 4761.

[46] L. C. Howe and J. A. Krosnick, "Attitude strength," *Annu. Rev. Psychol.*, vol. 68, no. 1, pp. 327–351, 2017.

[47] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

[48] S. J. Breckler, "A comparison of numerical indexes for measuring attitude ambivalence," *Educ. Psychol. Meas.*, vol. 54, no. 2, pp. 350–365, 1994.

[49] P. S. Dodds et al., "Human language reveals a universal positivity bias," *PNAS*, vol. 112, no. 8, pp. 2389–2394, 2015.

[50] B. Mathew, N. Kumar, P. Goyal, and A. Mukherjee, "Interaction dynamics between hate and counter users on Twitter," in *Proc. 7th ACM IKDD CoDS 25th COMAD*, 2020, pp. 116–124.

[51] I. B. Schlicht and A. F. Magnossao de Paula, "Unified and multilingual author profiling for detecting haters," in *Proc. CLEF*, CEUR, 2021, pp. 1837–1845.

[52] I. Vogel and M. Meghana, "Profiling hate speech spreaders on Twitter: SVM vs. Bi-LSTM," in *Proc. CLEF*, 2021, pp. 2193–2200.

[53] A. Giachanou, B. Ghanem, E. A. Ríssola, P. Rosso, F. Crestani, and D. Oberski, "The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers," *Data Knowl. Eng.*, vol. 138, Feb. 2023, Art. no. 101960.

[54] S. Berkovsky, J. Freyne, and H. Oinas-Kukkonen, "Influencing individually: Fusing personalization and persuasion," *ACM Trans. Interactive Intell. Syst.*, vol. 2, no. 2, pp. 1–8, 2012.

[55] E. Karakolis, P. F. Oikonomidis, and D. Askounis, "Identifying and addressing ethical challenges in recommender systems," in *Proc. 13th Int. Conf. Inf., Intell., Syst. Appl.*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–6.

[56] A. Smith, "Social media use in 2018," Washington, D.C., USA: Pew Research Center, 2018.

[57] S. Atske, "Social media use in 2021," Washington, D.C., USA: Pew Research Center, Apr. 2021.

[58] N. T. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," Brookings Institute, Washington, DC, USA, vol. 2, 2019. [Online]. Available: https://shorturl.at/cxAFU