RESEARCH ARTICLE OPEN ACCESS

DETECTING AND PREVENTING HATEFUL COMMENTS ON SOCIAL MEDIA USING BACK PROPAGATION NEURAL NETWORK ALGORITHM

Senthil S, Somesshwari G S, Sarupradha S, Suvedha S

Department of Computer Science and Engineering

K Ramakrishnan College of Technology,

Trichy - 621112

somesshwari96@gmail.com, ssuvedha2002@gmail.com, ssarupradha@gmail.com

Abstract— Everyone has the right to express themselves freely. However, this right is being abused to discriminate against and hurt people, either verbally or physically, under the preteens of free expression. Such intolerance is referred to as hate speech. Hate The definition of speech as language that conveys animosity toward a an individual or a group of people because of character traits like race, religion, ethnicity, gender, nationality, handicap, and sexual orientation. It can take the shape of statements, writing, actions, or displays that single out a person because of their membership in a specific group. Both offline and online, vile language has been more prominent in recent years. On social media and other online platforms, heinous content grows and spreads, eventually resulting in hate crimes. The utilization of social media sites and also the exchange of information have greatly benefited humanity. of information. Nevertheless, this has brought forth a variety of problems, including the propagation of hate speech. In order to address this growing problem on social media platforms, recent studies combined a variety of machine learning and deep learning approaches using text mining method for detecting hate speech messages in real-time datasets. Therefore, the purpose of this study is to review the different hate speech detection algorithms and forecast the top algorithms for social media datasets. Additionally, hate speech detection in real-time social settings has been enabled via cell phone notifications.

Index Terms—Social media, Hate Speech, Machine learning, Deep learning, Text mining

1. INTRODUCTION

Social media is a widely used and, more importantly, simple way for people to communicate. communicate with others online and openly share their thoughts and opinions. It is become an essential element of daily life. It's a stage where people are more susceptible to abuse or harassment from others who exhibit hate in a variety of ways, including sexism, racism, politics, and other types. These social media platforms are increasingly being used for cybertronic, online annoyance, and blackmail. We can now easily interact with a variety of societies or organizations that interest us thanks to social networking sites (SNS). Due to the advancement of several technologies, including high-speed internet and portable gadgets, these websites have reached a sizeable portion of the population. In these networks, handlers predominately have ages under thirty. Researchers have conducted considerable research in a variety of subjects by utilizing the enormous volumes of data present on different social networking websites. Popular academic discipline called sentiment analysis makes extensive use of data from social media. The numerous sorts of social networking sites are depicted in Figure 1.

Fig 1: Social media types

2. RELATED WORK

Paula fortuna, et.al,…[1] presented a crucial overview of how the rapid identification of inciting hatred in text has progressed so over past years. First, we examined the concept of hateful speech in various contexts, ranging from social media platforms to other organisations. Based on our findings, we proposed a more unified and precise definition of this concept, which can aid in the development of a model for the rapid recognition of hate speech. In addition, we contrast and classification rules from the literature, as well as arguments for and against those rules. Our critical viewpoint pointed out that our definition of hate speech is more inclusionary as well as general than some other perspectives cited in the literature. This is due to we propose a certain subtle kind of discrimination just on internet and through social networks be detected as well. We also concluded that comparing hate speech to cyber bullying, abusive language, marginalisation, toxicity, flaming, extremism, and radicalization was important. Our comparison demonstrated how hate speech differs from all these key topics and assisted us in understanding the limitations and subtleties of its definition.

Zafer Al-Makhadmeh,et.al,…[2] implemented To analyse Twitter data, the KNLPEDNN-based hate system for speech recognition is used. Tweets are initially gathered from hurricane front and rest of the audience fewer datasets. An NLP approach is used to process the collected data. The NLP tokenization process removes data protagonists, hashtags, user information, and other unwanted details. The system evaluates Tweets in terms of statement and word structure before generating NLP features such as semantic, sentiment, unigram, and pattern features. Vectors are generated from the extracted features, and morals are assigned based on the patterns. Using an optimised function and weight trying to update process, the features are then filtered using an outfit machine learning classifier to determine whether replying Tweets will be classified as hate speech, hateful speech, or neither. The excessive use of social media over the last decade has resulted in a rise in hateful activities through social networks. Verbal abuse is considered to be the most risky of these activities, as such users must protect themselves from it on YouTube, Facebook, and Twitter, among other places. This paper describes a method for predicting hateful speech from social media sites that combines natural language processing with machine learning techniques. After collecting hate speech, steaming, token trying to split, character removal, and infection elimination are performed before beginning the hate voice recognition process. Following that, the collected data is investigated using a natural and powerful language processing model - based outfit deep learning technique (KNLPEDNN).

Rui Cao, et.al,…[3] proposed Deep Hate, a new deep learning framework that used mega text portrayals for automatic hateful speech detection. Deep hate was evaluated on three publicly available real - world datasets, as well as our extensive experiments revealed that it outperformed the best baselines. We also empirically examined the Deep hate prototype as well as provided insights into to the key features that aided in the detection of hateful speech throughout online social platforms. Our salient feature analysis helped to explain Deep hate's hate classification decision. In the future, we hope to integrate non-textual features into the Deep hate model and use more advanced techniques to improve the sentiment as well as topic representations of posts. Although existing methods, particularly deep learning methods, have shown impressive outcomes in fully automated hateful speech identification in social media, these models have limitations. For starters, most current methods have only taken into account single type texts, ignoring both these rich text data that could be used to enhance hate speech detection. Second, current deep learning techniques provided limited explanations for why a specific comment must be flagged as hateful speech.

Zeerak Waseem, et.al,…[4] presented a set of criteria for identifying racist and sexist racial epithets based on critical race theory. These could be used to collect more data but also address the issue of a small but prolific group of hateful users. Whereas the problem is not yet solved, we have discovered that employing the character n-gram approach gives us a solid basis. Apart from gender, demographic information improves little, but this may be due to a lack of coverage. On social media, verbal abuse in the shape of racist and sexist comments is common. As a result, many social media platforms address the issue of identifying hateful speech; however the definition of hatred speech differs widely and is largely manual. Despite these factors, NLP study on inciting hatred has been extremely limited, owing primarily to the lack of a broad definition of hateful speech, an examination of its population influences, as well as an enquiry of the most awesome features.

Thomas Davidson, et.al,…[5] implemented Lexical methods, while effective for identifying potentially offensive terms, are ineffective for trying to identify hateful speech; just a small proportion of

tweets blacklisted by the Hate base lexicon have always been classified as inciting hatred by human coders. 4 While robotic classification methods can accomplish relatively high precision in distinguishing between these various classes, a close examination of the results reveals that the inclusion or absence of specific aggressive or hate filled terms can both help and impede accurate classification. Verbal abuse is a tough thing to define because it is not uniform. Our definitions of hate speech frequently reflect with us own subjective biases.

Pinkesh Badjatiya, et.al,…[6] investigated the Neural network based architectures are being used to detect hate speech. We discovered that they outperformed the existing methods significantly. When deep neural model embeddings were combined with steepest descent boosted decision trees, the best accuracy values were obtained. The manual method of filtering out hatred tweets is indeed not scalable, so researchers are looking for automated solutions. In this paper, we look at the problem of determining whether a tweet is racist, sexist, or neither. Because of the inherent complexity of natural language constructs - different types of hatred, different types of objectives, multiple methods of representing a same meaning - the task is quite difficult. To demonstrate the mission nature of the word embedding, we show in Table 2 the top few similar terms for a few written pieces to use the original GloVe word embedding as well as embeddings did learn using DNNs. The "hatred" towards to the target words is clearly visible in the similar words procured using a deep neural network did learn embeddings, which is not visible in the similar language obtained using GloVe.

Muhammad Okky Ibrohim, et.al,…[7] discussed Detection of hateful speech as well as abusive language on Indonesian Twitter. We held a Focus Group Discussion (FGD) with staff from Direktorat Tindak Pidana Siber Bareskrim Polri, Indonesia's agency in charge of investigating cybercrime, to obtain a valid definition of hate speech, including hate speech characterization. The FGD findings are then incorporated into annotation guidance for the purpose of jotting down hate speeches. We conducted FGDs with Direktorat Tindak Pidana Siber Bareskrim Polri staff as well as discussions with an expert linguist to ensure that the annotator guidelines we developed were legitimate and easy to understand by an annotator who was not a linguistic expert. Furthermore, we created gold standard annotations to determine whether or not a potential annotator had also seen and comprehended the annotations guide. We then used tagging guidelines as well as gold standard annotations to create a dataset for identifying abusive language and hateful speech (including identifying targets, categories, and levels of hate speech). The dataset, including the tagging guidelines as well as gold standard annotations, is freely available to other researchers interested in researching hate speech as well as abusive language identity in Indonesian social media.

Ika Alfina, et.al,…[8] built a new dataset of tweets in the Indonesian language for hate speech detection and carried out a preliminary study in which they compared the performance of various features as well as machine learning algorithms. We manually classified the twitter posts into two categories: those usually contains hate speech and those that did not. The resulting dataset was 520 bytes in size, with 260 tweets for each "hate-speech" and "non-hate-speech" class. According to the results of the experiments, the superior F-measure was obtained while using word n-gram, especially when combined with RFDT (93.5%), BLR (91.5%), and NB (90.2%). We discovered that the word n-gram feature outperformed the character n-gram feature. The results also revealed that it was preferable to combine term unigram and word bigram rather than using word unigram alone. We also discovered that adding character n-grams and negative sentiment to feature sets was unnecessary. The number of people using social media is rapidly increasing nowadays. Facebook, the market leader, had 2 billion monthly active consumers in June 2017, which is more than one quarter of the world's population. This demonstrates how important social media has become as a communication medium today. If indeed the topic tends to attract public attention, social media technology allows the text to be sent quickly, spread widely, and even go viral. Sadly, this also implies that hate speech can spread easily and quickly, leading to conflicts between different groups in society.

Muhammad Okky Ibrohim, et.al,…[9] discussed the Indonesian abusive language on social media, which frequently stems from an uncomfortable condition or something repulsive and forbidden by religion. We also talked about the difficulties in detecting foul language on Indonesian social media, as well as the abusive words trying to write patterns on Indonesian social media. In this section, we create a new dataset but also conduct an experiment for detecting abusive language in Indonesian. The experiment results show that NB outperforms SVM and RFDT in all scenarios for categorising abusive language utilising our dataset. Word subsection (1 and

the combination of phrase n-gram produce better results than other features when extracted using NB, SVM, or RFDT. The test results also demonstrate that categorising the tweet in to the three labels (non-abusive language, exploitative but not offensive language, but also offensive language) is much more difficult than simply categorising the tweet as not abusive or abusive. The classifier humans used here has difficulty distinguishing if the tweet is violent but not insulting or offensive language.

Joni Salminen, et.al,…[10] needed to reduce toxicity in social media platforms. In this study, we tested various machine learning models (Logistic Regression, Nave Bayes, Support-Vector Machines, XGBoost, as well as Neural Network) for online hate detection and discovered that Boosting as a classified as well as BERT features as the strongest impactful portrayal of hateful social media posts produced the best results. The model's generalizability to various social media platforms is good, but it varies slightly between platforms. The findings lend support to the goal of creating more universal internet hate classifiers for various social media platforms. The model that we make public can be used in practical applications and further established by internet hate researchers. Furthermore, the absence of universal classifiers makes it difficult to compare results across studies as well as social media platforms. To summarise, the splitting of models as well as showcase representations complicates absolutely loathe detection across platforms and contexts.

3. EXISTING METHODOLOGIES

Social networking is a trendy and, most importantly, easy method for individuals to similar respect their thoughts and views even while interacting with the others online. It is now an essential part of everyday life. It is a time when individuals have been easily harassed as well as abused by others who express hatred in a variety of ways, including such sexism, racial prejudice, politics, and so forth. The use of social media sites for cyber tyrannical, online menace to society, as well as blackmail is also increasing. Social networking sites (Myspace) have made it simple to connect with a variety of societies or organisations in which we are interested. As a result of technological advancements including such elevated internet and handheld devices, these sites have reached a large number of people in society. The majority of these networks' handlers are younger than the age of thirty. Researchers have used the enormous amounts of information available on various social networking sites to undergo comprehensive studies in a wide range of fields. Emotion Analysis is a common study field that makes extensive use of social media data.

Each user has been assumed to work autonomously in content-based filtering. As a result, a material filtering system chooses elements based on the relationship between the items' content and the user preferences, as compared to a cooperative filtering system, which selects objects based on the correlation of people with similar preferences. While initial stuff on information filtering focused on electronic mail, subsequent papers have addressed a wide range of topics, including wire service articles, Internet "news" articles, and wider network resources. Because the documents processed throughout content-based filtration are mostly text - based in nature, content-based filtering is similar to text classification. Filtering can, in fact, be modelled as a case of a single tag, classification, partitioning inbound documents into pertinent and irrelevant categories. Multi-label text categorization is a more complex filtering system that automatically labels messages into partial themes and categories. The use of the ML paradigm in content-based filtering is primarily based on the use of a classifier that is instantly induced by having to learn from the collection of pre-classified examples. A remarkable range of related research has recently emerged, with differences in feature extraction techniques, model learning, and sample collection. The extracting features procedure converts text into the a small subset of its content and is used consistently during the training and generalization phases. Several experiments show that Bag of Words (BoW) approaches outperform more sophisticated text representations, which may have superior semantic meaning but lower statistical quality.

4. PROPOSED METHODOLOGIES

Online Social Networks (OSNs) have become one of the greatest popular interactive mediums for communicating, sharing, and disseminating a large amount of personal information. One fundamental problem in today's On-line Social Media networks (OSNs) is providing users with the capability to control this same continuation on their own privacy in order to avoid the display of unwanted content. Until now, OSNs have provided little assistance with this requirement. To fill the void, we propose in this paper a system that gives OSN users direct control over the messages on their walls. This is accomplished through an adaptable rule-based system that allows customers to personalize the filtering criteria used on their walls, as well as a Machine Learning-based soft classifier that labels messages automatically in support of material filtering. Deep learning (DL) text

categorization techniques are used to assign appropriate each short text message to one of several classifications based on its content. The majority of the work in developing a rigorous back propagation algorithm is focused on the extraction as well as selection of a set of distinguishing and characterising features. A dataset of the categorised words is created and used to verify the words for inappropriate words. If the message contains any vulgar words, they will be added to the Blacklists to be removed from the message. Finally, as a result of the material technique, the message will be posted on the user's wall without the indecent words. A system filters inappropriate messages using blacklists based on message content as well as text creator friendships and characteristics. The proposed framework comprises deep learning algorithms for filtering rules to fit better the considered domain, to assist users in the specification of Filtering Rules (FRs), and to expand this same set of features deemed in the classification process. The proposed framework as shown in fig 2.

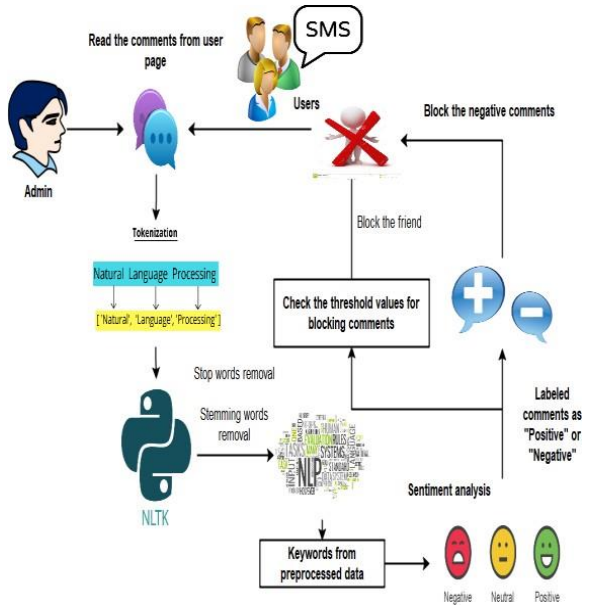


Fig 2: Proposed Framework

4.1 FRAMEWORK CONSTRUCTION

A social networking service (also known as a social networking site, SNS, or social media) is indeed an online platform that allows people to create social networking sites or social relationships with others who share similar personal or professional interests, activities, backgrounds, or real-life connections. The variety as well as evolving variety of hold and built-in social media platforms in the online environment presents a definitional challenge.

The term "social network" refers to human interaction in which people create, share, and/or transfer ideas and information in virtual networks and communities. Create a graphical user interface (GUI), which allows users to communicate with one another via integrated graphics and visual indicators. In this module, we can design the admin and user interfaces. The user can access the system as well as view the invitation. The images can be shared with friends by the user.

4.2 WORDS EXTRACTION

As social apps and websites proliferate, social networking is becoming an increasingly important part of everyday life online. Most conventional online media include different factors, such as user comment fields. Social networks is used in business to marketing, promote brands, connect with current customers, and foster new company. We can comment in an online social network using this module. Textual comments are welcome. The text can be unigrams, bigrams, or multigram. This component is used to collect feedback from social users. Comments can take many forms, including links, texts, and short texts. Comments are perused and forwarded to the server page.

TEXT MINING ALGORITHM

Most attempts in creating a powerful deep learning classifier are focused on the extraction and choice of a set of characterising and discriminating features. The text mining algorithm's steps are as follows:

- Tokenize text-based reviews as single terms
- Analyze unigrams, bigrams, and n-grams
- Remove stop words, analyses stemming words, and remove special characters
- Finally, extract key phrases
- Analyze extended words that can be substituted with right words

4.3 CLASSIFICATION

In this subsystem, we design a Filtered Wall (FW) automated system capable of filtering inappropriate messages from Online social network user walls. The architecture supporting OSN services is three-tiered. The first layer is commonly intended to provide basic OSN functionality (i.e., profile and relationship management). Furthermore, some OSNs include an additional layer that allows for the support of exterior Online Social Applications (SNA). Finally, the backed

SNA may necessitate the addition of an additional layer to accommodate the required graphical user interfaces. (GUIs). The majority of the work in developing a robust neural back propagation network (BPNN) is focused on extracting and selecting a set of characterising and discriminant features. The text classification is used to specify as well as enforce these constraints. From the perspective of BPNN, we reach the assignment by trying to define a hierarchical 2 strategy based on the assumption that it is preferable to recognize and remove "neutral" sentences before classifying "non-neutral" sentences even by class of interest rather than doing it all in one step.

DEEP LEARNING ALGORITHM

A database of categorised terms is formed here, and the words are then checked for any offensive words. If the user says any vulgar terms, it will be sent to the Watchlists, that will filter those phrases out. Finally, a text free of outrageous terms would be posted here on user's ceiling as a result of a content-based filtering technique. The recommended deep learning classifier is as follows:

Step 1: Initialize the neural network model

Step 2: Specify the layer type as input, hidden and output layer

Step 3: Activate the layers

Step 4: Specify the inputs and neurons

Step 5: Construct key terms as positive and negative

Step 6: Match with testing keywords

Step 7: Label as “positive” and “negative”

function INITBPNNMODEL (θ, [n1–5])

layerType = [input, hidden, output];

layerActivation = [smiliarity ()]

model = new Model();

for$i$=1 to 4 do

layer = new Layer();

layer.type = layerType[i];

layer.inputSize = ni

layer.neurons = new Neuron [ni+1];

layer.params = θi;

model.addLayer(layer);

end for

return model;

end function

A system employs registries to automatically reject unwanted messages based on relationship and characteristics of both the message content and the message author. The outgrowth of the acquisition of features evaluated inside the classification stage, a distinct definition for filtering rules to best fit the considered domain, to assist users with Filtering Rules (FRs) spec, and a distinct semantic and syntactic for screening rules to better fit the thought domain.

4.4 RULES IMPLEMENTATION

Users should be able to specify constraints on text creators through the filtering rules. Thus, creators to whom a filtration rule applies should be chosen based on a variety of criteria, one of which is impinging conditions on profile page attributes. In this way, we can define rules that apply only to young creators, creators with a specific religious/political viewpoint, or creators who we believe are not experts in a particular ground (e.g. by posing constraints on the work attribute of user profile). This entails filtering rules that identify messages based on restrictions on their contents. And block users who leave negative feedback more than five times, as well as send mobile notifications to users.

4.5 ALERT SYSTEM

BL's are managed directly by system, which ought to be capable of determining who is embedded in the BL as well as when consumer customer loyalty in the BL is complete. This knowledge is in the framework by a set of guidelines designed to improve flexibility; the rules on BL. The server generates rules for setting thresholds. We can block friends who leave negative comments based on threshold values. Finally, give users mobile notifications.

5. EXPERIMENTAL RESULTS

In this chapter, we can construct the social network using ASP.NET as front end and SQL SERVER as Back end. The performance of the system can be analyzed in terms of F-measure parameter.

The performance of the system is evaluated using Precision, Recall and F-measure.

$$\text{Precision} = \frac{TP}{TP+FP}$$

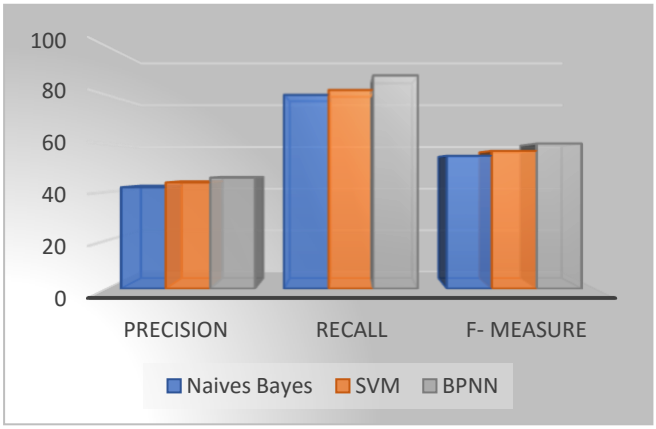
$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F measure} = 2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

The performance evaluation result is shown in following table 1 and shows in fig 3.

Algorithm/ Performance measures	Precision	Recall	F-measure
Naives Bayes	42	80	55
SVM	44	82	57
BPNN	46	88	60

Table 1: Performance Table



(a)

Fig 5: Performance chart

Based on the calculations above, the proposed neural networks algorithm has higher level F-measure value systems than existing Ensure that the required Bayes and SVM algorithms.

6. CONCLUSION

In this project, we proved a solution for filtering inappropriate messages from OSN walls. A DL soft classifier is used to impose a content-dependent filtration rules system that can be customised. The most time-consuming facets of developing a robust quick text classifier are the extraction but instead choice of a set of characterising but also discriminant features. Furthermore, handling BLs increases the system's flexibility in terms of filtration options. This is the first step inside a larger project. The promising early results of the classification technique inspire us to continue working on other projects that aim to improve classification quality. In this system, the DL soft classification model is used to filter out unwanted signals. BL is employed to increase the flexibility of the filtering system. We'll develop a more comprehensive mechanism for deciding when a user must be added to the BL. In addi - tion to that most, the scheme includes a powerful rule layer that constructs Filtering Rules (FRs) using a flexible language, allowing users to decide which data should not be presented on their walls. FRs can accommodate a variety of handling procedure that can be coupled and tailored to the user's needs. FRs specify the filtration criteria that will be used by leveraging user profiles, user connections, and the outcome of the DL classifier.

REFERENCES

[1] P. Fortuna and S. Nunes, ‘‘A survey on automatic detection of hate speech in text,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.

[2] Z. Al-Makhadmeh and A. Tolba, ‘‘Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach,’’ Computing, vol. 102, no. 2, pp. 501–522, Feb. 2020.

[3] R. Cao, R. K.-W. Lee, and T.-A. Hoang, ‘‘DeepHate: Hate speech detection via multi-faceted text representations,’’ in Proc. 12th ACM Conf. Web Sci., Southampton, U.K., Jul. 2020, pp. 11–20.

[4] Z. Waseem and D. Hovy, ‘‘Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,’’ in Proc. NAACL Student Res. Workshop, San Diego, CA, USA, Jun. 2016, pp. 88–93.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensive language,’’ in Proc. ICWSM, Montreal, QC, Canada, May 2017, pp. 15–18.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ‘‘Deep learning for hate speech detection in tweets,’’ in Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion), Perth, WA, Australia, Apr. 2017, pp. 759–760.

[7] M. O. Ibrohim and I. Budi, ‘‘Multi-label hate speech and abusive language detection in Indonesian Twitter,’’ in Proc. 3rd Workshop Abusive Lang. Online, Florence, Italy, Aug. 2019, pp. 46–57.

[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, ‘‘Hate speech detection in the Indonesian language: A dataset and preliminary study,’’ in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Jakarta, Indonesia, Oct. 2017, pp. 233–238

[9] M. O. Ibrohim and I. Budi, ‘‘A dataset and preliminaries study for abusive language detection in Indonesian social media,’’ Procedia Comput. Sci., vol. 135, pp. 222–229, Jan. 2018.
[10] J. Salminen, M. Hopf, S. A. Chowdhury, S.-G. Jung, H. Almerekhi, and B. J. Jansen, ‘‘Developing an online hate classifier for multiple social media platforms,’’ Hum.-centric Comput. Inf. Sci., vol. 10, no. 1, pp. 1–34, Dec. 2022