Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

DETECTING DISAPPROVAL SPEECH IN SOCIAL MEDIA WITH MACHINE LEARNING

S. VEERA SUDHAKAR[1], B.CHARISHMA[2]

[1]PG SCHOLAR (M.Tech., (CSE)), Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

[2]Associate Professor, HOD, Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

Abstract

In recent years, the increasing prevalence of hate speech in social media has been considered as a serious problem worldwide. Many governments and organizations have made significant investment in hate speech detection techniques, which have also attracted the attention of the scientific community. Although plenty of literature focusing on this issue is available, it remains difficult to assess the performances of each proposed method, as each has its own advantages and disadvantages. A general way to improve the overall results of classification by fusing the various classifiers results is a meaningful attempt. The proposed system has two level of classification, first level of classification is going to implement with fusion deep learning model using CNN-LSTM which is going to classify whether the message is normal or hate speech and second level of classification is going to implement with machine learning algorithms which will classify what type of hate speech.

Keywords— CNN, LSTM, SMN, SM, HATE SPEECH, CATEGORIZATION.

I. INTRODUCTION

Hate Speech is the used to harass, threaten, embarrass or target another person. With the advancement in the technology, the arena of social media has been prone to cyber crimes. Around 87% of today's youth have witnessed some form of hate speech. Hate speech can take different structures like sexual harassment hostile environment, revenge and retaliation. The offender is hidden to the victim, the problem statement gets complex. Hence Hate speech is an interesting field of research.

Social media networks (SMNs) are the fastest approach of communiqué as messages are sent and obtained nearly straight away. SMNs are the primary media for perpetrating hate speeches these days. In line with this, cyber-hate crime has grown signicantly in the previous couple of years. More research is being conducted to cut down on the rising cases of hate speeches in social media (SM). Different calls had been made to SM companies to later every comment before allowing it into the public domain . The impacts of hate crimes are already overwhelming due to sizable adoption of SM and the anonymity enjoyed via the online users . In this period of huge information, it is time- consuming and difcult to manually process and classify huge quantities of textual content records. Besides, the precision of the categorization of manual text can without difficulty be inuenced by human elements, inclusive of exhaustion and competence. To attain extra accurate and much less subjective results, it's miles benecial to apply machine learning (ML) techniques to automate the textual content classication procedures Social networks encourage the interactions between people to be more indirect and anonymous as a result presenting anonymity for some people making them feel more secure despite the fact that they express hate speech. It Can easily lead to disruptive anti-social outcomes if it remains unregulated and uncontrolled. Hate speech is therefore taken into consideration as a severe hassle internationally, and many countries and organizations resolutely resist it. The polarity detection of speech on structures is the rst step and is essential to government departments, social protection services, law enforcement and social media companies which expect to remove offensive content from their websites. Compared with guide ltering which is very time consuming, computerized identication of hate speech will enable the platform to hit upon the hate speech and cast off them a great deal greater quickly and efficiently. The problem of on-line hate speech detection has raised a hobby in each the scientic community and the business world. There had been many studies efforts aimed toward automating the technique which is usually modeled as a categorized classication problem. Recently, device getting to know technique that may study the extraordinary institutions between pieces of textual content, and that a specific output is anticipated for a selected input by using pre-categorized examples as schooling statistics is popular in systematic studies for hate speech detection. Among various device learning methods, deep learning which is a subset of machine getting to know, could be very prominent in classify the message for various categories. In recent times a potential and intense research awareness because of the hasty growth of social media which include blogs and social internetworking websites, in which individuals installed freely their perspectives on one of a kind topics. Researchers prove that people find it snug to opinionated and the messages are not categorized and we are unable to find out which type of hate speech it belongs and it is also not useful to the persons wheater it is private hate speech or public hate speech. However, not all statistics can be applicable; some may not have any impact on the end result and a few may have comparable meanings. A preprocessing phase is for this reason required to help make the dataset concise proposed system has two level of classification, first level of classification is going to implement with fusion deep learning model using CNN-LSTM which is going to classify whether the message is normal or hate speech and second level of classification is going to implement with machine learning algorithms which will classify what type of hate speech. The preprocessing manner consists of cleansing the statistics, tokenization, stop word elimination, etc. and detecting hate speech with categorizations.

II LITERATURE SURVEY

Theoretical underpinnings of online hate

Several concepts are commonly associated with the defnition of

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

online hate in the literature. As a phenomenon, online hate is cross-disciplinary; it has been studied using multiple theoretical lenses and conceptual frameworks, including social psychology,

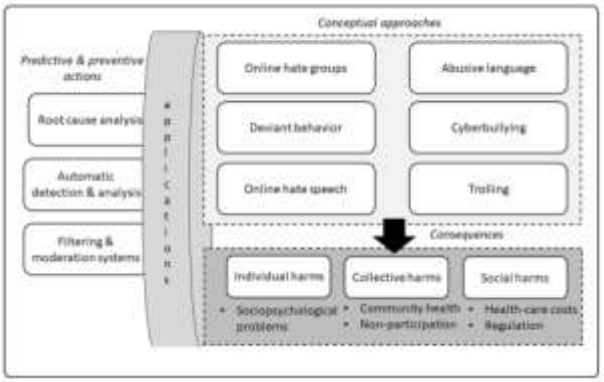

Fig 1: Conceptual Approach of Hatespeech

There are many approaches for detection of hate speech. But they differ from each other based on the output they obtained in Ref. 8 hate speech was classified into three classes race, nationality and religion. Ref. 8 uses sentiment analysis technique for detection of hate speech but just not detecting but they also classified into one of the three classes and also rate the polarity of speech. We found two survey papers for automatic hate speech detection [6],[14]. In Ref. 6 motivation for hate speech detection is shown and why it became necessary to develop more robust and accurate models for automatic hate speech detection. The problem of hate speech detection is more often researcher keep data private while collecting it and there are less open source code available which make it difficult for comparative study [6] . This degrades the progress in this field. Different features related to hate speech are described in Ref. 14, like simple surface feature which includes bag of words, unigrams or n-grams. Both training set and testing set need to have same predictive word but it is problem as detection of hate speech is applied on very small piece of text so to overcome this issue word generalization is applied [14]. Knowledge of annotator for hate speech was examined in Ref. 15. Authors produce some very good results in amateur annotation in comparison to expert annotations. Also, Waseem provide its own dataset and its evaluation. To penalize misclassification on minority classes weighted F1- score is suggested as an evaluation measure. Nowadays with development in deep learning, CNN can be used for hate speech detection [2],[1]. Word-vector also known as word embedding can be trained on relevant corpus of the domain. This pre trained word-vectors are used in CNN [2] . Most of machine learning models uses bag-of words which fails to capture patterns and sequences. It can be understood by the example in Ref. 2. if a tweet ends saying" if you know what I mean here each word can be considered as hate speech but it is most likely that this sentence is hate speech. This type of features cannot be handled by a bag of words which degrades the performance of traditional machine learning algorithms.

III. HATESPEECH OVERVIEW

Various authors in OHR cite the lack of commonly acknowledged definition for online hate [7, 16]. Instead of one shared defnition, the literature contains many definitions with distinct approaches to online hate

Evolution of online hate detection

Keyword based classifiers

In general, the evolution of online hate detection can be divided into three temporal stages: (1) simple lexicon or keyword-based classifers, (2) classifers using distributed semantics, and (3) deep learning classifers with advanced linguistic features. An example of the frst wave of studies is that used a list of profane words, being able to identify 40% of words that are profane and then correctly identifying 52% as hateful/not hateful.

Datasets Overview

We applied three criteria to select the datasets for this research: (a) the language is English, (b) the dataset was available at the time of conducting the research, and (c) the dataset and available details on the annotation procedure passed a manual evaluation (e.g., there was no high prevalence of false negatives/positives). Note that the previous research has found that online hate interpretation varies between individuals. For this reason, tends to apply aggregation methods such as majority vote, mean score, or consensus to determine if a comment is perceived as hateful or not. Tis precondition of “hateful on average” applies to all classifiers developed using this data. In the following sections, we briefly explain each dataset and how they were merged into one online hate dataset. Note that different authors use different terminology when referring to hateful online comments (e.g., “toxic”, “hateful”, “abusive”). Tese terms may have some nuanced conceptual differences, but for this study, the definitions provided by the authors of the chosen datasets are aligned with our operational definition presented in “Introduction” section. In this research, we refer to all these comments as hateful comments. When explaining the datasets, we will use the original authors’ terms and then explain how their terms overlap with ours. The case of hate speech and violent communication conducted over the internet can be referred as cyber-hate [5]. It is a narrow and specific form of cyber-bullying and it can be defined as “any use of electronic communications technology to spread racist, religious, extremist or terrorist messages” it is different from cyber-bullying in that hate speech can target not only individuals but it also has implications on whole communities [1]. Brown [6] has also defined hate speech as any textual or verbal practice that implicates issues of discrimination or violence against people in regard to their race, ethnicity, nationality, religion, sexual orientation and gender identity. According to Anis [7] hate speech can occur in different linguistic styles and several acts like insulting, provocation, abusing and aggression. However, according to Chetty and Alathur [8], hate speech can be categorized into the following categories:

Gendered hate speech

This category includes Any form of hostility towards particular gender or any devaluation based on person’s gender. This include any post that offense particular gender. Also it includes any form of misogyny. Moreover, clarify that sexism may come in two forms: Hostile (which is an explicit negative attitude) and Benevolent (which is more subtle).

Religious hate speech

This will include any kind of religious discrimination, such as: Islamic sects, calling for atheism, Anti-Christian and their respective denominations or anti-Hinduism and other religions. However, mentioned that religious hate speech is considered as a motive of crimes in countries with highest social crimes.

Racist hate speech

Lastly, this category includes is Any sort of racial offense or tribalism, regionalism, xenophobia (especially for migrant workers) and nativism (hostility against immigrants and refugees) and any prejudice against particular tribe or region. For instance, offending an

individual because he belongs to a particular tribe or region or country or favoritism of a particular tribe. Add to that, offending the appearance and color of individual.

Facebook dataset

To detect toxicity triggers (i.e., causes) of online discussions in facebook, the study developed a model that detects the toxicity in the comments posted on facebook communities (also denoted as facebook). Te dataset consists of relevance judgments specifying if a particular comment is hateful or not. Note the term “toxicity” as synonymous to “hateful”. Tey selected for crowdsource labeling a random sample of 10,100 comments from facebook (one of the largest facebook communities), which were obtained the API. Te designed labeling job asked workers to label a given comment as either toxic or non-toxic according to the toxicity definition provided by the Perspective API, which describes a toxic comment as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.12 Te labeling results showed that 81.57% of the comments in the collection were labeled as non-toxic, while the remaining 18.43% were labeled toxic. Te observed agreement between annotators was 0.85.

MACHINE LEARNING

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model Based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system.

HATE SPEECH

In antique instances, Hate Speech changed into restricted face to face conversations. But now because of the boom in social media systems the usage of hate speech is increasing. As human beings feel they're hidden on the net. Due to this, people feel secure to apply hate speech and it's human computive undertaking to identify hate speech on social media so we need some automatic strategies to come across hate speech. On the other hand, people are more likely to share their views online, thereby leading to the dissemination of hate speech. Given that this type of prejudiced contact can be particularly un favorable to society, policymakers and social networking sites may also profit from monitoring and prevention gadgets Hate speech is typically described as any touch that distorts a character or network on the basis of traits such as coloration, ethnicity, gender, sexual preference, nationality or faith. According to Paula Fortuna and Sergia Nunes Hate speech is language that assaults or diminishes, that incites violence or hate towards companies, primarily based on precise traits consisting of physical look, faith, descent, national or ethnic starting place, sexual orientation, gender identity or other. This sort of capabilities can't be treated with the aid of a bag of words which degrades the overall performance of traditional system learning algorithms.

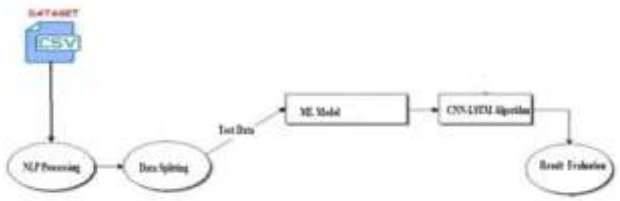

Fig 2: System Architecture

The model consists of following modules for detecting hate speech:

Data Acquisition: Extraction and importing of data.

Data Pre-processing: Cleaning of data and extraction of features.

Feature Extraction: Vectorizing the text.

Detection of hate speech: Detecting Hate Speech.

Output: Hate or Not Hate.

DATA ACQUISITION The process of collecting the data is called as data acquisition. The dataset used is Twitter data set obtained from Kaggle. It has two columns and 10490 rows. One column lists the tweets and the other column specifies whether the tweet is hate or not hate.

DATA PRE-PROCESSING This step involves cleaning our dataset by removing unnecessary parts of data that would have no role in the prediction task. Cleaning and Organizing of Raw Data to make it suitable for machine learning model We have performed the following stages of data preprocessing:

Tokenization: break the text present in the tweet into single words

Remove stop words: removing the common words like this, that etc. Eliminate punctuation marks: remove \, <,>, /, # etc.

FEATURE EXTRACTION In the present model, the feature is universal sentence encoder extracted using LSTM model. It provides better performance and efficiency. It can be applied to combination of sentences and paragraphs. It encodes the text present in the tweet into 512 high - dimensional vectors. These can be used for further classification.

HATE SPEECH DETECTION The aim is to get a classifier which best classifies the tweets into hate or not hate class using genetic programming approach. The module gets the extracted features from the previous module to perform further process. The operators are used to reduce complexity. The genetic programming approach selects the machine learning algorithm with best accuracy for the classification. The performance is evaluated for different dataset. LSTM model is used for result prediction

In the preprocessing step, the text is cleaned. Firstly, the emoticons are recognized and replaced by corresponding words that express the sentiment they convey. Also, all links and urls are removed. Afterwards, text is morphologically analyzed. In this way, for each resulting token, it is assigned. Then, the texts are represented as vectors with a word embedding model. We used pre trained word vectors in and proposed a model that consists in a LSTM at the word level as Figure 1 shows. At each time step t the LSTM gets as input a word vector with syntactic and semantic information, known as word embedding Afterward, an attention layer is applied over each hidden state. Finally, the presence of hate (or not) in a text is predicted by this final LSTM technique.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

IV. EXISTING ANALYSIS

With the advancement in technology, the internet has been a safe and secure sphere of communication, though the arena of social media has been prone to cyber crimes such as spamming, trolling and hatespeech. Although strict laws exist to punish hatespeech, there are very less tools available to effectively combat hatespeech. Detection of Hatespeech using Machine Learning Techniques is our problem statement.

IV. PROPOSED WORK

Proposed model uses based on deep learning and machine learning concepts. It uses Natural Language Processing (NLP) Techniques for pre-processing. First level classification is to classify normal or hate speech which is going to use deep learning fusion algorithms CNN-LSTM Second level classification is to categories Hate Speech which is going to use Machine Learning Algorithms Random Forest & SVM

Advantage of Proposed System

•The proposed system is aim to achieve more than 85% accuracy in classification results.

•The proposed system is important for below reasons.

oSocial media's ubiquity means that hate speech can effectively impact any one at any time or anywhere, and the relative anonymity of the internet makes such personal attacks more difficult to stop than traditional bullying.

oThe COVID-19 pandemic notably makes hate speech an increasingly worrying threat.

oOn April 15th, 2020, UNICEF issued a warning in response to the increased risk of hate speech during the COVID-19 pandemic due to wide spread school closures, increased screen time, and decreased face-to-face social interaction.

We propose to analyze various deep learning models and compare their performance metrics with a baseline model that combines the vectorization method with an LSTM classifier. Feature engineering is required with shallow models before fitting the data in the model, unlike deep learning models. However, in deep learning, a collection of nonlinear transformations can be included in the model where features are directly mapped to the outputs. Due to this added advantage, employing LSTM as well as CNNs based architectures prove to be beneficial. We will apply LSTM with a single dimension as well as multiple convolutions with filters of varied lengths and max-pooling layers. The Long Short Term Memory (LSTM) model is model which is commonly used which we will be using too. These trained modes will be used for evaluation on unseen data. These architectures will be evaluated alongside three resampling techniques—Random oversampling, The main metric that will be used for evaluation is Recall, instead of Precision, Accuracy or F1-score since Recall would give a measure of how correctly the model is identifying true positives (i.e., hate-speech). However, the model will be able to predict the minority class with higher accuracy, thus making it a more efficient method to correctly identify tweets containing hate speech. Model and Material which are used is presented in this section. Table and model should be in prescribed format.

V. EXPERIMENTAL RESULTS

Dataset

This system uses two data set first one to classify

1. Normal
2. Hate Speech

The second dataset is used to identify what type of Hate Speech, Categories are given below:

- Age;
- Ethnicity;
- Gender;
- Religion;
- Other type of cyberbullying;
- Not cyberbullying

Identical preprocessing steps have been applied in this paper for all models. We trained neural networks on the training set in order to get the highest recall value from the test data and reported the same. The final evaluation metric is recall since the objective is to accurately identify true positives (or hate-speech). We observe that the LSTM model have the highest recall values with no significant differences. The model has a lower recall value in general. However, among the resampling methods, The highest recall value when used along with LSTM. The following table summarizes the results obtained:

Model	TP	FN	FP	TN	F1-Score	Precision	Recall	Accuracy
Dense	7388	325	77	3908	0.91	0.89	0.94	0.91
CNN	7429	284	122	3863	0.92	0.9	0.95	0.92
M_CNN	7446	267	115	3870	0.93	0.91	0.94	0.93
LSTM	7476	237	76	3909	0.95	0.94	0.95	0.95

Table 1: Cumulative Metric Results Of The Models.

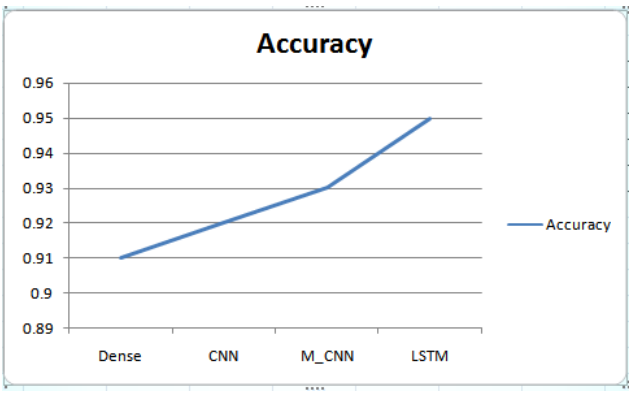

Figure 3: Accuracy Charts of various Models

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

Figure 4: Loaded Dataset for detecting hatespeech

Figure 5: Trained Dataset

Figure 6: Hate speech Detecting

Figure 7: Hate speech Detecting categorization wise

Figure 8: Accuracy for LSTM

CONCLUSION

Online hate detection is needed to reduce toxicity in social media platforms. In this research, we experimented with various machine learning models (CNN-LSTM) for online hate detection and found the best performance with CNN-LSTM as a classifier for categorizing the hate speech as the most impactful representation of hateful social media comments. Te generalizability of the model to multiple social media platforms is good but varies slightly between the platforms. Our findings support the goal of developing more universal online hate classifiers for multiple social media platforms. The model that we make publicly available can be deployed to practical applications as well as be further developed by online hate researchers. Text classification is the technique that has been primarily used to detect depression among social media users. It takes input text data which is depicted as a vector by the pre-training model. This vector is fed into multiple types of architectures for training and finally, the performance of these models is confirmed. For text classification existing models have proved useful, but there is still a long way to go and many domains remain unexplored. Moreover, different combinations of these models can be

studied with new machine learning algorithms. Numerous other methods that aim at fulfilling the lack of training data in supervised learning tasks are also worthy of being explored. Methods like transfer learning could also be beneficial because they examine the issue of adapting supervised models trained in a resource-rich context to a resource-scarce context. For instance, identifying whether features discovered from one particular hate class can be shifted to another, hence enhancing the training of each other, is a use case that can be investigated.

REFERENCE

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

[2]. Md Abul Bashar and Richi Nayak. Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019), 2019.

[3]. Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2):223–242, 2015.

[4]. Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pages 86–95, 2017.

[5]. Shimaa M Abd El-Salam, Mohamed M Ezz, Somaya Hashem, Wafaa Elakel, Rabab Salama, Hesham ElMakhzangy, and Mahmoud ElHefnawi. Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. Informatics in Medicine Unlocked, 17:100267, 2019.

[6]. Shimaa M Abd El-Salam, Mohamed M Ezz, Somaya Hashem, Wafaa Elakel, Rabab Salama, Hesham ElMakhzangy, and Mahmoud ElHefnawi. Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. Informatics in Medicine Unlocked, 17:100267, 2019.

[7]. Paula Fortuna and S´ergio Nunes. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4):1–30, 2018.

[8]. Purnama Sari Br Ginting, Budhi Irawan, and Casi Setianingsih. Hate speech detection on twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), pages 105–111. IEEE, 2019. [9]. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4):215–230, 2015.

[10]. Ammar Ismael Kadhim. Term weighting for feature extraction on twitter: A comparison between bm25 and tf-idf. In 2019 International Conference on Advanced Science and Engineering (ICOASE), pages 124–128. IEEE, 2019. [11]. Harpreet Kaur, Veenu Mangat, and Nidhi Krail. Dictionary-based sentiment analysis of hinglish text and comparison with machine learning algorithms. International Journal of Metadata, Semantics and Ontologies, 12(2- 3):90–102, 2017.

[12]. Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences, 10(1):1, 2020.

[13]. TYSS Santosh and KVS Aravind. Hate speech detection in hindienglish code-mixed social media text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pages 310–313, 2019.

[14]. Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, 2017. [15]. Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science, pages 138142, 2016.

[16]. Tingxi Wen and Zhongnan Zhang. Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic eeg multiclassification. Medicine, 96(19), 2017 [17]. Abro, S., Sarang Shaikh, Z. A., Khan, S., Mujtaba, G., & Khand, Z. H. Automatic Hate Speech Detection using Machine Learning: A Comparative Study. Machine Learning, 10, 6,2020.