An Analysis of COVID-19 related Twitter Data for Asian Hate Speech Using Machine Learning Algorithms

Sandeep Shah
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, United States
sshah@aggies.ncat.edu

Xiaohong Yuan
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, United States
xhyuan@ncat.edu

Zanetta Tyler
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, United States
zrtyler@aggies.ncat.edu

2022 1st International Conference on AI in Cybersecurity (ICAIC) | 978-1-6654-0043-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIC53980.2022.9896967

***Abstract*—Many people have suffered in this global pandemic of Covid-19 since March 2020. Some suffered from covid-19 disease whereas some suffered from hatred. There are posts and comments on Twitter that reflected Asian hatred blaming them for the coronavirus. Many people expressed their anger and racism through social media and even physically. It is an important task to understand the sentiments of public data to reduce hatred and racism through social media. Many machine learning algorithms can be used to classify Twitter data. In this research, we use two different machine learning classifiers to classify COVID-related Twitter data: Support Vector Machine (SVM) and Random Forest. We compare the performance of both learning methods according to their prediction accuracy, precision, recall, and F1-score. We then predict the tweet label using the data for each month from April to November 2020 using the SVM model. The label used in the dataset are Hate, Counter-hate and Neutral. We analyze the ratio of hate and counter-hate tweets and discuss their possible correlation with events that occurred in those months.**

***Keywords*—Support Vector Machine (SVM), Random Forest, Hate Speech, Sentiment Analysis, Text Classification, social network analysis, Twitter Data.**

I. Introduction

Cyberhate, cyberbullying, and cyberthreat are very common today as technology advances and the use of social media and other communication software increases. People tend to express good and bad feelings through social media. Researchers have become more concerned about the social behavior of people around the world. Although there are strict policies against hate speech through social media, it is hard to stop everyone from communicating disrespectful thoughts. In this research, we use Twitter data to analyze hate-related comments during the Covid-19 pandemic.

In this research, we use Support Vector Machine (SVM) and Random Forest classification algorithms to classify the tweeter data according to the hate-related text found in the tweet. We compare the classification performance of the two algorithms using metrics of accuracy, precision, recall, and F1-score. First, we preprocess the data to find the important features of the data to be used to classify the text. The dataset we used in this experiment is a collection of twitter data during the Covid-19 global pandemic and is publicly available [1]. The dataset contains tweets from March 2020 to Dec 2020. Since the data only contains the tweeted text and the date, we use a vectorizer to convert the text data to numerical values. We use Tokenizer to divide the texts into words or smaller sub-texts, which will enable the generalization of the relationship between the texts and the labels.

The main goal of this research is to use a classification model to classify hate-related tweeter data and analyze the trend throughout the period of 8 months from April to November. We use SVM since it has better classification performance than Random Forest. Since the Twitter data is text-based, it contains lots of words and phrases which are not needed for this study, and it should be removed before applying the algorithms for better performance result. Some of the examples includes emoji's characters, URLs, some other encoded characters and stop words such as: 'and', 'is', 'a', 'an', 'the' and so on. Those words are removed during the preprocessing of the data. The remaining words such as slang, racist, hatred, and bullying-related words are used for this research. In this research, we classify such words to detect hatred using the tweets during the Covid-19 pandemic. The dataset has labeled data, especially toward the Asian hate. The labels used to classify the data are Hate (refers to Asian Hate), Counter-Hate (anti-racism), and Neutral. The labeled data are used for training in this research. We analyze the ratio of hate and counter-hate tweets and discuss their possible correlation with events that occurred in those month.

The rest of the paper is organized as follows. Section II discusses the related work in text data classification and Twitter data processing. Section III describes the background for this experiment that includes the description of the algorithms we use and the dataset. Section IV describes the experiments conducted. Section V presents and discusses the results. section VI concludes the paper.

II. Related Works

Ziems et al. [1] has created a dataset related to hate and counter-hate and Covid-19 pandemic using the Tweeter data between January 2020 to March 2021. In this paper, the experiment is performed using BERT function to pave the path towards the use of public counter-hate messaging campaigns as a potential solution against hate speech on social media. This dataset was used in our research. Saha et al. [2] has introduced a machine learning model named "HateMonitor", which is used to identify hate speech and offensive content. In this research, the authors have used SVM and Gradient Boosting machine learning algorithms to classify the tweets separately. They have used three sets of dataset which includes Hindi, German and English language dataset. Gupta et al. [3] has studied the use of Bayesian Logistic Regression, Naive Bayes, SVM, and Artificial Neural Network to perform sentiment analysis of Twitter data. In their experiment they have used a publicly available dataset containing 16K tweet data which contains tweets containing political discussion. Georgios et al. [4] has used Recurrent Neural Network (RNN) classifier to distinguish between racism and sexism messages using twitter data. They have performed multiclass classification with three categories racism, sexism, and neutral. ELSherief et al. [5] has performed analysis on difference between directed and generalized hate speech. In this study, the authors has chosen two publicly available dataset on directed and generalized hate speech. The authors has performed two different analyses on the dataset: Lexical analysis and Semantic analysis. Badjatiya et al, performs an experiment using deep learning method for hate speech detection in tweets. They also have mentioned the used of TF-IDF vector and bag of words vector along with other vectors and stated the results of each vectorizer [10].

III. Background

This section introduces random forest and Support Vector Machine Learning algorithms, as well as the dataset used in this research.

A. Random Forest

Random Forest [7] is a machine learning algorithm used for classification and regression. It uses a set of decision trees to create a forest. A decision tree is simply a branched tree used to classify whether the given sample is true or false. The larger the size of data (i.e. number of rows), the higher the number of decision trees. While classifying the data, each of the data passes through the multiple decision trees and generates the classification with a high accuracy rate. Since the dataset we are using has a large number of text data, and Random Forest has been frequently used in classification with good performance, we assume random forest will be a suitable algorithm to predict the hate related texts.

B. Support Vector Machine

Support Vector Machine is also a machine learning technique that is used for the classification and regression of data. Support vector machine creates an imaginary hyperplane to separate the data depending upon their features which can be single dimensional or multidimensional depending upon the number of classes" [7]. SVM uses overfitting protection which does not depend on the number of features so it can handle large feature space. Since SVM avoids overfitting, we use SVM to classify the Twitter dataset which has many features. The text data is vectorized to numerical data in order to use SVM.

C. Dataset

The data was created by Ziems et al. [1]. It is the largest dataset of anti-Asian hate and counter-speech on Twitter which was created during the Covid-19 pandemic. This dataset contains 206 million tweets within the date range of January 2020 to March 2021. The authors have manually labeled certain data and classified the data as hate or counter-hate according to the nature of the tweets. 2291 tweets are manually annotated in three categories: Hate, Counter-hate, and Neutral tweets. In this dataset, Hate refers to Asian hate. The dataset contains a total of 206,348,116 tweets, after doing prediction using BERT function they retrieved 0.64% of the tweets as Hateful tweets, 0.55% of the tweets as counter-speech tweets, and the rest of the tweets as neutral tweets.

IV. Experiment

To perform the experiment, we conducted the following tasks: data preprocessing, feature extraction, training the classifier, and testing the model to check the performance of each model.

A. Data Preprocessing

After each tweet is hydrated it must be preprocessed. Here tweet hydration means getting complete details related to the tweets which also means getting all the fields of the tweets. Since we are using a dataset, which is already prepared, we do not have to use any hydration in this experiment. Through preprocessing each tweet is stripped to their most basic form before it is labeled. This is done by removing all odd characters, URLs, retweet syntax, extra whitespace, usernames, and stop words. This is followed by finding the lemmatized version of each word. This refers to the removal of endings such as 'ing', 'er', 's', etc. if it is just an additional ending. These steps are important in order to get a high accuracy classification . Stop words are words that are commonly used in a language such as 'are', 'is', 'a', 'an', 'the', and so on. Since stop words are often grammatical references, it won't affect the classification results. The use of vectorizer and tokenizer is also considered in this experiment to increase the accuracy which is further explained in the next sub-section. Table I shows an example of tweet before and after the prepossessing.

TABLE I
INSTANCE OF TWEET BEFORE AND AFTER PREPOSSESSING

Text Data	@OKAMIDUAN @Chamanpandey5 @realDonaldTrump It took you so long to https://t.co/lxQ7UYfeWk Were you learning english Mr chinese virus?China is endangering lives and u have to admit it.fucking your mother is not the only option I got.That too will result me suffering from the Chinese virus.
Data after removing URLs	@OKAMIDUAN @Chamanpandey5 @realDonaldTrump It took you so long to Were you learning english Mr chinese virus?China is endangering lives and u have to admit it.fucking your mother is not the only option I got.That too will result me suffering from the Chinese virus.
Data after removing Punctuation	OKAMIDUAN Chamanpandey5 realDonaldTrump It took you so long to Were you learning english Mr chinese virusChina is endangering lives and u have to admit itfucking your mother is not the only option I gotThat too will result me suffering from the Chinese virus
Data after removing stop words	OKAMIDUAN Chamanpandey5 realDonaldTrump long learning english chinese virusChina endangering lives admit itfucking mother option I gotThat result suffering Chinese virus

B. Feature Extraction

We use two vectorizers to extract the features of the tweets. Since machine learning models can only process numerical data, a vectorizer is used to convert text data into vectors. We use both TF-IDF vectorizer and count vectorizer to extract the features of the data. TF-IDF stands for "Term Frequency – Inverse Document Frequency". A TF-IDF vectorizer is used to learn the frequency of each word and return a score of each word present in the dataset. We also use a count vectorizer which simply counts the words creating a bag of words. It returns the integer counter value of each word present in the dataset.

For example, if we use a text as: "We are students at NCAT, and we are also Engineering Students",

the count vector returns the count of each word such as:

• Students: 2,

• Engineering: 1, and so on.

While TF-IDF vectorizer will return the frequency of occurrence of each word such as:

• Students: 0.485,

• Engineering: 0.242 and so on.

The TF-IDF is calculated multiplying following terms, *i.e. TF * IDF*.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = loge (Total number of documents / Number of documents with term t in it).

In this experiment, we use both vectorizers and compare the results to show which one works better in our tweet classification.

C. Classification and prediction of tweets.

After preprocessing and feature extraction is done, we then select the preprocessed data to classify hate-related data using SVM and Random Forest. For both classifications, we use both vectorizers to compare the accuracy, precision, recall, and F1-score. Since all the classifiers works differently, we need to use extra steps such as cross-validation. Cross-validation is used while performing Support vector machine and Random Forest. The performance metrics are defined below:

- **Accuracy** is the percentage of correctly classified records over the total number of records.
- **Precision** is the ratio of records correctly classified as hate (or counter-hate or neutral) over the total number of records classified as hate (or counter-hate or neutral).
- **Recall** is the ratio of records correctly classified as hate (or counter-hate or neutral) over the total number of actual hate (or counter-hate or neutral).
- **F1-score** is the harmonic mean (in percentile) of precision and recall.

V. RESULTS

We used 80% of the whole labeled dataset to train the classification algorithms, and used 20% of labeled dataset to predict and evaluate the accuracy of the models. We use scikit-learn splitting function to split the dataset into training and test set.

```
[[165  43   4]
 [ 72 113   1]
 [ 41  23   2]]
              precision    recall  f1-score   support

           0       0.59      0.78      0.67       212
           1       0.63      0.61      0.62       186
           2       0.29      0.03      0.05        66

    accuracy                           0.60       464
   macro avg       0.50      0.47      0.45       464
weighted avg       0.56      0.60      0.56       464
```

Fig. 1. Confusion Matrix and Performance Metrics using Random Forest using TF-IDF Vector.

```
[[152  60   0]
 [ 65 121   0]
 [ 39  27   0]]
              precision    recall  f1-score   support

           0       0.59      0.72      0.65       212
           1       0.58      0.65      0.61       186
           2       0.00      0.00      0.00        66

    accuracy                           0.59       464
   macro avg       0.39      0.46      0.42       464
weighted avg       0.50      0.59      0.54       464
```

Fig. 2. Confusion Matrix and Performance Metrics using Support Vector Machine using TF-IDF Vector.

```
[[167  43   2]
 [ 54 120  12]
 [ 26  24  16]]
              precision    recall  f1-score   support

           0       0.68      0.79      0.73       212
           1       0.64      0.65      0.64       186
           2       0.53      0.24      0.33        66

    accuracy                           0.65       464
   macro avg       0.62      0.56      0.57       464
weighted avg       0.64      0.65      0.64       464
```

Fig. 3. Confusion Matrix and Performance Metrics using Random Forest using Count Vector.

First, we used TF-IDF in both of the classifiers. Fig. 1 and FIg. 2 shows the results of Random Forest and SVM applying TF-IDF vectorizer respectively. Similarly in the next phase we have used count vectorizer on both Classifiers. Fig. 3 shows the confusion matrix and the prediction performance metrics using Random Forest classifier and Fig. 4 shows the confusion matrix and prediction performance metrics using SVM. Here SVM produces better results than Random Forest in predicting hate, counter-hate and neutral tweets. Seeing all the results from the figures, we found that count vector produces better accuracy over TF-IDF vectorizer.In the figures, ”0” represents the ”Hate” class, ”1” represents the ”Neutral” class, and ”2” represents the ”Counter-hate” class.

```
[[165  35  12]
 [ 46 124  16]
 [ 13  19  34]]
              precision    recall  f1-score   support

           0       0.74      0.78      0.76       212
           1       0.70      0.67      0.68       186
           2       0.55      0.52      0.53        66

    accuracy                           0.70       464
   macro avg       0.66      0.65      0.66       464
weighted avg       0.69      0.70      0.69       464
```

Fig. 4. Confusion Matrix and Performance Metrics using Support Vector Machine using Count Vector.

Fig. 5 shows the graph with the total percentage of hate tweets and counter-hate tweets for each month from April to November 2020. We can see the trend of tweets for each month and can analyze that people continued to post hate speech and counter-hate speech throughout the year. We can also see the trend that the counter-hate speech increases with the increase of hate speech.

TABLE II
ATTACK CATEGORIES AND ITS SUB-CATEGORIES

Months	**Classification**	**Values**
April	Hate	20114
	Neutral	121281
	Counter-Hate	1355
	Hate Percentage	14.0 %
	Counter-Hate Percentage	0.95 %
May	Hate	36674
	Neutral	345030
	Counter-Hate	3929
	Hate Percentage	9.51 %
	Counter-Hate Percentage	1.02 %
June	Hate	97140
	Neutral	942476
	Counter-Hate	8959
	Hate Percentage	9.26 %
	Counter-Hate Percentage	0.853 %
July	Hate	101353
	Neutral	932190
	Counter-Hate	15032
	Hate Percentage	9.66 %
	Counter-Hate Percentage	0.333 %
August	Hate	66155
	Neutral	927867
	Counter-Hate	8447
	Hate Percentage	9.31 %
	Counter-Hate Percentage	0.818%
September	Hate	77654
	Neutral	809953
	Counter-Hate	16341
	Hate Percentage	11.55 %
	Counter-Hate Percentage	1.194%
October	Hate	94988
	Neutral	700756
	Counter-Hate	9375
	Hate Percentage	13.63 %
	Counter-Hate Percentage	1.96 %
November	Hate	101789
	Neutral	934651
	Counter-Hate	12135
	Hate Percentage	16.71%
	Counter-Hate Percentage	2.157 %

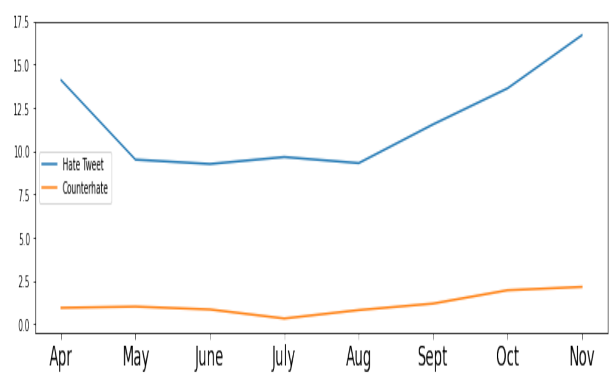

Fig. 5. Percentage of Hate and Counter-hate from April to November, 2020.

Table II. shows the number of tweets classified as hate, counter-hate, and Neutral. It also shows the calculated hate and counter-hate percentage for each months from April to

November. We can see that April, October, and November has slightly higher percentage of hate and counter-hate tweets than rest of the months.

Fig. 6 shows the number of tweets related to hate and counter-hate from April to November 2020. For each month we have selected different number of tweets and the results shows that the hate speech and counter-hate speech have increased with the increase of total number of tweets. Due to the different number of tweets each month, we have calculated the percentage of both hate and counter-hate as shown above in Fig. 5. We can also see that June, July, October, and November has the higher number of hate and counter-hate tweets.

We wonder whether certain events might have triggered more hate-related tweets in June, July, October and November. We have found several events during the months where the hate speech is greater. We have also retrieved covid-19 daily cases report for the months mentioned above globally. According to the New York Times coronavirus tracking [8], we have found the following data as shown in Table III and Fig. 7.

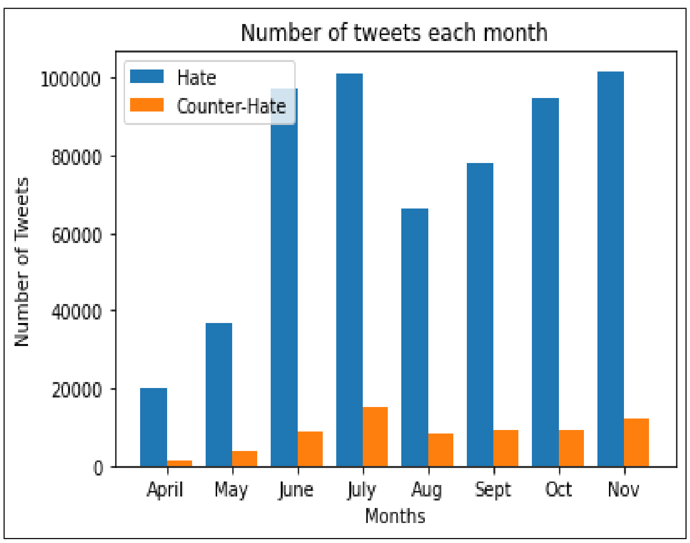

Fig. 6. Number of Hate and Counter-Hate Tweets from April to November 2020.

Table III shows the new coronavirus cases on the first and last day of each month.[8] It also has the calculated percentage change in new cases of each month. Fig. 7. particularly shows the graph of percentage change in new cases each month from April to November. Analyzing the graph, we can see that the number of cases is increasing rapidly in the month of June, July, September, October, and November. Referring to the predicted results in Table III, we may assume that the change in coronavirus cases may be one of the triggers of increase in hate speech around the world.

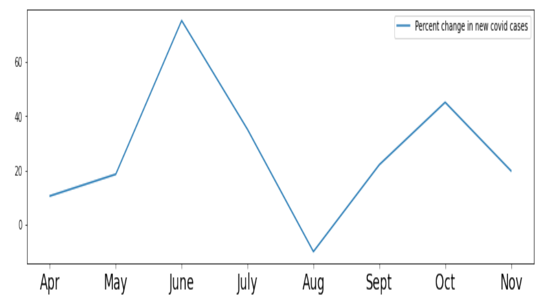

Fig. 7. Percentage(%) Change in new cases each month[8].

TABLE III
NUMBER OF NEW CASES EACH DAY FROM BEGINNING OF APRIL TO END OF NOVEMBER[8].

Months	Num of new cases worldwide		Percent Change
	First of the month	End of the Month	
April	77,475	85,603	10.49
May	88,975	105,462	18.52
June	104,266	182,580	75.10
July	216605	292304	34.95
August	245235	218466	-10.91
September	269070	328199	21.97
October	319287	465071	45.66
November	432396	517779	19.74

We also compare this prediction with some major events around world as shown in Table IV. Some of these events are major cause of spike in coronavirus cases, due to which people expressed their thoughts through social media. We can assume that the rise in hate speech in social media may be related to these events around the world.

VI. Conclusion

In this research, we have observed that the SVM classifier performs better in classifying hate-related tweets. We also observed that count vectorizer works better than Tf-idf vectorizer for classifying tweets. After predicting the hate-related tweets using SVM model for the Covid-19 related dataset collected from April to November 2020, we have found that the hate and counter-hate speech during the month of June, July, October, and November are higher than the other four months. We analyzed that there is some correlation between the number of coronavirus cases and events that caused high number of coronavirus cases and hate-related speech. It is predictable that the number of cases can effect the hate speech in social media. After analysis all the possible scenarios, we can assume that hate speech in social media will keep on increasing and it should be restricted to use offensive language in social media. It is recommended that classifiers with good performance be used to predict the level of hate in the text data. We will also be using different classifier in future to compare their results.

TABLE IV
NUMBER OF NEW CASES EACH DAY FROM BEGINNING OF APRIL TO END OF NOVEMBER [9].

Date	Events
June	• US protest for Black Life Matters • India had 6th and Italy had 5th highest number of cases • Yemen death rate doubled as of May • Tension between India and China rose after violent clashes in Himalayas. • 20 Indian soldiers died, and some Chinese soldiers also died (number unrevealed).
July	• Northern Spain announced lockdown after spike in covid-19 cases. • Australia announced isolation in Victoria. • Protest erupted in Belgrade Serbia due to lockdown in the country.
October	• Second lockdown in Paris • Belgium covid cases rises twice as of August. • Second Lockdown in Wales.

ACKNOWLEDGMENTS

This work is partially supported by National Science Foundation (NSF) under the grant CNS-1900187. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

REFERENCES

[1] He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., and Kumar, S. (2021, November). Racism is a virus: anti-asian hate and counter-speech in social media during the COVID-19 crisis. In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 90-94).

[2] Saha, P., Mathew, B., Goyal, P., and Mukherjee, A. (2019). Hatemonitors: Language agnostic abuse detection in social media. arXiv preprint arXiv:1909.12642.

[3] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., and Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 29-34.

[4] Dimitrakopoulos, G. N., Vrahatis, A. G., Plagianakos, V., and Sgarbas, K. (2018, July). Pathway analysis using XGBoost classification in Biomedical Data. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence (pp. 1-6).

[5] ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018, June). Hate lingo: A target-based linguistic analysis of hate speech in social media. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1).

[6] Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., and Galesic, M. (2022). Impact and dynamics of hate and counter-speech online. EPJ Data Science, 11(1), 3.

[7] Shah, S., Muhuri, P. S., Yuan, X., Roy, K., and Chatterjee, P. (2021, April). Implementing a network intrusion detection system using semi-supervised support vector machine and random forest. In Proceedings of the 2021 ACM Southeast Conference (pp. 180-184).

[8] Coronavirus World Map: Tracking the Global Outbreak Available: https://www.nytimes.com/interactive/2021/world/covid-cases.html

[9] https://www.infoplease.com/current-events/2021

[10] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).

[11] Garland, J., Zahedi, K., Young, J., Hébert-Dufresne, L., and Galesic, M. (2020). Impact and dynamics of hate and counter-speech online. ArXiv, abs/2009.08392.

[12] Vishwamitra, N., Hu, R. R., Luo, F., Cheng, L., Costello, M., and Yang, Y. (2020, December). On analyzing covid-19-related hate speech using bert attention. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 669-676). IEEE.

[13] Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., and Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. arXiv preprint arXiv:2006.01974.

[14] Bilbao, I., and Bilbao, J. (2017, December). Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks. In 2017 eighth international conference on intelligent computing and information systems (ICICIS) (pp. 173-177). IEEE.

[15] Manju, N., Harish, B. S., and Prajwal, V. (2019). Ensemble feature selection and classification of internet traffic using XGBoost classifier. International Journal of Computer Network and Information Security, 10(7), 37.

[16] Singh, G., Kumar, B., Gaur, L., and Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (pp. 593-596). IEEE.

[17] Liu, X., Dai, Y., Zhang, Y., Yuan, Q., and Zhao, L. (2017, May). A preprocessing method of AdaBoost for mislabeled data classification. In 2017 29th Chinese Control And Decision Conference (CCDC) (pp. 2738-2742). IEEE.

[18] Jadhav, S. R., Rokade, A. D., Sable, A. N., and Gade, V. B. (2021). PUBLIC HATE SPEECH DETECTION USING MACHINE LEARNING. INTERNATIONAL JOURNAL, 6(7).

[19] Bartlett, J., Norrie, R., Patel, S., Rumpel, R., Wibberley, S.: Misogyny on twitter. Demos (2014)

[20] Bhadoria, R. S., Sharma, N., Pandey, M. K., and Pandey, B. (2020). Asymptomatic Community Spread Of Coronavirus Disease 2019 (COVID-19) Outbreak Prediction Using Logistic Regression. European Journal of Molecular & Clinical Medicine, 7(06), 2020.

[21] Waseem, Z., and Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).