Improving Hate Speech Detection Using Machine and Deep Learning Techniques: A Preliminary Study

Jawaid Ahmed Siddiqui[1], Siti Sophiayati Yuhaniz[2], Zulfiqar Ali Memon[3], Yumna Amin[4]

[1]Sukkur IBA University Airport Road, Sukkur (Sindh) Pakistan
[2]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia
[3,4]FAST National University of Computer and Emerging Sciences, Pakistan
*jawaid@iba-suk.edu.pk

Article history

Received:
12th October 2021

Received in revised form:
18th October 2021

Accepted:
28th October 2021

Published online:
11th November 2021

*Corresponding author
jawaid@iba-suk.edu.pk

Abstract

The increasing use of social media and information sharing has given major benefits to humanity. However, this has also given rise to a variety of challenges including the spreading and sharing of hate speech messages. Thus, to solve this emerging issue in social media, recent studies employed a variety of feature engineering techniques and machine learning or deep learning algorithms to automatically detect the hate speech messages on different datasets. However, most of the studies classify the hate speech related message using existing feature engineering approaches and suffer from the low classification results. This is because, the existing feature engineering approaches suffer from the word order problem and word context problem. In this research, identifying hateful content from latest tweets of twitter and classify them into several categories is studied. The categories identified are; Ethnicity, Nationality, Religion, Gender, Sexual Orientation, Disability and Other. These categories are further classified to identify the targets of hate speech such as Black, White, Asian belongs to Ethnicity and Muslims, Jews, Christians can be classified from Religion Category. An evaluation will be performed among the hateful content identified using deep learning model LSTM and traditional machine learning models which includes Linear SVC, Logistic Regression, Random Forest and Multinomial Naïve Bayes to measure their accuracy and precision and their comparison on the live extracted tweets from twitter which will be used as our test dataset.

Keywords: *Hate speech, machine learning, Classification, Categorization, Random Forest, Logistic Regression and Multinomial Naïve Bayes.*

1. Introduction

Freedom of speech if the ultimate right of the people all over the world and over the years there are multiple emerging social media platforms to reach out your thoughts to other people and take note what the world is saying. Among many popular social media platform, twitter is the most popular micro blogging social media website which is used by 330 million active users' month as of 2019. It is used by celebrities and many other influential people. Daily approximately 500 million tweets are posted [1].

When millions of data is posted each day for expression of thoughts, there are people trying to spread negativity by sharing offensive and hateful content due to

* *Corresponding author. jawaid@iba-suk.edu.pk*

which there is a destructive impact on the target audience. Hate Speech is defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic [2]. Due to increasing hate content all over the social media on daily basis, it often leads towards the spread of hate crimes. Usually hate is targeted towards a group of people based on religion, ethnicity, class, gender, sexual orientation, disability and nationality. Due to this reason countries are introducing laws against hate speech and spreading of offensive material on social media. Many countries including France, United Kingdom and Canada have laws prohibiting hate speech and have imposed heavy fines and imprisonment also in some cases [3]. To overcome this issue of spreading hate content online, social media companies such as Twitter, Yahoo and Facebook have updated their terms of condition to forbidden content that is hateful, offensive, threatening and incite violence [4]. They are now focusing towards detecting Hate speech.

In this paper, an approach has been proposed to detect the hateful content of the most popular social media micro blogging website i.e. twitter using the hate dictionary provided at hatebase. Hatebase is defined as a service built to benefit organizations and online communities monitor, detect and analyze hate speech. The algorithms investigate public discussions using a broad vocabulary based on religion, gender, nationality, ethnicity, sexual orientation, class and disability, with data across 95+ different languages and 175+ countries of the world.

In this paper, we have trained a dataset obtained from Davidson et al and Founta et al.. That data is obtained from twitter and it is already classified into Hate or not hate category. For training of data we have used LSTM. For testing out tweets from the twitter obtained between the specified dates ranges are tested from the LSTM model, trained on the training dataset.

2. Literature Review

Every day, massive amount content is generated by the users over different social media networks therefore it is becoming vital to early detect hate content. Early discovery of this content can help to limit its propagation over the web and to fight against misogyny and xenophobia. [5] Hateful tweets, in twitter are considered those that comprise abusive speech which are targeted towards individuals (a celebrity, cyber-bullying, a politician, a product) or certain groups (LGBT, a country, gender, a religion, an organization, etc.). It is important to detect such speech for evaluating public sentiment towards a group of individuals and in the direction of another group, and for discouraging any hate related activities. [6] Detecting abusive language is a bit challenging than one as there is the disruptiveness of the data in combination with a requirement for world knowledge not only makes it a thought-provoking task to automate but also potentially a difficult assignment. [7] Twitter has a limitation of total characters allowed in a tweet (just 288 characters allowed on a single tweet). This makes its users frequently using rough words and abbreviations when they tweet. Therefore, extractions of features are mandatory to understand the meaning of the tweets and classify it as hate content or not. [8] Although this area is emerging but there are several good papers related to hate speech detection. Most of the researchers have worked only on binary classification i.e. hate speech detection.

Vega et al. [5] have identified aggressive hate speech or not and their target as individual or group from Spanish and English Language Tweets by linguistically driven features and numerous kinds of n-grams (functional words, words, characters, punctuation symbols, POS, among others). Support Vector Machine (SVM) was trained for aggressive speech detection using a combinatorial framework, whereas for identifying target group, a multi-labeled approach was used by applying the Random Forest classifier. Their methodology achieved the maximum F1-score in sub-task A for the language of Spanish.

Badjatiya et al. [6] have done a comparison of deep learning methods and the n-gram methodology on 16K annotated tweets which results that TF-IDF method is better than the character n-gram method. The best method identified is LSTM + Random Embedding + GBDT" where tweet embedding were modified to random vectors, LSTM was trained using back-propagation, and then learned embedding were used to train a GBDT classifier. Therefore deep neural networks outperformed the existing methods.

Silva et al. [19] proposed a similar study which we are going to do in this paper. He recommended “Systematic measurement study of the main targets of hate speech in online social media.” Twitter and Whisper datasets are used in this research and Sentence structure (Knowing hate words or target apriori) is used to detect hate words. Hate words are used from hate base database. They were categorized into 8 categories out of which top 3 categories are race, behavior, and physical. Ribeiro et al. [20] proposed methodology to characterize and identify hate speech with emphasis on content posted in Online Social Networks (OSNs). They used 5-fold cross validation for the two proposed approaches GradBoost and GraphSage). Nalini et al. [21] proposed “contextual and word level features based framework to detect offensive content.” Different Classification algorithms which includes Random Forest (RF), J48 (WEKA’S C4.5 execution) and Sequential Minimal Optimization were applied and their performance was compared.

Agarwal et al. [26] proposed the solution of identification of malevolent videos promoting hate and extremism through YouTube. They present a focused crawler based approach for numerous tasks. Extremist groups use YouTube as a medium to spread hateful and extremist content through its videos as this platform is accessible by every internet user. YouTube being the top site for videos consists of large amount of videos and it is accessed by millions of users every day. More than 100 hours of video is uploaded every minute. This paper scrutinize the application of a focused crawler (best-first search) based approach for retrieving YouTube user-profiles supporting extremism and hate. To gather the training data set, manual exploration and visual assessment is performed on the metadata of YouTube channels and 35 channels were initially identified which were promoting extremism and hate. Features such as titles, comments, shared, favored and user profiles are extracted from them to form the training data set using YouTube API. Character n-gram based language modeling approach is applied in this paper.

Binary classification is performed to identify the relevant user channel and features are extracted. Best first approached is used based on the similarity of training data set. Training dataset of 35 YouTube channels consists of 612 videos. Training dataset is obtained by manual keyword searches. 10 random hate channels were selected for testing the analysis.

3. Methodology

In this study, several methodologies were used for hate speech detection and its classification into multiple categories and sub-categories. To identify the hate words and its categories, hate words are extracted from Hatebase.org API. Hatebase.org is an online repository of multilingual terms classified identified as hate terms. This web-based application provides data through open API and web interface. Terms can added in the database by anyone using this form provided by hatebase. Hatebase.org data is retrieved through API [22] using the API token keys. The code to retrieve the data is written in Python. Language filter was set to 'en' to only retrieve English hate terms. Data retrieved through API consists of vocabulary, its meaning, offensiveness, and the hate category in which it belongs i.e. ethnicity, nationality, religion, gender, sexual orientation, disability and class. There are total 1530 hate terms in English language retrieved through API. These hate terms are used to retrieve the twitter data set from twitter for a specified date range 01-01-2019 to 07-06-2020. The hate dictionary was divided into list of 25 words and for each list around 1500 tweets were extracted. As there are millions of tweets but in this research, a limit has been placed for the retrieval of data as it is a time taking process of extracting tweets from the twitter. Approximately 239615 tweets were retrieved.

4. Proposed Model

Long Short Term Memory networks, commonly called "LSTMs", were introduced by Hochreiter and Schmiduber. These have extensively been used for speech recognition, language modeling, and sentiment analysis and text prediction. LSTMs have an benefit over straight feed-forward neural networks and RNN in many ways. This is because of their property of selectively memorizing patterns for long durations of time. LSTM have three different gates; Input, output and forget gate.

Along with LSTM, multiple machine learning models were used for the classification of tweets into categories and sub-categories which includes Linear Support Vector Classification (LSVC), Random Forest (RF), Multinomial Naive Bayes (MNB) and Logistic Regression (LR). There are several features used in this research which are combined with the models. Those features include Bag of Words (BOW), term frequency–inverse document frequency (Tf-idf) and N-grams. In N-gram, number of maximum words is set to be 3 in our experiment.

5. Dataset

5.1 Training Dataset

For the training of model, a combination of two different dataset is used. One dataset is by Davidson et al. and other is obtained by Founta et al. The data set gathered by Davidson et al. consists of 24,478 tweets which are classified into hate, offensive and none category. Labels to the dataset were categorized through manual identification and labeling by CrowdFlower (CF) workers. CrowdFlower is a company which cleans up disorganized and incomplete data using an online workforce. Typical users of CrowdFlower are data scientists who use the software to create training data, build models and train machine learning algorithms.

This data set will be used as the training data set for the training of our model. The data set was not balanced as it has fewer records of hate speech only as compared to the overall tweets. Out of 24,478 tweets, 4993 tweets were classified as hate by at least one CrowdFlower worker. In our model we have only classified tweet into hate and not hate category so we have classified tweets which are in offensive category and if at least 3 CrowdFlower workers have categorized it in offensive category into Hate category. Therefore 14996 records which were marked as offensive by atleast 3 CrowdFlower workers were included in the Hate dataset.

Another Data set is combined with the Training dataset to increase the training data as there is more than 235k testing tweets. The data set gathered by Founta have 41467 finalized tweets when tweets are mapped with the labels and irrelevant data was removed. The data was classified into Hate, abusive, spam and normal. Spam tweets were removed as it is of no use. Abusive tweets were included with Hate tweets and classified into one as they are closely related. Out if 41467, 15266 tweets were classified as hateful. Both the datasets were obtained in excel form and consolidated according to the requirements This dataset will be used to train the models for hate and not hate binary classification and further experiments will be done for its categorization into main and sub categories.

5.2 Test Dataset

The test data set has been extracted from twitter using twitter library GetOldTweets3 Library by providing date range. The date which we have provided for this experiment is from 01-01-2019 to 07-06-2020. Using the hate vocabulary list from Hatebase, we have divided that list into chunks of 25 per list and then extract the data as it is a time consuming process to extract the tweets. For each vocabulary list, a limit of 1500 per list has been defined. Therefore a total of 239615 tweets were extracted for our experiment.

5.3 Data Cleaning and Pre Processing

Raw text cannot be directly inputted into deep learning models. Text data must be encoded as numbers to be used as input or output for machine learning and deep learning models. For data cleaning process, read the csv file of tweets and process those tweets one by one and append it into the cleaned list. Only textual words are kept in the data and other special characters, hash tags and numbers are removed using python regex. Convert the texts into lower case letters and split the tweets into words. Stop words are then removed from the list of words by using NLTK Corpus stop words dictionary. English stop words which are found in this dictionary are removed from the cleaned list of words. All these words are then joined to again format the textual tweet.

5.4 Training Data Processing

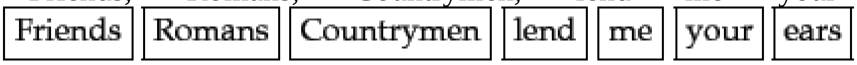
Original data used by Davidson et al. and Founta et al. doesn't consists of classification into hate speech categories, therefore in order to make the training the data according to the needs of our analysis, hate tweets are classified into ethnicity, nationality, religion, gender, sexual orientation, disability, class and none based on the hate terms and their categories retrieved by Hatebase.org. Both the dataset sheet which contains tweets and hatebase vocabulary and its categories were loaded into python data frame. Every tweet in data frame row is splitted into words list and each word from the list is matched with the hatebase vocabulary column.

If any word exists in the vocabulary list then its category is read and that tweet is assigned that category. In this way all the tweets are assigned with categories. Similarly for assigning sub-categories to the training data a dictionary has been created manually from the words list in different papers and by the meaning obtained from the hatebase dictionary list. Each word has meaning provided in the hatebase dictionary so it is tagged to the related sub-category and a new vocabulary list is combined for the sub-category classification. This list is used and if the word of tweet matches with it then its sub-category id assigned.

6. Implementation Detail

Pre-processed and cleaned data of twitter will be used as our training dataset to train the Long Short Term Memory Networks (LSTM) model. LSTM is a model for the short-term memory which can last for a long period of time. An LSTM is well-suited to classify, process and predict time series given time lags of unknown size and duration between important events.

After cleaning the data and removing all the stop words using NLTK stop words library, special characters using regex, cleaned tweets are then split into tokens.
Given a character sequence and a defined document unit, tokenization is the task of splitting it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization: [36]

Input: Friends, Romans, Countrymen, lend me your ears;
Output: 

Friends Romans Countrymen lend me your ears

Therefore using Keras tokenizer library from Keras pre-processing of text, all the cleaned tweets are then split into tokens. After tokenization, texts are then converted into sequence using keras library. This sequence is then padded using pad sequence so that the sequence of all the words in all the tweets should be of same length. It is used to ensure that all sequences in a list have the same length. By default this is done by padding 0 in the beginning of each sequence until each sequence has the same length as the longest sequence. For tokenization number of words size is kept 32. Number of words is defined as the maximum number of words to keep, based on word frequency. Only the most common words of the defined limit will be retained. The words are separated on the basis of empty spaces in a string of words.

For Categorization and sub categorization of tweets into main categories features are extracted using tfidf and unigram, bigram and trigram were extracted. TF-IDF to extract text features from text documents which is the most popular and widely used method for extraction of keywords. TF-IDF stands for "Term Frequency, Inverse Document Frequency". It is a way to score the significance of words (or "terms") in a document based on how often they appear across multiple documents. It calculates frequency for given word in the document, the value rises proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. So, the best features are obtained from TF-IDF as the features returned were not phrases but single words.

6.1 Training on the Model

For the training of the model for LSTM, the dataset of the training tweets was first trained with 20% of total 69160 tweets i.e. 13832 tweets. The batch size was kept 20 and total 3 epochs were run to train the data. The tweets were processed in the form of tokens by using Keras Tokenizer and then they were padded with sequence to form a uniform shape of features. The shape of features was 1348. The accuracy score obtained by the model was 0.924

7. Experiment and Results

7.1 Binary Classification

For Binary classification of Hate and Not Hate by traditional machine learning methods, the features of tweets are extracted by TF-IDF. In our combined test and training data set, more than 160k tweets can be classified as not hate and 140k tweets are considered as hate tweets. In this experiment, we have used trigram and find the most correlated words.

Table 1. Correlation of Words N-gram

Unigram	Bigram	Trigram
Bitches	bull dyke	piccaninnies watermelon smiles
Fucking	hern monkeys	police auction rings
Bitch	plastic paddies	uk police auction

After the experiment and testing 80% tweets, the accuracy score obtained by 4 different models are in the table below:

Table 2. Evaluation of Machine Learning models

Model	Accuracy
Linear SVC	0.70
Random Forest	0.72
Logistic Regression	0.74
Multinomial MNB	0.47

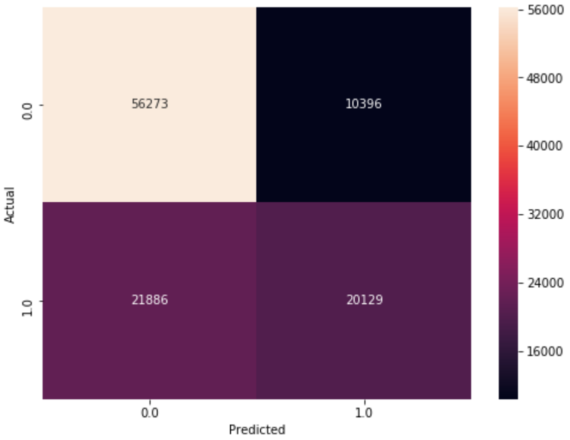


Figure 1. Prediction of Linear SVC

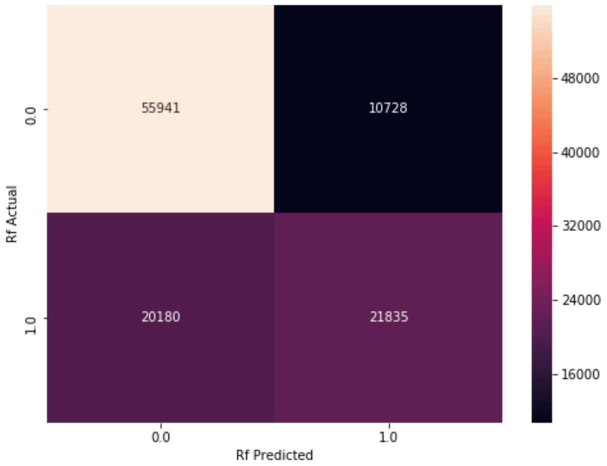


Figure 2. Prediction of Random Forest

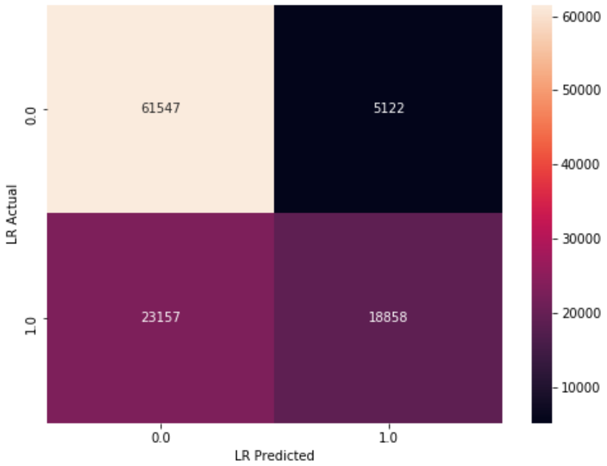


Figure 3. Prediction of Logistic Regression

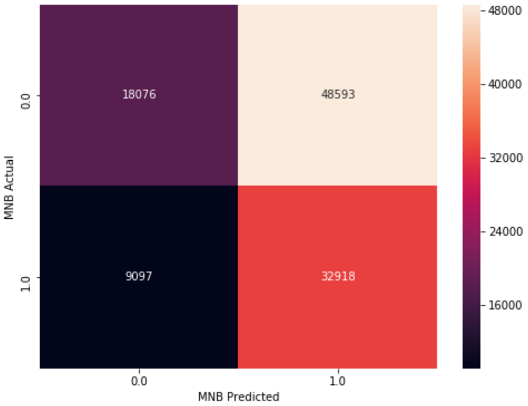
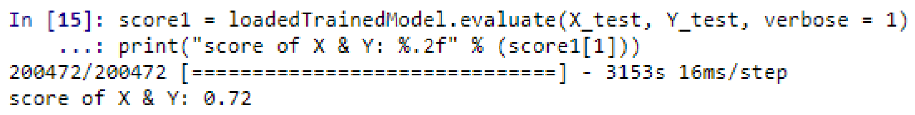


Figure 4. Prediction of Multinomial Naive Bayes

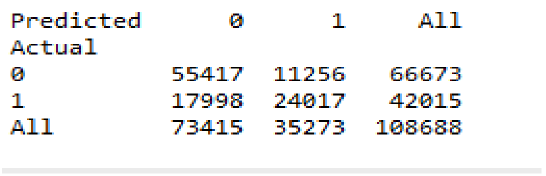
By performing testing on the extracted data using trained model on LSTM, an accuracy of 0.72 is achieved when batch size was kept 6. When we try to increase the batch size to 20, training accuracy was 0.92 but test accuracy was 0.55 as the data extracted from twitter was high in volume.



```
In [15]: score1 = loadedTrainedModel.evaluate(X_test, Y_test, verbose = 1)
    ...: print("score of X & Y: %.2f" % (score1[1]))
200472/200472 [==============================] - 3153s 16ms/step
score of X & Y: 0.72
```

Figure 5. LSTM Test Result

The confusion matrix of the results obtained by testing 100k tweets is below.



```
Predicted      0      1     All
Actual
0          55417  11256   66673
1          17998  24017   42015
All        73415  35273  108688
```

Figure 6. Confusion Matrix LSTM Binary Classification

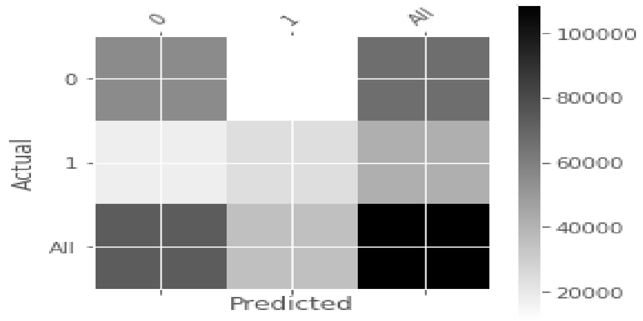


Figure 7. Confusion Matrix LSTM

8. Categorization

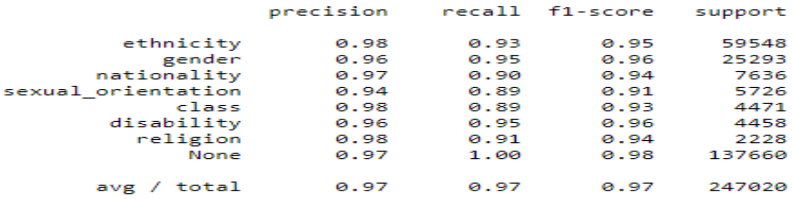
The categorization of tweets into the eight Categories of Religion, ethnicity, class, nationality, gender, Sexual Orientation, Disability and none. The test data initially doesn’t have any classification so by using the Hatebase vocabulary list, if the word of tweet exists in that list then it is classified as hate and its category is also annotated with it. Main Categories e.g. religion, ethnicity is also tagged with every word in hatebase so same approach is used for identifying the main categories. The sub category is annotated by Sub category list which is made from this hatebase list. That list consists of meanings of every hate word so it is manually classified into sub categories. Therefore if the word of tweet matches with the sub category list, then it is annotated with that category.

In order to extract features of the Hate Category, TF-IDF is used to extract the features and by applying N-gram correlated features of every hate category is extracted.

Table 3. Correlated Words of Class

Unigram	Bigram	Trigram
Div	bad boujee	millie bobby brown
Spides	skeet shooting	andrew hicks body
Yobes	im boujee	redneck kludge hillbilly

Majority of the tweets can be in any of the 8 categories can be categorized into ethnicity which is 59548 tweets followed by gender which is 25293 tweets. Almost 50% of the tweets cannot be categorized to any group.



```
                    precision    recall  f1-score   support

          ethnicity       0.98      0.93      0.95     59548
             gender       0.96      0.95      0.96     25293
        nationality       0.97      0.90      0.94      7636
 sexual_orientation       0.94      0.89      0.91      5726
              class       0.98      0.89      0.93      4471
         disability       0.96      0.95      0.96      4458
           religion       0.98      0.91      0.94      2228
               None       0.97      1.00      0.98    137660

        avg / total       0.97      0.97      0.97    247020
```

Figure 8. Evaluation of Categorization of Tweets

By performing testing with 80% test and 20% for training, the score of different models is listed below.

Table 4. Experimental Results

Model	Score
Linear SVC	0.85
Random Forest	0.74
Logistic Regression	0.81
Multinomial Naïve Bayes	0.12

Therefore for categorization of hate speech, Linear Support Vector Classification gave best results even though Random forest gave slightly better classification of tweets for ethnicity category instead of LSVC and gave categorized more tweets in any category than classifying it into none category. For Sub Categorization, in ethnicity Africans are targeted the most.

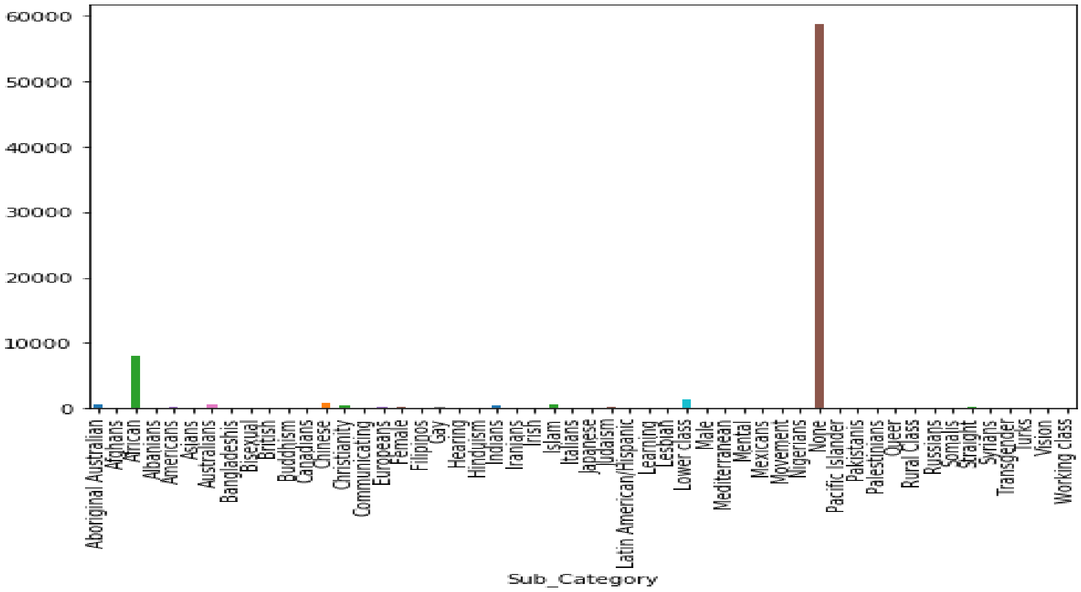


Figure 9. Sub Categorization of Ethnicity

The results of the evaluation of models of subcategories are in the below table.

Table 5. Sub Categorization evaluation

Model	Score
Linear SVC	0.92
Random Forest	0.95
Logistic Regression	0.88
Multinomial Naïve Bayes	0.85

Random forest classifier gave the best results for sub-categorization followed by LSVC

9. Conclusion and Future Work

Hate speech detection in social is important as it often leads towards hate crimes and it is targeted towards an individual or group of people who feel offended, bullied and harassed by such material. Twitter is one of the most used micro blogging website used by millions of people every day including many influential people and celebrities also. Due to limited length of tweet, people convey their messages and discuss on make their voice herd to everyone. In our research, tweets are extracted from the twitter by specifying the date range and some vocabulary list. The extracted tweets are more than 235k on which different experiments are performed to detect hate speech which includes deep learning model i.e. LSTM and traditional machine learning models including linear Support Vector Classification (LSVC), Random Forest (RF), Logistic Regression (LR) and Multinomial Naïve Bayes (MNB). For hate speech detection Logistic Regression performed the best 0.74 followed by Random Forest 0.72. For Hate speech Categorization LSVC performed the best and gave 0.85 accuracy but RF gave slight better categorization results. For sub categorization, Random Forest gave best result i.e. 0.95 accuracy. Furthermore work can be done in this area using deep learning models to get better accuracy and results and a lot more work is needed in the area of sub categorization of data so that specific group can be identified in target audience e.g. In ethnicity, Africans are identified to be targeted the most. In religion, Muslims are identified to be targeted the most. Work should be done to proceed with the further categorization of the target groups.

References

[1] Oberlo Blog https://www.oberlo.com/blog/twitter-statistics#:~:text=Here's%20a%20summary%20of%20the,daily%20active%20users%20on%20Twitter.

[2] Nockleby, J. T. (2000). Hate speech. Encyclopedia of the American constitution, 3, 1277-79.

[3] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Eleventh international aaai conference on web and social media.

[4] Warner, W., & Hirschberg, J. (2012, June). Detecting hate speech on the World Wide Web. In Proceedings of the Second Workshop on Language in Social Media (pp. 19-26). Association for Computational Linguistics.

[5] Vega, L. E. A., Reyes-Magaña, J. C., Gómez-Adorno, H., & Bel-Enguix, G. (2019, June). MineriaUNAM at SemEval-2019 Task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 447-452).

[6] Badjatiya, P., Gupta, S., Gupta, M., & 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).

[7] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

[8] Ibrohim, M. O., Setiadi, M. A., & Budi, I. (2019, November). Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. In Proceedings of the International Conference on Advanced Information Science and System (pp. 1-5).

[9] Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215-230.

[10] Charitidis, P., Doropoulos, S., Vologiannidis, S., Papastergiou, I., & Karakeva, S. (2020). Towards countering hate speech against journalists on social media. Online Social Networks and Media, 17, 100071.

[11] Sureka, A., & Agarwal, S. (2014, September). Learning to classify hate and extremism promoting tweets. In Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint (pp. 320-320). IEEE.

[12] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In Proceedings of the 24th international conference on world wide web (pp. 29-30).

[13] Ibrohim, M. O., Setiadi, M. A., & Budi, I. (2019, November). Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. In Proceedings of the International Conference on Advanced Information Science and System (pp. 1-5).

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

[15] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835.

[16] ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018, June). Hate lingo: A target-based linguistic analysis of hate speech in social media. In Twelfth International AAAI Conference on Web and Social Media.

[17] Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1-10).

[18] Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651.

[19] Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016, March). Analyzing the Targets of Hate in Online Social Media. In ICWSM (pp. 687-690).

[20] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., & Meira Jr, W. (2018, June). Characterizing and detecting hateful users on twitter. In Twelfth international AAAI conference on web and social media.

[21] Nalini, K., & Sheela, L. J. (2014). A survey on datamining in cyber bullying. International Journal on Recent and Innovation Trends in Computing and Communication, 2(7), 1865-1869.

[22] Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 5(1), 11.

[23] Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In 13th International Workshop on Semantic Evaluation (pp. 54-63). Association for Computational Linguistics.

[24] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48(12), 4730-4742.

[25] Ting, I. H., Wang, S. L., Chi, H. M., & Wu, J. S. (2013, August). Content matters: A study of hate groups detection based on social networks analysis and web mining. In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on (pp. 1196-1201). IEEE.

[26] Agarwal, S., & Sureka, A. (2016, August). But I did not mean it!—intent classification of racist posts on tumblr. In Intelligence and Security Informatics Conference (EISIC), 2016 European (pp. 124-127). IEEE.

[27] Mondal, M., Silva, L. A., & Benevenuto, F. (2017, July). A Measurement Study of Hate Speech in Social Media. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (pp. 85-94). ACM.

[28] Erin Griffith. 2013. with 2 million users, “secrets app” Whisper launches on Android. http://pando.com/2013/05/16/with-2-million-users-secrets-app-whisperlaunches-on-android/. May 2013).

[29] Twitter team. 2017. The Streaming APIs. https://dev.twitter.com/streaming/overview.

[30] Unsvåg, E. F., & Gambäck, B. (2018, October). The effects of user features on twitter hate speech detection. In Proceedings of the 2nd workshop on abusive language online (ALW2) (pp. 75-85).

[31] Anzovino, M., Fersini, E., & Rosso, P. (2018, June). Automatic identification and classification of misogynistic language on twitter. In International Conference on Applications of Natural Language to Information Systems (pp. 57-64). Springer, Cham.

[32] Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018, June). Large scale crowdsourcing and characterization of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media.

[33] Zhang, H., Wojatzki, M., Horsmann, T., & Zesch, T. (2019, June). ltl. uni-due at SemEval-2019 Task 5: Simple but Effective Lexico-Semantic Features for Detecting Hate Speech in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 441-446).

[34] Lee, Y., Yoon, S., & Jung, K. (2018). Comparative studies of detecting abusive language on twitter. arXiv preprint arXiv:1808.10245.

[35] Abuzayed, A., & Elsayed, T. (2020, May). Quick and simple approach for detecting hate speech in arabic tweets. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 109-114).

[36] NLP Stanford https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html