International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 3 | Issue: 5 | Sep - Oct 2022 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v3i5.16

A Multimodal Hate Speech Classification Process Using Dual Feature Extraction Techniques

Chibuike Onuoha[1*], Ikerionwu Charles[2] and Obi Nwokonkwo[1]

[1]Department of Information Technology, Federal University of Technology, Owerri, Nigeria
[2]Department of Software Engineering, Federal University of Technology, Owerri, Nigeria

E-mail: onuoha.chibuike@futo.edu.ng; charles.ikerionwu@futo.edu.ng; obi.nwokonkwo@futo.edu.ng

*Corresponding author details: Chibuike Onuoha; onuoha.chibuike@futo.edu.ng

ABSTRACT

Racist and ethnic violence, fabricated persecution, and some form of intimidation are all risks associated with hate speech, which is a concern with natural language processing. Given the sensitivity of hate speech in our society, it is essential to classify speeches into hate and non-hate categories in real time to minimize its risks. The main objective of this work is to investigate selected supervised machine learning algorithm model for the classification of hate speech on social media. The term frequency-inverse document frequency (TF-IDF) and bag of words (BOW) models were used by the model to extract features. Porter's stemming model and WordNet for lemmatization are used during the preprocessing step. The datasets were trained using logistic regression, naive Bayes, and random forest, and logistic regression was also utilized to create the classifier. For training purpose, 80% of the datasets was used to train the model and 20% was used for testing the model. Results obtained from the application of Logistic Regression algorithm revealed 98% accuracy and 98% F1-score. These scores indicate high accuracy in hate speech detection and classification.

Keywords: NLP; hate speech; classification; accuracy

INTRODUCTION

Although social media is a key component that has aided social engineering in the 21st century, it is not without disadvantages. Recently, nations on the receiving end have risen to police social media, purposely to easily detect hate speech and comments considered offensive. For example, hate speech has been propagated through social media and used to incite the populace against established authority. According to [1], hate speech is an offensive language that could be aggressive, insulting, provocative etc., targeted at a person or group of people. To raise the awareness and nip the propagation of hate speech at bud, social media platforms have requested their respective users to shun such acts. Because of the varieties of hate speech witnessed from different societies, [2] opined that there is no general acceptable concept of hate speech. Although hate speech is a controversial concept, what is considered a hate speech in a specific environment might not be seen as such in another environment.

Hate is an indication of an emotional state or opinion, and therefore distinct from any manifested action. Speech: any expression imparting opinions or ideas– bringing a subjective opinion or idea to an external audience [3]. It can take many

forms: written, non-verbal, visual, or artistic, and can be disseminated through any media, including the internet, print, radio, or television.

Based on the two highlighted component of the term 'Hate Speech', Hate speech is an expression of hate towards a person or group of people.

With regards to this definition, hate speech in the context of this project work is an opinion or idea that is emotional, directed to an individual or group.

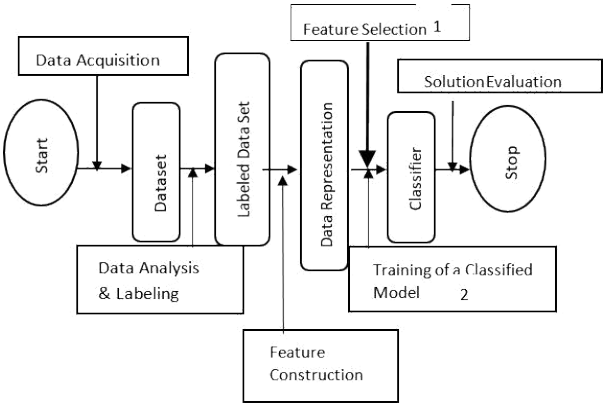

FIGURE 1: Text classification Pipeline Process

It could originate from different forms namely written, artistic or visual and could be distributed by different sources such as television, Internet, etc. Since hate speech is recent and has seen a surge in application, its detection is giving rise to recent research interest [4]. Manual detection of hateful texts is a tedious work and as such filled with a lot of biasness which is the reason researchers are finding automated ways to detect hate speech on the web. Due to the fact that speeches follow a natural language processing classification problem, there exist complications in terms of grammar and sentence structure involve with online media communities.

The different process of a text classification as identified by [5] is seen in Figure 1. Though different machine learning algorithm has been applied to carefully classify and predict hate speech, such as convolutional Neural Network (CNN), Support Vector Machine (SVM), Random Forest (RF), etc. Natural Language processing techniques have been applied and they have given

different accuracy in the domain of social media platform detection and classification of hate speech. Thus, this paper will try to classify hate speech by training the dataset through different machine learning algorithms in other to classify and differentiate speeches into hate or not hate, then, followed by a comparative analysis of different models selected by calculating their confusion matrices, recall values, and F-scores. The major challenge of text hate speech classification as observed by [6] is the accuracy of the classifier and high dimensionality of feature space. However, this approach is gradually gaining impact and popularity. In Nigeria, for instance, a new bill was passed for punishing online social media users using Hate speeches. History has shown that hate speech can be used by different ethnic or racial groups [7] to induce hatred and arouse anger, which can lead to ethnic wars. Other instances include kidnappers and bandits who run large cartel networks and use hate speeches in discussions to subdue their victims. companies and institutions boost over their reputation as they usually received ratings from the perceived image of users, and as such cannot make their platform known as a hate site. Thus, there is a slow rate in the identification of hate speech. To solve this problem, this paper focuses on using a hybrid feature selection procedure to extract context from a text document and training our result using a different classification algorithm.

The primary objective of this study is to develop an enhanced model for social media hate speech classification. To achieve this objective, the following specific objectives would be addressed: design an architectural framework to train hate speech datasets using multiple models, develop a classifier using the best- selected model, and extract features from dataset using term frequency-inverse document frequency (TF-IDF) and Bag of words (BOW).

RWVIEW OF RELATED WORKS

Our aim is to carefully identify words that has been labeled hateful by a large group of people so that we could use that to illustrate variation of hateful words [8]. Hate speech classification has been widely applied across the social networking sites and spam email detection over the years. Following the recent applaud in text mining and Natural language processing (NLP), a lot of researchers are continually building applications that can aid text classification [9].

According to [10], the n-grams, POS, TF-IDF, mentions, hastags, length, readability, sentiment, mispellings, emojis. They got an F1 score of 94% by applying SVM, Convolutional Neural Network, on a Twitter dataset. [11], got a precision value of 1 by using a rule-based model and sentence structure to classify hateful speeches. [12] used Logistic Regression (LR) to classify hateful words using tweet length, gender of the author, length of user description, location and word n-grams to extract features. Their research produced a 73% F1 score. [13], uses char n-grams, word n-grams, skip-grams, tweet length, author gender, clusters, POS, Author Historical Salient Terms (AHST) for feature extraction and applied LR to record an F1-score of 91.5%. [14], used Random Forest, Support Vector Machine, Gradient Boost Decision TreeBDT, Logistic Regression, Deep Neural Network, and Convolutional Neural Network. The study produced a precision, recall, and F1-score of Precision Value of 93%, 93%, 93% respectively. [15] extracted features from their collected dataset by applying word2vec embeddings, random embeddings, char n-grams, and the model is trained using CNN.Their model yielded 86% as precision value, 72% as recall value,and finally 78% as F1-score. [16], applied char embeddings, word embeddings to extract features, and CharCNN, WordCNN, and HybridCNN were employed for model training and also testing. They got the following as their findings as precision value; 83%, recall value: 83%, F1-score: 83%. [17] applied Logistic Regression, Decision Tree, Random, Forest, Adabost, & SVM on n-grams, semantic and syntactic, TF-IDF, word2vec embeddings, doc2vec and they achieved an F1-Score of 96%. [18], achieved Precision Value; 93%, Recall value: 80.5%, and F1-score: 91.2% on Bag-of-Words, Term Frequency-Inverse Document Frequency, Word2Vec.

EXPERIMENTAL SETUP

A. Design Architecture

The architecture uses two feature TF-IDF and BOW as feature extractors in other to assign tags to text. In the data preprocessing phase, the tweets/text were reduced to their rootforms by applying word stemming process. The training process was setup using Naive Bayes, Logistic regression and Random forest.

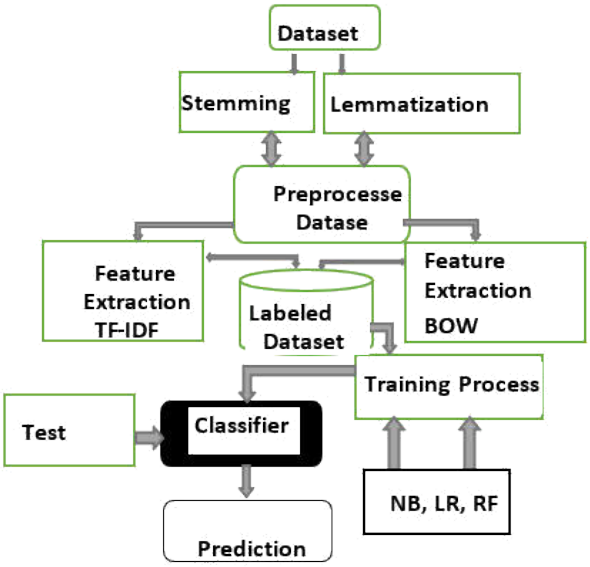

FIGURE 2: Proposed Model Architecture

From the training process, we observed that logistic regression did performed better than Naïve Bayes and Random Forest. The Classifier component was built using Logistic regression as its classification algorithm as observed in Figure 2

B. Deployment Architecture

The algorithm with the best score of accuracy was used to design the model. The entire phase from stemming, feature extraction and the supervised classification algorithm was stored in the pickle file which is then configured in flask framework to interface with web browsers. The web structure was designed using HTML and styled using cascading style sheets. When user enters a particular text in the text area provided on the webpage, he clicks on the predict button which calls the saved model pickle file to classify the text and send back a response to the user via the webpage.

The flask handles the server request via localhost 127.0.0.1 and port of 5000. A clear picture was seen in Figure 3 below.

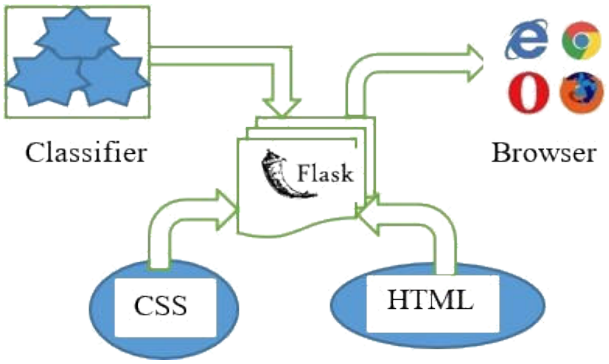

FIGURE 3: Proposed Deployment Architecture

C. Datasets

In other to achieve supervised training procedure, we used a Twitter dataset consisting of 24,000 tweets obtained from a repository from machine learning Kaggle community. The dataset is made up of Tweet ID, Hate sentiment and Non-hate Sentiment. Pre-cleaning process reduced the dataset to 23, 475 as blank spaces and non-English tweet lines were removed. Forthe purpose of over fitting, 80% of the dataset (18,740) was used for model training while 20% (4735) was used for model testing.

D. Computational Resource

We ran the experiment using Jupyter notebook running Anaconda 2020. The computer is a Core i7 HP EliteBook computer with Ubuntu 20.04 as OS, 1TB HDD and 12GB of RAM. Also, we use the following packages to achieve our research goal; matplotlib, Keras, Pandas and Skit-learn, etc.

E. Training Process

First, we made our datasets to have equal number of tweets labeled as hate and non-hate. Before training, we clean the datasets using stemming, stop words removal and lemmatization. For preprocessing steps, we employed Bag of words (BOW) and also Term Frequency-Inverse Document Frequency (TF-IDF). We split the datasets into corresponding containers for X_train and X_test. And lastly, we use Sklearn models, to import the different models used for training our datasets.

RESULTS AND DISCUSSIONS

We processed some samples of clean datasets which we have classified as hateful and non-hate. In Fig4, it shows a word matrix (8X12) of non-hateful tweets. Here we cluster maximum of 100 negative sentiments from the list of datasets into a word cloud of words arranged in a (8X12) matrix.

(a) Non-hateful words (b) Hatefulword cloud

FIGURE 4: Word cloud of sentiments

A. Model Evaluation

We used the confusion matrix to evaluate the performance of each of the algorithms on the 4735 of test datasets. From the confusion matrix, we evaluated the sensitivity, specificity, F1 ratio and the accuracy of the trained algorithm on the test datasets.

a. Using Random Forest (RF) on TF-IDF

$$\begin{bmatrix} 500 & 434 \\ 301 & 3500 \end{bmatrix} \quad (1)$$

$$Accuracy = \frac{TP+TN}{(TN+TP+FN+FP)} \quad (2)$$

$$= \frac{500+3500}{(500+3500+301+434)} \quad (3)$$

$$= 0.8450 \times 100$$

$$= 84.5\%$$

TABLE 1: Summary of Confusion Matrix for different Models

TF-IDF Feature Extraction Technique				
	Recall	Precision	F1score	Accuracy
Logistic Regression	0.99	0.98	0.98	98%
Naïve Bayes	0.82	0.95	0.88	80%
Random Forest	0.89	0.92	90%	84.5%
BOW Feature Extraction Technique				
	Recall	Precision	F1score	Accuracy
Logistic Regression	0.99	0.98	0.98	98%
Naïve Bayes	0.97	0.78	0.86	79%
Random Forest	0.94	0.92	0.93	89%

Summary of Results:

i. Using BOW and TF-IDF shows little variation when used with Logistic Regression.

ii. Random Forest enjoys some percentage gains with BOW over TF-IDF.

iii. Recall value of Naïve Bayes increases while the precision value reduces as we move from TF-IDF to BOW.

B. Comparative Analysis

The comparative evaluation is done to check rate the performance of the proposed model against published papers on hate speech classification and detection. The proposed model gives 98% accuracy as observed in Figure 5.

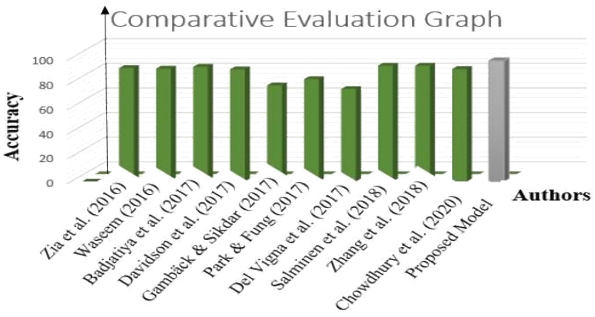

FIGURE 5: shows a graph of accuracy plotted against the developed

CONCLUSION AND FUTURE WORKS

We used the Hate speech datasets for a classification task. Due to the low rates of false positives, we were able to conclude from the results that cleaning was crucial in this study. From a modeling perspective, logistic regression is a fantastic model that works with both TF-IDF and BOW. On both feature extraction methods, Random Forest performed better than Naive Bayes. Future study should concentrate on a real-time intelligent model that can recognize the context and determine the author's mood and tone.

REFERENCES

[1] Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. gyo, Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. Human-Centric Computing and Information Sciences, 10(1), 1–34. https://doi.org/10.1186/s13673-019-0205-6

[2] Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018, April 2019, 61–66. https://doi.org/10.1109/BRACIS.2018.00019

[3] Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. Law and Philosophy, 36(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

[4] Biere, S. (2018). Hate Speech Detection Using Natural Language Processing Techniques. Vrije Universiteit Amsterdam, 30.

[5] Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

[6] Kumbhar, P. (2016). A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. International Journal of Science and Research (IJSR), 5(5), 1267–1275. https://doi.org/10.21275/v5i5.nov163675

[7] Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2016). Preprocessing Techniques for Text Mining -An Overview. 5(1), 7– 16.

[8] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PLoS ONE,14(8), 1–16. https://doi.org/10.1371/journal.pone.0221152

[9] Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. Sensors (Switzerland), 19(21), 1–37. https://doi.org/10.3390/s19214654

[10] Ziqi, Z., Robinson, D., & Jonathan, T. (2019). Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 11816 LNAI (1), 2546–2553. https://doi.org/10.475/123_4

[11] Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. 88–93. https://doi.org/10.18653/v1/n16-2013

[12] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. 26th International World Wide Web Conference 2017, WWW 2017 Companion, 2, 759–760. https://doi.org/10.1145/3041021.3054223

[13] Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. August, 85–90. https://doi.org/10.18653/v1/w17-3013

[14] Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. 41–45. https://doi.org/10.18653/v1/w17-3006

[15] Chowdhury, A. S., Mahamud, A. H., Nur, K., & Zabir Haque, H. M. (2020). Predicting behavior trends among students based on personality traits. ACM International Conference Proceeding Series. https://doi.org/10.1145/3377049.3377068

[16] Djuric, N., Zhou, J., Morris, R., Grbovic, M., & Vladan Radosavljevic, N. B. (2015). Hate Speech Detection with Comment Embeddings. 31st International Conference on Machine Learning, ICML 2015, 4, 2931–2939.

[17] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, 512–515.