4th International Engineering Conference (IEC 2023)
Federal University of Technology, Minna, Nigeria

Hate and Offensive Speech Detection Using Term Frequency-Inverse Document Frequency (TF-IDF) and Majority Voting Ensemble Machine Learning Algorithms

* Okechukwu, C. [1], Idris, I.[2], Ojeniyi, J. A.[3], Olalere, M.[4] & Adebayo O. S. [5]
[1,2,3,4,5] Department of Cyber Security Science, Federal University of Technology, PMB 65 Minna Niger State, Nigeria.
[5] Islamic University in Uganda.
*Corresponding author email: Anthony.pg208286@st.futminna.edu.ng +2348068573605

ABSTRACT
The advancement in technology especially the internet has opened new frontiers to criminality and abuses of information. Social media have given racists and extremists a platform for carrying out their criminalities and attacks on legitimate users' information. Thus, there is need to give adequate attention to the communications on social media so as to curtail these malicious acts before they materialize into causing physical harms. Hate speeches has been blamed for various degrees of violence experienced in the real world. A lot of research efforts have been put in detecting hate speeches using various techniques with varying degrees of accuracy and F-Measure. Term Frequency-Inverse Document Frequency (TF-IDF) with a majority voting ensemble learning classification Models were used for the detection of hate speech and a performance of 95% accuracy and 0.95 F-Measure were recorded.

Keywords *Ensemble Machine Learning, Hate Speech Detection, Majority Voting, Term Frequency-Inverse Document Frequency (TF-IDF).*

1 INTRODUCTION

United Nations in 2019 (United Nations, 2019) gave the definition of hate speech as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor". Generally, hate speech aims to generate hatred and intolerance thereby inciting divisions which can on the long run lead to xenophobia, racism and violence. Widespread adoption of Social Media (SM) makes the impact of hate speech overwhelming as a result of the anonymity enjoyed by users (Mullah and Zainon, 2021).

Tackling hate speeches according to United Nations is crucial towards deepening progress across the United Nations agenda by helping in preventing armed conflicts, terrorism and atrocity crimes. It will equally help in putting an end to violence against women, other grave violations of human rights, and fostering peace and an inclusive just society.

Recently hateful speech crimes have been on the increase in Nigeria both online and face-to-face. Anonymity of the internet among other factors has contributed to the rise in online hate speeches. The anonymity of the internet has given a voice to all and sundry to air their views and opinions without fear of any legal consequences.

In reactions to the threat of online hate speech, the government of Nigeria has recently placed a ban on Twitter social media website citing ethno-religious hate and violent speeches being spread through the medium and the refusal of Twitter to take down certain tweets the government sees as being hateful and inciting.

All over the world, different governments have been torn between tackling this crime specifically determining what is hateful and maintaining the citizens' right to free speech, hence the need to employ the unbiased nature of machines in solving this problem.

Among the key commitments of United nations in (United Nations, 2019) towards tackling this menace is the use of Technology by keeping up to technological innovations and encouraging more researches on what relates to misusing the social media and the internet in disseminating hateful contents like speeches and the factors that drive individuals towards being violent.

The rest of this paper is organized as follows: the review of related literatures is done in section two while the methodology to provide solution is given in section three, the results of the experiments are provided in section four while section five is used to conclude the paper.

2 RELATED LITERATURE

In the research by Fortuna and Nunes (2018), in which they surveyed automatic detection of hate speeches in texts conducted after analyzing hate speech concept in different contexts, from social networks platforms to other organizations, proposed a clearer and unified definition of the concept which can enable the building of machine learning models for automatic hate speech detection. They studied various techniques and approaches used by different researchers in detecting hate speeches and recorded a highest accuracy of 91% from all the papers reviewed. Different researchers have over the years proposed several techniques for detecting hate speeches

online. Each researcher achieved a different level of accuracy with their proposed models. Priyadharshini (2020) in her work" Detection of Hate Speech using Text Mining and Natural Language Processing" used TF-IDF for feature extraction and four different classifiers: Logistic Regression, Random Forest, Naïve Bayes and SVM in classifying tweets to be either hate or offensive speeches or being neither of the two using multi-class classifier and achieved a maximum accuracy of 90%.

Four convoluted neural network models were trained by Gambäck and Sikdar (2017) using character 4-grams and randomly generated word vectors to classify tweets and recorded according to them 78% F-score. The performance of the models was not evaluated using accuracy. In their work "Using Machine Learning for Detection of Hate Speech and Offensive Code-Mixed Social Media text", Pathak et al. (2020) used different classification and regression based machine learning algorithms to classify twitter messages into being "offensive" and "not offensive" in Indo-European Languages. They used Term Frequency-Inverse Document Frequency (TF-IDF) based feature modelling. Their model obtained highest F1 score of 0.87. Arabic Hate speech tweet problems were investigated using several neural networks (RNN) and convoluted Neural Networks (CNN) models in Alshalan and Al-Khalifa (2020) . A best performance of F1-score of 0.79 was obtained. In their work, Abro et al. (2020) proposed that the bigram features with support vector machine gave the best performance of 79% accuracy after comparing three feature engineering techniques and machine learning algorithms. They. For hate speech classification of tweets, Badjatiya et al. (2017) experimented with three different neural network architectures; CNN, LSTM and FastText. The best performance of 0.93 F1-score was recorded by the team.

Sari and Ginting (2019) worked on hate speech detection on twitter using Multinomial Logistic Regression. Their model's optimal performance is 84%. The authors recommended improving the feature extraction model to improve the model's performance.

Pariyani et al., (2021) in their work, "Hate Speech Detection in Twitter using Natural Language Processing" observed that it is not just enough to have high accuracy but to work to improve the F1-score of the model.

3 METHODOLOGY

In order to detect hate speech, the procedures depicted in figure 1 are followed from dataset collection, splitting, transformation to training the model and to the model's performance result evaluation. The dataset was first obtained from a public repository in kaggle.com, split into training and test sets. The training set is then vectorized using TF-IDF. The vectors were used to train an ensemble Machine Learning Algorithm after which the test set is equally vectorized and fed into the model to make predictions and results compared to the original class to get the performance result of the model.

TABLE 1: DATASET DESCRIPTION

SN	Feature Name	Data Type
0		Integer
1	Count	Integer
2	hate_speech	Integer
3	offensive_language	Integer
4	Neither	Integer
5	Class	String
6	Tweet	String

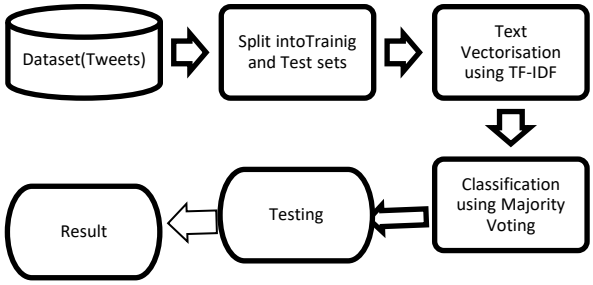

Figure 1: Hate Speech Detection Procedure.

DATA COLLECTION AND PREPROCESSING

A publicly available dataset from Kaggle.com (*Hate-Speech-Data-Analysis | Kaggle*, 2020) which contains 24783 instances is used for this work. The dataset has seven attributes as shown in Table 1 above which shows the features and their data types.

TABLE 2: SAMPLE DATASET

	count	hate_speech	offensive_	neither	class	tweet
0	3	0	0	3	0	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya …

FEATURE SELECTION

From the dataset, it was observed that the features; Hate_speech, Offensive_language, neither and count are

mere frequencies, the highest of which determines the class label. But TF-IDF as a text mining algorithm only needs the text in a feature to vetorise, weight and transform the text into a new feature set for training the model, thus the justification for using only the tweet text feature to determine their relationships with the class label.

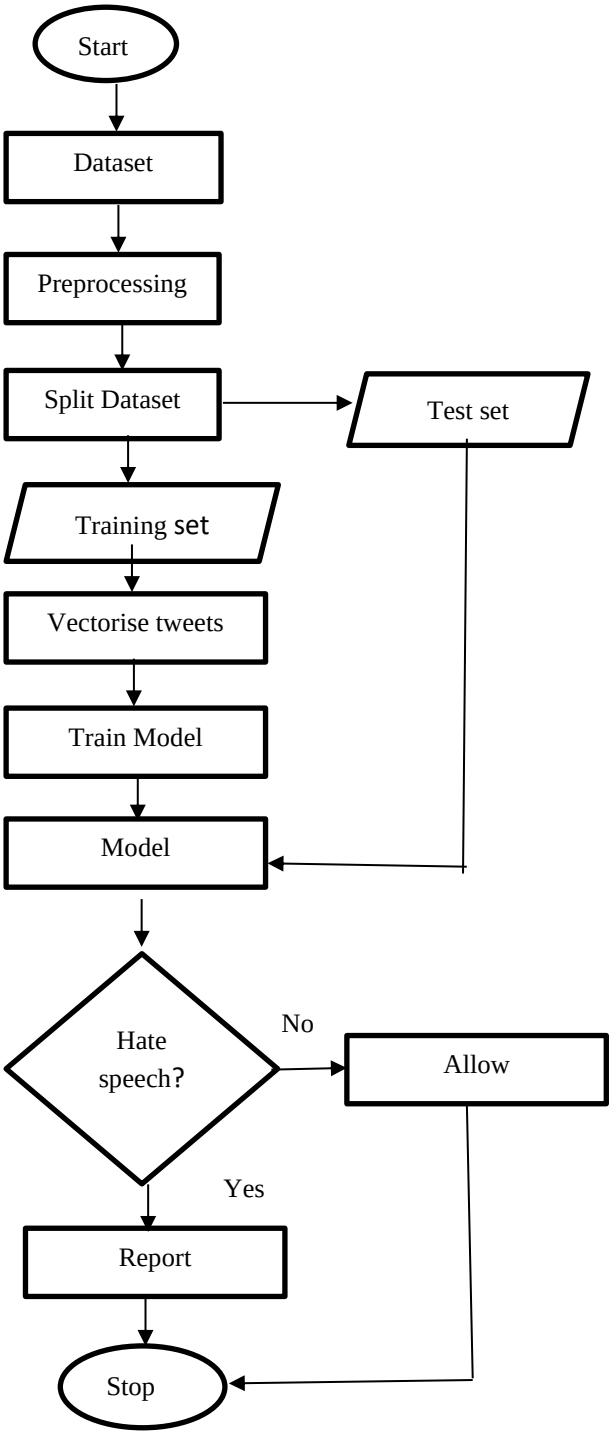

Figure 2: Flowchart illustration of the model.

FEATURE TRANSFORMATION

To improve classification accuracy since the focus is on detecting hate and offensive speeches, the class attribute which originally is made up of three classes; Hate, Offensive and None were transformed to only Hate or None Hate (1s and 0s respectively) speeches. For better performance, common words that impact little or no meaning on the tweet were removed. The filtered tweet attributes were passed to a TF-IDF algorithm which vectorized the tweets and weighted them according to their term and inverse document frequencies. The TF-IDF produced a transformed dataset which is further transformed to a sparse matrix for improved computation time. Figure 2 shows the flowchart diagram of the hate speech detection model from the dataset to the prediction from the model trained. It starts with acquiring and preprocessing of the dataset through which the dataset is made ready and optimum for training a machine Learning model. For this work, we split the dataset in the ratio of 70 / 30% for training and test sets respectively. The training set is vectorised using TF-IDF and the output used to train an ensemble Machine Learning Algorithms. Similarly, the test set is equally preprocessed and vectorised then used to test the model. If a tweet is detected as hate or offensive, it is classified as such and vice versa.

3.1 ALGORITHMS USED

For the research aim to be achieved, different algorithms were used to transform the data, extract features and implement the ensemble model. They are listed and explained subsequently.

3.1.1 TERM FREQUENCY — INVERSE DOCUMENT FREQUENCY (TF-IDF)

This is a technique used to transform a text or document into vectors. It is equally referred to as Word2Vec or Doc2Vec. It quantifies words in a text or document to determine their overall impact to the document in question. It is used in text mining.

$$TF - IDF = TF * IDF \quad (1)$$

Where = TF = Term Frequency and IDF = Inverse Document Frequency.

TERM FREQUENCY is the measure of the frequency of a word in a document. It is given by:

$$tf(t, d) = (\text{count of } t \text{ in } d)/(\text{number of words in } d) \quad (2)$$

where d is the document or text and t is the term.

DOCUMENT FREQUENCY (DF) is the number of documents d in which a term/word appears in, given by;

$$df(t) = \text{frequency of occurrence of } t \text{ in } N \text{ documents} \quad (3)$$

INVERSE DOCUMENT FREQUENCY (IDF)

This is the inverse of the document frequency. IDF is a measure of the informativeness of term/word t in a document (Scott Williams, 2019 B.C.E.).

$$\text{idf(t)} = N/\text{df} \quad (4)$$

where N is the total count of corpus

corpus = the total number of document sets

When a term that isn't in the vocab appears during the query, the df will be 0. We smooth the value by adding 1 to the denominator because we can't divide by 0 (Scott Williams, 2019 B.C.E.)

$$\text{idf(t)} = \log\left(\text{N}/(df+1)\right) \quad (5)$$

Finally, if we take the multiplicative value of TF and IDF, we get the TF-IDF score in equation (6).

$$\text{tf} - \text{idf(t, d)} = tf(t,d) * \log\left(\text{N}/(df+1)\right) \quad (6)$$

3.1.2 SUPPORT VECTOR MACHINE (SVM)

SVM is a machine learning algorithm which employs the supervised approach for finding solution to classification and regression problems. It is, however, mostly employed to solve classification problems. Each and every data item or feature is mapped as a point in n-dimensional space (where n is the total number of features in the dataset). Each feature's value is the value of a certain coordinate in the SVM algorithm. SVM maps the input vectors into a high-dimensional plane constructing a maximal hyperplane to separate each class.

In this work, three SVM models with varying kernels (linear, polynomial and radial basis function(rbf) were used in the ensemble model. Each of the models performed well individually with linear kernel having the lowest accuracy of 94%.

3.1.3 K-NEAREST NEIGHBOUR (KNN)

This is a non-parametric algorithm used in classification developed by Evelyn Fix and Joseph Hodges in 1951. The output of this algorithm is a class membership. The plurality of votes from the neighbour of an object determines its class. If the k nearest neighbour of an object belong to class A, that object is assigned class A of its k nearest neighbour.

In this work, of all the models used in the ensemble model, KNN had the worst performance of 88% individually but being in the ensemble, the final performance of the models is enhanced.

3.1.4 LOGISTIC REGRESSION

This machine learning algorithm used for classification problems is based on the probability concept. Logistic regression is a linear regression algorithm that uses a more complex cost function known as the logistic or sigmoid function instead of a linear function. The sigmoid function limits the output of the function between 0 and 1 thereby perfect for classification.

3.1.5 MAJORITY VOTING ENSEMBLE MODEL

This model improves performance of classifiers by training and combining predictions from sub-models to solve same classification problem. A meta learner is used to combine the prediction votes from sub-models and make the final prediction based on the majority of predictions by the individual sub-models, thus improving the classification accuracy by reducing the variance in predictions made by the sub-models. The ensemble combines the individual predictions from the models and make the final prediction based of the majority of the predictions by the models. The performance of the individual models in the final model is improved by reducing the prediction variance and the biasness.

The algorithm for the overall model is given in figure 3 while the flowchart is shown in figure 2. and the internal implementation of the ensemble model is depicted in figure 2. Three SVMs with linear, rbf and polynomial kernels were used with the addition of KNN and Logistic Regression. To avoid a tie, five sub-models were used in this work.

Start
Input: *x* ← tweets
Y ← Class
Output: Prediction – Class of a tweet
Foreach *tweet in tweet matrix* **do**
Foreach term/word in tweet **do**
Calculate_TF-IDF (*tweet*)
score ← *term/word*
end
append_term_score_to_matrix(term, score)
end
convert_TF-IDF_matrix_to_sparse_form
fit_data_to_train_models (*matrix, Y)*
ensemble_prdeictions_from_models
get_final_class_prediction
stop

Figure 3. The model algorithm

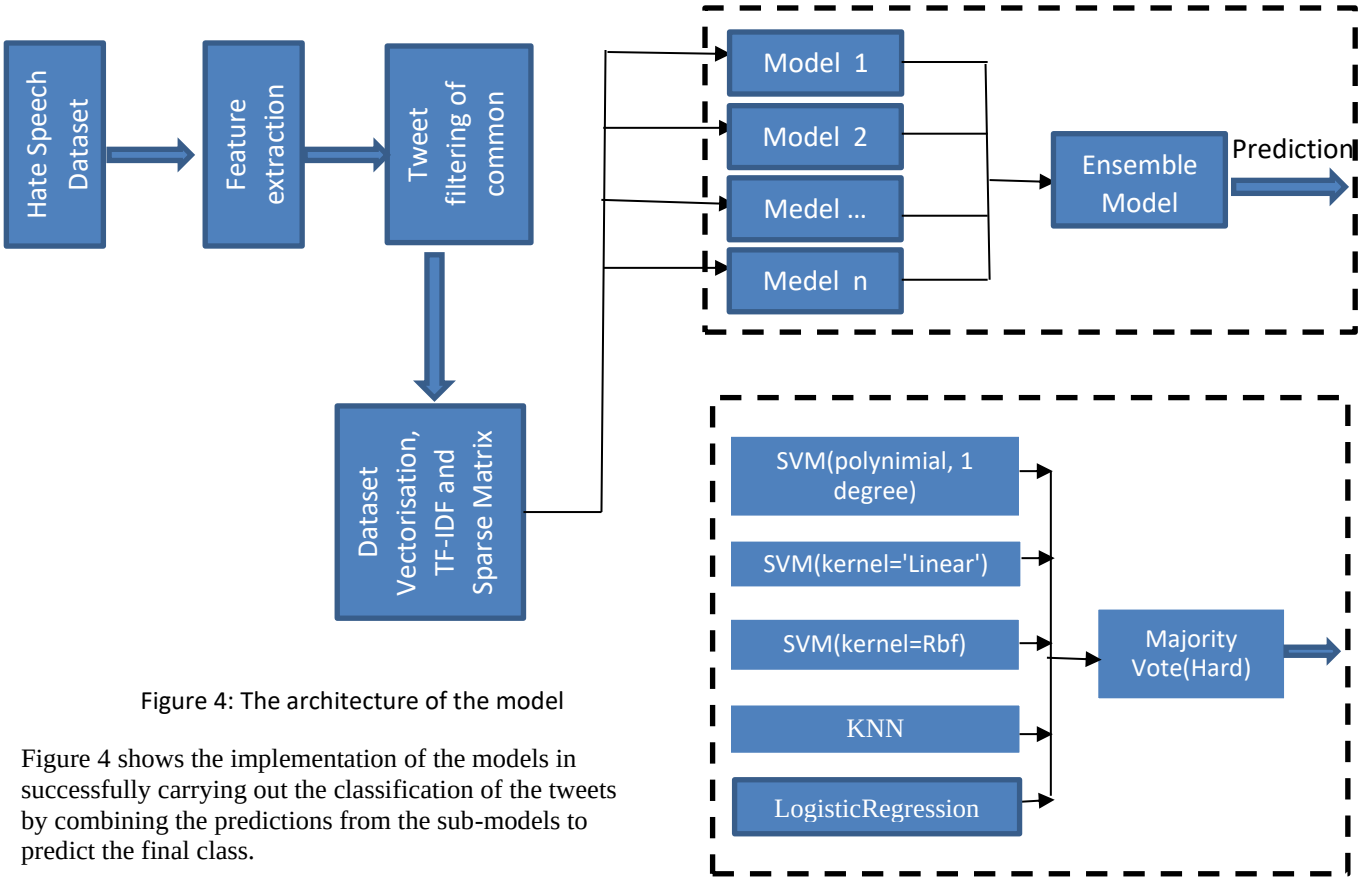

Figure 4: The architecture of the model

Figure 4 shows the implementation of the models in successfully carrying out the classification of the tweets by combining the predictions from the sub-models to predict the final class.

EVALUATION MATRICES

For evaluation purposes, the confusion matrix in Table 3 was used. False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN) were computed by comparing predicted and true values of the class label. Then, precision (P), recall (R), accuracy, and F1-measures were equally used in this work.

TABLE 3: CONFUSION MATRIX OF RESULTS

	Predicted positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

TP is the correct positive prediction.

FP are positive predictions that are incorrectly predicted to be positive but are not.

FN are the negative predictions that are not actually negative.

TN are the negative predictions that are actually negative.

Precision (P): This is the proportion of positively predicted values which are actually positive.

$$P = TP/(TP + FN) \quad (7)$$

Recall (R): It is the proportion of actual positives which are predicted positive.

$$R = TP/((TP + FP) \quad (8)$$

Accuracy: This is the proportion of correctly classified instances.

$$\text{Accuracy} = (TP + TN)/((TP + TN + FP + FN + FP)) \quad (9)$$

F1-Measure: This is the precision and recall harmonic mean.

$$F1 - Measure = 2 \times ((P \times R))/((P + R)) \quad (10)$$

Error rate: this is the number of all incorrect predictions divided by the total number of occurrences in the dataset, often given as 1-Accuracy.

4 RESULTS AND DISCUSSION

Table 4 compares the performance of our model; the majority vote ensemble model with those of other researchers. While some researchers reported only the F-measure score of their models, others reported only their accuracies. The model developed in this work has shown better performance as clearly depicted in table and subsequently in the figures 5. Ensemble model reduces the biasness of the individual models if they were trained as stand alone. Equally, the variance in the individual predictions of the models is improved upon by the ensemble model.

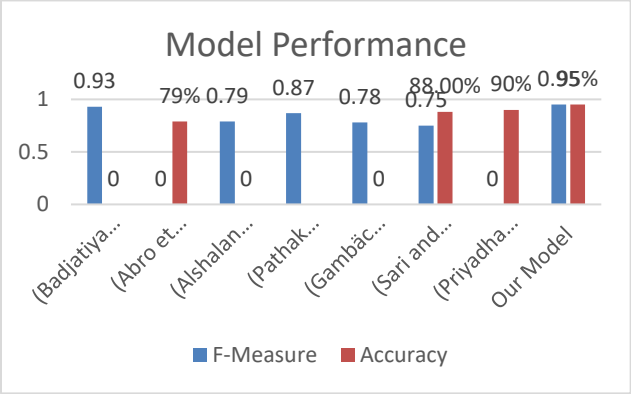

Figure 5: The performance of reviewed works.

TABLE 4: THE PERFORMANCE COMPARISM OF DIFFERENT MODELS

Author	Classifier	Precision	Recall	F-Measure	Accuracy
(Badjatiya et al., 2017)	LSTM+Random Embedding+G	-	-	0.93	-
(Abro et al., 2020)	TF-IDF+SVM	-	-	-	79%
(Alshalan and Al-Khalifa, 2020)	RNN+CNN		-	0.79	-
(Pathak et al., 2020)	TF-IDF+CNN			0.87	
(Gambäck and Sikdar, 2017)	CNN	-	-	0.78	-
(Priyadharshini, 2020)	Text mining + NLP	-	-	-	90%
(Sari and Ginting, 2019)	Multinomial Logistic Regression	80.02	82%	-	88%
Our Model	TF-IDF+ Majority vote (SVM+KNN+LR)	95%	95%	0.95	95%

Table 4 displays the results from authors who have worked on hate speech detection with the various techniques/models used. After transforming the text vectors from the tweets obtained using TF-IDF, majority voting ensemble model of Logistic Regression (LR), K-Nearest Neighbour (KNN) and SVM were used to train on the optimized dataset and an accuracy of 95% and F1-Measure of 0.95 were obtained. The model outperformed the classifications from the reviewed literatures with (Priyadharshini, 2020) getting 90% accuracy and

(Badjatiya et al., 2017) getting a score of 0.93 F1-Measure .

5 CONCLUSION

Online hate speech has been a source of concern to various governments all over the world even to the United Nations as it has been on the rise and is capable of destabilizing regional and world peace. Hate speeches are used to intimidate, abuse and promote violence against person or a group of people targeting their sexual orientation, race, gender, and socio-political affiliations which can lead to physical harm hence the need to identify those hateful speeches as early as possible. This work has successfully used TF-IDF and majority voting ensemble model to identify hateful and offensive tweets with an F1-Measure of 0.95 and a classification accuracy of 95%.

REFERENCE

Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491. https://doi.org/10.14569/IJACSA.2020.0110861

Alshalan, R., & Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences (Switzerland)*, *10*(23), 1–16. https://doi.org/10.3390/app10238614

Badjatiya, P., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets Pinkesh. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*(2), 427–431. https://doi.org/10.18653/v1/e17-2068

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676

Gambäck, B., & Sikdar, U. K. (2017). *Using Convolutional Neural Networks to Classify Hate-Speech*. *7491*, 85–90. https://doi.org/10.18653/v1/w17-3013

Hate-Speech-Data-Analysis | Kaggle. (2020). https://www.kaggle.com/tarushi89/hate-speech-data-analysis

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Pariyani, B., Shah, K., Shah, M., Vyas, T., & Degadwala, S. (2021). Hate speech detection in twitter using natural language processing. *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, *January 2018*, 1146–1152. https://doi.org/10.1109/ICICV50876.2021.9388496

Pathak, V., Joshi, M., Joshi, P., Mundada, M., & Joshi, T. (2020). KBCNMUJAL@HASOC-Dravidian-CodeMixFIRE2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. *CEUR Workshop Proceedings*, *2826*, 351–361.

Priyadharshini, G. (2020). Detection of Hate Speech using Text Mining and Natural Language Processing. *International Journal of Engineering Research & Technology (IJERT)*, *9*(11), 2018–2021. www.ijert.org

Sari, P., & Ginting, B. (2019). *Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method*. 105–111.

Scott Williams. (2019 B.C.E., February 15). *TF-IDF from scratch in python on real world dataset. | by William Scott | Towards Data Science*. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

United Nations. (2019). United Nations Strategy and Plan of Action on Hate Speech. *United Nations Report*, *May*, 1–5.