ARID ZONE JOURNAL OF ENGINEERING, TECHNOLOGY & ENVIRONMENT
AZOJETE June 2024. Vol. 20(2):491-508
Published by the Faculty of Engineering, University of Maiduguri, Maiduguri, Nigeria.
Print ISSN: 1596-2490, Electronic ISSN: 2545-5818
www.azojete.com.ng

HATE SPEECH IDENTIFICATION IN WEST AFRICA, USING MACHINE-LEARNING TECHNIQUES

A. A. Sosimi[1], O. Ipinnimo[1], C. O. Folorunso[1]*, B. A. Adim[1], E. Onoyom-Ita[2]

[1]*Department of System Engineering, University of Lagos, Lagos, Nigeria*
[2]*Department of Electrical and Electronics Engineering, University of Cross River State, Idim Ita, Calabar, Nigeria*

**Corresponding author's email address: cfolorunso@unilag.edu.ng*

ARTICLE INFORMATION

Submitted 19 January, 2024
Revised 25 February, 2024
Accepted 25 February, 2024

Keywords:
Hate speech
machine learning
natural language processing (NLP)
social media

ABSTRACT

The tremendous rise in social media usage over the past ten years has resulted in an extraordinary spike in hate speech activities in West Africa. Because of this, her unity is constantly in peril. This study combines relevant natural language processing techniques and machine learning classifiers to create a hate speech detection model using hate speech from West African countries, including Pidgin English, on Twitter, now 'X'. The data was pre-processed using word embedding, CountVectorizer, and Term Frequency-Inverse Document Frequency (Tf-Idf) to extract useful characteristics from the cleaned dataset. Five machine learning classifiers were used to train the dataset, these include Logistic Regression (LR), Naïve Bayes (NB), Extreme Gradient Boost (XGBoost), Deep Neural Network (DNN), and Bidirectional Long and Short-Term Memory (Bi-LSTM). The Bi-LSTM fitted on Global Vectors (GloVe) embedding produced the best experiment results, with an accuracy of 92% and an F1-Score of 83% when assessed on a test set. The machine learning models generally demonstrated strong performance on test data, suggesting that they had internalised the knowledge from the training set and could use it to analyse new data.

1.0 Introduction

Hate speech is becoming a bigger problem across the globe, especially in West Africa. The widespread use of social media and other digital platforms has made it easier for hate speech to spread quickly, escalating tensions and endangering national and local unity. Hate speech has been shown to widen differences between various ethnic, religious, and social groups and to inspire violence in West Africa as well as other regions of the world (Laub, 2019). This phenomenon can be especially unstable in nations where the population is heterogeneous and intergroup conflicts have historically occurred. In West Africa, efforts to combat hate speech frequently combine legislative actions, such as passing legislation prohibiting it, with educational programmes designed to foster tolerance, understanding, and respect for variety. Furthermore, community-based initiatives and forums for discussion can be extremely important in promoting improved social cohesiveness and lessening the negative effects of hate speech. Coordination of reactions to hate speech and the encouragement of regional collaboration in tackling this issue may also be the responsibilities of regional organisations such as the Economic Community of West African States (ECOWAS) (Daka, 2023).

In recent years, hate speech in West Africa has intensified tensions and posed a serious threat to unity in the continent (Akanji, 2017). This has prompted the government to think about

passing laws that will punish anyone discovered to have used hate speech (Asogwa and Ezeibe, 2022). For instance, in Nigeria, the National Commission for the Prohibition of Hate Speech was created to assist with investigations and criminal prosecutions (Independent National Commission, 2019). It is much simpler to keep an eye on the hate speech that appears in traditional mainstream media like television, radio, and print than it is to keep an eye on online, on sites like social media and microblogging services. This is mostly caused by the significant amount of daily content produced in internet media that needs to be checked.

Global efforts are also being made to address online hate speech as it develops and broadens. The 2019-launched United Nations Strategy and Plan of Action against Hate Speech serves as a prime illustration (Machirori, 2022). However, security authorities were given the authority to monitor the conversations and posts of well-known social media users as part of the process of holding hate mongers to account due to the escalating incidences of online hate speech and the associated concerns it poses to national security (Opusunju, 2018). Hate speech is proliferating alongside the growth of online information. Unfortunately, hate speech is nothing new in our culture. However, social media and other online communication tools have increased the prevalence of hate speech to alarming levels, fuelling a range of hate crimes and other social evils. Social media's anonymity and movement make it difficult to identify suspects since some users hide under aliases and bogus addresses, which makes it easy for them to procreate and disseminate hate messages.

Social media on the other hand, refers to a new generation of tools that individuals can use to interact with information and share it. These technologies include computers, smart phones, and text messaging capabilities on mobile devices (Auwal, 2018). Applications exist that naturally create interactive connections between people and information.

So many researches have been carried out on hate speech. For instance, (Asogwa and Ezeibe, 2022) looked at how state regulation of hate speech affects the development of sustainable democracy in Africa. They argued that state regulation of hate speech presents an opportunity for fostering order, advancing national cohesion, reducing hate speech, and promoting inclusive governance for all, regardless of age, sex, disability, race, ethnicity, origin, religion, and economic status. This was achieved by using the qualitative dominant mixed methods approach and data generated from Nigeria and Kenya. The article concluded that in order to reduce hate speech and the harms it causes as well as to advance sustainable democracy in Africa and beyond, it is important to enact laws against it in addition to alternative non-legal, dialogue-based, egalitarian, voluntary approaches.

In addition, Sunday (2020) addressed extremism, hate speech, and false news in the social media age in West African communities. He considered the non-Sahel region of West Africa, which includes Benin, Cote d'Ivoire, Ghana, Guinea, Liberia, Nigeria, Sierra Leone, and Togo, as case studies. He noted that unscrupulous propagandists and mischief-makers are using extreme ideas and actions, fake news, and hate speech to spread false, disparaging, and divisive information that impedes the socio-political and economic development of the region and to incite fear and ethnic and religious crises. In essence, the study suggested potential solutions to the dis-infodemic as it impacts the chosen nations. The research evaluates how active social media users cannot avoid the direct effects and influence of social media, and by extension, mass media, using the notion of magic bullet to construct the analyses. The study employed the quantitative approach to produce a list of doable actions to combat the disinformation epidemic. These actions include creating a fact-checking algorithm, requiring media professionals with training to combat misinformation with facts, and requiring media information literacy to be taught in schools.

Moreso, Chukwuebuka (2015) investigated the impact of hate speech on violence in Nigeria before, during, and following elections. The study was based on interviews with leaders of civil

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

society organisations, student leaders, religious leaders, and traditional leaders who were chosen from Nigeria's six geopolitical zones. The information gathered from the interview was enhanced by secondary data and observation. Discourse and content analysis were also used. The author made the following claim that political leaders and their supporters who are based on ethnicity, region, or religion are mostly responsible for hate speech in Nigeria. As long as hate speech helps Nigerian political leaders seize and hold onto power, they ignore its inflammatory characteristics. The paper suggested that simply denouncing hate speech verbally is insufficient. However, the Nigerian rules governing public speech and electoral campaigns should be enforced against persons and organisations that violate them, according to the Independent National Electoral Commission and other civil society organisations.

Furthermore, Ojakorotu *et al.* (2019) investigated the degree to which the dissemination of hateful and harmful utterances on social media poses a threat to Nigerian security. They presented a precise conceptual clarification of the phrases "hazardous" and "hate speech" due to the conceptual ambiguity of hate speech. Additionally, it was stated that dangerous statements put people and organisations at risk of violence. They found out that hate speech can divide and belittle people, and that it can also set off dangerous reactions that may eventually result in bloodshed. They concluded that hate speech and harmful communication should be restricted in Nigerian online because they both encourage unhealthy competition and make victims and target groups feel less safe.

Also, Ngwainmbi (2021) discussed the sociology of hate speech, which makes the argument that hate speech is a by-product of racism, which took over our world when humans first entered the cosmos. It examines various racial philosophers' conceptions of national space and identifies modern media as the primary global advocate of Caucasian nationalism. They demonstrated how anti-other attitudes and a certain egocentric pro-nationalist agenda are fostered among Caucasian groups in the Global North, particularly in Germany and America, using social media data and Information Technology (IT) platforms. convinced by Sigmund Freud's claim that unconscious conflicts nearly usually drive violent behaviour in people. The way that people with diverse racial, ethnic, sexual, gender, or other combinations are portrayed in the media is explained by the concepts of reverse psychology and group communication. They identified global initiatives to prevent hate speech and reveals how media firms and governments can manage hate speech. They concluded that hate speech is a threat to world peace and demonstrated that counter speech is a better approach of stopping hate speech's dignitarian harm

Furthermore, Rasaq *et al.* (2017) used a critical discourse analysis method to demonstrate how common hate speech is in Nigeria, especially in light of the country's 2015 general election, which may have resulted in a wave of violence. It is evident from the debate that hate speech served as the campaign's main focus and tool. As a result, the parade of hate speech found in a number of the publications under analysis demonstrated how politicians exploited the media to incite violence and hatred between various ethnic and political groupings both in the run-up to elections and in everyday life. They concluded that in order to avoid supporting hate speech, media organisations should constantly analyse the messages that politicians convey, evaluate the language that they use, carefully consider the facts and claims that they make, and assess the intention and potential influence on society.

In addition, Garbe *et al.* (2023) applied structural topic modelling on a corpus of news items worldwide (N = 7,787) mentioning hate speech and fake news in 47 African nations to estimate the significance of debates of legislative and technological approaches to content regulation. they discovered that, in particular, governments with less regard for media freedom and less legislative restraints tend to prioritise discussions of technology strategies. Their findings highlighted the need for a deeper comprehension of how regime-specific factors influence

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

regulatory decisions and indicate that the state is the primary factor influencing content regulation among African nations.

Moreover, Elliott *et al*. (2016) presented a key concept paper on hate speech which discusses the following: a comprehensive talk about the relationship between hate speech and freedom of speech; a quick look at worldwide laws and definitions of hate speech; a quick look at hate speech in the Media Conflict and Democratisation (MeCoDEM) project countries which include: South Africa, Kenya, Serbia, and Egypt.

1.1Review of hate speech in some West African countries

1.1.2. Hate Speech in the Republic of Benin

In the Republic of Benin, using hate speech is a serious offense, and there are laws in place to penalise offenders. Although the country's Constitution provides freedom of speech and the press, it also outlaws hate speech and violent provocation. Benin's government enacted legislation in 2018 making hate speech illegal, with jail term and fines as possible penalties. A statement or action that "expresses or incites hatred, discrimination, or violence against a person or group of persons on the basis of their origin, ethnicity, religion, race, gender, or sexual orientation" is referred to as hate speech under the law (Sunday, 2020).

1.1.3 Hate Speech in Côte d'Ivoire

In the Republic of Cote d'Ivoire, hate speech is an important problem. More than 60 different ethnic groups make up the population of the nation, and in the past, conflicts between these groups have resulted in violent outbursts. Hate speech has increased in frequency during the past few years, especially on social media platforms. The Ivorian government has taken action in response to the threat that hate speech poses. A law was established in 2019 that makes hate speech and other types of hate crimes illegal. If someone is found guilty of using hate speech, the law offers prison terms and penalties (Sunday, 2020).

1.1.4 Hate Speech in Ghana

In Ghana, hate speech is a significant problem that can result in violence, discrimination, and other types of harm. The freedom of expression is guaranteed by Ghana's constitution; however, it is not unrestricted and may be limited to protect others from harm. The Criminal Offences Act, 1960 (Act 29) and the Internet Transactions Act, 2008 (Act 772) both make certain forms of hate speech illegal, and the Ghanaian government has taken efforts to combat it. Additionally, the National Media Commission (NMC) has the authority to impose sanctions on media outlets that disobey its standards on responsible journalism (Ghana News Agency, 2011).

1.1.5 Hate Speech in Nigeria

In Nigeria, hate speech is a severe problem that has persisted for a long time. Any phrase or conduct intended to incite violence or discrimination against a particular person or group of people based on their identity, such as their race, religion, ethnicity, or sexual orientation, is referred to as hate speech. Political, religious, and ethnic unrest in Nigeria has generated hate speech, which has also been amplified by the growth of social media outlets. Politicians have used hate speech to rally their supporters and further their political interests, while religious and ethnic leaders have employed it to incite hatred and hostility among their adherents (Williams, 2020).

The Hate Speech Bill, which makes hate speech illegal and imposes severe penalties on offenders, was passed by the Nigerian government in 2019 as part of its efforts to combat hate speech in the nation. However, some have criticised the measure for having the ability to suppress free speech and for having ambiguous definitions of what constitutes hate speech. Human rights organisations and civil society organisations have also been trying to promote tolerance and respect for diversity as well as to increase public awareness of the hazards of

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

hate speech. There have also been requests for media outlets to report on delicate topics responsibly and without sensationalising them (Asogwa and Ezeibe, 2022).

In summary, in all the literature reviewed, no work was carried out on the use of a machine learning classifier and pertinent natural language processing techniques to build a hate speech detection model. Hence, the objective of this paper is to present a hate speech identification system using machine learning. Hate speech was identified on Twitter, now 'X', solely taking into account tweets using hate words that are conveyed in West African English and Pidgin English. The use of slang, memes, audio, and video in tweets was not taken into account. A new dataset was created between November 2019 and April 2020, by gathering tweets from Twitter that contained frequent hate speech used in West Africa. The tweets were then divided into two categories: hate and non-hate speeches.

2.0 Materials and Methods

This study focuses on combining relevant natural language processing techniques and machine learning classifiers to create a hate speech detection model using hate speech from West African countries, including Pidgin English, on Twitter. The flow diagram for the methodology is presented in Figure 1.

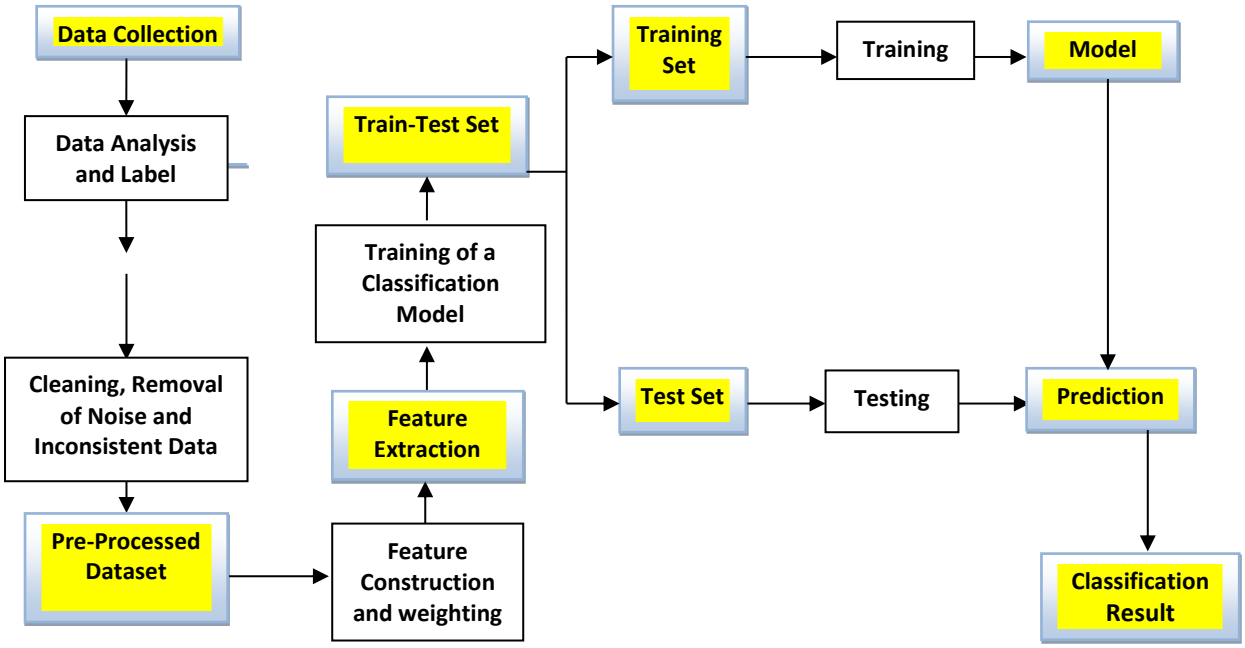

Figure 1: Methodology Flow Process

2.1 Dataset Description/Data Collection

Using Twitter's API, the dataset used in this study was generated from Twitter. The tweets were selected from Twitter users who used terms from PeaceTech Lab's list of terms associated with hate speech in their replies between November 2019 and April 2020. This was accomplished by first creating a Twitter developer account, after which an API ID and an API TOKEN were obtained in order to scrape the Twitter API. The generated data was saved in Microsoft Excel format. Data cleaning and analysis was carried out to get a clean dataset used for this study.

2.2 Pre-Processing

Text data is unreliable and usually contains a lot of noise. Removing irrelevant information from the data is a crucial step in order to reduce the possibility of dealing with noisy or inconsistent data. Information such as punctuation, special characters, numerals, and terms with minimal meaning within the text context are removed in order to prepare the raw text for mining. Pre-processing's main objective is to prepare data for use by machine learning algorithms. A well pre-processed classifier performs better and is more efficient. In addition,

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

it requires less time to train and yields a feature space with higher quality during features extraction. Initially 28,000 tweets were collated but after various pre-processing stages were carried out on the tweets, a total of 20,176 cleaned tweets were generated. The following methods are used in the pre-processing steps:

2.2.1 Removal of Twitter Handles (@user): The Twitter accounts that were concealed as @user because of privacy issues were taken down as they don't say anything about the substance of the tweet.

2.2.2 Removal of Punctuations, Numbers, and Special Characters: Numbers, punctuation, and even special characters were eliminated since they were ineffective in distinguishing between different types of tweets. Hashtags with spaces, however, were kept since they offer some helpful information.

2.2.3 Removal of Short Word: The majority of the lesser terms are not very valuable. Words like "he', "is", "it", and "a", for instance, that had two (2) letters or fewer were eliminated from the data.

2.2.4 Removal of URLs and HTML Tags: Since they don't provide any value and we don't want the classifier to pick up features from these tokens, URLs and HTML tags were eliminated.

2.2.5 Removal of Random Patterns (Randos): Because they are irrelevant for training models, any leftover patterns that may have developed or remained throughout the data cleaning process are likewise eliminated.

2.2.6 Changing all letters to Lowercase: This is to prepare the unprocessed text for extraction.

***2.2.7 Tokenization*:** To protect the tweets, they are divided into individual words or tokens.

2.2.8 Lemmatization: In this case, words with suffixes like "ed", "ing", "ly", "es", "er" and "s" were eliminated. Examples of distinct word variations used in the same context are "kill", "killer", "killed", "kills", and "killing." Reducing the overall number of unique words in our data while retaining a substantial quantity of information is the aim of stemming.

2.3 Feature Extraction Methods

Feature extraction, the process of selecting and/or combining variables to create features, greatly minimises the amount of data that must be processed while accurately and completely characterising the original dataset. Pre-processed data needs to be turned into features in order to be analysed with techniques that minimise dimensionality and superfluous features. Three feature extraction methods were used in this study, they are: Term Frequency-Inverse Document Frequency Vectorizer (TfidfVectorizer), CountVectorizer and Global Vectors (GloVe) embedding.

2.3.1 Term Frequency-Inverse Document Frequency Vectorizer (TfidfVectorizer)

2.3.1.1 Term Frequency (tf): Term frequency is the total frequency of a phrase in a given document. Thus, it is unique to a document and it is represented by Equation (1) (Unnikrishnan, 2021),

$$\mathrm{tf(t)} = \text{Count of instances of term in a document} \qquad (1)$$

2.3.1.2 Inverse Document Frequency (idf): Idf value indicates how common or uncommon it is within the whole corpus of documents. It appears in every document. The normalised *idf* value of a word will approach 0 if it is frequent and appears in a lot of texts, and it will approach 1 if it is unusual. Equations (2) and (3) shows *idf* formula for default behaviour, that is when *idf* is 'True' and for when *idf* equals "False" respectively (Unnikrishnan, 2021),

$$\text{idf}(t) = \log_e [(1+n) / (1 + \text{df}(t))] + 1 \text{ (when smooth\_idf = True)} \quad (2)$$

$$\text{idf}(t) = \log_e [n / \text{df}(t)] + 1 \text{ (when smooth\_idf = False)} \quad (3)$$

2.3.1.3 Term Frequency-Inverse Document Frequency (tf-idf): The tf-idf value in a document is equal to the aggregate of its *tf* and *idf*. The more relevant the term is in that document, the higher its *tf-idf* value (Unnikrishnan, 2021).

The 'TfidfVectorizer' class of the sci-kit learns package was used in this study to create TF-IDF features extraction methods in order to extract word TF-IDF from unigram to trigram, with a minimum document frequency of three (min_df =3). Unicode, which removes accents from any character, was selected for accent removal (strip_accents ='unicode').

***2.3.2 CountVectorizer*:** A class in scikit-learn creates a matrix of token counts from a set of text documents. It converts textual data into a numerical representation that machine learning algorithms may use (Pratyaksh, 2021). N-gram words spanning from unigram to trigram were extracted using the sci-kit-learns package to execute CountVectorization. Unwanted terms were added to the CountVectorization stop list, which already contained a list of English stop words (stop_words = 'english').

2.3.3 Global Vectors (GloVe) embedding: GloVe learns vector representations in a high-dimensional space with the goal of capturing the semantic meaning of words. Based on the global statistics of word co-occurrences in a sizable corpus of text, these representations are learned. GloVe directly optimises a global objective function that aggregates word co-occurrence statistics over the entire corpus, in contrast to several other word representation techniques (such as Word2Vec). GloVe's main tenet is that one may deduce a word's relationship from the distributional data of its co-occurrences. Since words with comparable meanings are likely to come together in similar circumstances, their vector representations should also be similar (Ganegedara, 2019). The Keras package was used to implement the GloVe embeddings.

2.4 Train and Test set

The machine learning model is trained using 18,158 (90%) data from the dataset. While a 2,108 (10%) data was used to assess how well the trained model performs. The input data and matching target labels make up the test set as well, but the model's predictions based on this data are evaluated for accuracy, precision, recall, F1 score, and other performance metrics by comparing them to the true labels.

2.5 Model

2.5.1 Logistic Regression (LR) Classifier

A logistic function is used as the principal means of designating a binary dependent variable in statistical representations called logistic regressions. LR seeks to establish a relationship between a group of independent variables and the result by applying an appropriate model. The confidence of an outcome in the LR model is estimated using the logistic function Q(x), which is reliant on independent factors. The logistic function, also known as the sigmoid function, produces an output in the interval [0,1], which is frequently expressed as a probability, given real values as input. The outcome demonstrates the validity of the model in

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

the categorization, with values closer to 0 denoting the other class and values closer to 1 designating the first class (Sperandei, 2014). The expression is shown in Equation (4),

$$Q(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

2.5.2 Naïve Bayes Classifier

The Naïve Bayes Classifier is a simple and efficient classification method that is used to quickly build machine learning systems with predictive capabilities. This supervised learning approach is mostly used for text classification using a large-scale training dataset. It is a probabilistic classifier that is used to solve classification issues. It is based on the Bayes theorem, as demonstrated in Eq. (5). Given a sample's set of feature values, the Bayes Theorem generates likelihood to assess the likelihood that the sample belongs to a specific class (Yang, 2018).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

The Bayes theorem states that there is a chance that A will occur when B does. In this case, B is the proof, while A is the hypothesis. It is assumed that the features/predictors in this case are independent. That's why it's referred to be Naïve—the existence of one particular quality does not interfere with the other. The parameters/features of size 'n' are represented by variable B.

2.5.3 Extreme Gradient Boost (XGBoost)

The gradient boosting framework is used by Extreme Gradient Boost (XGBoost). The goal of this supervised learning technique, which is based on decision trees, is to accurately predict a target variable by combining the predictions of several simpler, less accurate models. When it comes to prediction tasks involving unstructured data, artificial neural networks (ANNs) typically beat out all other methods. Nevertheless, decision tree-based algorithms are considered to be very good for small-to-medium structured/tabular data prediction issues. XGBoost is a supervised learning strategy that uses a variety of regularisation techniques along with function approximation through optimisation of particular loss functions (Aydin & Ozturk, 2021). The objective function (also known as the loss function and regularisation) of XGBoost at iteration 't' can be expressed mathematically as shown in Equation (6)

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t=1} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

2.5.4 Deep Learning

Deep Learning is a branch of machine learning that uses artificial neural networks (ANNs) to simulate how the human brain works to solve complex data-driven problems. Artificial neural networks (ANNs) are devices designed to simulate how a human brain might do a certain task or function. The term "deep learning" originated in the early 2000s when advances in computer development and skill training made it possible to train even larger and deeper networks (Biere, 2018). Prior to then, NLP techniques were dominated by machine learning techniques that used linear models that were learned over high dimensional but extremely sparse feature vectors. Non-linear neural networks have shown remarkable effectiveness recently when used with dense inputs. When used to hate speech and related fields like sentiment analysis, deep learning techniques have been shown to yield the greatest results (Hemker, 2018). There are two categories of neural networks: recurrent (Kadlaskar, 2021) and feedforward neural networks (example: Deep Neural Networks) (Kyrykovych, 2020).

2.5.5 Bidirectional Long Short-Term Memory (BI-LSTM)

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

Regular LSTMs are continued into bidirectional LSTMs, which can improve model performance on sequence classification tasks. The extraordinary RNNs known as Long Short-Term Memory (LSTM) networks are able to overcome the drawbacks of conventional RNNs. The most basic building blocks of LSTM networks are memory cells. Three gate units, which extract or input information into the memory cell state and maintain temporal information, are present in memory cells. In situations where every step of the input sequence is available, bidirectional LSTMs train twice on the input sequence instead of just one LSTM. While the second input sequence is trained as the inverse of the first, the first is taught in the original form of the input sequence. This can help the network learn the problem more quickly and thoroughly since it gives it additional context (Aggarwal, 2019).

2.6 Evaluation Metrics

Metrics are often used in evaluating the performance of classification model. These measures values such as True positives (tp) which represent the number of correctly classified positive instances. True negatives (tn) denote the number of correctly classified negative instances. False positives (fp) mean the number of incorrectly classified positive instances, while false negatives (fn) are the same for negative instances. *tp*, *tn*, *fp*, and *fn* help to find the Precision (Equation (7)), Recall (Equation (8)), F_1 Score (Equation (9)) and Accuracy (Equation (10)) of a system (Brownlee, 2020).

$$Precision = \frac{tp}{tp + fp} \tag{7}$$

Precision measures the ability of the classifier to correctly label samples.

$$Recall = \frac{tp}{tp + fn} \tag{8}$$

Recall measures the ability of the classifier to find all the relevant samples.

$$F_1 = 2\frac{Recall\ .Precition}{Recall + Precition} (9)$$

F_1-Score is used for a better overall evaluation of the classifier performance.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{10}$$

Accuracy measures the total correct predictions as a percentage of the total instances.

2.6.1 Confusion Matrix

A confusion matrix is a tool used to gauge how successful a classification programme is. Count values are used to quantify the proportion of accurate and inaccurate forecasts, broken down by class. The classifier's errors and their categories are displayed in the matrix. It is a performance metric for classifying problems in machine learning, where the result can belong to two or more classes. Four distinct combinations of the actual and anticipated values are shown in a table. Two

classes are the study's output.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

(11)

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

3. Results and Discussion

The outcome of the dataset after the pre-processing is shown in Table (1). The two class datasets utilised to build the classification model were Hate Speech (HS), which is the positive class with label 1, and Non-Hate Speech (NHS), which is the negative class with label 0. The experiment was analysed using three different features extractors: GloVe Embedding, Tf-Idf, and CountVectorizer.

Table 1: First Five Rows of Dataset, Before and After Cleaning

b'id'	b'message' Raw Data	b'message' Clean Data	b'polarity'
1	@user @user @user Problem is you\'re not important to Arabs as USA / Israel, the same people u call "Arne".\n\nSuch a shame Arabs considers Arnes worthy of their respect than "Northern Nigeria"	problem youre not important arabs usa israel the same people call arne such shame arabs considers arnes worthy respect than northern Nigeria	1
2	DSS Arrests Creator Of Video Depicting A Fake Buhari Marriage https://t.co/BDqrWX6EIV #TheTrent https://t.co/AtFU2Ps0jq	dss arrest creator video depict fake buhari marriage	0
3	@user May God punish you and your almajiri generations for reminding us of this pain of brain dead man.	may god punish you and your almajiri generation for remind this pain brain dead man	1
4	"HOW are there up to FIVE 'truckfuls' of Almajiri in Cross River State??? https://t.co/F3Wp5FMvR1"	how are there five truckful almajiri cross river state	0
5	"# fulani graduated and train # criminals the best thief all over the world stealing BIAFRANS money # NigeriaTheZoo # NigeriaTheZoo # NigeriaTheZoo # NigeriaTheZoo don't forget to ask Google who country is a zoo # NIGERIA @user @user @user @user @user https://t.co/o73fXgZpux"	fulani graduate and train criminals the best thief over the world steal biafrans money nigeriathezoo nigeriathezoo nigeriathezoo dont forget ask google who country zoo Nigeria	1

Five different machine learning models were used, namely: Logistic Regression (LR), Naïve Bayes (NB), Extreme Gradient Boost (XGBOOST), Deep Neural Networks (DNN), and Bidirectional Long Short-Term Memory (LSTM). The performance metrics used in this study to evaluate the models are accuracy, precision, recall, F1 score and the confusion matrix. The accuracy of all models is shown in Table 2, while the precision scores are presented in Fig. (2). In addition, the recall and F1 score are presented in Table 3 and Fig. (3) respectively. While the confusion matrix of each model according to the feature extractor used to fit the model are shown in Figures (4) to (8) respectively.

Table 2: Accuracy Scores of all the models

MODELS	TF-IDF (%)	COUNTVEC (%)	GLOVE (%)
LR	86	87	-
NB	83	68	-
XGBOOST	87	87	-
DNN	90	90	-

Bi-LSTM	-	-	92

According to Table 2, Bi-LSTM with GloVe Embeddings produced the best accuracy of 92% followed by DNN with 90% on both Tf-Idf and CountVectorizer feature extractors. Additionally, XGBOOST performed well on Tf-Idf vectorizer and CountVectorizer, scoring 87% respectively. With the CountVectorizer feature extractor, NB obtained the lowest accuracy score of 68% out of all the states.

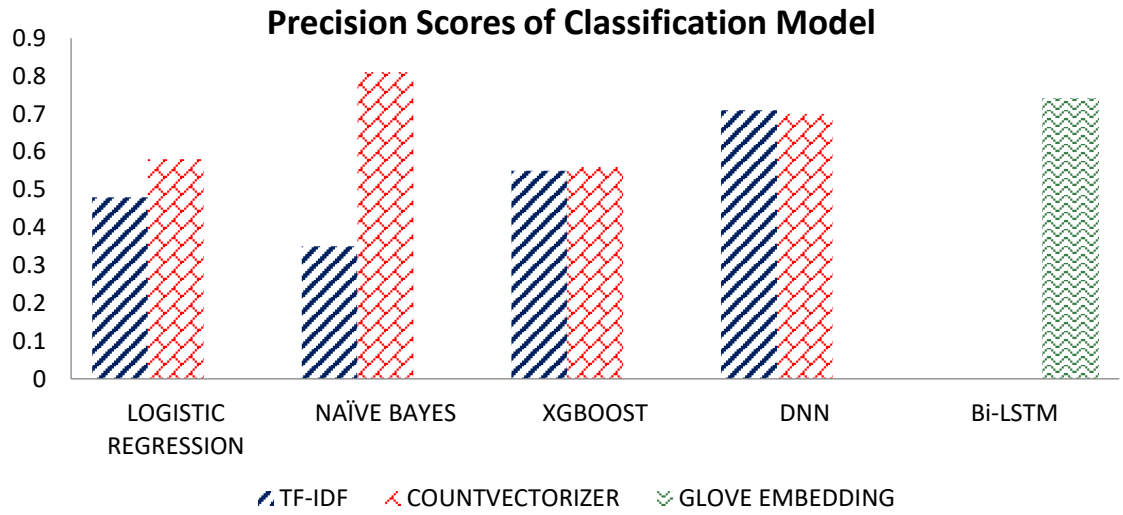

Figure 2: Precision Scores for all the models

With a precision score of 0.81 the NB model on CountVectorizer outperformed the Bi-LSTM model, which came in second with 0.74 on GloVe embedding as shown in Fig. (2). Additionally, DNN scored well on Tf-Idf and CountVectorizer, with precision values of 0.71 and 0.70 respectively. However, NB model obtained the lowest precision scores of 0.35 on CountVectorizer.

Table 3: Recall Scores of all the models

MODELS	TF-IDF	COUNTVEC	GLOVE
LR	0.88	0.82	-
NB	0.81	0.41	-
XGBOOST	0.83	0.84	-
DNN	0.72	0.70	-
Bi-LSTM	-	-	0.95

Table 3 demonstrates that on GloVe Embedding, Bi-LSTM had the maximum recall of 0.95. With respect to the Tf-Idf and CountVectorizer feature extractors, LR also shown strong recall performance, scoring 0.88 and 0.82 respectively, while XGBOOST demonstrated similar recall performance, scoring 0.83 and 0.84 respectively.

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

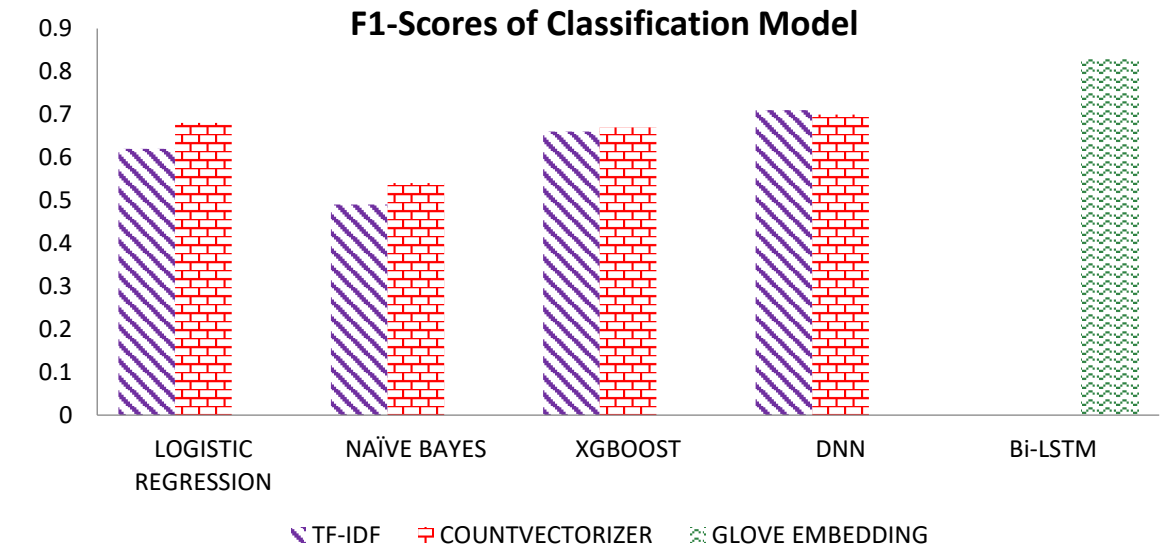

Figure 3: Chart Showing F1-Scores of all the models

From Figure 3, Bi-LSTM on GloVe Embedding achieved the best performance with F1-Scores of 0.83, followed by DNN with F1-Scores of 0.71, and 0.70 on both Tf-Idf and CountVectorizer feature extractors respectively. NB had the worst performance on both Tf-Idf and CountVectorizer feature extractor with 0.49 and 0.54 respectively.

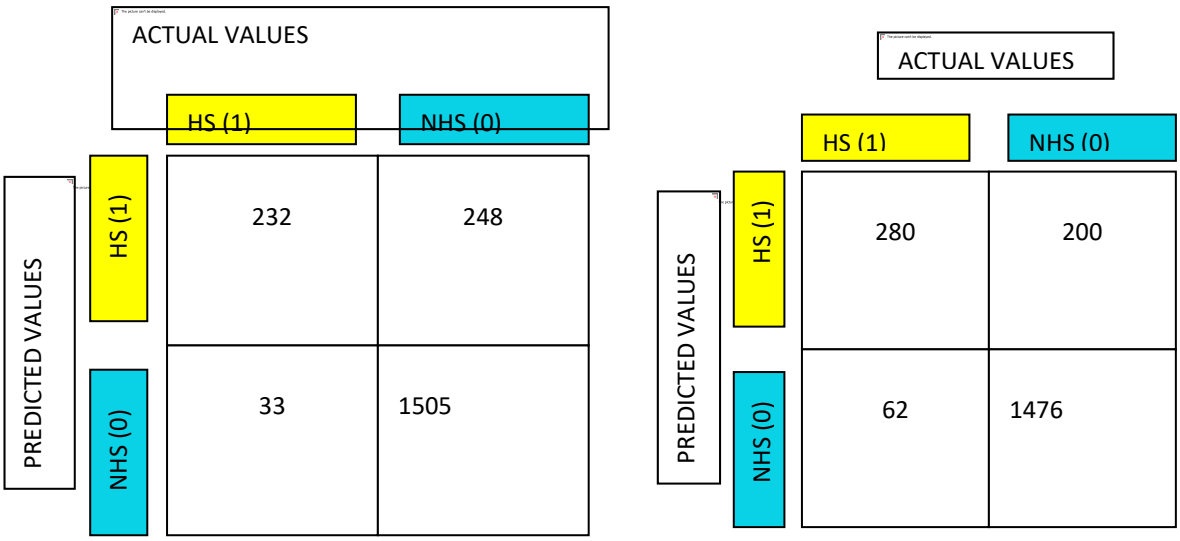

Figure 4: Confusion Matrix Logistics Regression (LR) Model Based on Extracted Features

Figure (4) illustrates the confusion matrix of LR model. LR with Tf-Idf classify 232 samples as HS and 1505 samples as NHS correctly, and also 248 samples of HS and 33 samples of NHS were misclassified. LR with CountVectorizer classify 280 samples of HS and 1476 samples of NHS correctly while 200 samples of HS and 62 samples of NHS were incorrectly classified.

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

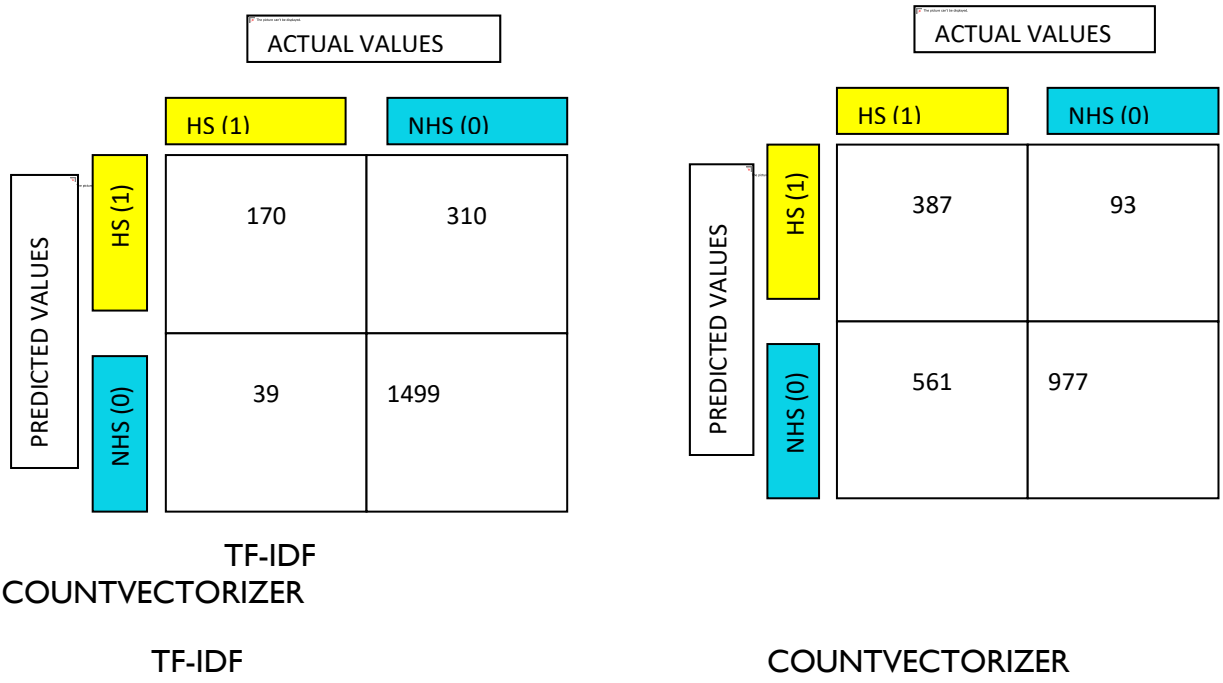

Figure 5: Confusion Matrix Naïve Bayes (NB) Model Based on Extracted Features

Figure (5) illustrates the confusion matrix of NB model. NB with Tf-Idf correctly classify 170 samples as HS and 1499 samples as NHS, and also 310 samples of HS and 39 samples of NHS were wrongly classified. NB with CountVectorizer classify 387 samples of HS and 977 samples of NHS correctly while 93 samples of HS and 561 samples of NHS are incorrectly classified.

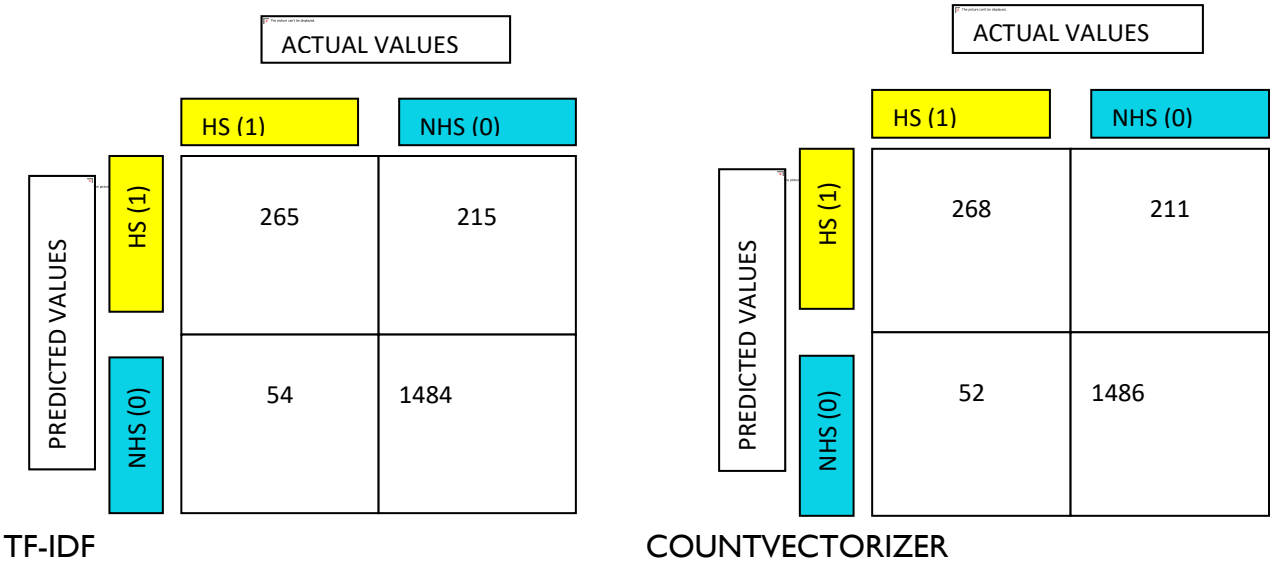

Figure 6: Confusion Matrix Extreme Gradient Boost(XGBOOST) Model Based on Extracted Features

Figure (6) illustrates the confusion matrix of XGBOOST model. XGBOOST with Tf-Idf classify 265 samples as HS and 1484 samples as NHS correctly, and also 215 samples of HS and 54 samples of NHS were misclassified. XGBOOST with CountVectorizer correctly classify 268 samples as HS and 1486 samples as NHS and wrongly classify 211 samples as HS and 52 samples as NHS.

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

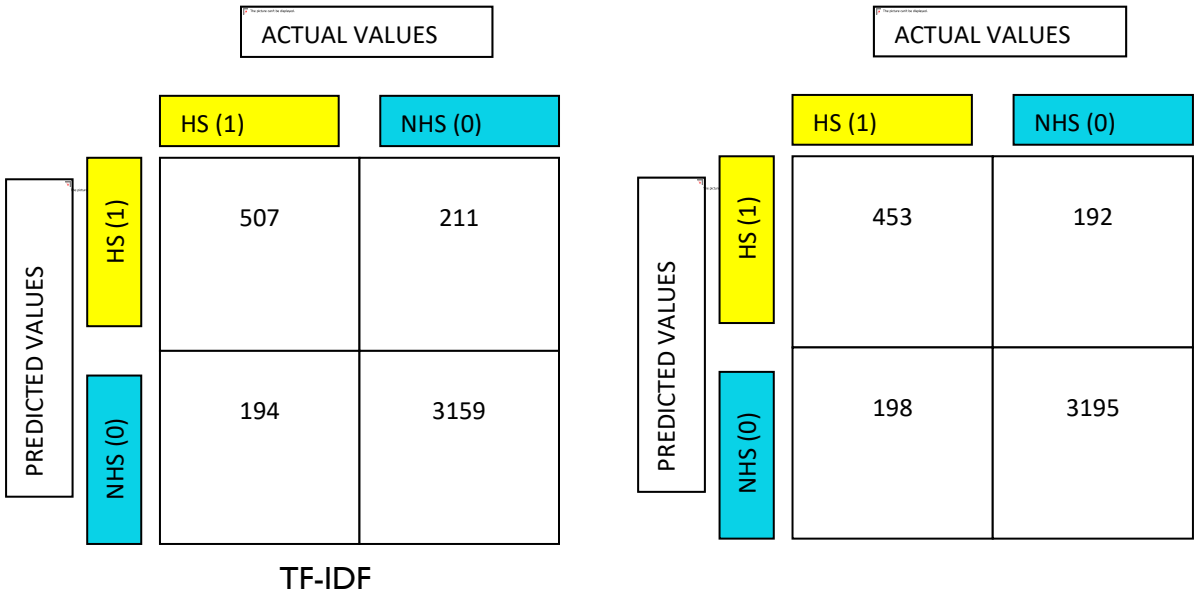

Figure 7: Confusion Matrix Deep Neural Network (DNN) Model Based on Extracted Features

Figure (7) illustrates the confusion matrix of DNN model. DNN with Tf-Idf correctly classify 507 samples as HS and 3159 samples as NHS, and also 211 samples of HS and 194 samples of NHS were wrongly classified. DNN with CountVectorizer correctly classify 453 samples as HS and 3195 samples as NHS and wrongly classify 192 samples as HS and 198 samples as NHS.

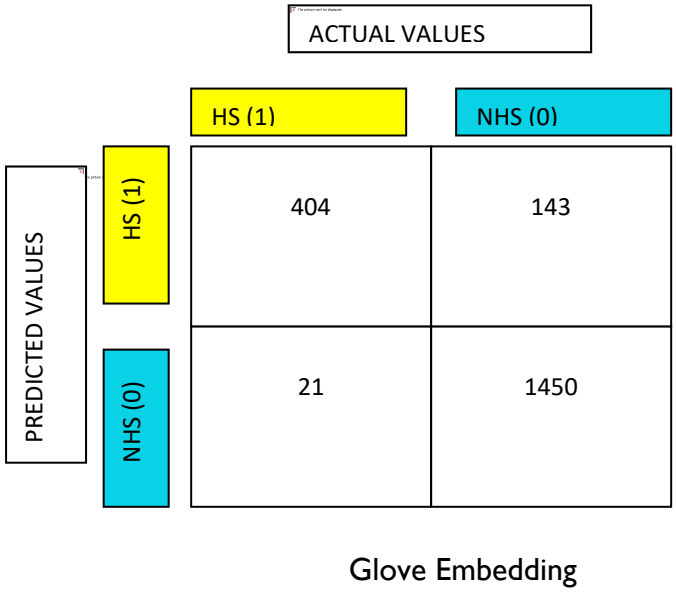

Figure 8: Confusion Matrix Bidirectional Long Short-Term Memory (Bi-LSTM) Model Based on Glove Embedding

Figure (8) illustrates the confusion matrix of Bi-LSTM model. Bi-LSTM with GloVe Embedding classify 404 samples HS and 1450 samples as NHS correctly. 143 samples of HS and 21 samples of NHS were misclassified.

With higher F1-Scores than traditional models (LR, NB, and XGBOOST), deep learning models (DNN and Bi-LSTM) performed better overall. With an F1-Score of 0.68 (68%) on CountVectorizer, the LR model outperformed the other traditional models. XGBoost came in second with F1-Scores of 0.67 (67%) and 0.66 (66%) on both CountVectorizer and Tf-Idf,

respectively. Moreover, it could be inferred that the classical model outperformed the Tf-Idf feature extractor with higher F1 scores when using the CountVectorizer feature extractor. On Tf-Idf and CountVectorizer, LR and XGBOOST, as well as NB on Tf-Idf, all displayed high recall and low precision values. The bulk of the contentious tweets in the dataset were accurately categorised as hate speech by the classifiers, as evidenced by the high recall scores. However, the poor accuracy value suggests that some noisy, inconsistent, and irrelevant tweets were also categorised into the hate speech category by the classifiers as seen in the confusion matrices.

Using the GloVe Embedding feature extractor, the Bi-LSTM model outperformed DNN in terms of deep learning models, yielding a high accuracy of 0.92 (92%) and F1 scores of 0.83 (83%). In contrast, DNN produced an accuracy of 0.90 (90%) on both CountVectorizer and Tf-Idf feature extractors, as well as an F1-Score of 0.71 (71%) on CountVectorizer and 0.70 on Tf-Idf. The reason for this is that neural network models perform optimally on embedding. Furthermore, on large datasets, embedding feature extractors outperform vector representation feature extractors because they retain some of the input's semantics by grouping semantically related inputs in the embedding space, whereas vector representation (Td-Idf vectorizer and CountVectorizer) lose word order and context and is unable to represent semantic similarities between words. High recall and precision rates were also achieved by the Deep Learning Models, suggesting that the majority of the contentious tweets were accurately categorised as hate speech with minimal noise.

3.1 Comparing the Performance of our Dataset with the State-of-the-Art Dataset

It was necessary to compare our dataset with the most recent hate speech dataset in order to assess the performance of the former. To do this, the Analytics Vidhya dataset for hate speech detection presented by Toosi (2019) was used. Out of the 32,000 tweets in the sample, 2242 (7%) were found to contain hate speech. Hate speech in this context refers to tweets that are racist or sexist in nature. The 29720 tweets (93%) of the total, were deemed to be free of hate speech. The same classifiers and feature extractors that were employed to develop our detection model were applied to the state-of-the-art dataset presented by (Toosi, 2019). Table 4 displays the F1-Scores and accuracy results for our dataset (West African Hate Dataset), here is highlighted in bold while the dataset (Toosi, 2019) was asterisked and italised.

Table 4: Results of accuracy and F1-Scores of our dataset and the state-of-the-art dataset

Classifiers	**Feature Extractors**	**Accuracy Scores**	***Accuracy Scores***	**F1-Scores**	***F1-Scores***
LR	TF-IDF	0.86	*0.95*	0.62	*0.43*
	COUNTVEC.	0.87	*0.96*	0.68	*0.65*
NB	TF-IDF	0.83	*0.95*	0.49	*0.45*
	COUNTVEC	0.68	*0.88*	0.54	*0.48*
XGBOOST	TF-IDF	0.87	*0.95*	0.66	*0.54*
	COUNTVEC	0.87	*0.95*	0.67	*0.56*
DNN	TF-IDF	0.90	*0.93*	0.71	*0.54*
	COUNTVEC	0.90	*0.93*	0.70	*0.52*
Bi-LSTM	GLOVE	0.92	*0.84*	0.83	*0.64*

Except the Bi-LSTM classifier, all classifiers outperformed the West African Hate Dataset in terms of accuracy, as shown by the results in Table 4. Unfortunately, because the Hatred Tweets Dataset tends to predict more non-hate speech—the majority class—than hate speech—the minority class—its accuracy cannot be used as a valid metric of model effectiveness. The F1-Scores are a superior metric measure for imbalance categorisation, such as is used in this paper. This is because the F1 scores account for both precision and recall,

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

generating a single score that combines the two issues into a single figure (Brownlee, 2020). Table 4 shows that in terms of overall F-Measure performance across all classifiers, our dataset (West African Hate Dataset) fared better than the State-of-the-Art Hatred Tweets Dataset. This illustrates how valuable the special dataset is for detecting hate speech in West Africa.

4. Conclusion

The use of machine learning techniques to detect and suppress hate speech in West Africa is an important first step towards tackling the growing problems of verbal hostility both offline and online. Through an examination of the region's varied language landscape, cultural quirks, and socio-political circumstances, this research has explored the complexity of hate speech. An encouraging option for the automatic detection of hate speech is machine-learning models tailored to the languages and geographical quirks of West Africa. The availability of extensive and varied datasets that capture the variety of language usage throughout the region is crucial to the efficacy of these models. The continuous endeavour to select datasets that encompass diverse linguistic expressions and contextual elements is crucial to ensure the resilience and applicability of hate speech identification models.

A total of 20176 tweets were manually divided into two categories for this study: hate speeches and non-hate speeches. Of the total number of tweets, 4801 (or 24%) were categorised as hate speech, whereas 15375 (or 76% of the corpus) were classed as non-hate speech. After that, the data were pre-processed to eliminate inconsistent and noisy data in order to make them ready for usage with machine learning techniques. The models were created using LR, NB, XGBOOST, DNN, and Bi-LSTM based on the dataset. Test precision, recall, accuracy, F1-Score, and confusion matrix were used to evaluate the models, which were extracted using the Tf-Idf, CountVectorizer, and GloVe Embedding features. The studies' results were encouraging, demonstrating the efficiency of the created models in identifying anti-West African hate speech. In comparison to the traditional models (LR, NB, and XGBOOST), the deep learning models (DNN and Bi-LSTM) outperformed them. On the test dataset, the LSTM had the best accuracy (92%), as well as the highest F1-Score (83%). In general, the machine learning models performed well on test data, demonstrating that they had learned from the training data and were able to apply that knowledge to fresh data when it was required to be examined using inference on user-generated data. This enables the machine learning models to automatically identify hate speech with a West African origin on social media, which can significantly reduce the harmful consequences of online hate speech on our community.

The dataset can be annotated with additional information than just a binary classification in order to improve the results. The dataset can also be expanded to incorporate fresh writing styles, subjects, and trends. In the future, it might also be looked into how hostile statements are used in non-textual material that users share, such as movies, images, and emoticons. Tf-Idf, CountVectorizer, and GloVe Embedding could be studied in addition to other feature extractors such as BERT, ELMo, FastText, etc. Additional features that could be included in word embedding are user gender, geography, and age. More machine learning classifiers, especially deep learning classifiers, can be researched. To develop a more trustworthy model.

Reference

Aggarwal, R. 2019. Bi-LSTM. Accessed on January 5, 2024, available at Https://Medium.Com/@raghavaggarwal0089/Bi-Lstm-Bc3d68da8bd0.

Akanji, A. 2017. Hate speech: Wrong narrative for national discourse, integration. The Cable. Accessed on January 5, 2024, available at https://www.thecable.ng/hate-speech-wrong-narrative-national-discourse-integration/.

Asogwa, N. and Ezeibe, C. 2022. The state, hate speech regulation and sustainable democracy in Africa: a study of Nigeria and Kenya. African Identities, 20(3): 199–214.

Auwal, AM. 2018. Social media and hate speech: Analysis of comments on Biafra agitations, Arewa youths' ultimatum and their implications on peaceful coexistence. Nigeria Media and Communication Currents, 2(1): 54–74.

Aydin, ZE. and Ozturk, ZK. 2021. Performance analysis of XGBoost classifier with missing data. Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS), 2(2): 1-5.

Brownlee, J. 2020a. Failure of Classification Accuracy for Imbalanced Class Distributions. Machine Learning Mastery. Accessed on January 4, 2024, available at https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

Brownlee, J. 2020b. How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Machine Learning Mastery. Accessed on January 4, 2024, available at https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

Chukwuebuka, C. 2015. Hate Speech and Electoral Violence in Nigeria. Two-Day National Conference: 1–35.

Daka, M. 2023. Challenging Draft Bills and the Role of the ECOWAS Court. Oxford Human Rights Hub. Accessed on January 4, 2024, available at https://ohrh.law.ox.ac.uk/challenging-draft-bills-and-the-role-of-the-ecowas-court/.

Elliott, C., Chuma, W., Gendi, Y. E., Marko, D. and Patel, A. 2016. Hate Speech, Key concept paper. Working Paper. MeCoDEM, ISSN 2057-4002 (Unpublished).

Garbe, L., Selvik, LM., and Lemaire, P. 2023. How African countries respond to fake news and hate speech. Information, Communication & Society, 26(1): 86–103.

GhanaNewsAgency. 2011. NMC chairman worried about hate speech and insults. GhanaWeb TV. Accessed on January 4, 2024, available at https://www.ghanaweb.com/GhanaHomePage/NewsArchive/NMC-chairman-worried-about-hate-speech-and-insults-206587.

Hemker, K. 2018. Data Augmentation and Deep Learning for Hate Speech Detection. Master's thesis, Imperial College London. Accessed on January 4, 2024, available at https://www.semanticscholar.org/paper/Data-Augmentation-and-Deep-Learning-for-Hate-Speech-Schuller/da7b 92f7a3010 f9e1 6ecaeec21f2e03c4cfdb928.

IndependentNationalCommission. 2019. A Bill for an Act to Provide for the Prohibition of Hate Speeches and for Other Related Matters. National Cohesion and Integration Act. Accessed on January 4, 2024, available at https://placng.org/i/wp-content/uploads/2019/12/Hate-Speech-Bill.pdf

Kadlaskar, A. 2021. Time Series Analysis Recurrence Neural Network in Python. Accessed on January 6, 2024, available at Https://Www.Analyticsvidhya.Com/Blog/2021/06/Time-Series-Analysis-Recurrence-Neural-Network-in-Python/.

Kyrykovych, A. 2020. Deep Neural Networks. Accessed on January 6, 2024, available at Https://Www.Kdnuggets.Com/2020/02/Deep-Neural-Networks.Html.

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng

Laub, Z. 2019. Hate Speech on Social Media: Global Comparisons. Council on Foreign Relations. Accessed on January 6, 2024, available at https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

Machirori, F. 2022. Tackling online hate speech in Africa and beyond: "We can't trust Big Tech to abide by its own rules." Association for Progressive Communications. Accessed on January 6, 2024, available at https://www.apc.org/en/news/tackling-online-hate-speech-africa-and-beyond-we-cant-trust-big-tech-abide-its-own-rules

Ngwainmbi, E. 2021. Hate Speech and Caucasian Nationalism in Western Democracies: Flashing Light on the Invisible Ghost. European Journal of Development Studies, 1(3): 33–42. https://doi.org/10.24018/EJDEVELOP.2021.1.3.39

Ojakorotu, V, Nnanwube, EF. and Ani, KJ. 2019. An evaluation of the concepts of dangerous and hate speeches, and their security implications in the social media era Nigeria. Gender and Behaviour, 17(1): 12417–12428.

Opusunju, O. 2018. Nigerian government begins monitoring social media to tame hate speech. ITedgenews.Africa. Accessed on May 7, 2024, available at https://www.itedgenews.africa/nigerian-government-begins-monitoring-social-media-tame-hate-speech/

Pratyaksh, J. 2021. Basics of CountVectorizer. Accessed on May 7, 2024, available at https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c

Rasaq, A., Udende, P., Ibrahim, A., and Oba, LA. 2017. Media, politics, and hate speech: A critical discourse analysis. E-Academia Journal, 6(1): 240-252.

Sperandei, S. 2014. Understanding logistic regression analysis. Biochemia Medica, 24(1): 12–18.

Sunday, O. M. 2020. Disinfodemic" in West Africa Communities: Tackling Extremism, Hate Speech and Fake News in Social Media Age. Resisting Disinfodemic Media and Information Literacy: 191–201.

Ganegedara, T. 2019. Intuitive Guide to Understanding GloVe Embeddings. Accessed on May 7, 2024, available at Https://Towardsdatascience.Com/Light-on-Math-Ml-Intuitive-Guide-to-Understanding-Glove-Embeddings-B13b4f19c010.

Toosi, A. 2019. Sentiment Analysis: Detecting hatred tweets. Kaggle. Accessed on May 28, 2024, available at https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech

Unnikrishnan, CS. 2021. How sklearn's Tfidfvectorizer Calculates tf-idf Values. Accessed on May 28, 2024, available at Https://Www.Analyticsvidhya.Com/Blog/2021/11/How-Sklearns-Tfidfvectorizer-Calculates-Tf-Idf-Values/.

Williams, EE. 2020. Effectiveness of Social Media Platforms in Combating Extremism, Hate Speech, and Fake News in Nigeria. Resisting Disinfodemic Media and Information Literacy: 7–21.

Yang, FJ. 2018. An implementation of Naïve Bayes Classifier. IEEE International Conference on Computational Science and Computational Intelligence (CSCI): 301–306

Corresponding author's e-mail address: cfolorunso@unilag.edu.ng