NIGERIAN JOURNAL OF ENGINEERING RESEARCH UNICROSS (NJERU)

Volume 1, Issue 2, 1st June, 2024

Hate Speech Identification in West Africa, Using Machine-Learning Techniques

Bassey A. Adim [a*], Comfort Folorunso [b], Olufemi Ipinnimo [c], Emmanuel Onoyom-Ita [d]

[abc] Department of Systems Engineering, Faculty of Engineering, University of Lagos.
[d] Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Cross River.

[*]Corresponding author: phera4u@yahoo.com

Received: 7th July, 2024 Reviewed: 20th July, 2024 Accepted: 2nd August, 2024

Abstract:

West Africa has witnessed an unprecedented surge in hate speech activities as a result of the sharp increase in social media usage over the past decade. Her unity is constantly in jeopardy because of the tense climate this has created. The existing efforts by security agencies to monitor hate speech on social media by employing human monitors and site spiders to determine what constitutes hate speech are inadequate. This study suggested using machine-learning techniques to create a detection model as a solution to this issue. In order to extract valuable features from the cleaned dataset, the data was pre-processed using word embeddings, Count Vectorizer, and Term Frequency-Inverse Document Frequency (Tf-Idf). The dataset was trained using five different classifiers: Logistic Regression (LR), Naïve Bayes (NB), Extreme Gradient Boost (XGBOOST), Deep Neural Network (DNN), and Bidirectional Long and Short-Term Memory (Bi-LSTM). The experiment's best result was an accuracy of 92% and an F1-Score of 83% when the Bi-LSTM fitted on GloVe embedding was evaluated on a test set. In general, the machine learning models performed well on test data, indicating that they had learned from the training set and could apply that information to the analysis of fresh data.

Keywords: Hate Speech, Machine Learning, Natural Language Processing (NLP), Social Media, Text Classification, West African Hate Words.

1. INTRODUCTION

The definition of "hate speech" is contested worldwide. Depending on the cultural, political, and legal settings, as well as the opinions and values of certain people or groups, the meaning of hate speech might change. Hate speech is typically defined as communication that targets one person or group of people because of that person's race, ethnicity, national origin, religion, gender, sexual orientation, or other traits. Nonetheless, what constitutes hate speech might differ significantly depending on the setting and the society in which it is used.

Some people think that all hate speech should be outlawed or prohibited, while others think that everyone has the right to their opinions, regardless of how hurtful or divisive they may be. How to strike a balance between the need to safeguard people and groups from the harm brought on by hate speech is a

topic of constant discussion. It's important to keep in mind tha/t hate speech can be illegal and subject to legal penalties in some nations. In other nations, the right to free speech often extends to hate speech, unless it encourages violence or other unlawful activity.

There is no single definition of hate speech that is accepted by all organizations, governments, or individuals. Hate speech is, for instance, described by the United Nations as "any sort of communication in speech, writing, or action that insults or employs derogatory or discriminating terms with reference to a person or a group on the basis of who they are, including their religion, ethnicity, nationality, race, colour, descent, gender, or other identity-related characteristics." Nonetheless, the legal meaning of hate speech may vary from one nation to another [1]. Hate speech encompasses a variety of expressions, including harsh language, that aim to denigrate, dehumanise, or stir up hatred for a particular group or class of people [2]. Any behaviour, expression, speech, writing, or display that might encourage violence or other unfavourable behaviour is prohibited [3].

In recent years, hate speech in west Africa has intensified tensions and posed a serious threat to unity in the continent. This has prompted the government to think about passing laws that will punish anyone discovered to have used hate speech [4]. For instance, in Nigeria, the National Commission for the Prohibition of Hate Speech was created to assist with investigations and criminal prosecutions [5].

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 [6]. However, security authorities were given the authority to monitor the conversations and posts of well-known social media users as part of the processes of holding hate mongers to account due to the escalating incidences of online hate speech and the associated concerns it poses to national security [7].

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, fueling 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.

The term "social media" 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 [8]. Applications exist that naturally create interactive connections between people and information.

II. THEORETICAL ANALYSIS

A. Constitutive Element of Hate Speech.

Hate speech uses derogatory slurs to disparage and dehumanise individuals based on their race, sexual orientation, or other group memberships. Any words, actions, behaviours, writing, or displays that could provoke violence or unfavourable behavior are prohibited. As accurately stated, hate speech is defined as any speech that disparages an individual or a group because of their race, color, ethnicity, gender, sexual orientation, nationality, religion, or any other feature. It mainly denotes inciting violence or prejudice against a person or a group and can take the shape of any speech, behavior, writing, or exhibition. Some of the constitutive element of hate speech include:

i. ***Target Group:*** The target group is the first aspect of hate speech. Hate speech targets a certain group of individuals who are singled out and condemned due to their individual traits. Race, ethnicity, religion, sexual orientation, gender, or any other attribute that is seen as different from the speaker or the

dominant group might be used to characterize this group.

ii. ***Dehumanising:*** Dehumanising the target group by depicting them as less than human, animalistic, or undeserving of respect or empathy is a common goal of hate speech. This can be accomplished by use derogatory words, caricatures, and stereotypes.

iii. ***Stereotyping:*** Another frequent component of hate speech is stereotyping. It entails distilling a complex and varied population down to a handful of distilled, judged unwanted, or undesired, traits. Using stereotypes to defend prejudice and violence towards the target group is common.

iv. ***Negative Emotion:*** Hate speech aims to incite hostile feelings toward the target group, such as fear, rage, and disgust. Inflammatory rhetoric, emotionally charged language, and appeals to bias can all be used to arouse unfavourable feelings.

v. ***Incitement to harm:*** Incitement to harm is the last component of hate speech. Violence and prejudice against the target group might result from hate speech. Also, it can foster an atmosphere of hatred and intolerance that makes it simpler for others to justify and carry out harmful actions.

There are three main reactions to hate speech, according to the Isola [9]: legalistic, non-legalistic, and no response. The majority of African states continue to support laws and legislation that would kill, imprison, or penalise offenders despite the success of combining legalistic and non-legalistic ways to restrict hate speech in Europe and Kenya. The majority of African states rely on juridical solutions. As a result, hate speech is now prohibited in at least 29 African countries [10].

B. 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 time 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 [10].

C. Hate Speech in Côte d'Ivoire.

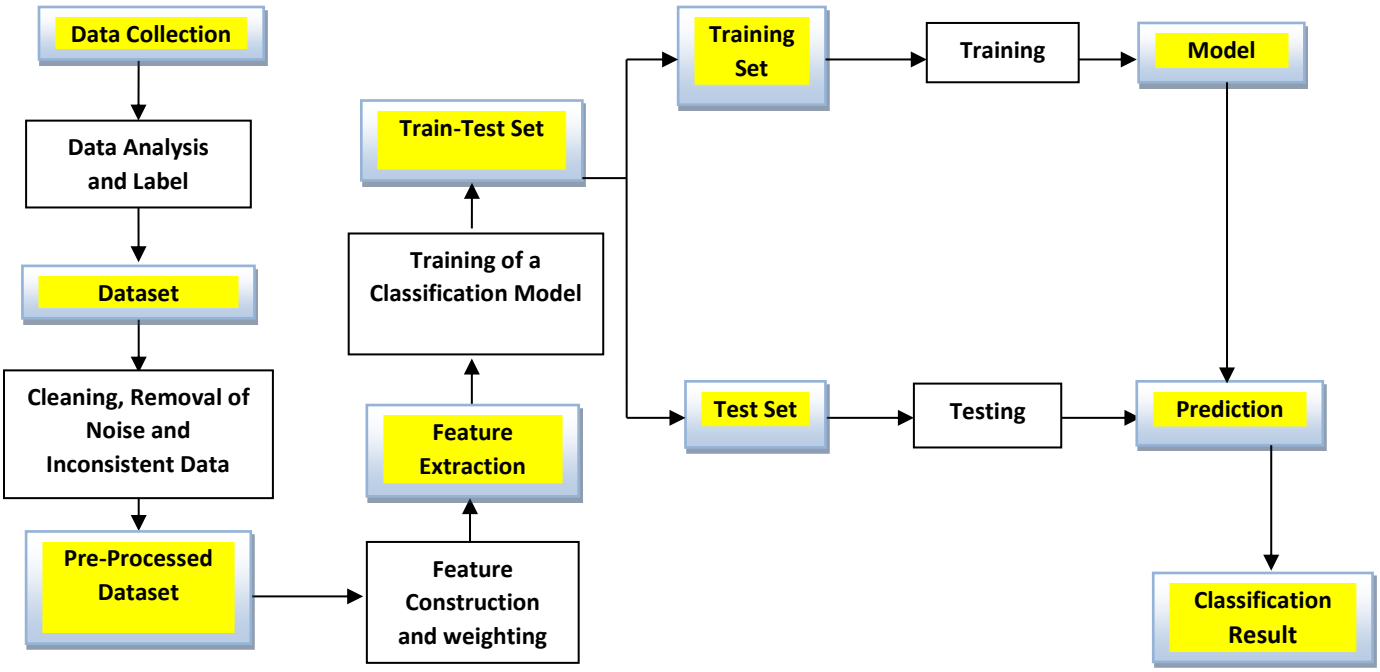
In the Republic of Cote d'Ivoire, hate speech is an important problem. More than 60 different ethnic groups make up the population of the nation, and in the past, conflicts between these groups have resulted in violent outbursts. Hate speech has increased in frequency during the past few years, especially 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 found guilty of using hate speech, the law offers prison terms and penalties [10].

D. 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 [11].

E. 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 [12].

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. But 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 hate speech. There have also been requests for media outlets to report on delicate topics responsibly and without sensationalising them [4].

This study focuses on identifying hate speech on Twitter, 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 May 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. The study uses a machine learning classifier and pertinent natural language processing techniques to build a hate speech detection model.

III. METHODOLOGY

The aim of this research is to develop a Natural Language Processing Machine Learning framework in Python to detect hate speech in Twitter using

West African hate words. A number of experiments were implemented using Python environment to determine the best combination of machine learning algorithm and features extraction to build the hate speech model. The experimental dataset was created from twitter using a collection of hate speech terms given by PeaceTech Lab. The dataset was created through the use of Twitter API, where a developer's account was set up with Twitter, after which an API ID and an API TOKEN was received for scraping the Twitter API. A Python script which makes use of functionalities provided by Tweepy was written to scrape the Twitter API. The gathered tweets were restricted to those originating within a radius of 500 miles using Abuja as a central point on Google map to cover West Africa. Functions were supplied for reading the data in CSV format, so that it can be viewed on MS Excel and labeled manually to create a corpus for training the Machine Learning models.

A. Experimental Setup.

In the experiment, various feature extraction and classification method combinations are compared.

As can be seen in Fig. (i) each test thus follows a straightforward procedure. To eliminate noise and inconsistent data, the collected text data samples go through preprocessing. Tokenization and the elimination of phrases and keywords with little meaning are part of this process. The preprocessed data are then transformed into features to facilitate analysis through the use of methods that lower the number of redundant features and dimensionality.

The Hate Speech models were implemented using a number of development tools and programmes. It is tested and implemented in this work using the Python programming language, from the data preprocessing phase to the model building phase, as well as to evaluate classifiers model. Python is used because it is the preferred programming language for developers, researchers, and data scientists who need to work with machine learning models [13]. It offers a broad range of ML and NLP libraries.

B. Hate Speech Detection Modeling.

The stages involved in creating the hate speech detection model are covered, starting with pre-processing, where the dataset is cleaned, followed by feature extraction, where the dataset is vectorized in order to be trained for the classification step. The list of hate words used in West Africa, together with their definitions, users, and intended audience, is shown in Appendix 1. Also mentioned are the assessment measures and confusion matrices used to rate the effectiveness of the detection models.

C. Pre-Processing.

Figure i: *Methodology Flow Process*

Text data typically contains a lot of noise and is unreliable. Cleaning the data to remove information that isn't important for identifying hate speech is necessary to lower the likelihood of working with noisy and inconsistent data. To prepare the raw text for mining, information like punctuation, special characters, digits, and terms that have little meaning in the context of the text are eliminated. Preparing data for machine learning algorithms to use is the core goal of pre-processing. A classifier works better and is more effective when properly pre-processed. It also gets a better-quality feature space during features extraction and takes less time to train. The pre-processing procedures employ the following techniques:

i. ***Removal of Twitter Handles (@user):*** The Twitter handles that were masked as @user due to privacy concerns were removed. Also, they don't convey much information about the nature of the tweet.

ii. ***Removal of Punctuations, Numbers, and Special Characters:*** Punctuations, numbers and even special characters were gotten rid of because they wouldn't help in differentiating different kinds of tweets. However, hashtags with spaces were left because they provide some useful information.

iii. ***Removal of Short Word:*** Most of the smaller words do not add much value. For example, words with length two (2) or less, such as 'he', 'is', 'it' 'a' were removed from the data.

iv. ***Removal of URLs and HTML Tags:*** URLs and HTML tags were removed because they do not add any useful information and we do not want the classifier to learn features from these tokens.

v. ***Removal of Random Patterns (Randos):*** Left over patterns that might be formed or left in the process of data cleaning are also removed because they are not relevant in training models

vi. ***Changing all letters to Lowercase:*** This is to make the raw text ready for mining.

vii. ***Tokenization:*** The tweets are then split into individual words or tokens to protect them.

viii. ***Lemmatization****:* Here, suffixes such as 'ed', 'ing', 'ly', 'es', 'er' and 's' were removed from words. For example, "kill", "killer", "killed", "kills" and "killing" are the different variations of the word 'kill' used in the same context. The purpose of stemming is to reduce the total number of unique words in our data without losing a significant amount of information.

D. Feature Extraction Methods.

The method of choosing and/or combining variables into features, known as feature extraction, significantly reduces the quantity of data that needs to be processed while properly and thoroughly characterising the original dataset. For preprocessed data to be examined using methods that reduce dimensionality and redundant features, it must be transformed into features.

i. ***Dimensionality Reduction:*** When dealing with a large number of variables there is a need for Dimensionality Reduction which can significantly improve a learning algorithm's performance.

Given a set of features F, as shown in Eq.(1),

$$F = (f_{1,\dots,} f_{i,\dots,} f_n) \quad (1)$$

The goal is to find a subset that "maximizes the learner's ability to classify patterns". ***F'*** maximizes the scoring function.

ii. ***Redundant features****:* These can be effectively reduced through the extraction of relevance features. However, there are two degrees of relevance: strong and weak relevance. A feature is relevant if it is strongly or weakly relevant otherwise it is irrelevant which means it is redundant.

iii. ***Strong Relevance of a variable/feature:*** Equation (2) is a set of all features except *fi*, denoted by *si* a value-assignment to all features in *Si* as shown in Eq. (2),

$$Si = \{f1, \dots, fi-1, fi+1, \dots fn\} \quad (2)$$

A feature *fi* is strongly relevant, if there exists some *xi*, *y* and *si* as shown in Eqs. (3) and (4), for which

$$p(fi = xi, \quad Si = si) > 0 \quad (3)$$

Such that

$$p(Y = y \mid fi = xi; Si = si) \neq p(Y = y \mid Si = si) \quad (4)$$

This means that removal of *fi* alone will always result in a performance deterioration of an optimal Bayes classifier.

iv. ***Weak Relevance of a variable/feature:*** A feature fi is weakly relevant, if it is not strongly relevant, and there exists a subset of features Si'of Si for which there exists some xi, y and si' as shown in Eqs.(5) and (6) with

$$p(fi = xi, Si' = si') > 0 \quad (5)$$

Such that

$$p(Y = y \mid fi = xi; Si' = si') \neq p(Y = y \mid Si' = si') \quad (6)$$

This means that there exists a subset of features *Si'*, such that the performance of an optimal Bayes classifier on *Si'* is worse than *Si'* **U** { *fi* }.

E. Classification.

A supervised learning challenge for modeling and predicting categorical variables is classification. Here, a certain set of data - whether structured or unstructured - is divided into classes. A training dataset with several instances of inputs and outputs is necessary for classification so that a classifier can gain new knowledge. Predicting the class of the input data points often referred to as the target, label, or categories - is the first step in the process. Finding the class or category that the new data will belong to is the key objective.

F. 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 (Eq. (7)), Recall (Eq. (8)), F_1 .Score (Eq. (9)) and Accuracy (Eq. (10)) of a system [15].

$$Precision = \frac{tp}{tp + fp} \quad (7)$$

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

$$Recall = \frac{tp}{tp + fn} \quad (8)$$

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

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

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

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad (10)$$

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

IV. RESULTS AND DISCUSSION

Each hate speech model's performance was assessed in comparison to the test set. The experiment demonstrates that fitting Bi-LSTM into GloVe embedding resulted in the best results. Additionally, the created dataset was contrasted with cutting-edge datasets. Hate Speech, which is also the positive class with label 1, and Non-Hate Speech, which is also the negative class with label 0, were the two class datasets used to create the classification model. Three separate features - CountVectorizer, GloVe Embedding, and Tf-Idf - were used to analyse the experiment. Accuracy, precision, recall, and F1 scores were provided as the outcomes of each categorization model that was examined as shown in Table I. The results of precision scores for LR, NB, XGBOOST, DNN and LSTM are presented in Fig. ii.

TABLE I

ACCURACY SCORES OF CLASSIFICATION MODELS

MODELS	**TF-IDF**	**COUNTVEC**	**GLOVE**
LR	0.86	0.87	-
NB	0.83	0.68	-
XGBOOST	0.87	0.87	-
DNN	0.90	0.90	-
Bi-LSTM	-	-	0.92

From, Table I, the best accuracy 0.92 (92%) was achieved in Bi-LSTM with GloVe Embeddings followed by DNN with 0.90 (90%) accuracy on both Tf-Idf and CountVectorizer feature extractors. XGBOOST also had a good accuracy of 0.87 (87%) on both Tf-Idf vectorizer and CountVectorizer. The least accuracy score of 0.68 (68%) was gotten in NB using the CountVectorizer feature extractor.

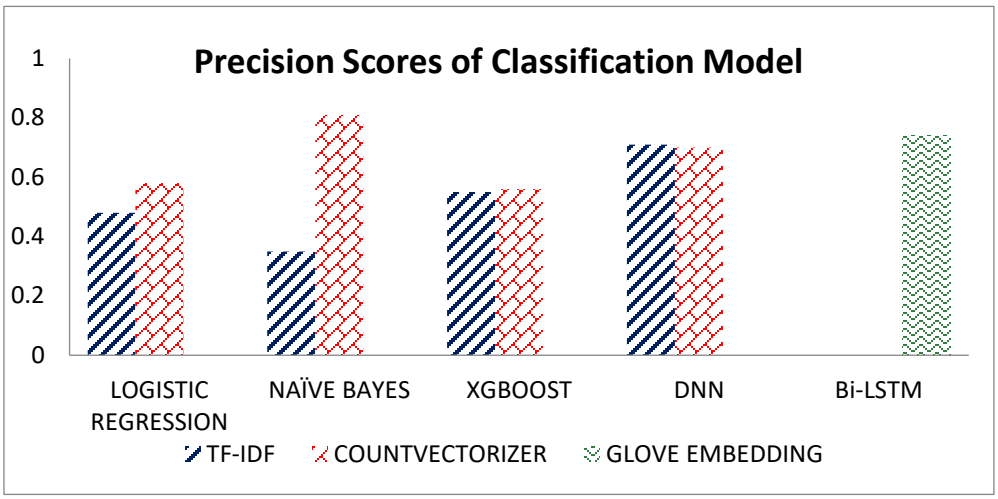


Figure ii: Chart Showing Precision Scores of Classification Model

From Fig. (ii), NB model on CountVectorizer achieved the best precision score of 0.81 (81%), followed by Bi-LSTM model with 0.74 (74%) on GloVe embedding. DNN also had a good precision score of 0.71 (71%) and 0.70 (70) on both Tf-Idf and CountVectorizer respectively. However, NB model on CountVectorizer also achieved the least precision scores of 0.35 (35%). CountVectorizer features extractor performed better than Tf-Idfin all classifications where they were used except for DNN classification where it fell behind slightly by 1%. GloVe embedding also performed well. The results of recall scores for LR, NB, XGBOOST, DNN and LSTM are presented in Table II. The results of F1-Scores for LR, NB, XGBOOST, DNN and LSTM are presented in Fig. iii.

TABLE II

RECALL SCORES OF CLASSIFICATION 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 II, show that Bi-LSTM had the highest recall of 0.95 (95%) on GloVe Embedding. LR also had a good recall score of 0.88 (88%) and 0.82 (82%) respectively on both Tf-Idf and CountVectorizer feature extractors

and XGBOOST had recall scores of 0.83 (83%) and 0.84 (84%) respectively on both Tf-Idf and CountVectorizer.

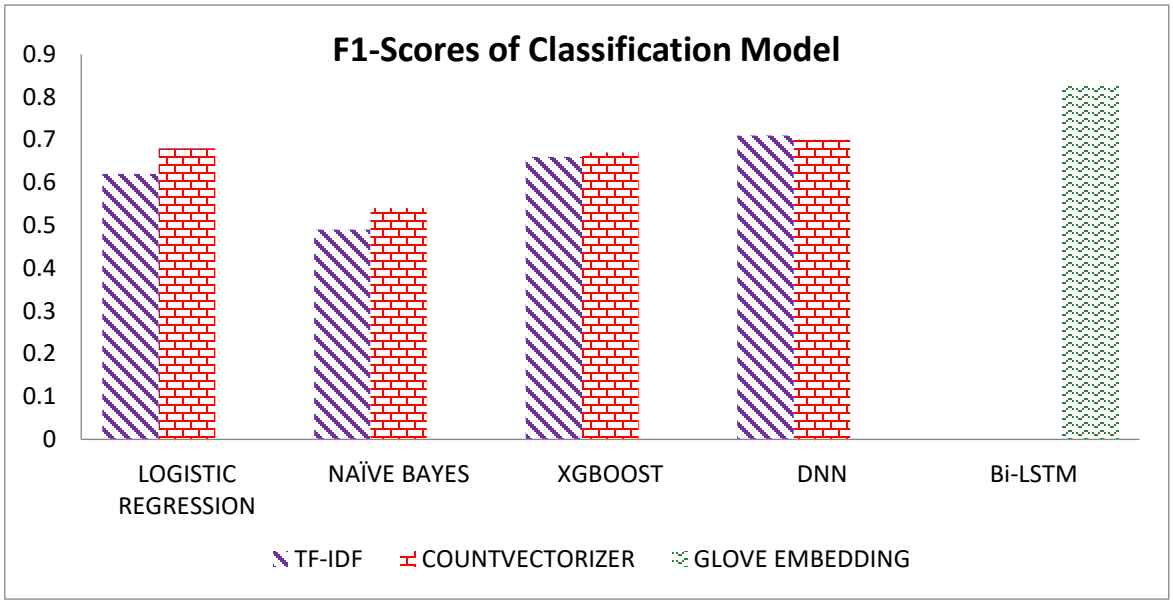


Figure iii: Chart Showing F1-Scores of Classification Model

From Fig. iii, Bi-LSTM on GloVe Embedding achieved the best performance with F1-Scores of 0.83 (83%), followed by DNN with F1-Scores of 0.71 (71%) and 0.70 (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 (49%) and 0.54 (54%) respectively. Table 4.4 shows that deep learning models outperformed classical models.

A. Comparing performance of the West African Based Hate Dataset with state-of-the-art dataset

It was required to evaluate the performance of the West African-based hate dataset by contrasting it with the most recent dataset on hate speech. For this, the Analytics Vidhya dataset for detecting hate tweets was examined. The dataset includes 32,000 tweets, 2242 of which - or 7% - were identified as hate speech. Here, tweets that contain racist or sexist content are considered hate speech. 93% of the tweets, or 29720, were classified as not being hate speech [14]. The cutting-edge dataset will be trained using the same classifiers and feature extractors that were used to create the West African hate speech detection model. Results of accuracy and F1-Scores of West African hate dataset in bold and that of hatred tweets dataset in asterisk as shown in Table III.

TABLE III

RESULTS OF ACCURACY AND F1-SCORES OF WEST AFRICAN HATE DATASET IN BOLD AND THAT OF HATRED TWEETS DATASET IN ASTERISK.

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

As can be seen from the findings in Table III, all classifiers performed better than the West African Hate Dataset in terms of accuracy, with the exception of the Bi-LSTM classifier. Unfortunately, the accuracy of the Hatred Tweets Dataset cannot be used as a valid indicator of model performance since it tends to predict more non-hate speech, which is the majority class, than hate speech, which is the minority class. The better metric measure for imbalance classification such as the one employed in this work is the F1-Scores. This is because, the F1-Scores takes into consideration the precision and recall and also provides a single score that balances both the concerns of precision and recall in one number [16]. Comparing the newly built West African-Based Hate Dataset with State of the Art Hatred Tweets Dataset as seen in Figure 3, the West African-Based Hate Dataset outperformed the State-of-the-Art Hatred Tweets Dataset it in terms of overall F-Measure performance across all classifiers. This demonstrates the unique dataset's usefulness in identifying hate speech in West Africa.

From the overall results, it can be seen that Deep learning models (DNN and Bi-LSTM) outperformed conventional models (LR, NB, and XGBOOST) with higher F1-Scores. Among the traditional models, LR model on Countvectorizer performed best with an F1-Score of 0.68 (68%) followed by XGBOOST with F1-Scores of 0.67 (67%) and 0.66 (66%) on both Countvectorizer and Tf-Idf respectively. It could also be deduced that the classical model performed better on Countvectorizer feature extractor, than on Tf-Idf feature extractor with higher F1-Scores. LR and XGBOOST on both Tf-Idf and CountVectorizer as well as NB on Tf-Idf all had a high recall values and low precision values. The high recall values means the classifiers correctly classified majority of the controversial tweets as hate speech in the dataset. On the other hand, the low precision value signifies that the classifiers also classified some noisy (irrelevant and inconsistent) tweets into the hate speech class.

For deep learning models, the Bi-LSTM model with Glove Embedding feature extractor performed better than DNN with high accuracy of 0.92 (92%) and F1 scores of 0.83 (83%), as against DNN which gave accuracy of 0.90 (90%) on both Tf-Idf and CountVectorizer features extractors and an F1-Score of 0.71 (71%) on Tf-Idf and 0.70 on CountVectorizer. This can be attributed to the fact that neural network models work best on embeddings. Also, embeddings feature extractors perform better than vector representation feature extractors on large datasets, as embeddings captures some of the semantics of the input, by placing semantically similar inputs close together in the embedding space, unlike in vector representation (Td-Idf vectorizer and CountVectorizer) where the order and context of words are lost and semantic similarities between words cannot be represented. The Deep Learning Models also produced high recalls and precisions indicating that most of the controversial tweets were classified correctly as hate speech with little or less noise.

V. CONCLUSION

It might be difficult to automatically identify hate speech in online media, particularly when it uses hateful expressions with a West African accent. This is due to the fact that neither a public dataset for hate speech detection based in West Africa nor any prior research of hate speech recognition employing specific hate words from that country exist. The goal of this research was to use machine learning and natural language processing to create a system to identify hate speech with a West African origin on social media. As part of the development process, tweets are gathered to create the dataset, annotation rules are created, preprocessing is done, features are extracted using Tf-Idf, CountVectorizer, and GloVe Embedding, and models are trained and tested using LR, NB, XGBOOST, DNN, and Bi-LSTM classifiers.

In this study, 20176 tweets in total were manually classified into two categories: hate speeches and non-

hate speeches. While 15375 tweets, or 76% of the corpus, were classified as non-hate speech, 4801 tweets, or 24% of the corpus, were classified as hate speech. In order to prepare the data for the use of machine learning algorithms, the data were then pre-processed to remove noisy and inconsistent data.

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 research on racial epithets in West Africa is still in its infancy and has a lot of room for development. To improve the findings, the dataset can be annotated with more than just a binary classification. Additionally, the dataset can be increased to include new trends, themes, and writing types. In the future, it may also be investigated how hateful expressions are expressed in non-textual content like films, photos, and emojis posted by users. Other feature extractors like BERT, ELMo, FastText, etc. could be examined in addition to Tf-Idf, CountVectorizer, and GloVe Embedding. Word embeddings could also include characteristics like location, age, and user gender. It is possible to investigate additional machine learning classifiers, particularly deep learning classifiers. To create a more reliable model with improved performance, the classifiers can also be combined.

REFERENCES

[1] Sękowska-Kozłowska, K., Baranowska, G., & Gliszczyńska-Grabias, A. (2022). Sexist hate speech and the international human rights law: towards legal recognition of the phenomenon by the United Nations and the Council of Europe. International Journal for the Semiotics of Law-Revue Internationale de Sémiotique Juridique, 35(6), 2323–2345.

[2] Heller, B., & Magid, L. (2019). Parent's and Educator's Guide to Combating Online Hate Speech. Retrieved January 18, 2020, from ConnectSafely website: https://www.connectsafely.org/hatespeech/

[3] Mrabure, K. O. (2016). Counteracting hate speech and the right to freedom of expression in selected jurisdictions. Nnamdi Azikiwe University Journal of International Law and Jurisprudence, 7, 160–169.

[4] Asogwa, N., & 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.

[5] IndependentNationalCommission. (2019). A Bill For An Act To Provide For The Prohibition of Hate Speeches and for other related Matter. Retrieved March 9, 2023, from National Cohesion and Integration Act website: https://placng.org/i/wp-content/uploads/2019/12/Hate-Speech-Bill.pdf

[6] Machirori, F. (2022). Tackling online hate speech in Africa and beyond: "We can't trust Big Tech to abide by its own rules." Retrieved March 9, 2023, from Association For Progressive Communications website: https://www.apc.org/en/news/tackling-online-hate-speech-africa-and-beyond-we-cant-trust-big-tech-abide-its-own-rules

[7] Opusunju, O. (2018). Nigerian government begins monitoring social media to tame hate speech. Retrieved March 24, 2021, from ITedgenews.africa/ website: https://www.itedgenews.africa/nigerian-government-begins-monitoring-social-media-tame-hate-speech/

[8] Auwal, A. M. (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.

[9] Isola, O. (2018). Hate Speech Law Proposal in Nigeria: When Beheading is the antidote for a Headache. Retrieved January 08, 2024, from

wilsoncenter website: https://www.wilsoncenter.org/blog-post/hate-speech-law-proposal-in-nigeria-when-beheading-is-the-antidote-for-a-headache

[10] Oyeleke, M. S. (2020). "Disinfodemic" in West Africa Communities: Tackling Extremism, Hate Speech and Fake News in Social Media Age. West-Africa-MIL-Week-Celebration-2020-Resisting-Disinfodemic-Media-and-Information-Literacy-for-Everyone-by-Everyone-Selected-Papers-UNESCO-Abuja-Regional-Office.Pdf, 190–201.

[11] GhanaNewsAgency. (2011). NMC chairman worried about hate speech and insults. Retrieved March 9, 2023, from GhanaWeb TV website: https://www.ghanaweb.com/GhanaHomePage/NewsArchive/NMC-chairman-worried-about-hate-speech-and-insults-206587

[12] Williams, E. E. (2020). Effectiveness of Social Media Platforms in Combating Extremism, Hate Speech, and Fake News in Nigeria. Resisting Disinfodemic Media and Information Literacy, 7–21.

[13] Melat, F. A. (2022). Hate Speech Detection for Amharic Language on Facebook Using Deep Learning. Retrieved from https://ir.bdu.edu.et/bitstream/handle/123456789/14487/Melat Fissha fInal.pdf?sequence=1&isAllowed=y

[14] Toosi, A. (2019). Sentiment Analysis: Detecting hatred tweets. Retrieved March 9, 2023, from Kaggle website: https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech

[15] Brownlee, J. (2020). Failure of Classification Accuracy for Imbalanced Class Distributions. Retrieved March 9, 2023, from Machine Learning Mastery website: https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

[16] Brownlee, Jason. (2020). How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Retrieved March 9, 2023, from Machine Learning Mastery website: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalancedclassification/

APPENDIX 1

PHRASE/ WORD	RELATED REFERENCE(S)	USER (S)	TARGET(S)	HATE MEANING
Zoo	Zoo country /Zoo people/Zoo Republic/ West Africa is a zoo; all animals in it must die/animals zoo	IPOB / Biafran	West African	Relates West Africans to animals that are senseless, illiterate, and wild."
Aboki	abokis/aboki cow/malu	Southerners	Northerners (Hausa's Fulani's)	Illiterate, unintelligent, foolish destitute, lower-class, without a future and uncivilized.
Arne	arna/aruna/ar'na/kirdi	Muslims	Christians	Infidels, unbelievers, pagans.
Boko Haram	Bokoharamists/the Jihadists/Islamist Boko Haram/Islamist Fundamentalists	Southerners	Northerners	Terrorist
Parasites	bloodsuckers/vultures/pests	Southerners	Northerners	North is poor and depends on oil revenue from the South

Inyamiri	Inyammiri Dodon Doya/nyamiri/inyamirin/ inyamuri/yanmiri/yamiris	Hausa	Igbo	A cheat, cunning, loves money too much and stupid Igbo man
Almajiri	almajirai/almajiris	Igbos, Yorubas	Hausas	Beggars
Biafrat	Biafrat zombies/Biafarats/Biafra/Biafran agitators/Biafrog/Biafraud/Biafrauds	Yorubas, Hausas and South South	Igbos	Rats, frogs and fraudulent people
Herdsman or Herdsmen	Buhari herdsmen/Fulani herdsman/stubborn Fulani herdsmen/Fulani herdsmen are kidnappers and criminals/herdsmen vampires/ herdsmen leeches/killer herders	Southerners	Fulanis	Lawless, disorderly, an invader, a parasite and living off the land they do not own
Product of baby factory	Baby factory/baby factory products/Ya'yan karuwai Wanda ake buga su a kamfani/born throwaway	Hausas and Yorubas	Igbos	Lower social status, irresponsible or lacking reproductive control, infertile, engaging in cultic activities by selling rituals, or lacking morality
PDPigs	PDP	Non PDP Political Parties especially APC	PDP	Pigs

.

COMMON HATE WORDS AND STEREOTYPES USED IN WEST AFRICA