ENHANCED HETEROGENEOUS STACKED ENSEMBLE MACHINE LEARNING MODEL FOR DETECTING NIGERIAN POLITICALLY MOTIVATED CYBERHATE

MULLAH NANLIR SALLAU

UNIVERSITI SAINS MALAYSIA

2023

ENHANCED HETEROGENEOUS STACKED ENSEMBLE MACHINE LEARNING MODEL FOR DETECTING NIGERIAN POLITICALLY MOTIVATED CYBERHATE

by

MULLAH NANLIR SALLAU

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

April 2023

ACKNOWLEDGEMENT

My sincere and profound gratitude goes to God almighty for His enabling grace and mercy over my life throughout this program. His presence made this dream of mine a reality. I remain grateful to God in every aspect of my life.

To my wife, thank you! This is the second time God is permitting me to leave you and the children in Nigeria. I sincerely appreciate your patience and hard work to keep the family going. Thank you for your understanding and patience. To Zhanfa Nanlir and Co, thank you all!

To my main supervisor, Associate Professor Dr Wan Mohd Nazmee Wan Zainon, thank you for being there for me all the time. You respond to my messages as promptly as possible, thank you. You gave me all I need to succeed in the PhD journey, thank you. You play the roles of both the father and the supervisor to me. God will reward you greatly for your humility and patience.

To the co-supervisor, Dr Mohd Nadhir Ab Wahab and the review team, I am sincerely grateful for critiquing the work right from the proposal stage to the final stage of this work. I am grateful and wish all of you God's blessings.

I am thankful to the Federal College of Education Pankshin management team under the leadership of Dr. Amos Bulus Cirfat. A special thanks to my Dean, Dr Solomon Mangvwat who stood by me during the storm and trying moment. Grateful to Mr Mbwas Caleb and the entire GSE department for nominating me for the TETFund sponsorship.

To the Dean, Professor Dr Bahari Belaton, thank you for your encouragement. To all staff of the School of Computer Sciences, I appreciate you all. To my former Dean, Professor Dr Rosni Abdullah, thank you. You did everything humanly possible to make us better researchers. It is worthy of emulation that even at the point of retirement you still dedicate your time to give your best to make us independent researchers. God bless and keep you in good health. Thank you.

To my former supervisor at Coventry University, United Kingdom, Dr Ali Niknejad, I say thank you. The mentor-mentee relationship that started in 2015 still lasts to date. Thank you for your encouragement and financial assistance. Remain blessed!

Dr Gwangtim T. Poyi, I sincerely appreciate the encouragement to push harder during the storm at FCEP. Architect Chris Gamde, thank you for your efforts to see that I succeed. I sincerely appreciate the efforts of Mr Sunday Gomo for his encouragement. To Dr Abrar Noor Akramin Kamarudin and Dr Haziqah Shamsudin, you people made me feel at home in Malaysia, thank you.

The space is not enough to list every person's name here. God will bless everyone who has assisted me in one way or the other during my PhD journey. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ABC	AdaBoost Classifier
ACHR	American Convention on Human Rights
ACO	Ant Colony Optimisation
AdaB	AdaBoost
APIs	Application Programming Interfaces
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional LSTM
BoW	Bag of Words
CNN	Convolutional Neural Networks
DL	Deep Learning
DMs	Distributional Models
DT	Decision Tree
EAC	Encyclopaedia of the American Constitution
ECHR	European Convention on Human Rights
ELMo	Embedding from language models
EU	European Union
FGN	Federal Government of Nigeria's
GAs	Genetic Algorithms
GBC	Gradient Boosting Classifier
GloVe	Global Vectors for Word Representation
GNB	Gaussian Naive Bayes
GPT	Generative Pretrained Transformer
HSE	Heterogeneous Stacked Ensembles
ICA	Independent Component Analysis

ICCPR	International Covenant on Civil and Political Rights
KNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LR	Logistical Regression
LR	Logistic Regression
LSTM	Long Short-term Memory
Mac avg	Macro Average
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLA	Machine Learning Algorithm
MNB	Multinomial Naive Bayes
NB	Naïve Bayes
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PSO	Particle Swarm Optimisation
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SA	Simulated Annealing
SBS	Sequential Backward Selection
SFM	Select from Model
SFS	Sequential Forward Selection
SMIs	Social Media Influencers
SMNPs	Social Media Network Platforms
SVM	Support Vector Machine

TFE	Texts Feature Extraction
t-SNE	t-Distributed Stochastic Embedding
UDHR	Universal Declaration on Human Rights
UN	United Nations
VSMs	Vector Space Models
W Avg	Weighted Average
XGB	XGBoosting

MODEL PEMBELAJARAN MESIN ENSEMBEL BERTINDAN HETEROGEN (HSE) DIPERTINGKAT UNTUK MENGESAN KEBENCIAN SIBER BERMOTIFKAN POLITIK NIGERIA

ABSTRAK

Ucapan kebencian adalah masalah sejagat sejak dahulu lagi. Penggunaan media sosial (SM) yang tinggi telah menjadikannya masalah ini berkadaran amat besar semasa pilihan raya di Nigeria. Penggunaan tanpa nama yang dinikmati oleh pengguna adalah sebab utama kebencian siber yang tinggi di ruangan media sosial Nigeria. Ahli politik biasanya menyebarkan mesej kebencian bermotifkan politik yang berbeza pada media sosial semasa pilihan raya. Walaupun begitu, pendekatan kecerdasan buatan (AI) yang berbeza seperti model pembelajaran mesin telah dibangunkan untuk menangani masalah dengan kejayaan yang munasabah. Namun begitu, masalah ini berterusan dan membawa kepada kadar jenayah kebencian siber yang tinggi di Nigeria. Masalah utama ialah kekurangan penyelidikan untuk membina model bagi menangani keadaan unik di Nigeria. Masalah ini menjadikan model sedia ada tidak berupaya di ruang siber Nigeria. Untuk menyelesaikan jurang penyelidikan yang dikenal pasti dari sudut penyelidikan pembelajaran mesin, masalah ini dimodelkan sebagai tugas pengelasan teks. Untuk mencapai objektif utama, kajian mencadangkan untuk mempertingkatkan satu teknik yang dipanggil kaedah ensembel bertindan. Kaedah yang dicadangkan dipanggil ensembel bertindan heterogen (HSE). Kajian ini menggabungkan teknik pengurangan ciri dan teknik pengesahan silang untuk meningkatkan prestasi. Pengesahan silang K-lipatan berstrata telah dibina ke dalam algoritma pembelajaran HSE untuk membolehkan pengelas belajar secara saksama

daripada pengagihan kelas yang tidak seimbang. Berdasarkan amalan terbaik antarabangsa, analisis perbandingan dibuat sebagai bukti konsep. Kajian itu membandingkan prestasi model HSE dengan kaedah terkini seperti pembelajaran mendalam (BiLSTM) dan pembelajaran pemindahan (BERT). Model HSE adalah lebih baik daripada pembelajaran pemindahan (BERT+CNN) sebanyak 0.4 dan 0.17 menggunakan skor F1 dan metrik MCC. Hasil daripada tiga set data penanda aras berbeza daripada Davidson, Yadav dan Mandl turut digunakan sebagai perbandingan. Pencapaian HSE lebih baik daripada Davidson, dan Yadav dalam F1-skor masing-masing sebanyak 2% dan 13%. Pencapaian HSE juga lebih baik daripada keputusan Mandl dalam kedua-dua purata makro skor F1 dan purata wajaran skor F1 masing-masing sebanyak 4% dan 2%. Model ini telah terbukti berkesan dalam kedua-dua tugas pengelasan binari dan berbilang kelas seperti yang ditunjukkan dalam penyelesaian masalah yang berbeza.

ENHANCED HETEROGENEOUS STACKED ENSEMBLE MACHINE LEARNING MODEL FOR DETECTING NIGERIAN POLITICALLY MOTIVATED CYBERHATE

ABSTRACT

Hate speech is a universal problem from time immemorial. The high adoption of social media (SM) has made it a problem of gigantic proportions during elections in Nigeria. The anonymity enjoyed by the users is the main reason for the high volume of cyber hate in Nigeria's social media space. Politicians usually circulate different politically motivated hate messages on social media during elections. Though, different artificial intelligence (AI) approaches such as machine learning models have been developed to address the problem with reasonable success. Nonetheless, the problem persists and leads to a high rate of cyberhate crime in Nigeria. The main problem is the lack of research to build models to address peculiarities in Nigeria. These problems made existing models incapacitated in Nigeria's cyberspace. To solve the identified research gaps from the vantage point of a machine learning researcher, the problem was modelled as a text classification task. To achieve the main objective, the study proposed to enhance a technique called the stacking ensemble method. The proposed method is called the heterogeneous stacked ensemble (HSE). The study incorporated the feature reduction technique and cross-validation technique to increase performance. A stratified K-fold cross-validation was built into the HSE learning algorithm to enable the classifier to learn equally from the imbalanced class distribution. Based on international best practices, comparison analyses were made as proof of concept. The study compared the HSE model performance with state-of-art methods such as deep learning (BiLSTM) and transfer learning (BERT). The HSE

model is better than transfer learning (BERT+CNN) by 0.4 and 0.17 using the F1-score and MCC metrics respectively. Three results from different benchmarked datasets from Davidson, Yadav and Mandl were also used for comparison. HSE is better than Davidson, and Yadav in F1-score by 2% and 13% respectively. HSE is better than Mandl's result in both the F1-score macro-average and F1-score weighted-average by 4% and 2% respectively. The model has been proven effective in both binary and multiclass classification tasks as demonstrated in different problems solved.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Humans express their thought, feelings, and opinions or generally communicate through verbal speech, sign language, or written messages (Bouazizi & Ohtsuki, 2019). The key characteristic of any democratic society is to constitutionally guarantee the freedom of speech to all citizens without exception (Hill, 2020). However, absolute freedom is not obtainable in any society and thus the call for legal boundaries for the operation of the right to free speech. The most important of these legal boundaries is hate speech, which infringes on the rights of another person or group (Brown, 2017). This can also be considered inimical to democratic stability in any society. In the 'new normal', people communicate more on social media than in the physical world. Therefore, there is a need to harness the power of artificial intelligence to implement legal boundaries on social media for peaceful coexistence. Social media networks have favoured, inter alia, communications and ease of information sharing across the globe (Kapoor et al., 2018).

Social media network platforms (SMNPs) play vital roles in everyday activities and schedules (Hegazi et al., 2021; Vashistha & Zubiaga, 2021), and political discourse is no exception (Gorrell et al., 2018). SMNPs are new virtual communities that everyone is scampering to belong to for various reasons: (i) ease of use – SMNPs are favouring communications and easing information sharing across the globe (Szabó & Kovács, 2018). (ii) Cheap cost – The cost of disseminating information via SMNPs is also next to nothing, and an internet-enabled device is enough (Albarrak et al., 2020). (iii) Instant posting – This virtual society is a user-driven Web 2.0 application (Lee-

won et al., 2020), as one can create and share content dynamically (Gitari et al., 2015; Vrysis et al., 2021) and almost instantaneously (Hefler et al, 2019; Kim & Hastak, 2018). (iv) Removal of boundary and distance – People can be connected around the globe via SMNPs irrespective of location (Burnap & Williams, 2016; Kapoor et al., 2018). (iv) For business: many people and companies use social media platforms for advertisement, buying and selling of goods and services (Huang & Benyoucef, 2013). (v) Another interesting reason for everyone to be on social media is based on the 'new normal', which discourages physical gatherings due to the COVID-19 pandemic (Zafri et al., 2021). Over 4 billion users (Datareportal, 2022)[1] of SMNPs can share their opinions, feelings, current thoughts and anything of interest in just a few clicks (Tartir & Abdul-Nabi, 2017; Voorveld, 2019). The most popular among these SMNPs include, inter alia, Twitter, Facebook, YouTube, WhatsApp and Instagram, with billions of active users (Parviainen, 2020), as summarised in Figure 1.1.

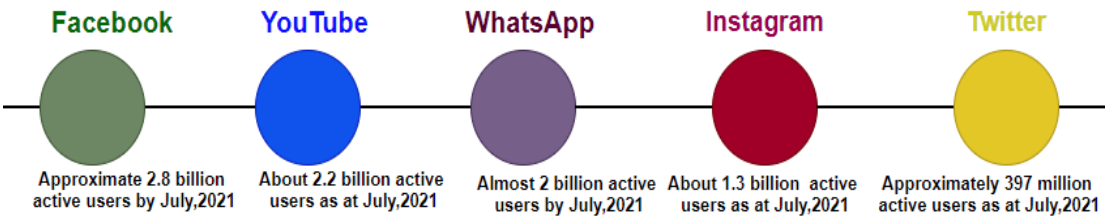

Figure 1.1 Most Popular Social Media Network Platforms (Statista, 2022)[2]

As presented in Figure 1.1, the most popular social media include Facebook, with almost 2.8 billion users (Guo & Johnson, 2020). Likewise, Twitter, is a useful marketing site (Parviainen, 2020) with approximately 397 million users, Instagram, with 1 billion users, and YouTube, with nearly 2.2 billion users, in 2021. Others include WhatsApp with approximately 2 billion subscribers, and Messenger has 1.3

[1] https://datareportal.com/social-media-users
[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

billion subscribers. WeChat is a popular application among the Chinese had over 1.06 billion active subscribers, LinkedIn with 660 million and Tumblr with 642 million. Hence, Figure 1.1 demonstrates that there has been a remarkably large number of users on social media globally.

Nigerian politicians usually leverage these advantages brought by SMNPs to disseminate their manifestos to electorates during electioneering. This is done through creating and sharing content on social media. Nonetheless, SMNPs content creation and sharing come with huge negative side effects on people (Lee-won et al., 2020), especially in countries practising democracy, which most times lead to electoral violence (Birch & Muchlinski, 2020). This disadvantage of SMNPs is tearing people apart. The devastating effects of sharing hateful and antagonistic posts created by users are a thing of serious concern and not new to social media users (Burnap & Williams, 2016; Lee-won et al., 2020; Vrysis et al., 2021). Many studies have revealed how people with corrupt minds towards others spread their hateful content targeting minorities or disadvantaged people on SMNPs (Tontodimamma et al., 2021). The concern regarding the rising cases of offensive or hate speech and its impacts during elections is alarming (Asogwa & Ezeibe, 2020; Kalampokis et al., 2013). According to Vrysis et al. (2021), "It is widely acknowledged that xenophobia, racism, gender issues, sexual orientation, and religion among others are topics that trigger hate speech." These are some variants of hate speech observed daily on SMNPs. All these variants of hate speeches are used by politicians during elections through social media.

Given that the interactions among the users on social media platforms can lead to valuable and insightful discussions, nevertheless, they have been progressively misused for the spread of hate speech mainly due to the anonymous identity enjoyed and wide adoption of these online platforms among the populace (Al-Maatouk et al.,

2020; Li et al., 2020; Touahri & Mazroui, 2021). Some content may be offensive or discriminatory, such as hate speech or abusive messages. Hate speech has caused much havoc to communities across the globe, especially in developing countries such as Nigeria. Hate speech is a common global concern (Bakalis, 2016), especially in countries with immature democracies (Asogwa & Ezeibe, 2020). Hate speech refers to text, post, comment or verbal speech that disparages, demeans a person or a disadvantaged or a minority group based on their colour, religion, nationality, sexual orientation, disability, gender race, ethnicity and other unique features associated with the person or group (Warner & Hirschberg, 2012; Zhang & Luo, 2018).

The viciousness ascribed to online hate speech has continued to increase worldwide, with mischievous users maliciously oppressing other people who they regard as their enemies or opposition either in personal life or politics. For example, in Nigeria, the use of hate speech become rampant during political activities, as different political parties usually go against each other not only on mass media platforms but also increasingly on social media (Jibril et al., 2017; Onimisi & Tinuola, 2019).

Recently, people who are fond of making hate speech or comments have changed tactics by shifting from ground attacks to positioning themselves in cyberspace to hide and create the possibility of being detected difficult (Al-Makhadmeh & Tolba, 2020; Narrain, 2017; Vidgen & Yasseri, 2020). The anonymity offered by the internet, coupled with the fact that comments are uncensored and non-restrictive, has made this problem continue to thrive. The mischievous users/perpetrators are regularly on Facebook, Twitter, YouTube, Instagram, and WhatsApp, to attack anybody whatsoever, especially their presumed enemies. They could post provocative messages through texts, audio, video, tape, cartoon, graphics,

and posters, on the net. These provocative messages can trigger problems that may lead to injury or defamation of character on an individual, and damage to either the targeted group of people, an organisation[3] or a government. Consequently, these can attract adverse reactions from the public and can lead to violence and hate crimes in society.

In addressing this problem of cyberhate comments or messages on social media, researchers have deployed various strategies. Efforts from diverse directions to manage these offensive contents on SMPs, by studies, (Burnap & Williams, 2015; MacAvaney et al., 2019; Nugroho et al., 2019), social media providers (Gonçalves et al., 2021), governments in different countries (Bakalis, 2016; Guo & Johnson, 2020; Wilson & Jibrin, 2019), and international organisations (Guterres, 2019). And many more have invested both time and financial resources to solve this problem. Additionally, repeated attempts have been made in the field of arts and humanities for decades with no solution in sight (Vidgen & Yasseri, 2020).

Now, the computing domain has come to rescue the situation (Vidgen & Yasseri, 2020). From a computing perspective, different methodologies have been proposed to address this problem through automatic detection methods such as canonical machine learning, deep learning (DL), ensemble approaches and transfer learning models. Conventional techniques, such as context-aware and statistical models, natural language processing (NLP) and feature engineering models that detect hate speech have also been proposed in the literature (Del Vigna et al., 2017; Gao & Huang, 2017; Robinson et al., 2018). Statistical methods were commonly used in the past to analyse these data for hidden trends that are useful for decision-making. Other

[3] https://blog.google/inside-google/company-announcements/commitments-racial-equity/

previous studies applied semantic content analysis techniques based on NLP (Schmidt & Wiegand, 2017) and machine learning (ML) (Burnap et al., 2015; Davidson et al., 2017; Gambäck & Sikdar, 2017) to build scalable machine learning models for cyberhate identification task. Although the automatic feature selection algorithm has shown a drastic reduction in feature space to detect malicious posts on social media (Robinson et al., 2018). An improved machine learning algorithm (MLA) could offer a better solution when the features are carefully identified based on context and other areas of interest (Robinson et al., 2018). This is what this work intends to achieve.

One of the articles closely related to this research is the current study by (Ahmed et al., 2022) and (D'Sa et al., 2020), which deployed deep learning and bidirectional encoder representations from Transformers (BERT) and fastText embedding for automatic detection of toxic speech. BERT is a technique for NLP pretraining developed by Google (Devlin et al., 2019). However, Geet et al. (2020) tried to distinguish between the terms hate, abusive, offensive, and toxic speech by performing binary and multiclass classification using a Twitter corpus. The proposed automatic classification of toxic speech using embedding representations of words embeddings as features and deep learning techniques did not capture context-based, coded, slang and out-of-vocabulary (OOV) terms used to propagate hate on social media. Hence, this proposed work intends to fill the gap of undetected hidden coded, slang, context-base and OOV terms in a speech on Twitter during a political debate.

1.2 Motivation for Study

The desire of humans to predict the future cannot be overemphasised (Chauhan et al., 2021). To keep the hope of predicting the future alive in this big data era, it is believed that social media is pregnant with information that can be mined (Ni et al.,

2017; Oikonomou & Tjortjis, 2018; Rousidis et al., 2020). Sadly, history has repeated itself time and again that hate speech can polarise societies along different divides and can also result in mass atrocities (Chauhan et al., 2021; Rousidis et al., 2020). It can also threaten the stability of democracy, especially in developing countries with weak economies, such as Nigeria (Asogwa & Ezeibe, 2020).

Nigeria as a country has been polarised along religious, cultural, political, ethnic, and geopolitical lines. This polarization along different divides has caused more harm than good to Nigerians (Asogwa & Ezeibe, 2020). As a result of the polarisation, hate speech has fertilised many civil unrests, and the resultant consequences, inter alia, are millions leaving in abject poverty, raising of extremists, banditry, kidnapping, and arm robbery, among others.

Social media worsens the situation, as cyberhate messages can be transmitted within seconds to all nooks and crannies of the world (Burnap & Williams, 2016). To police social media to keep Nigeria a once prosperous country, this study proposed to develop an ensemble technique approach as a solution to politically motivated cyberhate detection. This is very important now as current state-of-the-art methods cannot efficiently detect hate messages due to some peculiarities in language usage. Moreover, the cyberhate spread became more intense as political gladiators leveraged social media to gain cheap popularity among citizens. This research will dwell more on using machine learning to analyse political discourse in Nigeria's cyberspace.

Previous studies and organised bodies such as the United Nations (UN), (Guterres, 2019), European Union (EU) (Fortuna & Nunes, 2018) and the Nigerian government (Wilson & Jibrin, 2019) at different times have called for a robust solution to cyberhate on social media networks (Fortuna & Nunes, 2018; Guterres, 2019;

MacAvaney et al., 2019; Onimisi & Tinuola, 2019). Social media platforms have been used for propagating cyberhate, which has played an enormous role in election-related violence in Nigeria and some other countries (Asogwa & Ezeibe, 2020; Lai et al., 2019). For example, in Nigeria, hate speech is aggravated during elections (Febriana & Budiarto, 2019; Lai et al., 2019). During election periods, a particular party may campaign against another, including crafting deliberate crusades intended to vilify an opposing party.

The literature has shown that there are more cases of crises in Nigeria during the electioneering campaign than in non-election periods (Febriana & Budiarto, 2019; Lai et al., 2019). Research has also proven that most civil unrest in recent times is caused by cyberhate spread (Ajakaiye et al., 2019a; Bali & Desai, 2019; Guterres, 2019; Onimisi & Tinuola, 2019; Reed et al., 2020; Wahlström & Törnberg, 2019). Politicians and pundits have raised concerns over unchecked antagonistic posts on social media (Guo & Johnson, 2020). This symptom is common in countries whose democracies are still in the infant stages and are more vulnerable (Jibril et al., 2017; Onimisi & Tinuola, 2019). Mostly, these crises have led to hate crime, which equates to the loss of lives and wanton destruction of property before, during and after elections. Cyberhate has persisted in Nigeria, especially on Twitter and the trend is getting worse by the day. The Federal Government of Nigeria's (FGN) Twitter handle was suspended[4] June in 2021 due to cyberhate and FGN subsequently suspended Twitter from operating in Nigeria for about seven (7) months[5] (Anyim, 2021). Therefore, the need to leverage the power of machine learning models for accurate and timely detection of cyberhate among social media users is paramount. If the trend of

[4] https://edition.cnn.com/2021/06/04/africa/nigeria-suspends-twitter-operations-intl/index.html
[5] https://www.nytimes.com/2022/01/13/world/africa/nigeria-lifts-twitter-ban.html

cyberhate, especially on social media platforms, is not urgently addressed, it will grow uncontrollably in the nearest future.

1.3 Statement of the Problem

First, the answer to the question 'what constitutes a hate speech or cyberhate message in a political discussion' remains a very difficult one to answer (Martins et al., 2018; Vidgen & Yasseri, 2020). The difficulty arises from the variation in culture and tradition of people across the globe and the evolution of rare and new vocabularies (Schmidt & Wiegand, 2017). What is considered political cyberhate in a post in Nigeria could be seen as a non-hate message in the US/UK, Malaysia or other parts of the globe due to a lack of consensus definition (Mossie & Wang, 2019). The variation in culture is a sufficient problem to tackle since each social media platform has central control (Fortuna & Nunes, 2018). Managing everyone from different cultures under a single control means knowing the culture and tradition of every subscriber, which determines what constitutes hate speech. The current definition of hate speech used in the literature today does not consider the peculiarities of non-Western culture (Burnap & Williams, 2016).

Accurate detection of cyberhate messages must start with a clearer definition/explanation of cyberhate based on the culture and tradition of the people involved, in this case, Nigeria. Therefore, the need for a clear understanding of the cyberhate message on social media is paramount (Pereira-Kohatsu et al., 2019), especially for machine learning-based solutions. The definition of hate speech based on the Nigerian political context must be X-rayed, which will also significantly help in the process of building a guide for dataset annotation. The coding guides will enable the coders to see each message the same way to avoid ambiguity in the interpretation.

The usage of new words, slang and out-of-vocabulary (OOV) terms to construct hate speech messages can make identification of these messages difficult and hence the persistence of the problem of hate speech on social media (Alonso et al., 2020; Kovács et al., 2021; Vidgen & Yasseri, 2020). To make the case worse, most previous works remove slang and unknown words during the pre-processing stage. This can lead to the loss of valuable features for the training of the classifier.

Currently, there is a lack of insights into political hate speech in Nigeria's social media space. The current trend in solving problems is best through artificial intelligence approaches, such as machine learning techniques. Data are central to every machine learning-based problem-solving approach. Therefore, the need to build a dataset from social media to help investigate politically motivated cyberhate is important (Chauhan et al., 2021). On that note, to build a politically motivated hate speech detector, a large dataset from political discourse to help train the learning algorithm is required. The problem of benchmark dataset availability for training hate speech classifiers is a significant problem to address (Faris et al., 2020; Kapoor et al., 2018; MacAvaney et al., 2019). To the best of our knowledge, there is no dataset originating from Nigeria for training machine learning for hate speech detection. Circulation of hate speech on Twitter is a common problem in Nigeria, especially during elections. There are instances in Nigeria that call for this research due to hate speech and polarisations of citizens across different divides, such as ethnic, religious, political, and geopolitical locations. In 2019, a hate speech bill that proposed a death penalty for perpetrators of hate speech in Nigeria was proposed in the house of the senate (Wilson & Jibrin, 2019). However, this bill could not scale through due to public outcry from within and outside Nigeria. The use of hate speech on Twitter became too rampant that the Nigerian government banned Twitter usage on the 5th of

June 2021 (Anyim, 2021). However, the current global diffusion of social media makes data cheaply available for research. The rapid growth of social media networks and microblogging, such as Twitter, enables extensive and near real-time data sources through which the analysis of hate speech can easily be conducted.

The social media dataset generally has three common problems or characteristic attributes inherent in them and must be addressed along with the aforementioned problems. All real-life datasets, especially social media data have skewed or imbalanced class distribution (Dong et al., 2019; Luque et al., 2019) and data sparsity problems (Chen & Guestrin, 2016; Maimaiti et al., 2022) and informal styles of writing (Tartir & Abdul-Nabi, 2017). Data sparsity always leads to a curse of dimensionality (Nanni et al., 2019). Unfortunately, every learning algorithm tends to learn more about the majority class than the minority class (Luque et al., 2019). However, hate speech posts, which are the main target, belong to the minority class. The main goal is for the model to learn more about the properties of the minority class and be able to detect them effectively and efficiently. This is a necessary problem that must be addressed along with the specific ones listed.

1.4 Research Questions

The main goal of this study is to improve the prediction accuracy of the cyberhate detection model in a political discourse using an ensemble approach. Therefore, one main question this study will attempt to answer is, how to design a robust and efficient model with high prediction accuracy. This study is guided by the following investigative or research questions to help address the problem statement articulated in the last section:

1. What is a proper definition of cyberhate message to enhance identification and annotation of social media posts as politically motivated hate speech or non-hate in the context of Nigeria?
2. How to identify and annotate social media posts (especially) as politically motivated cyberhate or non-hate for training MLAs for effective and efficient detection?
3. What is the best approach in terms of feature reduction, representation, and extraction techniques to build an improved and robust ML model to effectively detect hate speech or cyberhate messages?

1.5 Research Aim and Objectives

The overall aim or main objective of this research is to enhance the performance accuracy of the ensemble approach for categorising Nigerian politically motivated messages on Twitter as hate or non-hate. To achieve this main objective, the following research sub-objectives were proposed:

- To investigate and review the definition and perception of cyberhate given in the literature by various researchers, social media providers, organised bodies, and government agencies. This will help determine what constitutes cyberhate and the causes of misclassification by previous models. The deliverable will be a derivation of a new definition that captures peculiarities in Nigerian usage and the formulation of a new cyberhate annotation guideline based on the Nigeria context.
- To develop a new dataset for building ML model for detecting politically motivated cyberhate during elections in Nigeria. The

expected deliverable in this objective is the building of a new dataset that can be used for training machine learning algorithms for identifying political hate speech.

- To propose and build a robust and efficient machine learning model based on an optimised heterogeneous stacked ensemble algorithm. The model will improve cyberhate detection performance given the high dimensionality and imbalanced class distribution of the dataset. The expected deliverable is an enhanced stacking ensemble model.

1.6 Expected Contributions from This Study

Implementing the model can provide many benefits to the government, political parties, organisations, and peace lovers in societies. The study will be beneficial also to the legal and security sectors, during any investigation, especially in the aspect of detecting cyberhate posts and propaganda against people in Nigeria. Furthermore, the study will also be helpful to social media companies such as Twitter, Facebook, and Instagram in addressing the persistent problem of hate post detection in Nigeria. In addition to being beneficial to societies, it also contributed to cyberhate detection and machine learning communities in the following ways:

- **Literature Review Contribution**

  The comprehensive literature will serve as a starting point for new scholars in this field of study or old scholars who wish to refresh their knowledge. Every standard research starts with a literature review. The literature review helps researchers to avoid duplication of efforts on the already solved problem and identification of research gaps.

- **Comprehensive Definition**

  To identify a post as hate or no hate, the researcher must understand in clear terms what constitutes hate speech. This is a very challenging task. A more comprehensive definition of hate speech that captures the uniqueness of the Nigeria scenario will be proposed. Secondly, a comprehensive annotator guide will be built to improve the coding of the dataset for machine learning training based on the Nigeria context. This guide can immensely help those carrying out related research.

- **Dataset Creation Contribution**

  The non-availability of the dataset which captured peculiarities in the Nigeria context fuelled this objective. Therefore, to address this objective, there is a need to build a new dataset based on political discourse to enable the training of the learning algorithm. This study intends to use any election in Nigeria as a case study to harvest the data.

- **Feature Space Reduction or Dimensionality Reduction and Hyperparameter Optimisation**

  Every short messaging analysis is confronted with an inherent problem of high dimensionality. The best feature reduction technique will be integrated as part of the algorithm to enable the use of the best features or most informative features for the training. This work will also optimise the classifier through hyperparameter tuning and a suitable learning rate for a robust, efficient, and balanced bias-variance model.

- **Model Selection**

  The choice of the learning algorithm for the classification is an important step in the machine learning pipeline. The goal is to build a robust and efficient model for detecting politically motivated hate speech on the Twitter platform. The study proposed to improve the stacking ensemble method. This work will evaluate the model using a benchmark dataset as a proof of concept.

1.7 The Significance of the Study

The study of hate speech in the past is normally conducted by scholars in the arts and legal domains. However, computer scientists have put their interests in this field of study in recent times due to the availability of large data. More researchers especially those in the field of data science have proposed different models to solve the cyberhate problem. Non-western countries have been ignored in this struggle (Burnap & Williams, 2016). The significance of this research is to propose a model that can solve the problem of politically motivated cyber-hate in Nigeria. This will go a long way to ensure a sense of lasting peace among Nigerians always. The goal of this research is to build a machine-learning model for improving prediction accuracy. The research is planned to achieve this goal by implementing an enhanced heterogeneous stacked ensemble technique. This work intends to use multiple baseline classifiers to build one robust and efficient hate speech detector. The meta-classifier of the ensemble model will add up the advantages of all the baseline models to one robust classifier.

1.8 Scope of the Study

The main aim of this research is to investigate and propose a robust and efficient algorithm to detect politically motivated hate posts on social media such as Twitter. For this research, cyberhate, hate speech, and any other abusive terms are used interchangeably. This study plans to use the Nigeria scenario as a typical example of a non-western country. A new dataset will be created from Nigeria's political discourse during the Nigerian election. The dataset will be manually annotated and only text messages will be considered for the research. However, it does not include hate posts on all social media networks. Likewise, the thesis focuses on the messages in the English language and no other languages. The cyberspace jurisdiction for this research is the Nigeria cyberspace and Twitter platform**.** The cyberspace jurisdiction for this research is Nigeria cyberspace. The data to be collected will be based on political discourse in Nigeria, and annotators are experts who are conversant with Nigeria's traditions and culture.

1.9 Thesis Organisation

This thesis has five chapters and each chapter contributes to the achievement of the research's main objective. The main aim is to improve the model performance accuracy for the identification of politically motivated cyberhate posts on Twitter.

Chapter 2 contains a comprehensive literature review of the necessary areas in this domain. Chapter 2 is organised in this way: first, the overview of the definition of hate speech by different researchers, organised bodies and social media providers. The remaining sections are reviewed based on the components of the text classification pipeline for detecting hate speech. Several research gaps are revealed and listed in section 2.8.

Details of the methodology used to achieve the research objectives are given in Chapters 3 and 4. Sections include research design, dataset annotation and creation, dataset description and exploration. Other sections include feature extraction, hate speech modelling, model development, ensemble method, hyperparameter optimization, experimentation setup procedure, and model evaluation.

In chapter 5, the study presents and analyses the results of the experiments conducted. The study also compares the results with the state-of-the-art and three benchmark datasets to prove the superiority of the proposed method. Finally, chapter 6 contains the conclusion, achievement of research objectives, the contribution of the research to the body of knowledge, and future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

A thorough work is done in this chapter to include the appraisal of other scholars' works in solving the cyberhate problem in the past. Through this review work, research gaps in the previous studies are identified. The reviews is done with the main objective in mind – to improve the predictive performance of a machine learning model for identifying politically motivated cyberhate. From a machine learning approach standpoint, cyberhate identification is considered a text classification task. Text classification pipeline involves the following: data collection & preparation, feature engineering, algorithm selection & training and model evaluation phases.

The definition of hate speech is first reviewed as this can vary based on tradition and cultural affiliation. This is important for data annotation for training the proposed MLAs. The study comprehensively reviewed the developmental trends from classical machine learning, deep learning, and the current pre-trained transformer-based models. All the research gaps identified during the review are summarised at the end of this chapter. The outline of the entire review process is displayed in Figure 2.1.

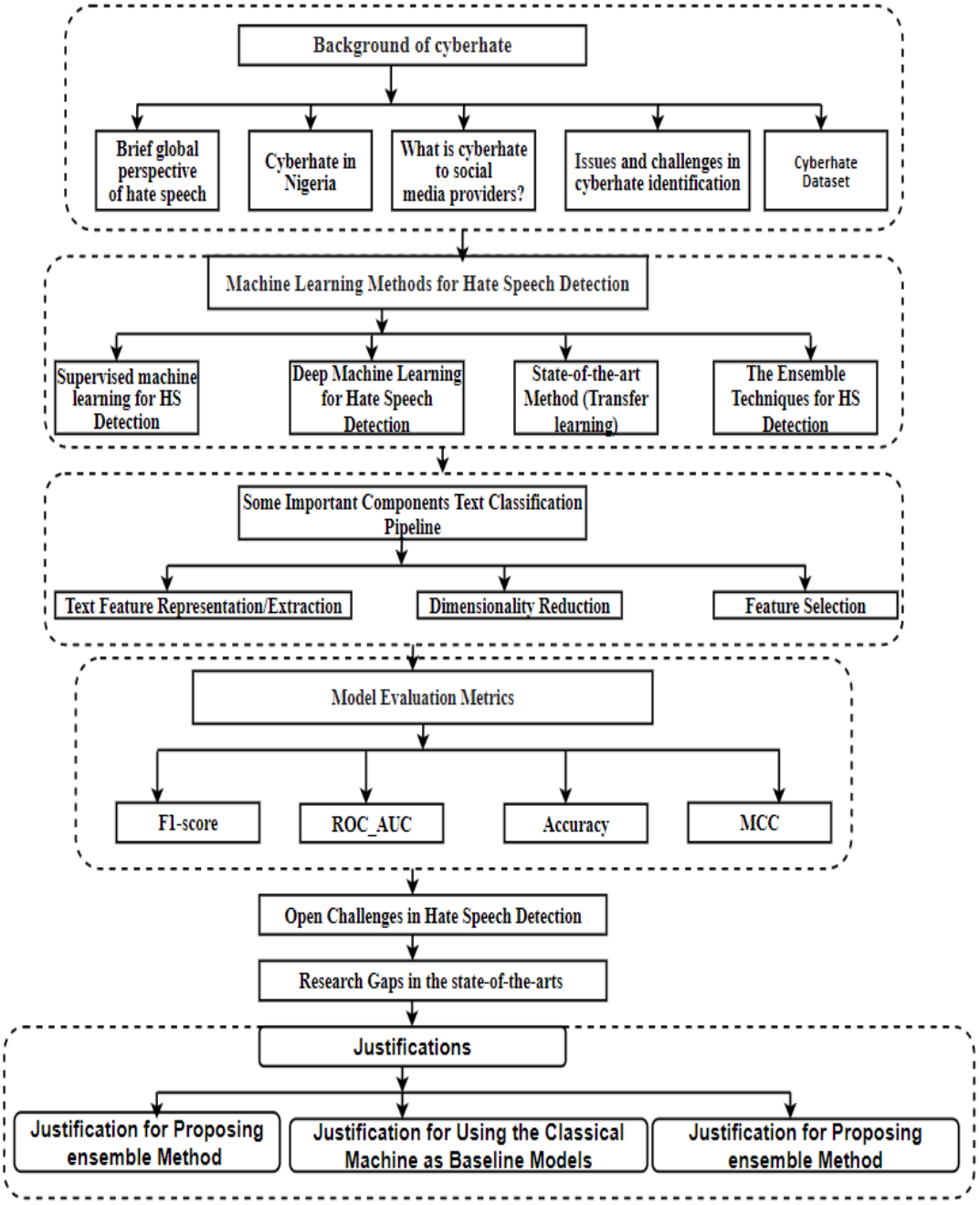

Figure 2.1 Literature Review Outline

2.2 Background of Cyberhate

No society is immune to cyber-hate. Different efforts in both human and financial resources have been expended towards managing this social ill. This section of the chapter will review from the global perspective, researchers' point of view, social media perspective and the Nigeria case study.

2.2.1 Brief Global Perspective of Hate Speech

Hate speech is as old as human existence, and different attempts have been made to put them under check (Briguglio et al., 2021; Haynes, 2019). The provision that guarantees the right of equality among human beings, such as the right to freedom from discrimination, is captured by article 1 of the Universal Declaration on Human Rights (UDHR), and the UN General Assembly adopted it in 1948, which stipulates that (Bukovska et al., 2010; Fattah & Fierke, 2009): "All human beings are born free and equal in dignity and rights (Assembly, 1948)." This principle prohibits any form of discrimination based on some non-exhaustive protected characteristics such as race, sex, the colour of skin and the like as contained in article 2 of the UDHR. Under article 2 of the UDHR, all humans should enjoy equal rights and freedoms.

After World War II, different regions adopted different strategies to manage hate speech based on the International Covenant on Civil and Political Rights (ICCPR) as the baseline (Haynes, 2019). American Convention on Human Rights (ACHR), African Charter on Human Peoples' Rights (ACHPR) and European Convention on Human Rights (ECHR) (Gelashvili, 2018). This was the common instrument deployed by Europe, Africa, and the US to fight hate speech.

In recent times, the wave of xenophobia in South Africa, the UK Brexit, and the United States election that brought Trump as president, all have refocused attention on cyberhate in the international discourse (Billingham & Bonotti, 2019). In Nigeria, for instance, the 2015 and 2019 general elections open the eyes of many to the fact that cyberhate is a timed bomb. The rise of social media draws national and international attention as it serves as the major means for carrying out cyberhate activities (Albarrak et al., 2020).

This cyber pandemic needs urgent attention in all countries as soon as possible. The UN, scholars and the Nigerian government, among other countries, have called for the need to censor or moderate hate speech on social media, which may soon tear many countries apart (Fortuna & Nunes, 2018; Iwuchukwu et al., 2019; MacAvaney et al., 2019). As cyberhate continues to torment individuals or groups in societies, there is an urgent need for a robust automatic cyberhate detection system. A better and more robust way of solving this problem is highly needed for peaceful coexistence.

In this review, the study is more interested in hate speech detection as part of offensive comments on social media networks. Offensive comments can be considered cyberbullying, aggressive, hate speech and other abusive comments. The attempt to automatically detect offensive comments on social media is a relatively new area of research in computer science (Fortuna & Nunes, 2018). The study used different nomenclatures to identify any derogatory or antagonistic message online. Table 2.1 shows some examples of these studies and their corresponding nomenclatures. For this research, the terms cyberhate, hate speech, offensive comments, or abusive messages were used interchangeably to mean the same. Any political post (text comments) which may stir up civil unrest belongs to this class.

Table 2.1 Studies on Offensive Messages

Nomenclature	Reference
Hate speech	(Warner & Hirschberg, 2012), (Kwok & Wang, 2013), (Burnap & Williams, 2014), (Djuric et al., 2015), (Wei et al., 2016), (Gambäck & Sikdar, 2017), (Saksesi et al., 2018), (Arango et al., 2019), (Plaza-del-Arco et al., 2021)
Cyberbullying	(Dinakar et al., 2011), (Slonje et al., 2013) (Whittaker & Kowalski, 2015), (Foong & Oussalah, 2017), (Tommasel et al., 2019), (Bozyiğit et al., 2021)
Profanity	(Sood et al., 2012), (Su et al., 2017) (Ratadiya & Mishra, 2019), (Yang & Lin, 2020), (Hahn et al., 2021)
Cyber-aggression	(Singh et al., 2018), (Tommasel et al., 2018), (Chatzakou et al., 2019), (Herodotou et al., 2020), (Herodotou et al., 2021),(Kumari et al., 2021)
Offensive	12 (Alakrot et al., 2018), (Ibrahim et al., 2020), (Pradhan et al., 2020), (Husain & Uzuner, 2021)
Toxic language	(Mohan et al., 2017), (Wijesiriwardene et al., 2020), (Sahana et al., 2020)
Cyber Harassment	(Winkelman et al., 2015), (Pearce, 2015), (Taylor et al., 2015)
Hostile messages	(Spertus, 1997)

2.2.2 Cyberhate in Nigeria

The problem of cyberhate and fake news in Nigeria is pervasive, and it is an issue of grave concern. For the past decade, cyberhate has attracted the attention of scholars and interest groups in Nigeria and around the world (Ajakaiye et al., 2019b). Experts in both arts and computer sciences are involved in this struggle.

Hate speech issues are also seen to be very common during an election period in Nigeria, and most cases translate into civil unrest where lives and property are lost (Asogwa & Ezeibe, 2020). There was a sharp increase in the number of deaths due to hate crime-related incidents in 2014 as preparation for the 2015 general election drew nearer (Bagga, 2019). This same trend repeats itself in 2018 as general election preparation was coming to an end, as reported in (Bagga, 2019). This statistic corroborates the fact that hate speech cases skyrocketed during electioneering in

Nigeria (Febriana & Budiarto, 2019; Lai et al., 2019). Countries whose democracies are still in the infant stages are more vulnerable, such as Nigeria (Jibril et al., 2017; Onimisi & Tinuola, 2019). This simply indicates that election is one of the motivators or trigger events for cyberhate, which generally translates into civil violence. The need to solve this problem using an automated approach is necessary. The best available option to do this is by employing MLA.

Leveraging the power of computers and MLAs to solve the problem of cyberhate is a relatively new field of research in the computer science domain (Fortuna & Nunes, 2018). Statistical methods were commonly used in the past to analyse these data for hidden trends that are useful for decision-making (Balaji et al., 2020; Li et al., 2020). As the population of Nigerians and the rest of the world continue to increase, the number of social media users also increases proportionally (Dilawar et al., 2018).

The exponential increase in data generated by social media renders the traditional statistical methods of analysing the data incapacitated (Syamala & Nalini, 2020). This massive data created by these users of social media networks is called big data (Ghani et al., 2019). Therefore, the need to leverage the power of computing algorithms to analyse these huge data becomes necessary through the use of MLAs (Balaji et al., 2020). By analysing the data, more insight was gained from the data generated due to users' interactions with each other (Han et al., 2019).

2.2.3 What is Cyberhate to Social Media Providers?

The most commonly used social sites are Facebook, Twitter, Instagram, WeChat, and YouTube (Omar et al., 2020) among others. Each social media site is aware that it is its legal duty to make cyberspace toxic-free for all and sundry to survive on it. This means that all social media sites need to accurately define hate speech in a

more comprehensive way such that all countries' citizens are protected. For comparison, the definition offered by each of the significant players is summarised in Table 2.2.

Table 2.2 Comparison of Definition/Explanation of the Term Cyberhate by the SM Providers

Social media	Definition/Explanation	Identifiable features
Facebook	"Facebook[6] defines hate speech as a direct attack on people based on what we call protected characteristics."	Religion, ethnicity, race, national origin, caste, sexual orientation, disability, sex, gender, and health status
Twitter	"Twitter[7] policy prohibits the promotion of violence against or directly attack or threaten other people based on protected characteristics."	Religion, ethnicity, race, national origin, caste, sexual orientation, disability, sex, age, gender, and severe health status
Instagram	"Instagram[8] removes credible threats of violence, hate speech, and the targeting of private individuals. We do not allow attacks or abuse based on protected characteristics."	Religion, ethnicity, race, national origin, caste, sexual orientation, disability, sexual orientation, gender, and health status
YouTube	"YouTube[9] does not allow content promoting violence or hatred against individuals or groups based on protected characteristics."	Religion, ethnicity, race, national origin, caste, sexual orientation, disability, sexual orientation, gender and health status, Immigration Status, Victims of a major violent event, age
WeChat	"In some jurisdictions, WeChat[10] classified personal information as sensitive or protected and are subject to stricter regulation than other types of Personal Information."	Religious inclination, health status, race, philosophical views, ethnicity

[6] https://www.facebook.com/communitystandards/hate_speech
[7] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[8] https://about.instagram.com/blog/announcements/instagram-community-guidelines-faqs#:~:text=Hate%20Speech%2C%20Bullying%20and%20Abuse,%2C%20religion%2C%20disability%20or%20disease.
[9] https://support.google.com/youtube/answer/2801939?hl=en
[10] https://www.wechat.com/en/privacy_policy.html