American Journal of Information Science and Technology
2024, Vol. 8, No. 2, pp. 34-44
https://doi.org/10.11648/j.ajist.20240802.12

Research Article

Development of Pidgin English Hate Speech Classification System for Social Media

Folake Oluwatoyin Adegoke[1, *], Bashir Tenuche[1], Eneh Agozie[2]

[1]Department of Computer Science, Prince Abubakar Audu University, Anyigba, Nigeria
[2]Department of Computer Science, University of Nigeria, Nnsuka, Nigeria

Abstract

With the widespread use of social media, people from all walks of life—individuals, friends, family, public and private organizations, business communities, states, and entire nations—are exchanging information in various formats, including text, messages, audio, video, cartons, and pictures. Social media also facilitates the distribution and propagation of hate speech, despite the immense benefits of knowledge sharing through these platforms. The purpose of this work was to construct a text-based, Pidgin English hate speech classification system (HSCS) in social media, taking into account the alarming rate at which hate speech is shared and propagated on social media, as well as the negative effects of hate speech on society. We used text data sets in Pidgin English that were taken from Twitter and Facebook (3,153). To train the Support Vector Machine (SVM) text classifier to identify hate speech in Pidgin English, 70% of the Pidgin English data set was annotated. The SVM classifier's performance was tested and assessed using the remaining thirty percent of the Pidgin English text data set. The test set findings' confusion matrix, as determined by the HSCS performance evaluation, was 62.04%, 64.42%, 0.7541, 0.6947, and 0.64 in terms of accuracy, precision, recall, F1-score, and Receiver Operating Characteristics (ROC) curve. When HSCS was compared to other Machine Learning (ML) classifiers, such as Logistic Regression (LR), Random Forest (RF), and Naive Bayes, the results showed that LR had accuracy and precision of 61.51% and 63.89%, RF had 54.88% and 50.65%, and Naive Bayes had 61.51% and 63.89%.

Keywords

Hate Speech Classification System, SVM Classifier, Machine Learning, Pidgin English

1. Introduction

Maravilla provides a succinct definition of hate speech as follows: Hate speech is any communication (text, image, video, etc.) that attacks, diminishes, incites violence or hate against individuals or groups, based on actual or perceived specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or any other" [20]. While other authors have given hate speech a variety of definitions, this study will focus on Maravilla's definition. Hate speech frequently results from a conceptual framework of "us versus them," where people distinguish between the "in-group," or the group they consider themselves to be, as opposed to the "out-group." In this analysis, hate speech directed towards members of the out-group could be divided into three main categories. The

*Corresponding author: Folakemiadegoke2022@gmail.com (Folake Oluwatoyin Adegoke)

Received: 16 March 2024; **Accepted:** 2 April 2024; **Published:** 14 June 2024

Copyright: © The Author(s), 2024. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

first, most frequently linked to hate speech, entails dehumanizing and demonizing members of the out-group [4]. The second involves violence, incitement, and even death against the out-group and shifts from the conceptual to the physical [5]. Out-groups are frequently the target of various negative speech forms, which are referred to as an early warning category, prior to these two more severe categories.

1.1. Effect of Hate Speech

Hate speech is widely available on social media. Because hate speech on these platforms is not adequately and firmly controlled, the trend is growing. Among the consequences of unchecked hate speech are:

Cyberbullying: The majority of kids have experienced cyberbullying at some point in their lives. Bullying on social media has become very simple because anyone can make a fake account and do anything without being discovered. It is possible to send rumors, intimidation messages, fake news, and threats to a large audience in an attempt to spread unrest and disorder throughout society.

Reputation damage: Hate speech on social media can quickly destroy someone's reputation by fabricating a tale and disseminating it widely. In a similar vein, companies may experience losses if their negative reputation is spread via social media.

Lenhart *et al*., 2016 claimed that social media has become a hub for the spread of hate speech [16], which has caused vulnerable people and groups to be reluctant to express themselves freely online and, in extreme situations, to completely withdraw from the online community. The frequency and volume of hate speech on social media are rising, creating an environment that is hostile to the targeted people and groups.

According to the definition, people's expressions in online platforms, policy statements, media coverage, and incorrect preferences, emotions, behaviors, and outright hatred now reflect exclusion, marginalization, and wrong preferences [30]. This has resulted in intolerance issues among individuals, political parties, and members of various tribes, communities, and religions. Conversations in various public forums have been weaponized for political, religious, and tribal games, with antagonistic language aimed at the vulnerable and oppressed who are the targets of repression and abuse.

Hate speech sows the seeds of distrust, hatred, and fear, weakening and destroying communities. If left unchecked, it may result in hate crimes or acts of violence directed towards the targeted individual or group. Hate speech has the power to sow the seeds of intolerance and hate, which in turn can justify hate crimes. Disagreement among friends, political parties, tribes, and religions can result from the propagation of hate speech on social media. Because hate speech incites people to fear attacks on their party, identity, tribe, or religion, it has a significant negative impact on the caliber of online discourse and contributions. If it is not controlled, it may result in violence or social disorder. Due to the prevalence of hate speech and abusive language on social media, these platforms are now tools for inciting conflict, fostering division, and inciting hatred among people in the community. Social media hate speech has made it more difficult for people to exercise their right to free speech. In certain cases, users who are exposed to hate speech may become radicalized as a result, which further erodes cultures and values. They radicalized people by convincing them that using violence against others—or even against oneself—to protect oneself or to dehumanize other groups is acceptable.

Social media is a low-cost communication tool that rapidly reaches millions of users, but hate speech on the internet can lead to vulnerable people experiencing anxiety and depression [35]. Numerous young women have experienced sexual harassment from young men on the internet [10], and other users have also experienced violence and harassment from other users [35]. Online hate speech perpetrators have occasionally encouraged others to harass them online by disclosing their target address (doxxing) (online harassment). Strong feelings of rage, guilt, shame, humiliation, fear, love, and hate can all be triggered by hate speech online [27].

Regarding the types of hate speech that are permitted on their platforms, Facebook and Twitter each have their own policies. The groups, however, are profit-driven organizations that compromise, which leads to uneven enforcement of the law prohibiting hate speech on their platforms. The 2012 Digital Terror and Hate Report revealed that there are approximately 15,000 problematic websites, social networking pages, forums, and newer online technologies like games and applications dedicated to inciting hatred based on ethnicity, race, or sexual preference. This is in stark contrast to the insignificant effort made by the above hate speeches on their platform. Many people are doing nothing to stop hate speech online [16]. The following are some benefits of the hate-controlled social media domain:

The goal of hate detection is to identify hate speech on Nigerian social media platforms early on and take appropriate action to stop hate speech and related vices. Aiding the government and agencies in the fight against hate crimes: One of the motivations behind the proposed system is that it would make it easier for the government and security agencies to spy on and apprehend criminals, which will aid in the fight against hate crimes.

Increased sales and reputation for the company would result from this suggested system's assistance in preserving goodwill. Positive feedback and word-of-mouth marketing can boost a company's sales and goodwill.

As there are many different religions and beliefs in the world, hate-free social media could aid in the development of a community where people can discuss and learn about these beliefs without harboring hatred. In a similar vein, people from various communities can get in touch to talk about and exchange similar non-hate ideas. Promotion: You can reach the widest audience by promoting your business, whether it is

online or offline. You have the entire world at your disposal to help them. Because advertising and promotion account for the majority of a business's expenses, this makes the businesses less expensive and more profitable.

By consistently and frequently using social media to interact with the appropriate audience, this can be reduced.

Positive Awareness: Social media also invents new ways for people to live and raises awareness. Social media has made it easier for people to find fresh, creative ideas that can improve their daily lives. Every member of society, from farmers to educators, students to attorneys, can profit from social media and its awareness factor. A model that can identify different types of hate speech in Nigerian languages on blogs, microblogs, and social networks is desperately needed in order to control and slow down the alarming rate of spread, given the volume and detrimental effects of hate speech on targeted individuals, groups of individuals, and society at large. This will ensure that everyone is included in public affairs by supporting the fundamental rights of communities and individuals. Because of this, the goal of this project is to create a hate speech classification system (HSCS) that can automatically identify hate speech in Pidgin English text on social media.

1.2. Machine Learning

The study of how a computer system can "learn" without being explicitly programmed is known as machine learning [39]. According to Juan and Roger 2017, computers learn from data [15]. According to Mitchell, the term "learning from experience" refers to a computer program's ability to improve its performance on tasks T as measured by P after it has gained experience [21]. This is the definition of machine learning that is most frequently cited; it was published in Tom A. Mitchell's 1997 book Machine Learning. There is close association between machine learning and five domains. specifically, statistics, generation, data mining, artificial intelligence, and optimization. Since both machine learning and artificial intelligence (AI) are thought to be techniques for imbuing machines with intelligence akin to that of humans, they are regarded as subsets of each other.

1.3. Supervised Education

The supervised approach to machine learning uses examples to learn; the data used to develop this kind of learning includes input objects and the intended output [23]. A function that converts inputs into intended outputs is produced by the different algorithms. The classification problem is a common structure for supervised learning tasks, where the learner must learn (to approximate the behavior of) a function that, by looking at the function's multiple input and output examples, maps a vector into one of several classes [25]. After that, the model forecasts using the hidden dataset. Given that the data is thought to be changing over time, the model is expected to complete the task reasonably, taking learning bias (also known as inductive bias) into consideration [37]. The process of supervised learning is as an algorithm for learning looks for a function with input and output spaces. The function is a component of a potential function space, also known as the hypothesis space. Using a scoring function that is defined as returning the value that results in the highest score can be convenient at times. Let F represent the scoring function space. Hyper-parameter tuning is what makes an algorithm useful in different problems, even though machine learning algorithms are applied in a wide range of fields and can solve multiple problems at once. The process of fine-tuning a machine learning algorithm's hyper-parameters to better generalize the problem without succumbing to over fitting is known as hyper parameter tuning [40]. follows: given a set of training data, where the feature vector and label of the data are, Thus, the process of initializing the values of the supervised learning (or unsupervised learning) algorithm prior to model training is referred to as the hyper-parameter. For machine learning or feature extraction, hyper parameter tuning is just as crucial as clean data. Generalization is the process by which an algorithm responds to new data after it has been trained. An overview of related literature Numerous authors have written about indigenous languages, as evidenced by the reviews of works in Table 1 below, which includes [9, 19, 28, 31, 34, 42]. Still, the hate speech detection systems on different social media platforms are not limited to Nigerian native languages. Thus, the focus of this work is on identifying Pidgin English hate speech, an indigenous language in Nigeria, on social media platforms. This creates the research void that this study aims to bridge. If this work is successful, hate speech in the indigenous language of Nigeria will no longer have a negative impact on social media or society as a whole.

1.4. Architecture of the Proposed HSCS

The text classification system input is Pidgin English text comments from social media users (Facebook and twitter) stored in CSV files. The various libraries and modules of python were engaged to carry out data cleaning, filtering, Natural Language Processing and feature extraction. The vectorized preprocessed text comments annotated were used to train the SVM classifier for text comment prediction. The generated text comments were classified as either hate or not hate for this binary classification. Hate text comments were blocked and retained in the hate database while the not hate were allowed to get to the intended social media users as shown in Figure 1.

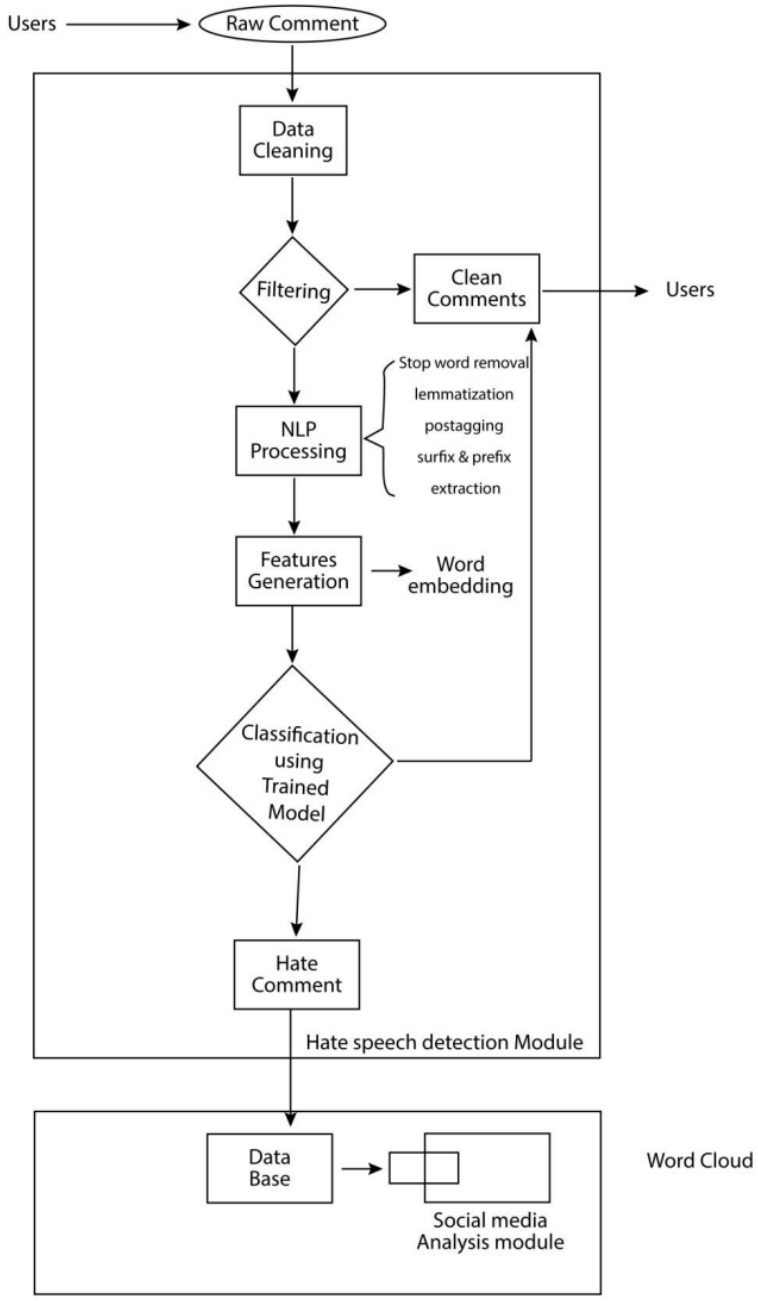

Figure 1. *Implementation Architecture for the proposed HSCS.*

Table 1. *Summary of Related Literature.*

S/N	Author(s)	Year	Title of Article	Contribution	Relevance to the study
1	[20]	2017	Detecting hate speech in Social Media	Hate Speech Detection	Discussion on Hate Speech
2	[7]	2017	Islamophobia	Monitor on Psychologists	Impact of anti- Muslim sentiment and exploring ways to prevent it
3	[34]	2017	A survey on hate speech detection using natural language processing	Detection on Hate Speech	A survey on Hate Speech Detection
4	[3]	2017	Deep learning for hate speech detection in tweets	Deep learning for Hate Speech Detection in tweets	Detection on hate speech deep learning

S/N	Author(s)	Year	Title of Article	Contribution	Relevance to the study
5	[2]	2017	Hate speech and foul languages	Hate speech and Foul language in the social media in Nigeria Research on the nature and extent of hate speech	Perception of Hate speech and Foul language in the social media in Nigeria
6	[24]	2016	Hate speech Report	Research on the nature and extent of hate speech	Report/ research on the nature of hate speech
7	[26]	2016	Abusive language	Online User Content	Abusive Language Detection in Online User Content
8	[32]	2017	Natural language processing	Natural Language Processing for Computer Mediated Communication	Measuring the Reliability of hate speech Annotations
9	[9]	2017	Hate speech detection	Automated hate speech detection	Hate speech detection and the problem of offensive language
10	[1]	2017	Hate speech wrong narrative	Wrong narrative of hate speech	Wrong narrative of hate speech for national discourse integration
11	[33]	2017	The menace of hate speech	The menace of hate speech	How the menace of hate speech can have effect on the society
12	[11]	2017	Causes of hate speech	The cause of hate speech in Nigeria.	The causes of hate speech in Nigeria and the best way to tackle it.
13	[12]	2017	Dataset and annotation	Legal framework, Dataset and annotation.	Dataset, Legal framework, and annotation schema for socially Unacceptable On-line discourse practices in Slovene in proceedings of the workshop on Abusive language online (ALW).
14	[38]	2017	Supreme court unanimously Reaffirms	Supreme court unanimously Reaffirms Hate speech:	There is no Hate speech exception to the first amendment.
15	[8]	2017	What' wrong with counter speech	The effect of counter speech.	The effect of counter speech to the society and how it can be eradicated
16	[42]	2018	Hate speech detection	A solved problem?	The challenging case of long tail on twitter
17	[31]	2018	Hate speech detection on twitter.	Feature engineering vs feature selection.	Hate speech detection on twitter using Feature engineering vs feature selection.
18	[19]	2017	Detecting Hate speech in social media.	Detection of hate speech in social media.	The first step in curbing the hate speech on social media is detection of hate speech and abusive words.
19	[3]	2017	Deep learning for hate speech.	Learning deep about hate speech.	Deep learning for hate speech control and prevention
20	[28]	2018	Detecting Offensive languages.	Detection Of Offensive languages.	The importance of offensive language detection and controls.
21	[18]	2019	Hate speech detection; Challenges and solutions.	Hate speech detection	Challenges of detecting hate speech.
22	[6]	2020	Hostility detection dataset in Hindi	Hindu Dataset for hostility.	Low resource language dataset
23	[22]	2021	Racism, hate speech and social media; A systematic review and critique.	Review of hate speech and social media	Hate speech review.
24	[17]	2017	Deep learning for natural language processing.	Learning deeply about natural language.	Deep learning for natural language processing against abusive and hate speech

S/N	Author(s)	Year	Title of Article	Contribution	Relevance to the study
25	[13]	2021	Cross-lingual offensive language identification for low resource languages; the case of Marathi	Identification of offensive language	Ways of identification of offensive language.
26	[36]	2022	L3Cube-Maha Hate; A tweet-based Marathi Hate Speech Detection Data- set and BERT Model	Hate speech detection	Text based hate speech detection
27	[14]	2022	Investigating the effect of preprocessing Arabic text on offensive language and hate speech detection	Preprocessing of data Input	Effect of preprocessing on data input
28	[29]	2022	AI technologies and application in the metaverse	AI technologies and application	Application of AI in test classification

2. Methodology

Analytical and experimental methods are combined in the research process. While machine learning, model training, and testing were conducted using analytical methodology, data collection, extraction, and preparation were done using experimental methods.

2.1. Data Source

The data set source consisted of text comments posted by users on social media platforms, specifically Facebook and Twitter, in the Pidgin English indigenous language of Nigeria. A link to a response collection page was posted on my social media page.

2.2. Data Collection

Using the twin library in Python, web scraping was used to gather the publicly available crowd sourced Yoruba and Pidgin English text comments, which were then imported into a CSV file from Facebook and Twitter. By establishing a secure connection with the source website (my Facebook page), where the contents were fetched into the document object module (DOM), access was created to be able to copy the necessary content. The DOM's content was sorted using value dash pairs, and relevant data were chosen and saved in a CSV file. To allow for convergence and prevent overfitting, the data sets that were collected were rich in size—1,420 for the Yoruba data set and 3,155 for the Pidgin English text data.

Sample Data Sets

1) dis niger politicians children and drugs Na 5 and 6 you no fit separate them. Thunder go fire dem as dey scatter dis country.
2) niger and kogi I no know which one worse kogi no companies na just hotel wey politicians build just full
3) dis aboki na just goat weh no de use sense and aboki no de joke with gold
4) na bandit him be nonsense mumu person wey dem go capture within few days just dey warm up to spend ur remaining days in kuje prisons
5) dem still dey kill inside fasting na wa oh enemy in disguise trust no one else you go see ur sef for mortuary ooooo
6) good percent of my friendship with people na bus stop be the end na enemy be that nor b friend at all
7) 35m make you murder person na poverty b this abi na greed. Rest on man I pray your killers play host to karma
8) na only aboki dey commit this crime abi later dey will come online nd be shouting mak dem give dem presidency dere elders dey see all diz now dem no talk people with dia oloshi sense, well na online dem sabi win dem no fitwin in real life.

2.3. Annotation of Data Sets

A statistical technique similar to that in scikit-learn was used to divide the data set into training and test sets. Three annotators used crowdflower to upload and distribute the 70% of the pidgin English text to a specific website made from an acquired domain. They then manually tagged the Pidgin English text comments as either hate or not hate (binary classification). The training set for the support vector machine (SVM) and the model parameter optimization for the Pidgin English text data were the manually annotated text comments. The remaining portion that was left untagged was used as text data, with 30% of the set for Pidgin English. All of the documents that included the text data set were categorized as either hate or not-hate.

2.4. Preprocessing Data Sets

By lowering dimensionality and eliminating extraneous information, preprocessing input text can greatly enhance text classification [14]. Natural language processing (NLP), a Python library that provides modules for processing text comments to remove bothersome and pointless elements from the text comments in Pidgin English, was used to pre-process the datasets in a number of ways.

2.5. Data Cleaning

Removals Usernames, URLS, hash tag, Gmail, symbols, xml, were removed, remaining only the harshly part the comments. Also removed were punctuations such as “,”, -, and? Stop words such as ‘a’. ‘the’, ‘and’, ‘is’, and Natural Language Tool Kit (NLTK) predefined set of English stop words were removed. Yoruba stop words such as ‘un’, ‘o’, ‘mo.’, ‘mi’; ‘a’, ‘ti’, etc. were removed to avoid delayed convergence. Special characters were also removed such as -, I, e, &, %, #, were also removed.

Translation

Pidgin words such as “don’t”, “can’t”, “n’t”, won’t replaced with “do not”, “cannot”, “not”, “will not”, respectively, Upper case letters were replaced with lower case and accented to unaccented eg ȧ, ė, ē with a, e, and n, respectively.

Tokenization

Comments were broken into tokens in a process known as tokenization. A meaningful unit of text is called token.

Stemming

A snowball stemmer was utilized to carry out the stemming technique, which is used to find the root of a word. "Go" is the root of "went," "offend" is the root of "offended," etc. Prefixes and surfixes were likewise eliminated.

Lemmatization

During the lemmatization process, all duplicate terms were eliminated using the Wordnet lemmatizer. Using the scikit-learn library, a bag of words (BOW) was created after processing each and every comment.

Vectorization

The comment was converted from text to a numerical format that the SVM could utilize. Using the word2vec program, each text comment is shown as a vector. Numerical data were generated from categorical attributes, namely hate and not hate. In this binary classification, hate was denoted by the number (0 1), and non-hatred by the number (1 0).

Data Scaling

There were more hate text data than not hate text data. Consequently to avoid attributes in high numerical strength dominate those in smaller numerical strength and avoid over fitting, the attributes were linearly scaled.

Kernel Selection

Given that the dataset can be divided linearly, a linear kernel was employed. In the higher dimensional space, a hyper plane with maximal margin and linear separation was produced by the support vector machine (SVM). K (x1, xj) =˘(x1) T˘(xj) is the kernel function.

System Main Menu

The main menu for the implementation of the classification system is shown in bellow:

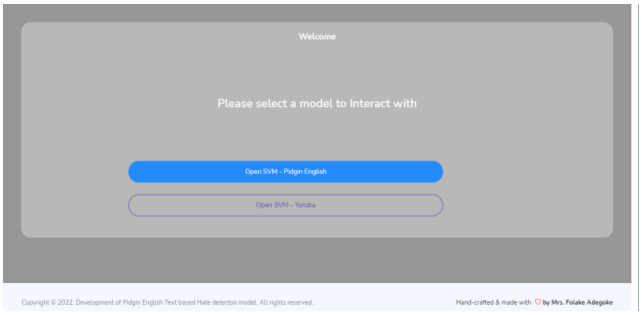

Figure 2. *HSCS Main Menu.*

All of the operational commands that users of the application will interact with and carry out in order to predict text comments in Pidgin English are contained in the main menu of the system. Essentially, it has a Pidgin English language section with features on the software dashboard like Run Task and Log out. A user can access any of these pages only after successfully registering with the system and logging in with the correct user name and password. The model buttons provide the user with the option to choose which language they would like to work in. Pidgin English is one of the indigenous language varieties. The process of making predictions begins when a specific language is taken into account. The workflow coordination is made possible by the Natural Language Tool Kit (NLTK), and the text comment prediction for the input text data is initiated by the Support Vector Machine (SVM) classifier.

System Sub Menu

Login Menu

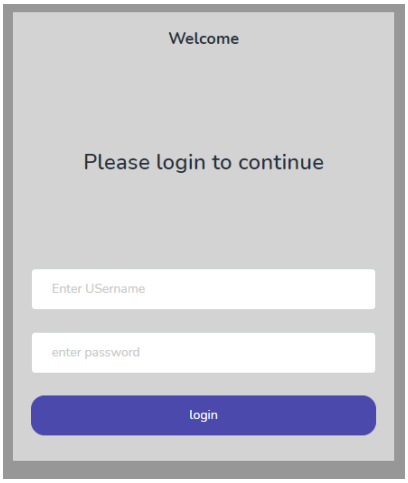

Figure 3. *User Login Page.*

On the user login page, the two main fields are the user

name and password. It is anticipated that the user will fill in these fields accurately, since he may have registered based on the screenshot that follows, which is displayed in Figure 4 below. The user will not be able to access the main application and utilize it if the correct details are not entered.

User Registration Menu

Users of the application must first be registered on the system before proceeding to perform any task. The user must provide full name, email and user will also supply the username and password to be used in login into the system. It is only after the registration on the software that users can access the services to be offered by the software.

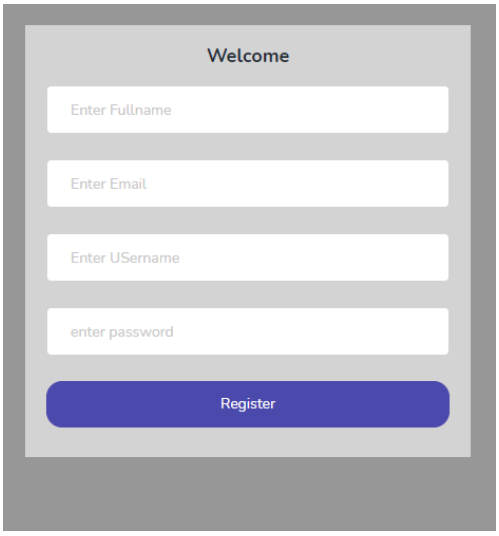

Figure 4. *User Registration Page.*

Figure 5 shows the hate words filter. For more precision and accuracy of comments processing and classification a sub menu was added to make registration of new hate comments to the application database to help make classification and text processing easier, accurate and more precise.

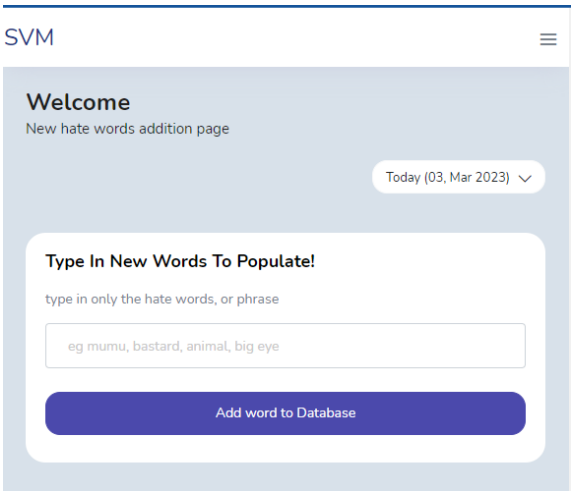

Figure 5. *Hate words filter page.*

3. Testing Software

Functional testing along with the Alpha testing strategy are used to test the Pidgin English language SVM text classifier. The Alpha testing method is utilized in the study to verify the developed application's efficient work-ability. This guarantees that the researcher will test the application and look for any potential problems. The functional testing, sometimes known as "black box" testing, consists solely of observing the output for specific inputs and applying a boundary value analysis value to it.

3.1. The Datasets

Table 2. *Pidgin English language text dataset distribution.*

	Hate	Not hate	Row Total
Dataset	1260	944	2204
Training Test	545	406	951
Column Total	1805	1350	3155

3.2. Prediction Table

Table 3. *Pidgin English Text Data Set Predicted.*

Test Data	SVM Predicted	SVM Predicted	Row
Labels	Hate	Not Hate	Total
Hate	411 (tp)	134 (fn)	545
Not Hate	227 (fp)	179 (tn)	406
Column Total	638	313	951

Pidgin English Text Data Set Confusion Matrix

$$\begin{pmatrix} 411 & 134 \\ 227 & 179 \end{pmatrix}$$

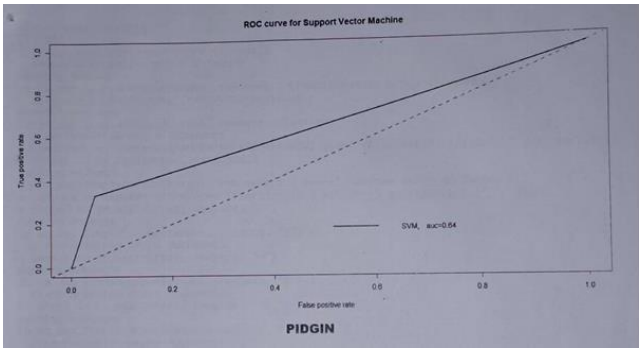

Figure 6. *Receiver Operating Characteristic curve for Pidgin English SVM Classifier.*

The area under the curve (AUC) as shown in Figure 6 represents the degree of separability. It shows how well the model can distinguish between hate and non-hatred. AUC improves evaluation over accuracy by utilizing the models' true positive and false positive rates.

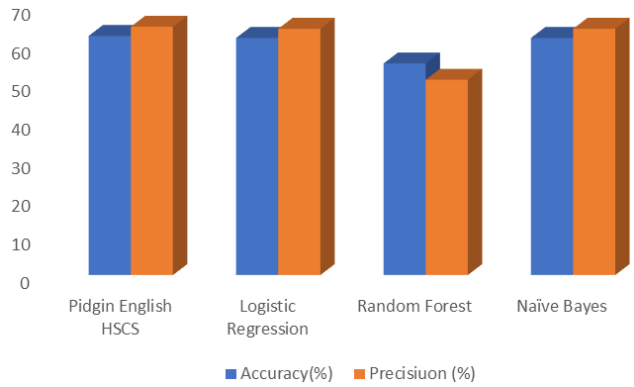

Figure 7. *Pidgin English SVM Classifier against other machine learning.*

4. Conclusion

Hate speech sows the seeds of distrust, hatred, and fear, weakening and destroying communities. If left unchecked, it may result in hate crimes or acts of violence directed towards the targeted individual or group. Hate speech has the power to sow the seeds of intolerance and hate, which in turn can justify hate crimes. Disagreement among friends, political parties, tribes, and religions can result from the propagation of hate speech on social media. Because hate speech incites people to fear attacks on their party, identity, tribe, or religion, it has a significant negative impact on the caliber of online discourse and contributions. If it is not controlled, it may result in violence or social disorder. This work has been able to developed HSCS that can identify and block hate text comment in Pidgin English for Social Media. The Type I and Type II errors noted can be reduced by improving the HSCS.

5. In Summary

Comparing the SVM classifier to other classifiers, its performance was superior. By using the model as a back-end for social media operators, hate speech in Nigerian indigenous language of Pidgin English could be less harmful to society and social media when properly controlled on Nigerian social media through automated detection.

Abbreviations

SVM	Support Vector Machine
AUC	Area Under Curve
Tp	True Positive
Fp	False Positive
Tn	True Negative
Fn	False Negative
NLTK	Natural Language Tool Kit
BOW	Bag of Words
HSCS	Hate Speech Classification Scheme
CSV	Comma Separated Values
DOM	Document Object Module
ML	Machine Learning
ROC	Receiver Operating Characteristics

Author Contributions

Folake Oluwatoyin Adegoke: Conceptualization, Software, Investigation, Methodology, Writing – original draft

Bashir Tenuche: Resources, Data curation, Formal Analysis

Eneh Agozie: Supervision, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Akanji A. (2017). Hate speech: wrong narrative for national discourse integration. Accessed 12/3/2018 https://www.thecable.ng/hate-speech-wrong-narrative-nationaldiscourse-integration

[2] Alakali, T. T., Faga, H. P., & Mbursa, J. (2017). Audience perception of hate speech and foul language in the social media in Nigeria: Implications for morality and law. *Academicus International Scientific Journal*, *15*, 166–183. https://dx.medra.org/10.7336/academicus.2017.15.11

[3] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* pp. 759-760. https://doi.org/10.48550/arXiv.1706.00188

[4] Bahador, B. (2012). Rehumanizing Enemy Images: Media Framing from War to Peace. In: Korostelina, K. V. (eds) Forming a Culture of Peace. Palgrave Macmillan, New York. https://doi.org/10.1057/9781137105110_9

[5] Bahador, B. (2020). Classifying and Identifying the intensity of hate speech. https//items.ssrc.org

[6] Bhardwaj, M., Akhtar, M. S., Ekbal, A, Das, A. and Chakraborpy, I (2020). Hostility detection dataset in hindi. *ACM Transactions on internet Technology* (TOIT) 20(2): 4-22. https://doi.org/10.48550/arXiv.2011.03588

[7] Clay R. A. (2017) Islamophobia. Psychologists are studying the impact of anti-Muslim sentiment and exploring ways to prevent it *American Psychological Association* 48(4): 34.

[8] Coustick-Deal, R. (2017). What's wrong with counter speech? Retrieved 13/3/2018 from https://mediumcom/@ruthcoustickdeal/medium-com-whats-wrong-with-counterspeech- f5e972b13e5e

[9] Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Eleventh International AAAI Conference on Web and Social Media 2017 11(1) https://doi.org/10.1609/icwsm.v11i1.14955

[10] Duggan, M. (2017) Online Harassment 2017 (Rep) Raw Research Centre. www.pewresearch.org

[11] Ezike, G. (2017).The cause of hate speech in Nigeria Retrieved on 16/2/2018 from https://www.linkedin.com/pa/se/cause-hate-speech-nigeria-genesis-ezike

[12] Fiser, D., Erjavee, T, and Ljubesic, N, (2017). Legal Framework, Dataset and Annotation Schema for socially Unacceptable On-line Discouse Practices in Slovene: In Proceedings of the Workshop on Abusive Language Online (ALW).

[13] Gaikurad, S. Panasingle, J., Zampreri, M. and Homan, C. M. (2021). Cross-lingual offensive language identification for low resource languages: The case of Marathi. https://doi.org/10.48550/arxiv2109.03552

[14] Husain, F and Uzuner, O. Investigating the effect of pre-processing Arabic Text on offensive language and hate speech detection *ACM Transactions on Asia and low Resources language information Processing* 2022 21 issue 4(73): 1-20. https://dx.doi.org/10.1145/3501398

[15] Juan, C. and Roger, G. M. Machine learning phases of matter. *Nature Physics 2017 13(5):* 431–434. https://doi.org/10.48550/arXiv.1605.01735

[16] Lenhart, A., Ybarra, M., Zickhur, K and Price – Feency M, (2016). Online Harassment, Digital abuse and cybertalking in American (Rep) New York, NY Data and society. https://www.datasociety.net/pubs/oh/online_Harrasment_2016

[17] Lopez, M. M. and Kalita, J. Deep Learning Applied NLP *arXiv preprint arXiv: 1703.03091* (2017).

[18] MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N., and Frieder, O. (2019) Hate speech detection: Challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal

[19] Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. Proceedings of the International Conference Recent Advances in Natural Language Processing, pp 467-472. https://doi.org/10.26615/978-954-452-049-6_062

[20] Maravilla, C. S. Hate Speech as a War Crime: Public and Direct Incitement to Genocide in International Law. *Tulane Journal of International & Comparative Law 2008 30(2): 523-548.*

[21] Mariette, A. and Rahul, K. (2015). Machine Learning: in Efficient Learning machine. Berkeley, https://doi.org/10.1007/978-1-4302-5990-9

[22] Matamoros-Fernandez, A. and Farkas, J. (2021) Racism, hate speech and social media: A systematic review and critique *Television and New Media* 20 (2): 205-224. https://doi.org/10.1177/1527476420982230

[23] Muhamedyev, R. I. (2015). Machine learning methods: An overview. Computer modelling and new technologies, 14-29.

[24] Nadium and Fladmoe (2016) Silencing Women? Gender and Online Harassment https://doi.org/10.1177/0894439319865518

[25] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons, pp. 51-62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

[26] Nobata, C., Tetreault, J. R., Thomas, A. O., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*.

[27] Online Harassment: Extremists Ramp up Trolling, Doxxing Effort (of A) Retrieved from https://www.adl.org/resources/blog/online-harassment-extremists-ramp-trolling-doxxing-efforts

[28] Pitsilis, K. G., Ramampiaro, R. and Langseth, H. (2018) Detecting offensive language in tweets using deep learning. https://doi.org/10.1007/s10489-081-1242-y

[29] Qiang, W. U., Xlieting, J. I. Linyuan and LYU (2020). AI technologies and applications in the metaverse *Chinese J. of Intelligent Science and Technology* 4(3): 324-334. https://doi.org/10.11959/j.issn.2096-6652.202241

[30] Resnick, B. (2017). The dark Psychology of dehumanization explained. Retrieved from https://www.vox.com/science-and-health/2017/3/7/14456154/dehumanization-psychology-explained

[31] Robinson, D., Zhang, Z. and Tepper, J. (2018). Hate speech detection on twitter: Feature engineering vs feature selection. https://doi.org/10.1007/978-3-319-98192-5_9

[32] Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In: Beißwenger M, Wojatzki M, Zesch T (eds.) Proceedings of NLP4CMC III, pp. 6–9; 2016. https://doi.org/10.17185/duepublico/42132

[33] Salihu, J. S. (2017). The menace of Hate Speech https://www.dailynigeria.com/feature-the-hate-menace

[34] Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Social Science Computer Review 39(2): 089443931986551 https://doi.org/10.18653/v1/W17-1101

[35] Tynes, B. M, Giang, M. T. Williams, D. R. and Thompson, G. N. (2008). Online Racial Discrimination and Psychological Adjustment among Adolescents. *J. of Adolescent Health* 43(6): 565-569. https://doi.org/10.1016/j.jadohealth.2008.08.021

[36] Velankar, A., Patil, H, Gore, A., Salinke, S and Joshi, R. (2022) L3Cube- maha Hate: A Tweet – based Marathi Hate:// speech Detection Dataset and BERT models, pp1-12. https://doi/org/10.4855/arXiv.2203.13778

[37] Vinicius, Z., David, R., Adam, S., Victor, B., Yujia, L., Igor, B., Pe. (2019). Deep Reinforcement Learning with Relational Inductive Biases. ICLR, London, UK: Deep Mind. Pp 1-18.

[38] Volokh, E. (2017). Supreme Court unanimously Re-affirms: There is no Hate Speech Exception to the First Amendment Washington.com Retrieved on July, 2020.

[39] Wehle, H. D. (2017). Machine Learning, Deep Learning and AI: What's the Difference. In Internationan Conference on Data scientist innovation day. Bruxelles, Belgium.

[40] Yang, L., and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 295-316.

[41] Young, T. Hazarika, D. Poria, S and Cambria, E (2018) "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," in IEEE Computational Intelligence Magazine, 13 (3), pp. 55-75, https://doi.org/10.1109/MCI.2018.2840738

[42] Ziqi Zhang, Z. (2018). Hate speech detection: A solved problem? the challenging case of long tail on twitter. https://doi.org/10.3233/SW-180338