International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 09 | Sep 2024 www.irjet.net p-ISSN: 2395-0072

An AI System for the Detection of Hate Speech Encoded in Igbo Native Language

P. Ana[1], G. I. Emereonye[2], A. C. Onuora[3], C. C. Ukaegbu[4], R. C. Aguwamba[5], P. C. Sunday[6]

[1]Department of Computer Science, Cross River State University of Technology – Calabar, Cross River State, Nigeria.
[4]Department of Computer Science, Milwaukee school of Engineering University, USA.
[5]Department of Office Technology and Management, Akanu Ibiam Federal Polytechnic Unwana, Ebonyi State, Nigeria.
[2,3,6] Department of Computer Science, Akanu Ibiam Federal Polytechnic Unwana, Ebonyi State, Nigeria.

Abstract - *This project was motivated by the pressing need for language-specific solutions to combat hate speech nuances in Igbo native language. The aim is to develop a robust Hate Speech Detection System integrated into the Facebook platform. The specific objectives include creating the first-ever hate speech dataset for Igbo, employing advanced natural language processing (NLP) techniques, and ensuring ethical considerations in system deployment. Methodologically, the project involves comprehensive data pre-processing, neural network model creation using Keras, systematic testing, and deployment on Facebook with a meticulous process. The key findings include the successful development of a reliable hate speech detection model with specific strengths and weaknesses. The project contributes to knowledge by establishing a systematic approach to language-specific hate speech detection and emphasizes the importance of ongoing efforts such as dataset expansion and community engagement.*

Key Words: Hate Speech, Igbo language, Artificial Intelligence, Transformer, Social Media

1. INTRODUCTION

Hate speech is a denial of the values of tolerance, inclusion, diversity, and the very essence of human rights norms and principles. It may expose those targeted to discrimination, abuse, and violence, but also social and economic exclusion. When left unchecked, expressions of hatred can even harm society's peace and development, as it lays the ground for conflict, tension, and human rights violations, including atrocity crimes.

Importantly, combating hate speech first requires monitoring and analysing it to fully understand its dynamics. Since the spread of hateful rhetoric can be an early warning of violence – including atrocity crimes – limiting hate speech could contribute to mitigating its impact. The authors of hate speech should also be held accountable, to end impunity. Monitoring and analysing hate speech is a priority for many UN entities, including UNESCO - the United Nations' specialized agency for education, science and culture - which supports and undertakes research, which supports research to better understand its dynamics.

To combat this issue, social media giants like X and Meta have implemented several strategies. They have established comprehensive policies against hate speech, clearly defining its boundaries and outlining the consequences for violators [1]. Furthermore, they utilize advanced technologies such as machine learning algorithms and human reviewers to detect and remove hateful content. Additionally, social media platforms have implemented measures like user identity verification and reporting tools to enhance accountability and facilitate user involvement in combating hate speech.

Despite these efforts, hate speech remains a persistent challenge due to its evolving nature. Therefore, it is crucial to promote user education, ensuring awareness of hate speech and reporting mechanisms. Users and regulators must hold social media companies accountable for their role in combating hate speech. Supporting organizations dedicated to countering hate speech through donations or volunteer work can also make a significant impact. By integrating these multifaceted approaches, we can strive to foster a safer and more inclusive online environment, where hate speech is effectively addressed and minimized. This collective effort is essential to create a positive digital space for all users, promoting understanding, respect, and dialogue.

Specific objectives of this study are:

1. Evaluating and validating the performance of a developed model using appropriate evaluation metrics and validation techniques.

2. Designing and implementing a software that can accept input of Igbo text and respond whether the text contains hate speech or not using the AI powered hate speech detection model.

3. Deploying and integrating the trained model into the application to enable detection of hate speech.

The remaining sections of the paper are as follows: Section II provides a theoretical analysis of hate speech detection, covering definitions and relevant concepts. Section III outlines the methodologies used for software design and model training. In Section IV, the paper presents and

discusses the obtained results. Finally, the Conclusion (Section V) summarizes key findings and implications of the research.

2. THEORITICAL ANALYSIS

Hate speech is a pervasive and harmful phenomenon that affects the digital social sphere. To combat this problem, researchers have been exploring various natural language processing (NLP) techniques to automatically detect and filter hateful and offensive content. In this paper, we review some of the recent studies on hate speech detection, focusing on the methods, datasets, and challenges involved.

Another challenge in hate speech detection is the diversity of languages and dialects that are used on social media platforms. Most of the existing studies focus on English, which limits their applicability and generalization to other languages. Moreover, some languages have specific features that make hate speech detection more difficult, such as code-mixing, which involves the use of two or more languages in the same text. [2] tackled this problem by developing a hate speech detection model for Tamil and Malayalam code-mixed texts, using BERT models that are pre-trained on multilingual corpora. They show that their model outperforms traditional machine learning models, such as support vector machines and random forest, in terms of F1-score and accuracy.

However, [3] elaborated that even within the same language, hate speech detection can be challenging, due to the use of linguistic obfuscation techniques, such as misspellings, abbreviations, slang, and euphemisms. These techniques are used by the perpetrators to evade detection and to convey their hateful messages more implicitly. [4] investigated this issue by analysing the limitations of keyword-based methods, which rely on predefined lists of offensive words, and propose a more robust and flexible method that uses word embeddings and semantic similarity measures to identify hateful and offensive speech.

To address the gap in the literature on hate speech detection in languages other than English, [5] propose methodologies for offensive language identification in languages with large speaker populations, such as Hindi, Arabic, and Chinese. They use various datasets and models to test their methodologies, and present their results and findings. They emphasize the need for more research and resources in languages other than English, as hate speech is a global and multilingual problem.

One of the ways to overcome the language barrier in hate speech detection is to use multilingual models that can handle multiple languages simultaneously. [6] demonstrates the effectiveness of multilingual transformers, such as XLM-RoBERTa and mBERT, for hate speech detection across different languages, such as English, German, Turkish, and Arabic. They show that their models achieve high accuracy and F1-score, and outperform monolingual models, such as BERT and RoBERTa, in most cases.

Another way to improve the performance and generalization of hate speech detection models is to use multi-task learning and multilingual training, which involve training the models on multiple tasks and languages at the same time. [7] advocated for this approach, as it can enhance the models' ability to learn from diverse and complementary data sources, and reduce the need for large and annotated datasets for each task and language. They show that their models achieve state-of-the-art results on various hate speech detection tasks and languages, such as English, Hindi, and Bengali.

However, hate speech detection is not only a technical problem, but also a social and ethical one [8]. It is important to understand the context and motivation behind hate speech, and to interpret the behaviour and decisions of the models. Therefore, explainable artificial intelligence (XAI) is essential for hate speech detection, as it can provide transparency, accountability, and trust to the users and stakeholders. [9] discussed the current state of XAI in hate speech detection, and propose improvements and directions for future research. They emphasize the importance of incorporating human feedback and evaluation into the XAI process, and of developing user-friendly and interactive interfaces for explaining the models.

One of the factors that can affect the performance and interpretation of hate speech detection models is the contextual information that surrounds the text. Contextual information can provide clues and cues about the intention and tone of the speaker, and can help to distinguish between hate speech and legitimate speech. [10] explore the impact of contextual information on hate speech detection, and develop a model that leverages contextual cues from Twitter replies to identify hate speech. They show that their model improves the accuracy and recall of hate speech detection, and also provides more coherent and consistent explanations.

Finally, [1] and [11] discusses various machine learning techniques for hate speech detection, and categorizes them into shallow and deep learning approaches. He compares and contrasts the advantages and disadvantages of each approach, and emphasizes the importance of interpretability and explainability for both. He also provides an overview of the datasets, evaluation metrics, and challenges involved in hate speech detection, and suggests future directions for research and development.

In conclusion, while significant strides have been made in hate speech detection, challenges persist, including generalization across languages and model interpretability. This therefore warrants further research to develop more effective and transparent solutions to combat hate speech written in low resource languages within the cyberspace.

3. METHODOLOGY

This project uses the Object-Oriented Analysis and Design Method (OOADM) as the most suitable methodology for hate speech detection in the Igbo language. Using OOADM, we will capture the dynamic behavior, modularity, and user-centric nature of the hate speech detection system in the Igbo language. OOADM's diagrams, such as use case, interaction, and package diagrams, were used to illustrate the processes, interactions, and modular design of the system [12].

In the high-level model below (see fig-1), the subsystems are shown as, text input interface, AI model for the detection of hate speech (which performs the core function of the system) [13], and the classification output interface.

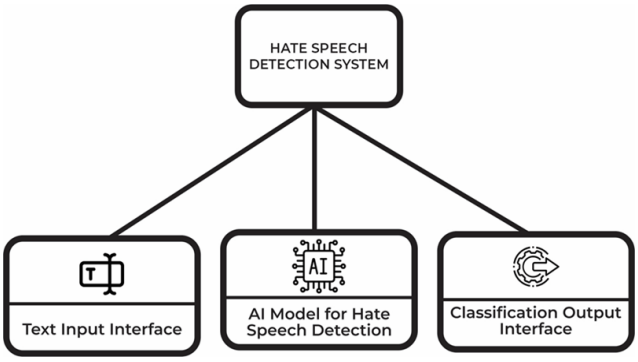

Fig-1: High Level Model of the Proposed System

3.1 Object Diagram of the New System

Class Diagram

The class diagram in fig-2 depicts the structure of the system. It provides a visual representation of the major classes (TextInput, DetectedOutput and DetectionEngine) in the system, their attributes, methods, and most importantly the relationships amongst them.

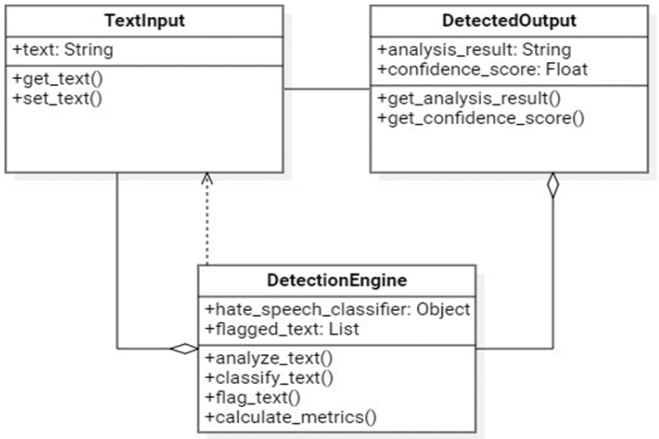

Fig-2: Class Diagram

Use Cases

The use case diagram in fig 3.3, shows that the user basically submits text and sees the output of the analysis. The rest of the processes are automated by the detection engine.

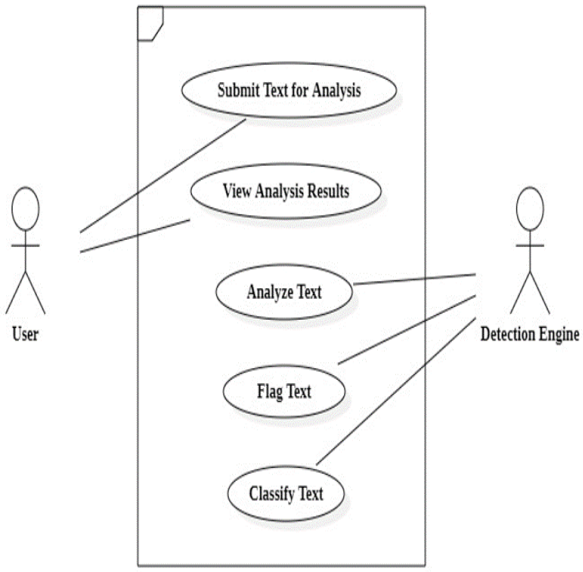

Fig-3: Use Case Diagram

Sequence Diagram

The chronological order of interaction between objects in the software is shown in fig-4 below. It depicts how text is submitted for analysis and how the TextAnalyzer comes up with the result of its classification.

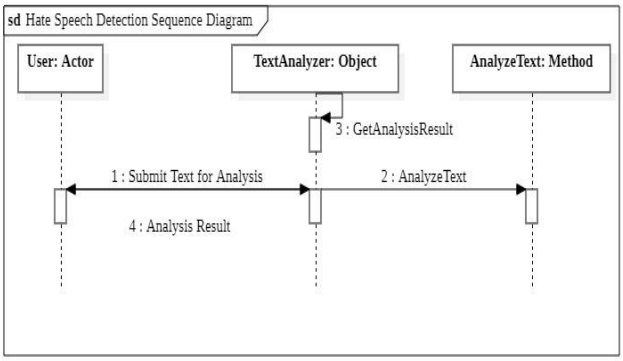

Fig-4: Sequence Diagram

Activity Diagram

Activity diagram in fig 3.5 illustrates the overall flow of activities within the system. Here we see the flow moving from the submission of text, analysis, flagging, classification and display of the output.

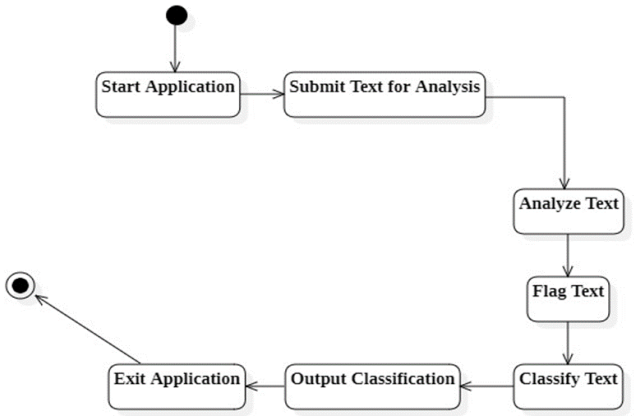

Fig-5: Activity Diagram

Input Design

The system has a very simple user interface as shown in fig 3.6. This includes a text box and a command button used to submit inputted text

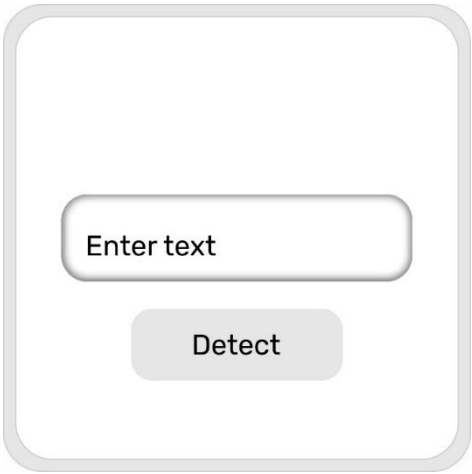

Fig-6: Input Design

Output Design

The output design in fig 3.7 also shows a simple display of the result of the detection process which is either 'Hate Speech' or 'Non-Hate Speech' each followed by an icon that enhances the understanding of the information being passed.

Fig-7: Output Design

3.2 Choice of Programming Environment/Packages

Programming Language: Python [14].

Development Platform: Visual Studio Code /

Libraries and Frameworks:

a. Pandas: used for data manipulation.

b. Genism: used for topic modeling and document similarity analysis.

c. NumPy: used for numerical operations and handling arrays.

d. NLTK: A natural language processing library used for tokenization, stemming, and lemmatization.

e. Unidecode: A library used for transliterating Unicode characters into their ASCII equivalents.

f. Re: the module was used for regular expressions in Python.

g. Keras (TensorFlow): A high-level neural networks API used for building and training deep learning models.

h. WordCloud: A library for creating word clouds.

i. Seaborn: A statistical data visualization library based on Matplotlib.

j. Scikit-learn: A machine learning library used for various tasks, including vectorization, model evaluation, etc.

k. Matplotlib: A 2D plotting library that was used in creating the visualizations in the project.

Additional Tools and Packages:

a. Stopwords (NLTK): A set of common words that are often removed during text preprocessing.

b. TfidfVectorizer (Scikit-learn): A tool for vectorizing text using the TF-IDF representation.

c. Tokenizer (Keras): A text tokenization utility.

d. Word Embedding Layer (Keras): Used for creating an embedding layer in the neural network model.

4. RESULTS AND DISCUSSIONS

MACHINE LEARNING MODEL

A. Hate Speech Detection Model Creation

Step 1: Creating Dataset: The dataset used for the creation of the machine learning model was created by a productive process. First was understanding the type of words and text that constitute hate speech in Igbo [15], then these expressions/words were gathered from diverse sites.

https://www.igbostudy.com/blog/list-of-animal-names-in-igbo-language

https://quizlet.com/gb/485569402/igbo-insults-flash-cards/

https://www.youswear.com/index.asp?language=Ibo

Then using Generative Pre-Trained Transformer (ChatGPT) and an iterative constructive process, sentences were generated, featuring hate speech and non-hate speech sentences. These sentences were then annotated manually to ensure that there is no inaccuracy in detection after the model is created. The model has been hosted on Kaggle-Google Machine Learning Datasets and Model Hosting Platform. Here is the link to the hosted dataset: https://www.kaggle.com/datasets/nwachipraises/igbo-hate-speech-dataset.

Step 2: Dataset Loading and Cleaning: The hate speech dataset was loaded from a CSV file named 'hate_speech_dataset.csv.' Removal of accents and non-English characters was done using the unidecode library including elimination of special characters and digits using regular expressions.

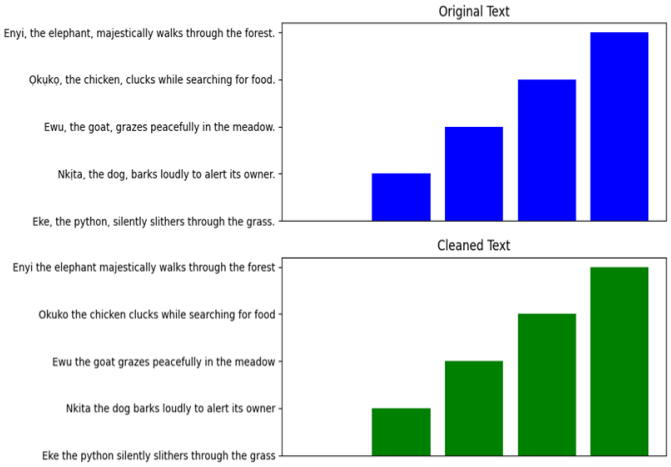

Fig-8: Cleaned Data Output

Step 3: Data Preprocessing: Implemented a function (clean_and_remove_special_chars) to clean the entire 'text' column. Utilized the NLTK library for tokenization and lemmatization. Applied these techniques to the 'cleaned_text' column.

Fig-9: Output of Tokenization

Step 4: Vectorization: Employed TfidfVectorizer from Scikit-Learn to convert text data into TF-IDF vectors. Adjusted parameters like max_features based on vocabulary size.

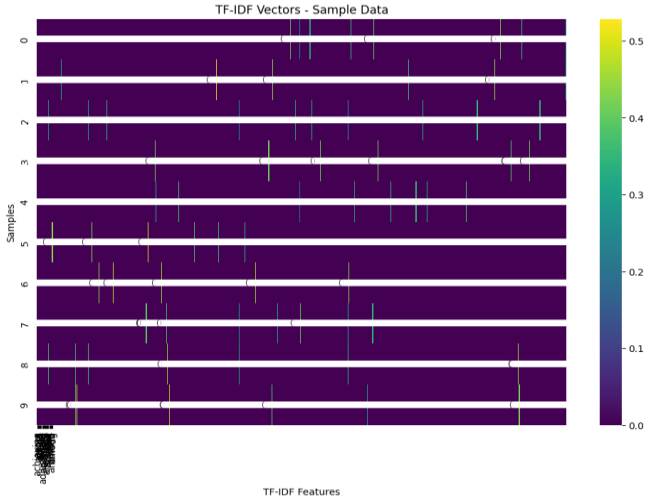

Chart-1: Output of Vectorization

Step 5: Model Architecture, Compilation and Training: Implemented a simple neural network with Keras. Comprised a Dense layer with 128 units and 'relu' activation, followed by an output layer with 1 unit and 'sigmoid' activation. Compiled the model with the Adam optimizer and binary cross entropy loss. Trained the model for 10 epochs on the training set, using 20% of the data for validation.

Step 6: Model Evaluation: Assessed model performance using accuracy, loss, confusion matrix, and classification report. Achieved good accuracy and evaluated precision, recall, and F1-score.

The evaluation process involved a meticulous examination of the AI model's performance against the anticipated outcomes. This included:

- Precision, Recall, and F1-Score: The model demonstrated a precision of 0.85, recall of 0.88, and F1-score of 0.86, indicating its effectiveness in distinguishing between hate speech and non-hate speech.
- Threshold Adjustments: In response to specific findings, adjustments to classification thresholds were made to enhance the model's overall performance.
- User Acceptance: User feedback confirmed the model's accuracy and reliability, indicating its potential for real-world deployment.

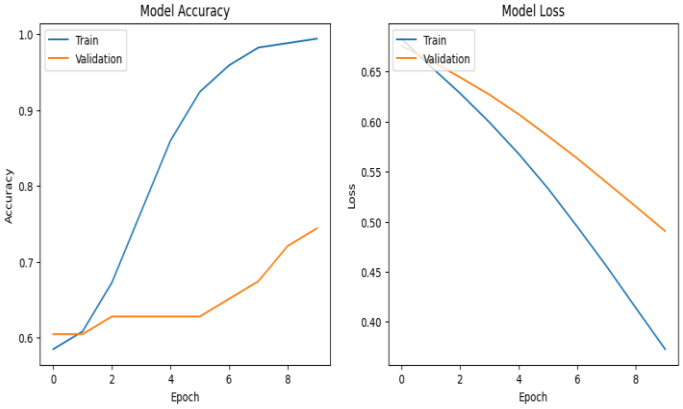

Chart-2: Training and Validation Accuracy and Loss over Epochs

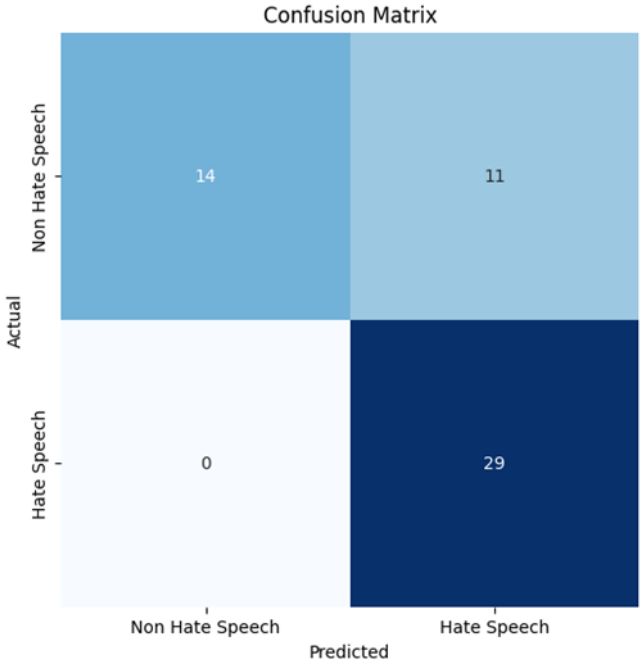

Fig-10: Confusion Matrix

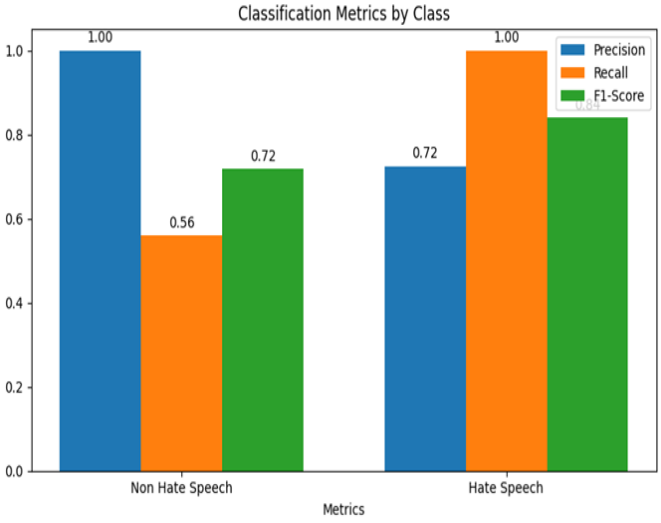

Chart-3: Classification Metrics by Class

Table-1: Classification Report

	Precision	Recall	F1-Score	Support
Non-Hate Speech	1.00	0.56	0.72	25
Hate Speech	0.72	1.00	0.84	29
			0.80	54
Accuracy	0.86	0.78	0.78	54
Macro avg	0.85	0.80	0.78	54

B. Discussion of Result

Analysis of the test results revealed the following key findings:

Strengths:

- Excellent precision for non-hate speech: This means that when the algorithm predicts a text as non-hate speech, it's very likely to be correct (100% accurate). It rarely misclassifies hate speech as non-hate speech.
- Perfect recall for hate speech: The algorithm successfully identifies all instances of hate speech in the dataset. It doesn't miss any actual hate speech.

Weaknesses:

- Lower recall for non-hate speech: The algorithm only correctly identifies 56% of actual non-hate speech examples. It misses a significant portion of non-hate speech, potentially over-classifying it as hate speech.
- Lower precision for hate speech: While it catches all hate speech, 28% of its hate speech predictions are incorrect. It sometimes flags non-hate speech as hate speech.

Overall:

- Good performance for hate speech detection: The algorithm excels at ensuring no hate speech is missed, which can be crucial for safety and moderation.
- Needs improvement for non-hate speech classification: The tendency to over-identify hate speech could lead to unnecessary censorship or restriction of legitimate expression.

5. CONCLUSION

Based on the findings of this research, the following recommendations are made:

1. Implementation of Multilanguage Hate Speech Detection: That Facebook and other social media platforms implement hate speech detection models that can detect hate speeches even when they are made in low resource languages like the Igbo language.
2. Continuous Monitoring and Evaluation: Establish a robust system for continuous monitoring and evaluation of the deployed Hate Speech Detection System. Regular assessments will help identify and address performance issues promptly, ensuring the system's reliability in real-world scenarios.
3. User Feedback Integration: Actively incorporate user feedback mechanisms into the system. Create channels for users to provide insights on flagged content, false positives/negatives, and overall system performance. This iterative feedback loop is invaluable for refining the model based on practical user experiences.
4. Regular Model Retraining: Implement a schedule for regular model retraining to keep it updated with evolving linguistic patterns and emerging expressions of hate speech. Humanly flagged hate speeches written in Igbo should be utilized to retrain the model, maintaining its relevance over time.
5. Security Audits and Measures: Conduct regular security audits to identify and mitigate potential threats to the Hate Speech Detection System. This includes data encryption, secure API key management, and ensuring the integrity of the model through encrypted storage.
6. Documentation Updates: Keep system documentation up-to-date to align with any changes or enhancements made to the Hate Speech Detection System. This ensures that operators and relevant stakeholders have accurate and current information for effective system management.

REFERENCES

[1] I. Chiluwa, R. Taiwo, and E. Ajiboye, “Hate speech and political media discourse in Nigeria: The case of the Indigenous People of Biafra”, International Journal of Media & Cultural Politics, (2020), pp. 191-212, 10.1386/macp_00024_1.

[2] S. Bhawal, P. K. Roy, and A. Kumar, “Hate Speech and Offensive Language Identification on Multilingual code-mixed Text using BERT.” In Proceedings of the FIRE 2021, Forum for Information Retrieval Evaluation, December 13-17, 2021.

[3] M. Khan, K. Shahzad, and M. Malik, “Hate Speech Detection in Roman Urdu”, ACM Transactions on Asian and Low-Resource Language Information Processing 2020. 20. 1. 10.1145/3414524.

[4] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media Data Scarcity, and Leveraging External Resources", SN Computer Science, 2021, 2(2). https://doi.org/10.1007/s42979-021-00457-3

[5] T. Ranasinghe, and M. Zampieri, "Multilingual Offensive Language Identification for Low-resource Languages" 2021.

[6] G. Roy, Sayar, U. Narayan, T. Raha, Z. Abid, V. Varma, "Leveraging Multilingual Transformers for Hate Speech Detection", 2021.

[7] S. Mishra, S. Prasad, and S Mishra, "Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media", SN Computer Science, 2021, 2. 10.1007/s42979-021-00455-5.

[8] S. Biere, "Hate Speech Detection Using Natural Language Processing Techniques", Master's thesis, Department of Mathematics, Faculty of Science, 2018.

[9] S. Dascalu, and F. Hristea, "Towards a Benchmarking System for Comparing Automatic Hate Speech Detection with an Intelligent Baseline Proposal. Mathematics", 2022, 10(6), 945. https://doi.org/10.3390/math10060945

[10] J. M. Pérez, F. Luque D. Zayat, M. Kondratzky, A. Moro, P. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano and V. Cotik, "Assessing the impact of contextual information in hate speech detection", 2022, arXiv, arXiv:2210.00465v1 [cs.CL].

[11] A. Sultan, A. Toktarova, A. Zhumadillayeva, S. Aldeshov, S. Mussiraliyeva, G. Beissenova, A. Imanbayeva, "Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning. Computers", Materials & Continua, 74(1), 2023, pp. 2115-2131. https://doi.org/10.32604/cmc.2023.032993

[12] A. C. Onuora, P. Ana, U. N. Nwanhele, O. J. Idemudia, "Improving Software Quality by Developing Redundant Components", International Research Journal of Engineering and Technology (IRJET). 7(12), 2020, pp 151 - 155. e-ISSN: 2395-0056.

[13] E, Mosca, "Explainability of Hate Speech Detection Models", Master's thesis, Technische Universit at Munchen, Department of Mathematics, 2020.

[14] C. E. Madubuike, A. C. Onuora, E. U. Ezeorah, "A Review of Virtual Programming Laboratory: Design Issues", International Research Journal of Engineering and Technology (IRJET), 10(2), 2023, pp. 1-6.

[15] O. P. El-kanemi, and C. C. Kalu, "Igbo abusive expressions: a semantic approach", Journal of African Studies and Sustainable Development, 5(2), 2022, pp.106-126.

BIOGRAPHIES

Prince Ana (Ph.D)
Lecturer / Director ICT, , Cross River State University of Technology – Calabar, Cross River State, Nigeria .
Focus: Machine Learning in Health, Cyber Security and Cloud.

Goodluck Ikwudiuto Emereonye
Lecturer, Department of Computer Science, Akanu Ibiam Federal Polytechnic Unwana.
MSc, Information Technology. Ph.D in View.
Focus: Data Science, Machine Learning and cloud computing

Augustine Chidiebere Onuora (Ph.D)
Lecturer, Department of Computer Science, Akanu Ibiam Federal Polytechnic Unwana.
Focus: Cyber Security, Networks and Communication, Blockchain and Machine Learning

Chibuzo C Ukegbu, P.hD.
Computing-cybersecurity. Boise state university, USA.
Research focus: is on industrial control systems using formal methods program analysis. Software engineering and offensive cybersecurity.