Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 116-123
https://doi.org//10.47514/kjcs/2025.2.1.009

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

A Deep Learning Technique for Detecting Offensive Hausa Language

Bashir Idris Suleiman[1] and Muhammad Aminu Ahmad[2]

[1]Department of Computer Science Kaduna State University

[2] Dean Faculty of Computing, Kaduna State University, Kaduna

bashir.idris@kasu.edu.ng[1], muhdaminu@kasu.edu.ng[2]

Abstract

This study explores the complexities of hate speech detection in social media, particularly in the context of the Hausa language. Social media has revolutionized online interactions, but the freedom of expression it offers has also led to the proliferation of hateful content. Hate speech is a multifaceted phenomenon with severe consequences, including emotional distress, offline violence, and psychological harm. This study highlights the challenges of defining and detecting hate speech, as well as the importance of addressing it as a public health issue. The use of Artificial Intelligence (AI) in content moderation is also discussed, including its benefits and concerns. The study provides a foundation for understanding hate speech in the Hausa language and culture, which is deeply rooted in Islamic traditions and characterized by a unique tonal system and vowel harmony.

Keywords: Cyberbullying, Hausa Language, Low-Resource Languages, Hate Speech Detection, Machine Learning

1.0 Introduction

Social media has revolutionized the way people interact, share information, and express themselves. It has become an indispensable platform for sourcing information, expressing opinions and feelings, as well as posting multimedia contents (Alsafari et al., 2020). The anonymity and freedom of expression offered by social media have enabled people to communicate with one another without geographical constraints. However, this freedom has also led to the proliferation of hateful content and online harassment, which is becoming a significant menace in these online communities Hate speech detection in Hausa, a low-resource language spoken in West Africa, is a critical task in promoting a safer and more inclusive online environment. The proliferation of social media has facilitated the spread of offensive content, including hate speech, which can have severe

consequences on individuals and communities. Developing effective detection mechanisms for Hausa hate speech is essential to address this issue. However, the task is challenging due to the linguistic nuances and complexities of the Hausa language, as well as the limited availability of resources and datasets. This research aims to bridge this gap by developing novel datasets and models for Hausa hate speech detection, ultimately contributing to the development of more responsible and ethical AI systems.

.

2.0 Related Works

Hate speech is a complex and multifaceted phenomenon that lacks a universally accepted definition. Despite this, various authors have offered similar opinions on what constitutes hate speech, highlighting its harmful and discriminatory nature (MacAvaney et al., 2019). A common thread among these definitions is that hate speech is always directed towards a target, whether it is an individual, group, or community (Ali et al., 2021). The consequences of hate speech can be severe and far-reaching, extending beyond the online realm to affect the victim's psychology and even physical well-being (Park, 2023). Hate speech can lead to feelings of anxiety, depression, and isolation, as well as physical symptoms such as sleep disturbances and cardiovascular problems. In some cases, hate speech has even been linked to physical violence and offline harm. The lack of a clear definition and the complexity of hate speech make it challenging to detect and mitigate. However, researchers and developers are working to improve hate speech detection using sentiment analysis and other machine learning techniques (Ali et al., 2021). Additionally, there is a growing recognition of the need to address hate speech as a public health issue, with some arguing that it should be considered a form of psychological violence (Park, 2023). The consequences of online hate speech and harassment can be severe, ranging from emotional distress to offline violence. Therefore, social media platforms have developed regulations to prohibit the posting of offensive and hateful contents (Chetty & Alathur, 2018; Alkiviadou, 2019). These regulations aim to create a safe and respectful online environment, where users can express themselves without fear of harassment or intimidation. To enforce these regulations, social media platforms have employed automatic and semi-automatic approaches using Artificial Intelligence (AI) to detect and remove offensive and hateful contents (Schneider & Rizoiu, 2023). These AI-powered systems can analyze vast amounts of data in real-time, identifying and flagging suspicious content for human review. Additionally, AI can be used to send warnings to users who post offensive content, educating them about the harmful effects of their actions and encouraging them to modify their behavior. The use of AI in content moderation has several benefits, including increased efficiency, accuracy, and scalability. However,

it also raises concerns about bias, transparency, and accountability. As AI systems become more pervasive in content moderation, it is essential to ensure that they are designed and trained to promote fairness, respect, and inclusivity online. The Hausa culture is deeply rooted in Islamic traditions, and its language reflects this influence. One of the distinctive features of the Hausa language is its tonal system, which consists of three basic tones: high, mid, and low. Additionally, the language employs vowel harmony, where vowels within a word tend to harmonize with respect to their tongue root positions (Lambu, I. B. 2019). This unique feature adds complexity and richness to the language.

3.0 Methodology

In the research process, analytical and experimental methods are effectively combined. Analytical methodology is employed for machine learning, model training, and testing, while experimental methods are used for data collection, data extraction, and data preparation. This combination of approaches enables a comprehensive and rigorous research process, allowing for a thorough examination of the research question from multiple angles as shown in figure 3.1.

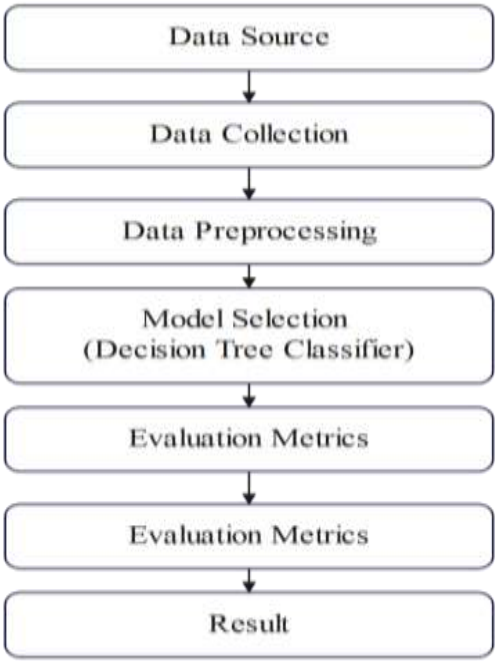

Figure 3.1 Data flow diagram

3.1. Data Source

The dataset source comprised text comments posted by users on social media platforms, specifically Facebook and Twitter, in the Hausa language of Nigeria. A link to a response collection page was posted on my social media page, soliciting comments related to Hausa.

3.2. Data Collection

The datasets Hausa Hate Speech were collected from two primary sources. Firstly, tweets were gathered from Twitter and then translated into Hausa. Secondly, data was manually collected from Hausa-speaking communities who openly expressed hate nuances. Once the data was collected, it was merged and converted into a readable CSV (Comma Separated Value) format using the panda's library. This process resulted in the generation of over 2000 datasets. Table 1 shows a sample of keywords used for each language.

Table 1. Samples of keywords used to collect tweets in Hausa Language.

Hate Speech	**Offensive Speech**	**Non-Offensive Speech**
Dakiki	Boko Haram	Ina zuwa
Arne	Wawa	Shikenan
Jahili	Inya'muri	Takalmi

3.3 Data Preprocessing

Before analyzing a dataset, it's essential to ensure it's free from issues that can impact performance. These issues may include missing values, incorrect data types, and irrelevant attributes. Research has shown that text preparation significantly improves classification results. To address these issues, various preprocessing techniques were applied to the dataset to remove noisy, irrelevant, and non-information data. This included converting tweets to lowercase for uniformity and normalization.

Additionally, pattern matching techniques were used to remove unwanted elements such as URLs, usernames, punctuations, hash tags, white spaces, and stop-words from the collected tweets. The tweets were then tokenized, which involves breaking down each tweet into individual words or tokens, and stemmed using a Porter stemmer, which converts each word to its root form (e.g., "hateful" becomes "hate"). Guidelines as follows;

Hate – *A tweet is hateful if it is offensive and based on target such as religion, race, gender or political affiliation.*

Offensive – *A tweet is offensive if it insults, abuses, threatens, mocks, disparages an individual.*

Non-offensive – *All tweets that are not offensive or hateful.*

3.4 Model Selection

In the context of cyberbullying detection in Hausa, employing a Decision Tree Classifier proves to be an effective strategy. This model is particularly advantageous due to its interpretability, which allows for clear insights into the decision-making process, and its ability to manage both categorical and numerical data. Key aspects to consider when selecting this model include maintaining an appropriate level of complexity to prevent overfitting, tuning hyperparameters like maximum depth and minimum samples per leaf, and evaluating performance through metrics such as accuracy, precision, recall, and F1-score. By focusing on emotional cues such as tone, pitch, and rhythm, the Decision Tree Classifier can accurately classify speech as offensive or non-offensive, thereby enhancing the detection and response to cyberbullying incidents within the Hausa language community.

4. Results and Discussion

The evaluation measures employed to appraise the model performance are:

- Accuracy: The proportion of correctly classified instances out of all instances in the dataset.
- Precision: The proportion of true positives among all positive predictions made by the model.
- Recall: The proportion of true positives among all actual positive instances in the dataset.
- F1 Measure: The harmonic mean of precision and recall, providing a balanced measure of both.

These evaluation metrics provide a comprehensive understanding of the model's performance, allowing for identification of strengths and weaknesses.

Table 2. Model results with accuracy score in Hausa.

Model	Accuracy
Accuracy (%):	61.01694
Precision:	0.57794
Recall:	0.61016
F1-Score:	Balanced mean of precision and recall

The Decision Tree Classifier demonstrated moderate success, achieving an accuracy of 61%. However, challenges such as linguistic nuances and the limited dataset affected performance.

5.0 Conclusion

This study contributes a novel dataset and evaluation of cyberbullying detection models for Hausa. While results indicate room for improvement, the research sets a foundation for future studies on low-resource language processing.

References

Ali, M. Z., Ehsan-Ul-Haq, Rauf, S., Javed, K., & Hussain, S. (2021). Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis. IEEE Access, 9, 84296–84305. doi: 10.1109/ACCESS.2021.3087827

Alkiviadou, N. (2019). Hate speech on social media networks: towards a regulatory framework? Information and Communications Technology Law, 28(1), 19-35. doi: 10.1080/13600834.2018.1494417

Alsafari, S., Sadaoui, S., & Mouhoub, M. (2020). Deep Learning Ensembles for Hate Speech Detection. In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 526-531). IEEE. doi: 10.1109/ICTAI50040.2020.00087

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. Aggressive and Violent Behavior, 40, 108-118. doi: 10.1016/j.avb.2018.05.003

Lambu, I. B. (2019). Hausanization of Nigerian Cultures. In S. D. Brunn & R. Kehrein (Eds.), Handbook of the Changing World Language Map (pp. 1–9). Cham: Springer International Publishing. doi: 10.1007/978-3-319-73400-2_53-1

Park, J. (2023). The Mental & Physical Health Argument Against Hate Speech. Journal of Cognitive Neuroethics, 9(1), 2023.

Al-Hassan, A., & Al-Dossari, H. (2019). Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science and Information Technology*, 9(2), 83–100. https://doi.org/10.5121/csit.2019.90208

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

Faris, H., Aljarah, I., Habib, M., & Castillo, P. (2020). Hate speech detection using word embedding and deep learning in the Arabic language context. *International Journal of Information Processing and Communication, 6*(2), 453–460. https://doi.org/10.5220/0008954004530460

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys, 51*(4), 1–30. https://doi.org/10.1145/3232676

Husain, F., & Uzuner, O. (2020). A survey of offensive language detection for the Arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(3), 1–32. https://doi.org/10.1145/3340922

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57. https://doi.org/10.18653/v1/W19-3510

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech detection using natural language processing and deep learning technologies. *Computational Intelligence, 39*(3), 421–456. https://doi.org/10.1109/CIT.2023.120931

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE, 14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Ndabula, J. N., Olanrewaju, O. M., & Echobu, F. O. (2023). Detection of hate speech code mix involving English and other Nigerian languages. *Journal of Information Systems and Informatics, 5*(4), 1416–1431. https://doi.org/10.51519/journalisi.v5i4.595

Conference Proceedings

Diallo, A. K., & Abainia, K. (2023). Offensive language detection in code-mixed Bambara-French corpus: Evaluating machine learning and deep learning classifiers. *Proceedings of the International Conference on Decision Aid Sciences and Applications (DASA)*, 121–125. https://doi.org/10.1109/DASA.2023.10286577

Kanessa, L. G., & Tulu, S. G. (2021). Automatic hate and offensive speech detection framework from social media: The case of Afaan Oromoo language. *Proceedings of the 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, 42–47. https://doi.org/10.1109/ICT4DA53266.2021.9672232

Mossie, Z., & Wang, J.-H. (2018). Social network hate speech detection for Amharic language. *Proceedings of the Computer Science and Information Technology Conference (CSIT)*, 41–55. https://doi.org/10.5121/csit.2018.80604

Ombui, E., Karani, M., & Muchemi, L. (2019). Annotation framework for hate speech identification in tweets: Case study of tweets during Kenyan elections. *Proceedings of the 2019 IST-Africa Week Conference (IST-Africa)*, 1–9. https://doi.org/10.23919/ISTAFRICA.2019.8764868

Tesfaye, S. G., & Kakeba, K. (2020). Automated Amharic hate speech posts and comments detection model using recurrent neural network. *Proceedings of the International Conference on Advances in Big Data Analytics (ABDA)*, 118–126. https://doi.org/10.21203/rs.3.rs-114533/v1

Tonneau, M., et al. (2024). NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data. *arXiv Preprint*. https://arxiv.org/abs/2403.19260