Predicting Xenophobic Attacks: Sentiment Analysis on South African Twitter Data

1st Sibusiso Ziqubu
Johannesburg, South Africa
ziqubu.sibusiso@gmail.com

***Abstract*—This study explores the impact of Twitter on online communication since its 2006 inception, acknowledging its role in political and corporate realms. However, the platform's acceptance is tempered by its potential for disseminating hate speech. Focusing on xenophobia, particularly in South African Twitter, the research addresses the need for automated offensive language detection. Leveraging machine learning, Support Vector Classification, and TF-IDF feature vectors, sentiments are classified into negative, positive, and neutral categories. The findings reveal a prevalent negative sentiment in xenophobia discussions, with temporal analysis exposing dynamic narrative shifts. Thematic concerns, extracted through word cloud and frequent word analysis, underscore the significance of terms like "foreigners" and "immigrants." The model attains a 61% accuracy rate, excelling in negative sentiment classification while revealing areas for improvement in positive and neutral sentiment detection. This study adds to a more sophisticated understanding of xenophobic sentiments by emphasizing continual model refining and natural language processing for complete public sentiment comprehension. The proposed automated model holds promise for social media monitoring, content moderation, and sentiment analysis, fostering a safer online environment.**

***Index Terms*—Twitter, machine learning, xenophobia, sentiment analysis**

I. Introduction

In October 2006, Twitter, an online social networking and micro-blogging service, was launched. It is a free real-time short messaging service that allows users to post and read messages (tweets) via the Twitter website, mobile app, and other desktop apps. Twitter users are limited to sending updates in just 140 characters, which is an essential feature. Despite the criticism leveled at the 140-character medium, Twitter is seeing rapid development and acceptance. For example, during his 2008 presidential campaign, Barack Obama used Twitter to communicate with the American people. Some companies, such as Dell, have found significant success in utilizing Twitter to educate their customers about product discounts and news. Many social media marketers and experts feel Twitter offers many commercial benefits. Marketers, in particular, may use Twitter to readily discover what people are saying in real time about their products [7].

People's communication has gotten faster and simpler as social media platforms such as Facebook, Twitter, Instagram, and Tik Tok have grown in popularity. People can convey their sentiments, criticism, opinions, accomplishments, and so on using various communication tools. However, social networks are frequently used to disseminate hate speech using harsh language. Offenses can be directed at a variety of factors, including ethnicity, sexism, economic class, religion, sexual orientation, and so on. Thus, the main issue with this is that the offense is demonstrated to individuals or groups, which might be potentially detrimental to them [4]. People who participate in these forums or social networks come from a variety of cultures and educational backgrounds. Disagreements in viewpoint can sometimes escalate to verbal attacks. Furthermore, unregulated free speech on the internet, as well as the anonymity provided by the internet, encourages individuals to use racist slurs or insulting phrases. People's self-esteem may suffer as a result, leading to mental illness and a detrimental influence on society as a whole. Furthermore, toxic language may take many forms, including cyberbullying, which was one of the leading causes of suicide. This issue has grown in importance over the last decade, and manually finding or deleting such information from the web is a time-consuming operation. As a result, there is a need to develop an automated model capable of detecting such harmful information on the internet [5].

The issue in automatic identification of offensive language is that the used language in social networks has a certain format that corresponds to the environment. Numerous word abbreviations are employed in this context, as well as different types of expression amplification and word alteration, such as numerous letter repeats (e.g., looooved, gooood) and excessive punctuation (e.g., i adored!!!!, what????). As a result, the original content must be adjusted during a critical preprocessing stage to obtain a version that retains the original sense while still being compatible with other comparable postings. A program that can identify these types of violations in a social network is a critical step in ensuring user security and mental wellness.

Xenophobia, a social phenomenon characterized by hostility and prejudice towards foreigners, has emerged as a significant concern in South Africa. This study uses Twitter data to analyze xenophobic sentiments, uncovering patterns and trends, and using predictive analytics to anticipate potential surges in xenophobic attacks. We propose an approach to devise a machine learning model which can differentiate between a positive, negative and neutral sentiment in a tweet. By using publicly available Twitter datasets, we train our classifier model using term frequency-inverse document frequency (TFIDF) as features.

II. Objectives

The Objectives of this study is to predict and understand xenophobic attacks in South Africa through sentiment analysis and predictive analytics applied to Twitter data. The Main objectives are:

1) Analyze sentiments expressed in South African Twitter data related to xenophobia
2) To classify tweets into sentiments using the Support Vector Machines (SVM) based classier called Support Vector Classifier (SVC). The model should learn to distinguish between the three classes based on the feature vectors derived from the tweet text using the TF-IDF representation.
3) To convert the tweet text into numerical feature vectors using the TF-IDF representation. TF-IDF stands for Term Frequency-Inverse Document Frequency and assigns weights to each term in a tweet based on its frequency in the tweet and inverse frequency across all tweets. The feature extraction aims to capture the important terms in the tweet text that are indicative of offensive content.

III. Literature review

The study of offensive language in social media is a relatively new research field, but it has generated a lot of interest, and an increase in related publications. Offensive content on social media platforms has become a growing concern in recent years, and the need for effective detection methods has escalated. According to Kwok and Wang [6], 86% of the time a tweet was labeled as racist because it contained offensive words. Given the relatively high incidence of offensive language and "curse words" on social media, hate speech detection is especially difficult. The distinction between hate speech and other offensive language is sometimes dependent on tiny grammatical nuances; for example, tweets using the term n*gger are more likely to be categorized as hate speech than tweets featuring the word n*gga. Other supervised techniques to hate speech categorization have, however, confounded hate speech with offensive language, making it impossible to determine how well they are recognizing hate speech [2].

For decades, machine learning (ML) approaches to natural language processing (NLP) issues have relied on shallow models (e.g., SVM and logistic regression) trained on very high dimensional and sparse data [1]. Despite the limitations encountered by machine learning (ML), we discover that numerous ML-based methods for automated detection have been presented and have demonstrated good results, such as the Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) classifier [8]. SVM-based classifiers have shown promising results in various natural language processing (NLP) tasks, including offensive language detection. Zampieri et al. [11] shown that n-grams may perform well for hate speech identification utilizing SVMs with several surface-level features, such as surface n-grams, word skip-grams, and word representation n-grams induced with Brown clustering. They also discovered that these features reached their limits for more difficult tasks, such as discriminating between vulgarity and hate speech. More detailed language qualities may be necessary in such tasks.

The study conducted by Pérez-Landa, Loyola-González, and Medina-Pérez [9] addresses the critical issue of accurately classifying xenophobic posts on Twitter. To achieve this, the study utilizes Natural Language Processing (NLP) to extract new features from a database of xenophobic tweets. The incorporation of an Explainable Artificial Intelligence (XAI) model enhances interpretability, providing insights into why a particular post is considered xenophobic. This approach is crucial for decision-makers seeking actionable insights to prevent real-world acts of violence spurred by xenophobic content on Twitter. Davidson et al. [3] generated a new dataset for Twitter data categorization as hate speech, offensive language, or neither using the Twitter API. They gathered 85.4 million Twitter samples from around 33 thousand Twitter users for their dataset. They then created a collection of 24k labeled twitter samples using features including bigram, unigram, and trigram that were weighted by their TF-IDF and utilized for the classification task. In addition, binary and count indications for hashtags, mentions, retweets, and URLs were incorporated. They put a wide range of classifiers to the test, including logistic regression, Naive Bayes, decision trees, random forests, and linear SVMs (Support Vector Machines). During their trials, they discovered that Logistic Regression and Linear SVM produced better results. Gaydhani et al. [5] proposed a machine learning solution for detecting hate speech and offensive language on Twitter using n-gram features weighted with TFIDF values and performed a comparative analysis of Logistic Regression, Naive Bayes, and Support Vector Machines on various sets of feature values and model hyperparameters. Logistic Regression performed better with the optimal n-gram range 1 to 3 for the L2 normalization of TFIDF.

IV. Ethical considerations

The analysis of sensitive topics like xenophobia entails ethical considerations that necessitate careful attention to mitigate potential biases and uphold the integrity of the research. In the context of data collection, biases may inadvertently emerge from the platform's user demographics, language, or algorithmic processes. Social media data inherently carries the risk of reflecting certain user groups more prominently than others, potentially skewing the representation of sentiments related to xenophobia. Moreover, biases may manifest during the analysis phase, influenced by the algorithms and models employed. The risk of perpetuating existing biases or inadvertently introducing new ones underscores the importance of transparency and ethical rigor. Researchers must critically assess the choice of models, features, and training data to ensure fairness and accuracy in sentiment analysis.

V. Methodology

This section provides an in-depth explanation of the research methodology, including the use of data, pre-processing, data cleaning, and the proposed method.

A. Data

The dataset contains five-year historical tweets of discussions about Xenophobia and Xenophobic attacks in South Africa. The 18,278 tweets were extracted using the Twitter API from January 2017 to July 2022. The tweets are classified into positive, negative, and neutral sentiments. The tweets were extracted based on the mentions of selected keywords regarding Xenophobia. These keywords are "Foreigners", "immigrants", "OperationDududla", "Nhlanhla Lux", "Take back SA", "Take back South Africa", and "Dudula". In Fig. 1, the sentiment distribution of tweets is illustrated. Among the total of 18,278 tweets, 10,273 exhibited a negative sentiment, 4,790 conveyed a neutral sentiment, and 2,765 reflected a positive sentiment, highlighting the presence of a class imbalance.

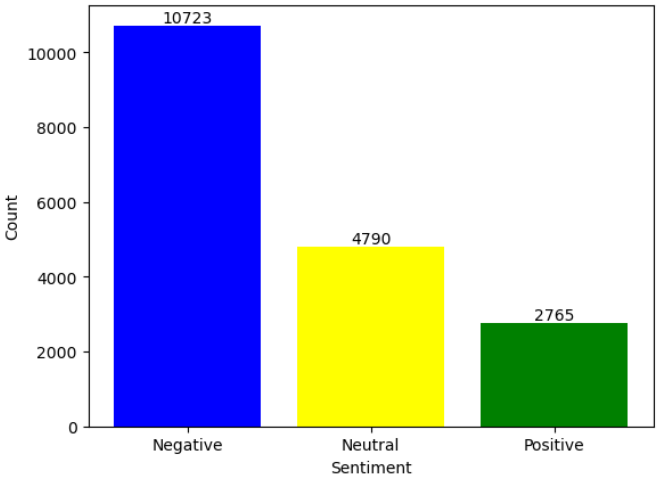

Fig. 1. Sentiment distribution

B. Pre-processing

The pre-processing methodology implemented involves a series of steps to refine and prepare textual data for analysis. Beginning with the removal of usernames using regular expressions, the process ensures the exclusion of Twitter handles, fostering anonymity. Subsequently, non-alphabetic characters are eliminated, promoting a focus on meaningful linguistic content. Tokenization was a crucial step, it breaks down the text into individual words, providing a foundation for further analysis. The removal of stopwords, common and non-informative words, contributes to reducing noise and enhancing the significance of the remaining words. Lastly, addressing extra whitespaces ensures a clean and consistent representation of the text, facilitating subsequent analyses. The aim of pre-processing was to enhance the quality and relevance of the text data, creating a more conducive environment for downstream tasks such as sentiment analysis.

C. Model development

The model is trained on the data provided in the CSV file. The tweet text is converted into feature vectors using the TF-IDF representation. An SVM classifier (SVC) is created and trained on the feature vectors and corresponding labels from the DataFrame. The algorithm assumes the CSV file has a column named "Sentiment" containing the labels for each tweet (0 for negative, 1 for positive, 2 for neutral). The *detect_sentiment* function takes a tweet as input and converts it into a feature vector using the same vectorizer. It then uses the trained classifier to predict the label for the tweet. The function maps the predicted label to either "Negative", "Positive" or "Neutral" and returns the result.

D. Performance evaluation metric

In this study, the accuracy, precision and recall are used to evaluate the performance of the model. Accuracy represents the proportion of correctly identified data among the entire set of testing data. Accuracy represents the proportion of correctly identified data among the entire set of testing data [10]:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

where P and N indicate the number of positive and negative samples, respectively. Precision measures the accuracy of the positive predictions made by a model. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives. Recall, also known as sensitivity or true positive rate, gauges the model's ability to capture all positive instances in the dataset. It is calculated as the ratio of true positive predictions to the sum of true positives and false negatives.

E. Experimental procedure

A stratified train/test split with an 80/20 ratio is employed to ensure representative training and testing datasets in a machine learning task. The training set is used to train the machine learning model. During this phase, the model learns the patterns and relationships present in the data. The testing set is reserved to evaluate the performance of the trained model. The model has not seen this data during the training phase. By assessing the model on unseen data, we can gauge its ability to generalize well to new, previously unseen examples. Splitting the data helps in providing an unbiased assessment of the model's performance.

VI. Results

A. Data exploration

Firstly, we conducted data exploration using the entire dataset to extract valuable insights. Table 1 shows the sentiment proportions. The sentiment proportions indicate a predominant negative sentiment, constituting approximately 58.67% of the analyzed data. Neutral sentiments follow, representing 26.21% of the sentiments, while positive sentiments make up the remaining 15.13%.

TABLE I
SENTIMENT PROPORTIONS

Sentiment	Percentage
Negative	0.586662
Positive	0.151275
Neutral	0.262064

Fig. 2 shows how the sentiment distribution changes over-time. The sentiment distribution over time narrates a story of xenophobia in South Africa, where the peaks of negative sentiment may surge in response to xenophobic incidents, and the positive sentiments may signify collective efforts toward understanding and uniting against xenophobia.

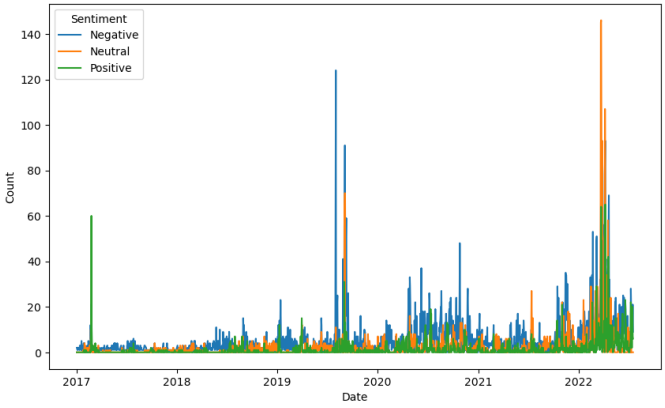

Fig. 2. Sentiment distribution over time

Fig. 3 shows the word cloud of the frequent used words our tweets. The bolder words like ”foreigner” and ”illegal immigrant” are frequent words used in our dataset. This word cloud gives us an insight of what people are saying about xenophobia in South Africa.

Fig. 3. Word Cloud of the most appearing words

Figures 4, 5, and 6 unveil a lexical panorama, showcasing the most frequently employed words within our dataset, providing insightful glimpses into prevailing sentiments and thematic concerns. Notably, the recurring appearance of words such as ”foreigners,” ”immigrants,” ”illegal,” and ”South” attests to the prominence of themes surrounding immigration and nationality. The persistent use of terms like ”foreigners” and ”immigrants” suggests a focal point on individuals not native to South Africa, indicating a heightened discourse around the experiences, challenges, or perceptions of those considered outsiders. The inclusion of ”illegal” in the frequently used words hints at a discourse on the legal status of certain groups, potentially shedding light on the complex interplay of legal frameworks within the broader discussion. Moreover, the recurrent mention of ”South” underscores a geographical context, framing discussions within the boundaries of the nation. This may reflect a strong national identity or highlight the local perspectives influencing conversations related to immigration.

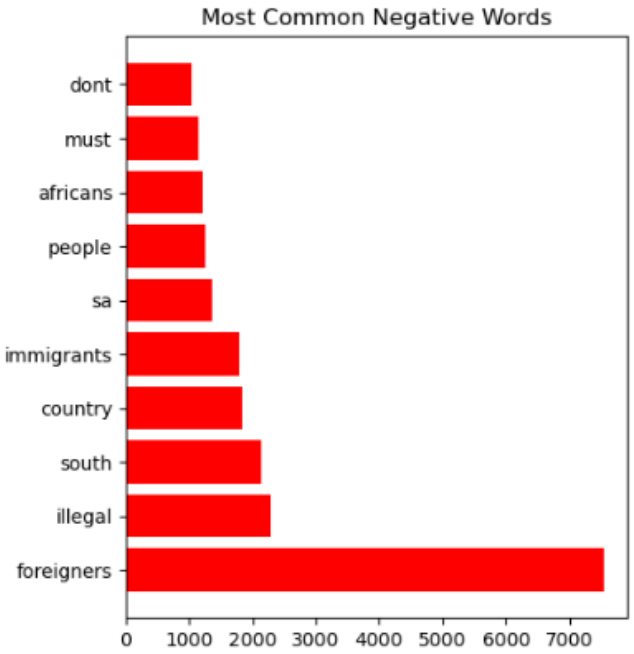

Fig. 4. Most common negative words

B. accuracy

Within the framework of this study, we employed the accuracy metric on the test set as a yardstick to evaluate the effectiveness of our sentiment prediction model. The outcomes revealed that the model attained an accuracy rate of 61%. This metric serves as a quantitative measure of the model's ability to correctly classify sentiments within the test dataset. The 61% accuracy unveils the model's proficiency in making accurate predictions, offering insights into its overall performance. While accuracy provides a holistic view of the model's success.

C. precision and recall

The model's performance in sentiment classification, as indicated by the precision and recall metrics, reveals varying levels of effectiveness across different sentiment categories.

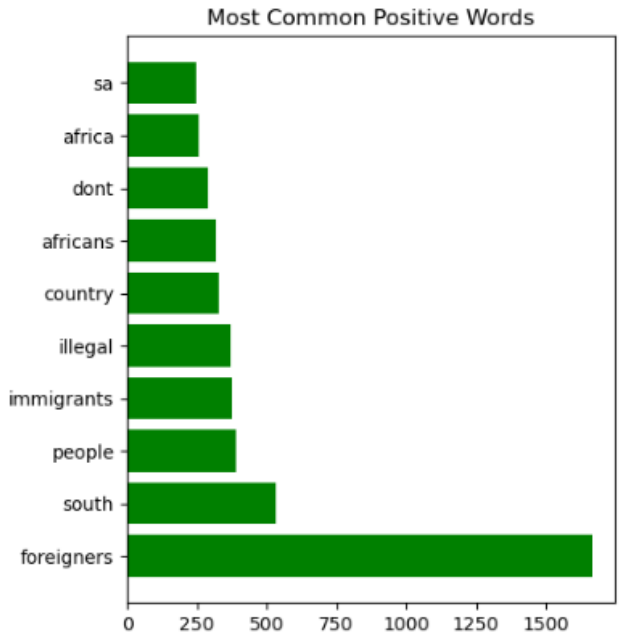

Fig. 5. Most common positive words

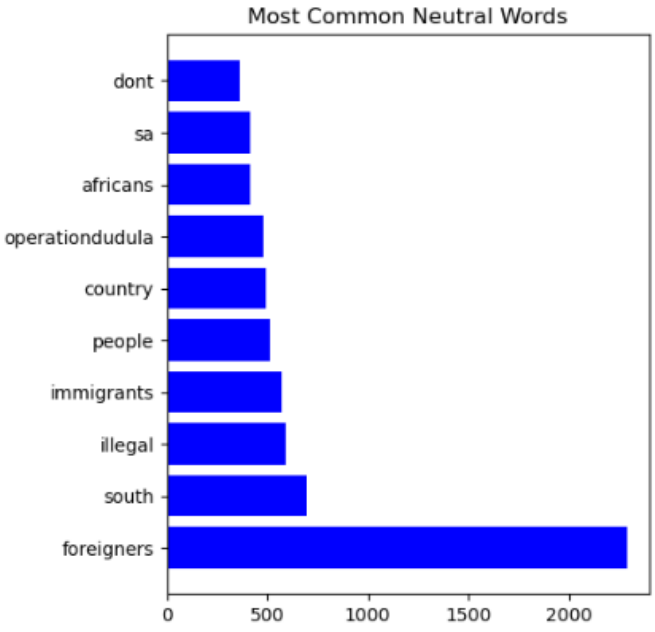

Fig. 6. Most common neutral words

For the negative sentiment class (0), the model demonstrates a relatively high precision of 0.61, signifying a considerable accuracy in correctly identifying instances classified as negative. The corresponding recall of 0.98 indicates the model's strong ability to capture a high proportion of actual negative instances. In contrast, the positive sentiment class (1) displays a lower precision of 0.43, suggesting a higher likelihood of false positives within the predicted positive instances. The recall for positive sentiment is notably low at 0.01, indicating that the model struggles to identify the majority of actual positive instances, leading to a high number of false negatives. For the neutral sentiment class (2), the model exhibits a precision of 0.67, suggesting a moderate accuracy in correctly classifying instances as neutral. However, the recall of 0.14 indicates a challenge in capturing a significant proportion of actual neutral instances.

TABLE II
PRECISION AND RECALL OF EACH SENTIMENT

Sentiment	Precision	Recall
Negative	0.61	0.98
Positive	0.43	0.01
Neutral	0.67	0.14

VII. Discussion

The literature review provides a comprehensive background on the study of offensive language in social media, emphasizing the challenges in detecting hate speech and offensive content. Various machine learning approaches, including SVM-based classifiers, have been explored to address these challenges. Additionally, the literature highlights the growing concern about offensive content on social media platforms, particularly in the context of xenophobia, necessitating effective detection methods. In alignment with the literature, our study focuses on sentiment analysis related to xenophobia in South Africa using Twitter data. The results of sentiment proportions reveal a predominant negative sentiment, indicating the prevalence of negative opinions and emotions in discussions related to xenophobia. The sentiment distribution over time sheds light on the evolving narrative, with peaks of negative sentiment corresponding to potential xenophobic incidents and positive sentiments signifying collective efforts against xenophobia. The word cloud and frequent word analysis further corroborate the thematic concerns in the dataset, emphasizing words such as "foreigners," "immigrants," and "illegal" that play a central role in discussions about xenophobia in South Africa. The recurrence of these terms underscores the prominence of immigration-related themes and legal considerations within the discourse.

Transitioning to the model's performance evaluation, the accuracy metric on the test set yields a 61% accuracy rate, indicating the model's proficiency in classifying sentiments. However, the precision and recall metrics unveil distinction in the model's effectiveness across sentiment categories. While the model excels in accurately identifying negative sentiments, there are challenges in both precision and recall for positive and neutral sentiments, suggesting areas for improvement in capturing positive sentiments and minimizing false positives. In summary, the integration of insights from the literature review with the study results offers a comprehensive understanding of sentiment analysis in the context of xenophobia on Twitter. The findings underscore the complexity of detecting

sentiments related to sensitive topics, emphasizing the importance of continual refinement in machine learning models for accurate predictions.

VIII. Conclusion

In conclusion, classifying tweets into negative, positive and neutral sentiment categories using the SVC model and feature vectors derived from tweet text using the TF-IDF representation can be an effective approach. The TF-IDF representation allows us to capture the importance of words in tweets, giving more weight to terms that are unique to a particular tweet while downplaying common terms. This representation helps to capture the context and relevance of words in the classification process.The SVC model, known for its ability to handle complex decision boundaries and high-dimensional data, is well-suited for tweet classification tasks. By learning from the feature vectors derived from the TF-IDF representation, the SVC model can effectively distinguish between negative, positive and neutral tweets.Training the SVC model on a labeled dataset of tweets enables it to learn patterns and relationships between the tweet content and their respective negative, positive or neutral labels. The trained model can then be used to predict the labels of new, unseen tweets based on their TF-IDF feature vectors.

It is important to note that the performance of the SVC model for tweet classification relies heavily on the quality and representativeness of the labeled dataset used for training. A diverse and balanced dataset that covers various types and contexts of negative, positive and neutral sentiment is crucial for achieving accurate and robust classification results. Additionally, pre-processing steps such as removing stop words, handling special characters, or stopwords can further enhance the performance of the SVC model by reducing noise and improving the representation of the tweet text. Overall, the combination of SVC with TF-IDF representation for tweet classification provides a powerful framework for distinguishing between negative, postive and neutral . It has the potential to contribute to various applications such as social media monitoring, content moderation, and sentiment analysis, helping to create a safer and more positive online environment.

This study contributes to the ongoing dialogue on xenophobia by employing machine learning model. The integration of literature insights with empirical findings enhances the contextual understanding of sentiment dynamics on social media. As we navigate the complexities of sentiment analysis in sensitive topics, this research underscores the continual need for model refinement and NLP approaches to capture the intricacies of public sentiment, fostering a more comprehensive understanding of the evolving narrative around xenophobia in South Africa.

The study highlights avenues for future research and actions to address the issue of xenophobia on social media. Further exploration of machine learning model enhancements, such as incorporating advanced natural language processing techniques, could improve sentiment analysis accuracy. In terms of actions, the findings underscore the importance of continuous monitoring and moderation of online content, particularly on sensitive topics like xenophobia. Collaborative efforts involving social media platforms, policymakers, and advocacy groups could implement strategies to mitigate the spread of offensive content and promote positive engagement. Educational initiatives aimed at raising awareness about responsible online communication and fostering inclusivity could contribute to a more respectful and understanding online environment. Overall, a different approach involving technological advancements, research, and proactive measures is essential to effectively address and counter xenophobic sentiments on social media.

References

[1] Nabil Badri, Ferihane Kboubi, and Anja Habacha Chaibi. “Combining FastText and Glove word embedding for offensive and hate speech text detection”. In: *Procedia Computer Science* 207 (2022), pp. 769–778.

[2] Pete Burnap and Matthew L Williams. “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making”. In: *Policy & internet* 7.2 (2015), pp. 223–242.

[3] Thomas Davidson et al. “Automated hate speech detection and the problem of offensive language”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1. 2017, pp. 512–515.

[4] Gabriel Araújo De Souza and Márjory Da Costa-Abreu. “Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata”. In: *2020 international joint conference on neural networks (IJCNN)*. IEEE. 2020, pp. 1–6.

[5] Aditya Gaydhani et al. “Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach”. In: *arXiv preprint arXiv:1809.08651* (2018).

[6] Irene Kwok and Yuzhou Wang. “Locate the hate: Detecting tweets against blacks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1. 2013, pp. 1621–1622.

[7] Ivy LB Liu, Christy MK Cheung, and Matthew KO Lee. “Understanding twitter usage: What drive people continue to tweet”. In: *14th Pacific Asia Conference on Information Systems, PACIS 2010*. 2010, pp. 928–939.

[8] Oluwafemi Oriola and Eduan Kotzé. “Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets”. In: *IEEE Access* 8 (2020), pp. 21496–21509.

[9] Gabriel Ichcanziho Pérez-Landa, Octavio Loyola-González, and Miguel Angel Medina-Pérez. “An explainable artificial intelligence model for detecting xenophobic tweets”. In: *Applied Sciences* 11.22 (2021), p. 10801.

[10] Alaa Tharwat. “Classification assessment methods”. In: *Applied computing and informatics* 17.1 (2020), pp. 168–192.

[11] Marcos Zampieri et al. “Predicting the Type and Target of Offensive Posts in Social Media”. In: *Proceedings of NAACL-HLT*. 2019, pp. 1415–1420.