Xenodet: Uncovering Xenophobic Sentiment

Joblin Kanjikuzhiyil James
joblinkj@vt.edu
Virginia Tech
Falls Church, Virginia, USA

Kidus Michael
kidusm128@vt.edu
Virginia Tech
Falls Church, Virginia, USA

Sona Grace John
sonagj@vt.edu
Virginia Tech
Falls Church, Virginia, USA

Abstract

Xenodet: Detecting Xenophobic Content in Social Media Posts aims to develop a robust tool for identifying and classifying xenophobic content on social media platforms, particularly Twitter. With the increasing prevalence of hate speech online, addressing xenophobia has become imperative to fostering healthier and more inclusive online spaces. This project bridges the gap between social media analytics and public welfare by enabling early detection and mitigation of harmful content. By focusing on the classification of posts into two categories—xenophobic and non-xenophobic—this study adopts a targeted approach to improving detection accuracy and prioritizing critical elements for identifying xenophobic discourse. The project combines various natural language processing (NLP) techniques and advanced machine learning models to enhance the classification task and contribute to a safer digital environment.

1 Introduction

In recent years, global migration patterns have undergone significant transformations, driven by various factors such as armed conflicts, environmental crises, and socio-political instability. These challenges have been further compounded by the COVID-19 pandemic, which has exacerbated economic disparities and heightened societal fears about migration and foreign populations. As a result, social attitudes toward immigrants and minority groups have shifted, with xenophobia becoming increasingly normalized in both offline interactions and online discourse. The spread of xenophobic sentiments on social media has fueled division, reinforced harmful stereotypes, and escalated the polarization of communities.

Addressing this pressing issue, the project, titled Xenodet: Detecting Xenophobic Content in Social Media Posts, seeks to develop an advanced tool for recognizing and classifying xenophobic content on platforms such as Twitter. The project is designed to empower social media companies, researchers, and policymakers by providing actionable insights into the nature and prevalence of xenophobia. By bridging the gap in the analysis of sentiments related to immigration and perceptions of individuals labeled as “foreign,” Xenodet seeks to equip stakeholders with the tools needed to foster a safer and more inclusive digital environment.

To achieve this, Xenodet employs a two-class prediction model that categorizes social media posts into xenophobic and non-xenophobic content. This classification approach leverages natural language processing (NLP) techniques and machine learning models to enhance the precision and reliability of xenophobia detection. The project goes beyond simple content moderation by adopting a nuanced methodology that captures the complexities of xenophobic rhetoric, enabling early identification and intervention.

By facilitating the detection of xenophobic content, Xenodet not only improves moderation capabilities but also contributes to the broader objective of promoting constructive online dialogue. It underscores the importance of using technology to mitigate the harmful effects of hate speech, foster inclusivity, and protect public discourse. Through this initiative, the project aims to counteract the normalization of xenophobia, encourage meaningful interactions, and support efforts to build more equitable and tolerant communities in the digital age.

2 Dataset Overview

For this project, we utilized the **HateXplain** dataset, which was introduced in the paper titled *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection* [Submitted on 18 Dec 2020 (v1), last revised 12 Apr 2022 (v2)], authored by Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Although the dataset is primarily designed for hate speech detection, xenophobia represents a critical subset of hate speech, making this dataset suitable for our project.

The dataset comprises social media posts collected from **Twitter** and **Gab** over a specific time period. Key fields in the dataset include:

- **text**: Contains the posts or messages.
- **label**: Categorizes posts into two classes: *“nhs”* (non-hate speech) and *“hs”* (hate speech).
- **target**: Indicates the group being targeted by the post, such as *women*, *religion*, or *ethnicity*.

The full dataset consists of **15,905 rows** and **5 columns**, but since this project focuses specifically on xenophobia, we filtered the dataset to include posts targeting *ethnicity*. This filtering resulted in:

- **5,801 rows** for posts where the `target` is *ethnicity*.

To balance the dataset and include non-xenophobic content, we added posts where the `target` is *null* (indicating no specific group was targeted). These rows account for:

- **4,850 rows** for posts with no target value.

After combining these subsets, the final dataset used for our project consists of:

- **10,651 rows** and **5 columns**.

2.1 Ensuring Class Balance

To ensure there is no class imbalance in the dataset, we verified the distribution of `label` values in the final data:

- *nhs* (non-hate speech): **5,387 rows**.
- *hs* (hate speech): **5,264 rows**.

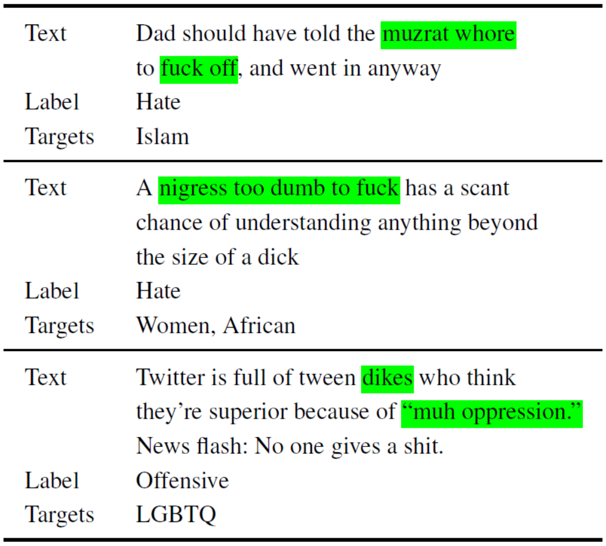
This near-equal distribution minimizes potential biases in the prediction model and ensures fair representation of both classes.

2.2 Dataset Curation Details

Several preprocessing steps were taken during the creation of the HateXplain dataset to ensure data quality and consistency:

- **Emojis**: Retained in the text as they can carry significant emotional or contextual meaning for hate speech detection.
- **Links, Pictures, and Videos**: Posts containing these elements were removed to avoid distractions and potentially harmful content for the annotators during crowdsourced labeling.
- **Usernames**: Replaced with `<user>` tokens to maintain anonymity while preserving sentence structure.
- **Reposts and Duplicates**: Excluded to avoid redundancy and ensure unique data points.

By focusing on posts targeting ethnicity and including posts with no specific target, this curated dataset forms the foundation for our xenophobia detection model. It captures a balanced view of xenophobic and non-xenophobic content, making it an ideal resource for this project.



Text	Dad should have told the muzrat whore to fuck off, and went in anyway
Label	Hate
Targets	Islam
Text	A nigress too dumb to fuck has a scant chance of understanding anything beyond the size of a dick
Label	Hate
Targets	Women, African
Text	Twitter is full of tween dikes who think they're superior because of "muh oppression." News flash: No one gives a shit.
Label	Offensive
Targets	LGBTQ

Figure 1: Example tweets from the dataset showing key attributes for each post

3 Data Preprocessing

Data preprocessing plays a pivotal role in preparing the dataset for effective model training and evaluation. Several steps were undertaken to ensure the dataset was clean, consistent, and optimized for the task of xenophobia detection. The detailed preprocessing pipeline is as follows:

3.1 Dropped Unwanted Columns

The original dataset contained multiple columns, some of which were irrelevant to the objective of xenophobia classification. To streamline the dataset and reduce computational overhead, only the essential columns were retained:

- **Text:** Containing the social media posts.
- **Label:** Indicating whether the post is xenophobic or non-xenophobic.
- **Target:** Specifying the group targeted in the post, such as ethnicity or religion.

All other unnecessary columns were dropped, simplifying the dataset structure.

3.2 Handling Missing Values

The dataset was checked for missing values across all selected rows and columns. Upon inspection, no null values were found in the data subset. Therefore, explicit handling of missing values, such as imputation or deletion, was not required.

3.3 Removal of Unnecessary Characters

To ensure the text data was clean:

- Symbols such as <, >, and other non-alphanumeric characters were removed, as they introduced noise into the data.
- Emojis, which can carry significant sentiment information, were retained but handled using their Unicode representations for compatibility with text processing tools.

This step ensured the text data was standardized and ready for feature extraction.

3.4 Stopword Removal

Stopwords, which are common words like *is*, *the*, and *and* that do not contribute significant meaning, were removed using SpaCy's predefined list of stopwords. In addition:

- A custom list of domain-specific stopwords was defined and removed. This included words like *make*, *anything*, *user*, and *someone*, which provided negligible value for xenophobia detection.
- A word cloud visualization was generated to identify high-frequency, contextually irrelevant words. Based on this visualization, additional custom stopwords were identified and excluded from the dataset.

3.5 Label Encoding

The `label` column in the dataset initially contained two string values:

- **nhs:** Non-hate speech.
- **hs:** Hate speech.

For binary classification, these labels were encoded into numerical values:

- **0:** Non-xenophobic posts.
- **1:** Xenophobic posts.

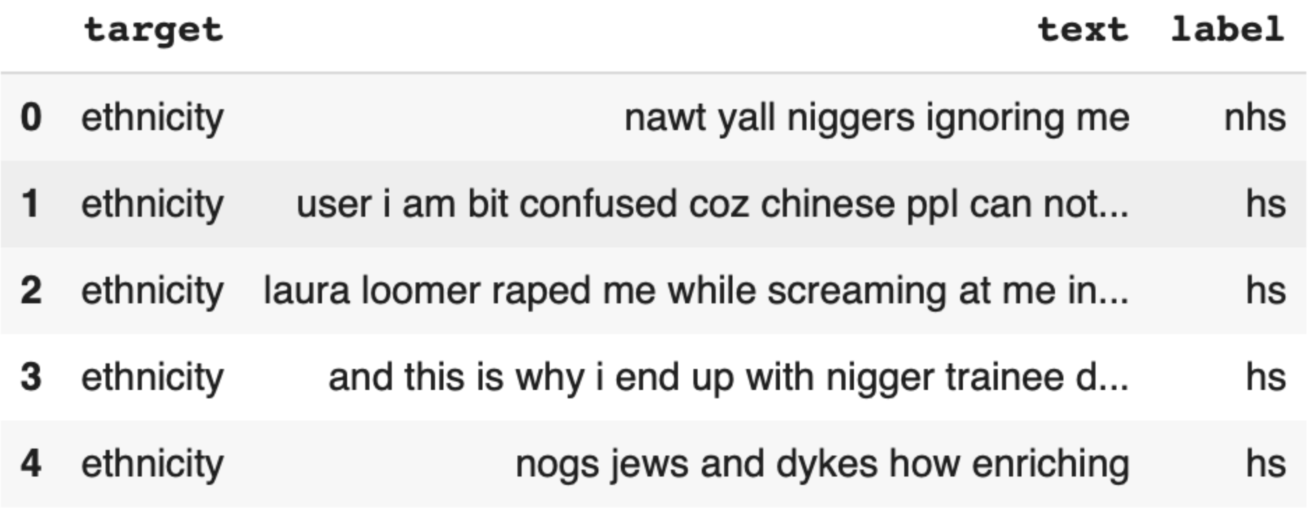
This numerical encoding ensured compatibility with machine learning algorithms.

3.6 Final Dataset Structure

After preprocessing, the dataset was clean and structured as follows:

- **Text:** Cleaned social media posts.
- **Label:** Numerical labels indicating xenophobia or non-xenophobia.

The processed dataset was ready for feature extraction and model training, providing a robust foundation for the classification task.



	target	text	label
0	ethnicity	nawt yall niggers ignoring me	nhs
1	ethnicity	user i am bit confused coz chinese ppl can not...	hs
2	ethnicity	laura loomer raped me while screaming at me in...	hs
3	ethnicity	and this is why i end up with nigger trainee d...	hs
4	ethnicity	nogs jews and dykes how enriching	hs

Figure 2: Example rows from the dataset after preprocessing.

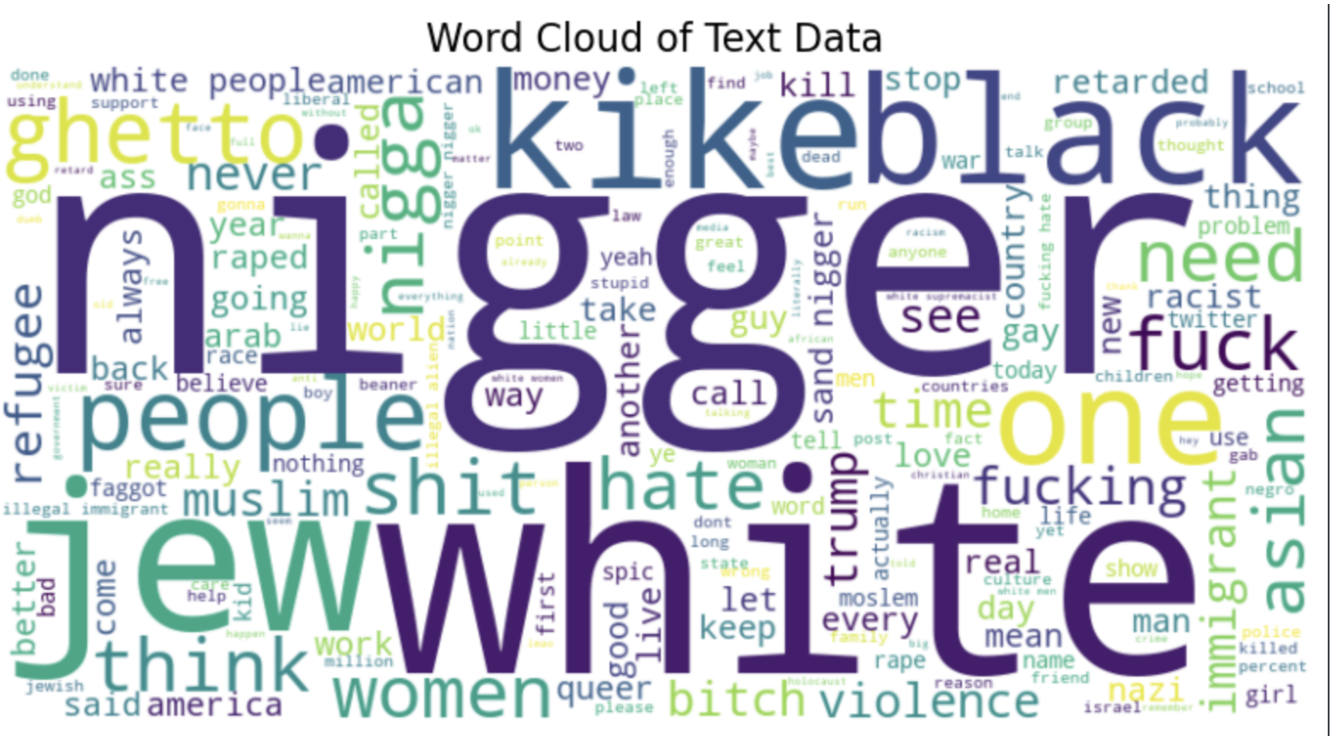


Figure 3: Word Cloud generated after preprocessing.

4 Multiple Approaches

To effectively detect and classify xenophobic content in tweets, a combination of natural language processing (NLP) techniques and machine learning models will be applied. Given the complexity of identifying nuanced sentiments and xenophobic language, a multi-faceted approach is essential to capture both the lexical and contextual features of the data. The following methods include preprocessing techniques to clean and prepare text, embedding models to represent the semantic relationships between words, and named entity recognition to highlight specific entities related to foreigners and immigration. These diverse approaches aim to improve classification accuracy and provide a robust framework for analyzing xenophobic discourse on social media.

4.1 Approach I: Frequency-Based Embedding

Frequency-based embedding techniques are one of the simplest and most effective ways to represent text data numerically. These methods rely on calculating the frequency of words or terms in the text, which provides insights into the importance of certain words or phrases within the corpus. Among frequency-based methods, **TF-IDF (Term Frequency-Inverse Document Frequency)** is widely used due to its ability to highlight significant terms while minimizing the impact of commonly used words.

4.1.1 *TF-IDF*.

TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to the entire corpus. It assigns higher weights to words that are frequent in a specific document but rare across the entire dataset. The formula for calculating TF-IDF for a word t in a document d is:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \cdot \text{IDF}(t)$$

Where:

- $\text{TF}(t,d) = \frac{\text{Number of occurrences of } t \text{ in } d}{\text{Total words in } d}$: Term Frequency.
- $\text{IDF}(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$: Inverse Document Frequency.

Compared to **Bag of Words (BoW)**, which considers only the presence or absence of terms without accounting for their relevance, TF-IDF provides a more nuanced representation. It discounts the weight of common words (e.g., "the," "is") and emphasizes rarer, context-specific terms, making it particularly suitable for detecting xenophobic language.

For this approach, the lemmatized posts were directly embedded using TF-IDF to capture the most relevant terms while ignoring unimportant words.

4.1.2 *Capturing Sentiment Meaning Using VADER*.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based sentiment analysis tool designed to capture the sentiment polarity of text. It assigns a compound score to each input, categorizing the sentiment as:

- **Positive:** Compound score > 0.05.
- **Neutral:** Compound score between -0.05 and 0.05.
- **Negative:** Compound score < -0.05.

The compound score, which ranges from -1 (most negative) to +1 (most positive), provides an overall sentiment measure. In the context of xenophobic tweets, VADER is crucial as sentiment often plays a significant role in conveying hostility or bias. For example, negative sentiment can reflect derogatory or offensive language, while positive sentiment in a xenophobic context might indicate sarcastic or mocking tones. Incorporating VADER scores allows models to better understand these subtleties and improve classification accuracy.

4.1.3 *Capturing Contextual Meaning Using POS Tags and NER Features*.

To enhance the contextual understanding of posts, linguistic features such as Part-of-Speech (POS) tags and Named Entity Recognition (NER) features were utilized:

Part-of-Speech (POS) Tagging: POS tagging involves identifying the grammatical role of each word in a sentence (e.g., noun, verb, adjective). By highlighting the syntactic structure of a sentence, POS tags help models focus on specific parts of the text that carry significant meaning. For example:

- Verbs and adjectives often convey strong sentiment in xenophobic posts (e.g., "destroyed," "disgusting").
- Nouns can identify the target group being referenced (e.g., "immigrants," "refugees").

Named Entity Recognition (NER): NER identifies and categorizes entities within the text, such as people, locations, organizations, or other proper nouns. In the context of xenophobia, NER helps pinpoint specific groups or individuals being targeted, such as ethnicities, religions, or countries. For instance:

- A tweet mentioning "immigrants from Mexico" explicitly identifies a group and location, providing context crucial for classification.

Combining POS tags and NER features allows models to capture both the structural and semantic components of a sentence, enabling a deeper understanding of xenophobic intent. This contextual insight is essential for detecting not just overtly xenophobic posts but also those that employ subtle or implicit biases.

4.2 Approach II: Word Embedding

Word embeddings are dense vector representations of words that capture their semantic and syntactic relationships. Unlike frequency-based embeddings like TF-IDF or Bag of Words, which represent words as sparse vectors, word embeddings map words into continuous vector spaces where similar words are located close to each other. This allows the model to understand the contextual meaning and relationships between words, making them particularly useful for tasks like xenophobia detection.

4.2.1 *Using GloVe and Word2Vec*.

GloVe (Global Vectors for Word Representation): GloVe is a word embedding technique that leverages the global co-occurrence statistics of words in a corpus. It constructs a co-occurrence matrix and factorizes it to generate dense vector representations. These embeddings capture both semantic and syntactic relationships between words. For example, "immigrant" and "refugee" might have similar vectors due to their frequent co-occurrence in similar contexts. GloVe embeddings are static, meaning each word has a fixed vector representation regardless of its context.

Word2Vec: Word2Vec is another word embedding method that uses neural networks to generate word vectors. Word2Vec focuses on local contexts, capturing word relationships based on proximity in the text. Unlike GloVe, Word2Vec generates embeddings that are better suited for capturing word relationships in smaller or domain-specific datasets.

For this approach, the lemmatized posts were embedded using both GloVe and Word2Vec to train the models. While both techniques effectively represented the semantic meaning of individual words, they struggled to capture sentence-level contextual nuances, which are crucial for identifying xenophobic language.

4.2.2 *Using GloVe and Word2Vec with POS Tagging*.

To enhance the contextual understanding provided by word embeddings, Part-of-Speech (POS) tags were incorporated as additional features. POS tagging assigns grammatical roles (e.g., nouns, verbs, adjectives) to words, which helps highlight the structural and syntactic aspects of a sentence.

Using POS Tags with Word Embeddings:

- Verbs and adjectives often carry strong sentiment and are crucial for detecting xenophobic tones (e.g., "blamed," "disgusting").
- Nouns provide information about the entities or groups being targeted (e.g., "immigrants," "foreigners").

By integrating POS tags, the embeddings were enriched with syntactic information, allowing the models to better capture the sentence's structure and its underlying meaning. For example:

- A sentence like "Immigrants are ruining the economy" highlights the noun "immigrants" as the target and the verb "ruining" as indicative of hostility.
- Combining the semantic representation from word embeddings with the grammatical structure from POS tags enables the model to interpret the sentence more holistically.

This approach showed improved results over using word embeddings alone, emphasizing the importance of contextual and structural features in detecting xenophobic content.

4.3 Approach III: Sentence Embedding

Sentence embeddings are advanced text representations that capture the semantic and contextual meaning of entire sentences rather than individual words. Unlike word embeddings, which generate fixed vectors for words regardless of their context, sentence embeddings create dense vector representations that consider the relationships between all words in a sentence. This makes them particularly useful for tasks requiring a deep understanding of sentence-level semantics, such as identifying xenophobic content that may involve complex linguistic patterns or implicit biases.

4.3.1 *S-BERT*.

Sentence-BERT (S-BERT) is a modification of the BERT architecture, specifically designed to generate high-quality sentence embeddings. By incorporating a Siamese network structure, S-BERT effectively computes sentence similarity by mapping sentences into a vector space where similar sentences are closer together. For this study, we employed the **all-MiniLM-L6-v2** model, which is a lightweight version of S-BERT optimized for computational efficiency.

The **all-MiniLM-L6-v2** model generates 384-dimensional embeddings for each post, capturing both contextual and semantic nuances. These embeddings were used as input features for three classification models: Support Vector Machine (SVM), XGBoost, and Logistic Regression.

Why S-BERT is Expected to Perform Well:

- **Contextual Understanding:** S-BERT captures the relationships between words and phrases in a sentence, making it highly effective for tasks involving implicit or nuanced language, such as xenophobia detection.
- **Efficiency:** By leveraging a lightweight architecture, S-BERT ensures a balance between computational efficiency and embedding quality, making it suitable for datasets of varying sizes.
- **Semantic Representation:** The embeddings generated by S-BERT are highly robust, enabling the models to identify both overtly hostile content and subtle biases present in the text.

4.3.2 *Universal Sentence Encoder (USE)*.

The **Universal Sentence Encoder (USE)** is another method for generating sentence embeddings, designed to provide universal representations for diverse text types. Unlike S-BERT, which is optimized for semantic similarity tasks, USE is trained on a combination of supervised and unsupervised data, making it a general-purpose encoder. USE produces 512-dimensional embeddings, which are larger than those of S-BERT, and are known for their scalability and performance on large, diverse datasets.

- **Scalability:** USE is optimized for high scalability, making it suitable for applications involving massive datasets.
- **Pretrained on Diverse Data:** Its training on a wide range of datasets ensures that USE embeddings are versatile and generalizable across different text types.
- **Efficiency:** Despite its larger embedding size, USE maintains computational efficiency, allowing it to be used in real-time applications.

5 Models

To classify xenophobic and non-xenophobic content effectively, we employed three machine learning models: **Support Vector Machine (SVM)**, **XGBoost**, and **Logistic Regression**. These models were selected for their complementary strengths in handling various data complexities and feature types.

5.1 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that works by finding the optimal hyperplane to separate data points into distinct classes. It is particularly effective for high-dimensional data, making it a strong choice for this task, where the feature space includes lexical, grammatical, and sentiment-based attributes. SVM's ability to utilize different kernels allows it to handle both linear and nonlinear relationships within the data.

5.2 XGBoost

XGBoost, or Extreme Gradient Boosting, is a powerful and efficient implementation of gradient-boosted decision trees. It is well-suited for tasks involving complex feature sets as it captures intricate interactions between variables. XGBoost is particularly adept at handling sparse data and reducing overfitting through regularization techniques, making it an excellent choice for datasets with diverse feature representations.

5.3 Logistic Regression

Logistic Regression is a simple yet effective linear model used for binary classification. It predicts the probability of a class using the logistic function and is known for its interpretability and computational efficiency. Logistic Regression serves as a strong baseline model and provides insight into the relationships between features and target variables.

5.4 Model Application

The combination of these models allowed us to leverage the strengths of each approach. SVM captured high-dimensional relationships, XGBoost handled complex feature interactions effectively, and Logistic Regression provided a reliable baseline for comparison. Together, these models form the foundation of our classification approach, offering a comprehensive understanding of xenophobic content in social media posts.

6 Metrics

To evaluate the effectiveness of the approaches in identifying and classifying xenophobic content, several key performance metrics will be used. These metrics will provide a comprehensive assessment of the models' ability to accurately detect xenophobic discourse, balance false positives and negatives, and generalize well to unseen data. Success will be measured through the following criteria:

- `Accuracy`
- `Precision`
- `Recall (Sensitivity)`
- `F1 Score`

These metrics together will offer a detailed evaluation of the model's performance, allowing for a thorough analysis of its strengths and weaknesses in detecting xenophobic content. Success will be determined by achieving a high F1 score, balanced precision and recall, indicating a reliable and accurate classification model.

7 Observations

This section provides a detailed analysis of the results obtained through various models and preprocessing approaches. The evaluation metrics—accuracy, precision, recall, and F1-score—are used to assess the performance of the models.

Table 1: Xenophobic scores for TF-IDF

	Acruaccy Score	Precision	Revall	F1-Score
SVM	0.85	0.89	0.81	0.85
XGBoost	0.86	0.90	0.81	0.85
Logistic Regression	0.85	0.90	0.79	0.85

Table 2: Xenophobic scores for VADER

	Acruaccy Score	Precision	Revall	F1-Score
SVM	0.85	0.90	0.81	0.85
XGBoost	0.86	0.90	0.81	0.85
Logistic Regression	0.85	0.90	0.78	0.84

7.1 Approach 1: Frequency Based Embedding

*7.1.1 **TF-IDF**.*

The first approach involved lemmatizing the posts using SpaCy and then applying TF-IDF (Term Frequency-Inverse Document Frequency) to extract features from the lemmatized text. These features were passed to the models—SVM, XGBoost, and Logistic Regression—for classification. As for the results seen in **Table 1**.

- **SVM:** The Support Vector Machine performed well with an accuracy of 85%, showcasing its ability to classify xenophobic and non-xenophobic posts effectively. It achieved a precision of 0.89 for xenophobic posts, but its recall for the same category was slightly lower at 0.81.
- **XGBoost:** XGBoost demonstrated the best performance among all models, with an accuracy of 86%. It maintained a good balance between precision (0.90 for xenophobic posts) and recall (0.81 for xenophobic posts), leading to a high F1-score.
- **Logistic Regression:** While Logistic Regression achieved the same accuracy as SVM (85%), its recall for xenophobic posts was slightly lower at 0.79. Despite this, it maintained a high precision of 0.90.

*7.1.2 **Capturing Sentiment Meaning using VADER**.*

In this approach, the lemmatized posts were analyzed to extract sentiment polarity using VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER assigns a compound score to the text, categorizing it into:

- **Positive:** Compound score > 0.05.
- **Neutral:** Compound score between -0.05 and 0.05.
- **Negative:** Compound score < -0.05.

The sentiment scores, combined with TF-IDF features, were used to train the models. The results, **Table 2**, however, showed minimal improvement compared to the TF-IDF-only approach. The following observations were made:

- **Support Vector Machine (SVM):** The SVM model maintained an accuracy of **85%**, similar to the previous results.

While the integration of VADER's sentiment scores was expected to improve performance, the results remained consistent. This could be due to SVM's inherent focus on identifying linear decision boundaries, where the additional sentiment information did not contribute significantly beyond the features already captured by TF-IDF.

- **XGBoost:** XGBoost achieved an accuracy of **86%**, again showing no substantial improvement despite incorporating sentiment polarity. XGBoost is highly effective at capturing feature relationships, and it appears that the existing TF-IDF features already provided sufficient information for classification, leaving little room for additional gains from sentiment scores.
- **Logistic Regression:** Logistic Regression retained an accuracy of **85%**, similar to SVM. However, a minor decrease in the recall for xenophobic posts was observed, dropping from **0.79** to **0.78**. This slight decline could be attributed to the sensitivity of logistic regression to changes in feature scaling and weighting, where the addition of sentiment scores might have slightly shifted the decision boundary.

Overall, while the inclusion of sentiment polarity provided valuable insights into the data, its impact on the model's predictive capabilities was limited. This suggests that the core features captured by TF-IDF and the balanced dataset played a more significant role in driving the model's performance, with sentiment scores offering marginal contributions.

7.1.3 *Capturing Contextual Meaning using POS Tag and NER Features*.

To enhance the contextual understanding of xenophobic content, Part-of-Speech (POS) tagging and Named Entity Recognition (NER) features were extracted and incorporated into the models.

- **POS Tagging:** The frequency of grammar roles (e.g., nouns, verbs, adjectives) was calculated for each post. While the expectation was that this syntactic structure would add nuanced insights, the models demonstrated similar performance to earlier approaches, indicating that the additional grammatical context did not significantly enhance the detection of xenophobic content.
- **NER:** Named entities such as locations, ethnic groups, and other specific mentions were extracted and included as features. These were aimed at highlighting targeted groups in xenophobic posts. Despite the hope for improved classification, the results remained comparable to those achieved with TF-IDF and sentiment polarity features. This suggests that the inherent linguistic patterns in the text, already captured by TF-IDF, were dominant in driving the model's predictions.

Although these techniques contributed valuable information for analysis, their integration into the models yielded results consistent with previous approaches. This consistency indicates that the models were already effectively leveraging the most impactful features, leaving limited room for further enhancement through these additional methods.

7.1.4 *Conclusion*.

Table 3: Evaluation Metrics for Xenophobic Posts: GloVe

	Accuracy	Precision	Recall	F1-Score
SVM	0.64	0.69	0.50	0.58
XGBoost	0.66	0.66	0.63	0.65
Logistic Regression	0.65	0.67	0.60	0.63

Table 4: Evaluation Metrics for Xenophobic Posts: Word2Vec

	Accuracy	Precision	Recall	F1-Score
SVM	0.67	0.67	0.63	0.65
XGBoost	0.68	0.67	0.67	0.67
Logistic Regression	0.65	0.66	0.61	0.63

Table 5: Evaluation Metrics for Xenophobic Posts: Glove+POS

	Accuracy	Precision	Recall	F1-Score
SVM	0.77	0.77	0.74	0.76
XGBoost	0.76	0.77	0.72	0.74
Logistic Regression	0.72	0.74	0.68	0.70

Table 6: Evaluation Metrics for Xenophobic Posts: Word2Vec+POS

	Accuracy	Precision	Recall	F1-Score
SVM	0.80	0.83	0.75	0.79
XGBoost	0.77	0.79	0.74	0.76
Logistic Regression	0.77	0.80	0.71	0.75

Since TF-IDF is a frequency-based embedding, it was observed that xenophobic content often contains "bar" or "slay" words—words that carry strong emotional weight or frequent usage in such posts. These terms became dominant features during training, leading the models to prioritize these patterns when making predictions.

This behavior reflects an intriguing aspect of xenophobic language: its reliance on specific repetitive terms that amplify hostility or bias. As a result, the models effectively learned to associate these frequent terms with xenophobic posts. While this improved classification performance for posts containing these patterns, it also introduced a risk of overfitting to linguistic styles dominated by such words. This underscores the importance of combining frequency-based features with context-aware techniques, such as word or sentence embeddings, to capture both lexical and semantic intricacies in the text.

Future enhancements could explore leveraging deep contextual embeddings like BERT or exploring hybrid models that balance frequency-based and semantic representations to achieve more robust and generalized detection of xenophobic content.

These observations highlight the strengths and limitations of each approach and pave the way for future enhancements in detecting xenophobic content.

7.2 Approach 2: Word Embedding

7.2.1 *Word Embedding: GloVe and Word2Vec*.

The initial experiments with word embeddings, namely GloVe and Word2Vec, were conducted by directly applying lemmatized posts to these models to understand their effectiveness in capturing contextual meaning.

Using GloVe embeddings, the models exhibited moderate performance, with SVM achieving an accuracy of 0.64, XGBoost slightly better at 0.66, and Logistic Regression at 0.65. In comparison, Word2Vec embeddings yielded slightly higher scores, with SVM achieving 0.67, XGBoost 0.68, and Logistic Regression maintaining a similar 0.65 accuracy.

These results suggest that while traditional word embeddings can provide a foundational understanding of individual word relationships, they are insufficient for tasks requiring deeper semantic comprehension, such as identifying xenophobic sentiment. Incorporating additional linguistic features, such as POS tags or exploring contextual embeddings, becomes essential to improve performance.

The detailed evaluation scores are presented in **Tables 3 and 4**, respectively.

Reason for Lower Scores: The low scores can be attributed to the inability of these embeddings to fully grasp the contextual meaning of sentences. Both GloVe and Word2Vec models rely on statistical co-occurrence of words in a corpus and do not capture intricate relationships or sentence-level semantics effectively. As a result, subtle nuances in xenophobic language, such as sarcasm or context-dependent meanings, were not well understood.

7.2.2 *Word Embedding with POS Tags*.

To improve contextual understanding, POS tags were incorporated alongside word embeddings such as GloVe and Word2Vec. This approach aimed to emphasize grammatical roles and patterns that could enhance the models' ability to detect xenophobic content.

Using GloVe embeddings with POS tags, the SVM model with an RBF kernel achieved an accuracy of 0.77, while XGBoost and Logistic Regression followed with 0.76 and 0.72, respectively. Similarly, incorporating POS tags with Word2Vec embeddings resulted in further improvement, with SVM (RBF kernel) achieving the highest accuracy of 0.80, followed by XGBoost at 0.78 and Logistic Regression at 0.77. The detailed evaluation scores are presented in **Tables 5 and 6**, respectively.

The improvement in scores with the inclusion of POS tags can be attributed to their ability to highlight key grammatical roles, such as verbs, adjectives, and adverbs, which often carry significant sentiment in xenophobic posts. By focusing on these parts of speech, the models were better equipped to identify patterns and contextual nuances within the text, leading to a more accurate classification.

Notably, Word2Vec performed better than GloVe in this approach. This is likely because Word2Vec creates embeddings based on word sequences and relationships, which aligns well with the structured grammatical information provided by POS tags. In contrast, GloVe, which relies on global co-occurrence statistics, might not have leveraged the additional POS information as effectively.

The highest score was observed with Word2Vec and SVM using an RBF kernel, achieving an accuracy of 0.80.

Reason for Improved Scores: The inclusion of POS tags improved the contextual understanding by emphasizing grammatical roles and syntactic relationships. For instance, verbs and adjectives often carry significant sentiment in xenophobic posts, and the models were better equipped to recognize these patterns when POS tags were incorporated. This enhancement allowed the embeddings to focus on more meaningful features, resulting in noticeable performance gains.

7.2.3 *Conclusion*.

This method highlights the importance of incorporating linguistic features, such as POS tagging, to capture the conceptual meaning behind xenophobic content. Among the various configurations tested, the SVM model with an RBF kernel, combined with Word2Vec embeddings enriched by POS tagging, emerged as the most effective. This approach achieved the highest accuracy, indicating its capability to identify nuanced patterns and relationships in the text that go beyond surface-level word associations.

The use of POS tagging in conjunction with word embeddings allowed the model to focus on syntactic and semantic roles within sentences, improving its ability to discern xenophobic intent. The RBF kernel further enhanced this by capturing non-linear decision boundaries, which are essential for distinguishing subtle differences in linguistic context. This method showcases a promising direction for detecting complex linguistic phenomena in social media data.

7.3 Approach 3: Sentence Embedding

7.3.1 *S-BERT*.

In this approach, the posts were embedded using Sentence-BERT (S-BERT), a sentence-level embedding model that builds upon BERT's transformer architecture. S-BERT generates dense sentence embeddings that effectively capture semantic relationships and contextual nuances in the text. For this experiment, embeddings with a dimension of 384 were used as input features for the models.

The performance of the models trained on S-BERT embeddings is as follows:

- **Support Vector Machine (SVM):** Achieved an accuracy of **0.86**, the highest among the tested models. Notably, it also recorded the highest recall across all approaches, with a value of **0.83**, indicating its robustness in identifying xenophobic posts.
- **XGBoost:** Scored an accuracy of **0.82**, demonstrating good performance but slightly lower than SVM.
- **Logistic Regression:** Achieved an accuracy of **0.83**, balancing precision and recall effectively.

The detailed evaluation scores are presented in **Table 7**, respectively. The use of sentence embeddings significantly improved the contextual understanding of xenophobic content. Unlike word-level embeddings, S-BERT captures sentence-level meaning, enabling the models to better interpret complex linguistic structures, implicit biases, and subtle variations in tone. This resulted in a marked improvement in recall for SVM, highlighting its ability to generalize well across diverse samples.

While XGBoost and Logistic Regression performed slightly below SVM, their results validate the effectiveness of S-BERT embeddings in this classification task. This approach emphasizes the importance

Table 7: Evaluation Metrics for Xenophobic Posts: S-BERT

	Accuracy	Precision	Recall	F1-Score
SVM	0.86	0.88	0.83	0.85
XGBoost	0.82	0.83	0.80	0.82
Logistic Regression	0.83	0.85	0.81	0.83

Table 8: Evaluation Metrics for Xenophobic Posts: USE

	Accuracy	Precision	Recall	F1-Score
SVM	0.82	0.84	0.79	0.81
XGBoost	0.82	0.85	0.79	0.82
Logistic Regression	0.80	0.83	0.78	0.80

of leveraging sentence-level semantic representations for detecting nuanced linguistic phenomena like xenophobia.

7.3.2 *Universal Sentence Encoder (USE).*

In this approach, the posts were embedded using the Universal Sentence Encoder (USE), which generates 512-dimensional sentence embeddings. USE is designed to provide high-quality sentence representations and is optimized for scalability, making it particularly effective for large-scale datasets and diverse text inputs.

The performance of the models trained on USE embeddings is as follows:

- **Support Vector Machine (SVM):** Achieved an accuracy of **0.82**, demonstrating consistent performance in this context.
- **XGBoost:** Also scored an accuracy of **0.82**, indicating that it effectively utilized the USE embeddings but did not surpass SVM.
- **Logistic Regression:** Recorded an accuracy of **0.80**, slightly lower compared to the other models.

Reason for Lower Performance Compared to S-BERT: Despite its high-quality embeddings and scalability for massive datasets, USE did not perform as well as S-BERT on this specific task. This is likely due to the relatively small size of our dataset, comprising only 10,651 posts. USE's strength lies in its ability to generalize across massive and diverse datasets, but for smaller datasets, such as ours, S-BERT's fine-tuned contextual representations provided a better fit for capturing nuanced xenophobic content.

The detailed evaluation scores are presented in **Table 8**. These results highlight the importance of selecting embeddings that align with the dataset characteristics. While USE is highly efficient and scalable, S-BERT offers superior performance on smaller, focused datasets by capturing more intricate contextual details.

7.3.3 *Conclusion.*

The comparison between S-BERT and USE highlights the importance of selecting sentence embeddings that align with the specific dataset characteristics. S-BERT demonstrated superior performance, achieving the highest accuracy and recall across all models. Its fine-tuned contextual representations effectively captured the nuances and semantics of xenophobic content, making it a better fit for smaller, focused datasets like ours.

In contrast, while USE performed consistently with an accuracy of 0.82 for both SVM and XGBoost, its generalized embeddings, optimized for scalability, were less effective in leveraging the contextual richness of our relatively small dataset. This underscores that while USE excels in large-scale applications, S-BERT provides a more precise and nuanced understanding of text when working with targeted datasets.

8 Conclusion

In this study, we implemented and evaluated three machine learning models: **Support Vector Machine (SVM)**, **XGBoost**, and **Logistic Regression**, to classify xenophobic and non-xenophobic content in social media posts. Each model demonstrated unique strengths in handling different types of embeddings.

The **SVM classifier**, configured with a radial basis function (RBF) kernel, provided robust performance, effectively managing the complexity of the data and achieving strong predictive results. **XGBoost**, known for its speed and predictive power, utilized log loss as the evaluation metric, delivering balanced performance across all metrics. **Logistic Regression**, while simplistic, served as a reliable baseline model for binary classification, highlighting the importance of linear relationships in the data.

The models were evaluated using metrics such as accuracy, precision, recall, and F1-score. Among the various embedding approaches, **S-BERT** demonstrated strong performance, significantly enhancing the classifiers' ability to distinguish between xenophobic and non-xenophobic content. The contextual understanding provided by S-BERT embeddings allowed the models to better interpret nuanced hate speech, with **SVM and XGBoost** showcasing particular strengths in accuracy and robustness. S-BERT's ability to provide efficient, scalable, and context-aware semantic representations made it an indispensable tool, outperforming frequency-based methods like TF-IDF and basic word embeddings such as GloVe and Word2Vec.

Interestingly, both **S-BERT and TF-IDF** embeddings achieved an accuracy of **86%**, a result that was not anticipated. The explanation lies in the nature of xenophobic posts, which often contain repetitive and high-frequency terms that are well captured by TF-IDF. However, TF-IDF struggles to encapsulate sentence-level contextual meaning, which is where S-BERT excels. For our dataset, S-BERT embeddings offered the highest recall value of **0.83**, making it the preferred choice for this study. Recall is a critical metric for this task as it measures the model's ability to identify all xenophobic posts without missing any, aligning with our priority of maximizing xenophobic content detection.

Based on these observations, we conclude that while TF-IDF performs well in identifying xenophobic content due to its strength in frequency-based representations, S-BERT provides a more robust solution for this specific classification task. Its ability to capture both contextual and semantic nuances makes it the optimal embedding method for ensuring comprehensive detection of xenophobic posts.

9 Future Work

This study has laid the groundwork for detecting xenophobic content in social media posts, but several avenues for future exploration remain to further enhance the robustness and effectiveness of the system:

- **Incorporation of Contextualized Language Models:** While S-BERT demonstrated strong performance, future work can explore more advanced transformer-based models like RoBERTa or BERT-large to capture deeper semantic relationships. Fine-tuning these models specifically on xenophobic datasets could further improve detection accuracy and recall.
- **Multilingual Support:** The current system is limited to English posts. Expanding to include other languages commonly used on social media, along with cross-lingual embeddings, could significantly broaden the scope of xenophobia detection.
- **Incorporating Image and Multimodal Data:** Many xenophobic posts on social media include images, memes, or videos that convey hostile content. Future work could integrate image and text analysis using multimodal approaches to detect xenophobia comprehensively.
- **Real-Time Implementation:** Deploying the system as a real-time monitoring tool for social media platforms could provide immediate insights and alerts for xenophobic content, aiding in faster moderation.

By addressing these areas, future iterations of this system can become more comprehensive, accurate, and applicable across diverse scenarios, ultimately contributing to the fight against xenophobia in online spaces.

10 References

- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the Eighteenth International AAAI Conference on Web and Social Media
- Khonzoda Umarova, Oluchi Okorafor, Pinxian Lu, Sophia Shan, Alex Xu, Ray Zhou, Jennifer Otiono, Beth Lyon, Gilly Leshed. 2024. Xenophobia Meter: Defining and Measuring Online Sentiment toward Foreigners on Twitter. Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM2024)
- Khonzoda Umarova, Oluchi Okorafor, Pinxian Lu, Sophia Shan, Alex Xu, Ray Zhou, Jennifer Otiono, Beth Lyon, Gilly Leshed. (2024).Xenophobia Meter Dataset. https://dataverse.harvard.edu/dataset.xhtml? persistentId = doi:10.7910/DVN/KYR4IY
- Jackie Swift. 2024. XENOPHOBIA METER AIMS TO TRACK ANTI-IMMIGRANT HATE SPEECH. https://cis.cornell.edu/xenophobia -meter-aims-track-anti-immigrant-hate-speech