Identification of Social Media Users that Perpetuate Xenophobic Attitudes and Hate Speech Narratives in South Africa

Carl du Plessis[1], Michael du Plessis[1], Koena Ronny Mabokela[2], Abiodun Modupe[1,3], Vukosi Marivate[1,3,4]

[1] Dept. Computer Science, University of Pretoria
[2] Applied Information Systems, University of Johannesburg
[3] Data Science for Social Impact
[4] Lelapa AI
u88534953@tuks.co.za

Abstract. Social media—particularly X (formerly Twitter)—has become a critical platform for political discourse. It shapes public opinion, influences voter behaviour, and provides real-time insight into contentious issues. Xenophobia, defined as the hostility, or hatred towards foreigners, is a polarising topic in South Africa, especially during election seasons. This paper analyses South African Twitter data from the 2016 and 2021 municipal elections, as well as the 2019 and 2024 national elections, with a focus on Xenophobia-related discourse. We develop a novel machine learning model to identify xenophobic tweets despite the removal of explicit hate speech by platform moderation. Using a labelled dataset of xenophobic tweets, we fine-tuned a transformer-based classifier that achieves over 95% F1-score in distinguishing xenophobic content. We then analyse the prevalence of xenophobic narratives over time, the peaks around election dates, and the user accounts most active in propagating xenophobia. Our results reveal thousands of Xenophobic tweets, peaking sharply during election periods, and show that over half of the top 20 xenophobia-spreading accounts appear affiliated with political figures or parties. We discuss implications for social media policy, election integrity, and community cohesion. We also address ethical considerations such as data privacy, anonymisation of users, and bias. This work contributes a framework for identifying harmful election-related discourse and insights for mitigating the impact of xenophobic narratives on social media.

Keywords: Xenophobia · Hate Speech · Social Media · South Africa · Large Language Models · Elections.

1 Introduction

In democratic societies, elections often heighten debates on identity, immigration, and security. In South Africa, xenophobia—hostility toward foreigners—has

periodically escalated into violence, with social media amplifying these sentiments [1]. Platforms like X (Twitter) provide space for political rhetoric to evolve into harmful narratives against immigrants. Since 1994, Xenowatch has recorded over 1,150 xenophobic incidents and 686 deaths [2]. Such attitudes, rooted in fears of outsiders "*altering traditions*" or "*threatening economic survival*," are often politicised during elections, underscoring the need to examine how social media perpetuates them in this context.

South Africa's social media landscape is significant but not ubiquitous. As of early 2024, X had about 4.10 million users in South Africa (roughly 6.8% of the population) [3]. This user base, although a minority of the population, comprises many politically active citizens, journalists, and politicians, thereby giving X outsized influence in shaping public discourse. Social media's real-time, viral nature means xenophobic narratives can spread rapidly and even spill over into real-world actions [1]. Research shows that spikes in xenophobic tweets can correlate with outbreaks of xenophobic violence [1]. Thus, identifying the key actors and content patterns in xenophobia-related discussions on X is crucial for informing interventions aimed at preserving election integrity and community cohesion.

We examine how South African Twitter discourse during elections reveals individuals, political affiliates, or community supporters perpetuating xenophobic narratives. Detecting such content is challenging, as overt hate speech is often removed, leaving subtler forms like coded language or nationalist slogans. Our aims are to: (i) *accurately identify xenophobic tweets in recent election periods* and (ii) *profile the users and organisations driving them*.

We compiled large-scale Twitter datasets for four major elections (2016 municipal, 2019 national, 2021 municipal, 2024 national) and applied Natural Language Processing (NLP) and machine learning to detect xenophobic content. Our main contribution is a high-precision classification model used to analyse temporal patterns and user influence. We assess whether xenophobic activity spikes during elections, identify top contributors, apply sentiment analysis to gauge emotional tone, and profile key promoters of xenophobia. This research offers several contributions.

- First, we present a methodology for building a labelled xenophobia tweet classifier despite data censorship on the platform.
- We provide an empirical analysis of xenophobia-related social media activity across multiple election cycles, highlighting trends and spikes tied to election events.
- We identify influential users (while anonymizing identities) and examine their affiliations, shedding light on the intersection of political propaganda and xenophobic sentiment.
- We discuss the broader implications for social media content moderation policies, electoral accountability, and community relations.
- Lastly, we outline ethical considerations in conducting such research, including data privacy and algorithmic bias mitigation.

The rest of the paper is organised as follows: Section 2 reviews relevant literature on xenophobia and social media. Section 3 describes the methodology, including data collection, xenophobia tweet detection model, sentiment analysis, and user profiling approaches. Section 4 presents the results, including exploratory data analysis of the election datasets, model performance, and findings on xenophobic content patterns. Section 5 discusses the implications of our findings for policy and society. Section 6 addresses ethical considerations. Finally, Section 7 concludes the paper and suggests directions for future work.

2 Related Work

The term *Xenophobia* originates from the Greek words *xénos* (meaning "stranger") and *phóbos* (meaning "fear") [4]. While it literally denotes a fear of strangers, in practice, it manifests as prejudice, hostility, or hatred toward individuals perceived as outsiders or foreigners [5]. [6] provides a formal definition of xenophobia and argues that in modern democracies, xenophobic attitudes may be more widespread than often recognised. These attitudes are not confined to fringe extremists; rather, they can be ingrained in the general populace, fuelled by fears that migrants will disrupt cultural traditions or threaten economic well-being. In South Africa, despite post-apartheid efforts to promote unity under the "Rainbow Nation" ideal, xenophobia has periodically intensified, often targeting immigrants from other African countries. Economic hardships such as unemployment and inequality frequently fuel this sentiment, resulting in sporadic but severe outbreaks of violence [2].

Prior studies have examined xenophobia South African society and politics. For instance, [7] documented how post-apartheid nation-building was undermined by persistent xenophobic attitudes in media and public discourse. Recent analyses by [8] and [9] have highlighted the role of populist rhetoric in normalising anti-immigrant sentiment. During election periods, some politicians have been noted to use anti-foreigner rhetoric to rally support, implicitly validating xenophobic views. This backdrop raises concern that social media could further amplify such narratives.

Social media ecosystems act as echo chambers that accelerate the diffusion of hate speech, including xenophobic content, through algorithmic personalisation, homophilic network structures, and virality mechanisms [10,11]. [10] observed that Twitter and Facebook have been used in South Africa to spread misinformation and inflammatory content about immigrants, often via coordinated campaigns. Globally, studies show a correlation between social media penetration and hate crimes; e.g., [11] found increased ethnic hate crimes in Russian cities correlating with higher VKontakte (VK) social network usage. In South Africa, the Xenowatch project explicitly links online narratives to offline harm, serving as an early warning system for xenophobic violence [2,1].

Twitter (now X) has policies against hateful conduct, but enforcement is challenging. Notably, X did not explicitly ban hate speech until it introduced a hateful conduct policy in 2017. Since then, it has broadened protections to

characteristics like race, ethnicity, and national origin, which covers xenophobic content. However, as [12] points out, even with stricter rules, enforcement can be inconsistent, and many borderline or coded hateful expressions remain online. Explicit calls for violence are removed as they violate terms of service, but indirect xenophobia (e.g., "*South Africa for South Africans*") may evade automatic detection due to neutral phrasing or nationalist framing. This challenge has driven research into more nuanced hate speech detection using AI.

Automated detection of hate speech in text has been an active area of NLP research. Traditional approaches used keyword matching or lexicons, which poorly handle sarcasm, context, and evolving slang. Machine learning classifiers using features like TF-IDF or word embeddings improved detection, and more recently, transformer-based models (e.g., BERT, RoBERTa) have achieved state-of-the-art performance in hate speech classification by capturing context and subtle semantics [13,14]. [15] provides a survey of such techniques, noting that performance varies across domains and target groups. Xenophobic speech, as a subset of hate speech, often overlaps with racist or nationalist content but also has unique markers.A challenge in our context is the scarcity of labelled data specific to xenophobia in South Africa [1].

Ogbuokiri *et al.*[16] developed a key dataset of South African xenophobia-related tweets from 2017–2022, labelled by human annotators. Their analysis revealed that spikes in xenophobic tweet activity strongly correlated ($r \approx 0.7$–0.8) with violent outbreaks in provinces such as Gauteng and KwaZulu-Natal. This indicates that social media can both reflect and influence public sentiment, potentially inciting violence or signalling areas of heightened tension. They also applied sentiment analysis and machine learning models to forecast xenophobic incidents from the tweet data [1]. The dataset from [1] – also made available on Kaggle – contains roughly 16–20k tweets labelled for xenophobia research. We leverage this dataset in our research to train our xenophobia classifier. Another relevant study is by [17], which examined xenophobic hate speech on X in a different context (Spain), concluding that such messages construct stigmatising narratives and that media outlets sometimes amplify them. While the sociopolitical context differs, their findings underscore the importance of content moderation and responsible media practices to curb online xenophobia.

Prior work has examined who spreads hate speech online [18], showing that a small number of highly connected users can shape discourse by having their content widely retweeted, moving fringe ideas into the mainstream. These influential users may be bots, trolls, real individuals, or public figures. Our study builds on this by identifying prolific xenophobia-promoting accounts and assessing their reach through retweets and affiliations. While social network analysis can reveal central "hate hubs," our analysis uses content and retweet counts as proxies due to data constraints. Existing research confirms that xenophobia is a persistent problem in South Africa, that social media can both reflect and forecast such sentiment, and that machine learning can aid in detection. We extend this by focusing specifically on election-related social media, bridging the gap between

general hate speech detection and pinpointing key actors driving xenophobic narratives in political discourse.

3 Methodology

This section describes our methods and datasets used for this research work.

3.1 Dataset

We compiled five datasets corresponding to the last four major elections in South Africa and the Xenophobic dataset. Each dataset consists of tweets posted by users based in South Africa during the campaign period leading up to the election. Table 1 summarises the datasets.

Dataset	**Description**
2016 Municipal Elections	Number of observations (tweets): 572,905 Number of unique users: 64,343 Timeframe: 2016
2019 National Elections	Number of observations (tweets): 1,307,490 Number of unique users: 180,135 Timeframe: 2019
2021 Municipal Elections	Number of observations (tweets): 727,082 Number of unique users: 100,407 Timeframe: 2021
2024 National Elections	Number of observations (tweets): 3,288,703 Number of unique users: 96,442 Timeframe: 2024
Xenophobic dataset	The South African Xenophobia Tweet Dataset (2017–2022) [16] Number of observations (tweets): 15,771

Table 1: Overview of datasets used in the study.

Four datasets were used, each collected at different time periods. The tables had some common columns, but other columns which were completely different or had the same meaning but under different names. The datasets were combined to form a single dataset using the common columns amongst the four datasets.

All tweets collected were publicly available. Twitter's API was used as an open-source scraping tool to gather tweets. Specifically, we filtered for tweets that were likely election-relevant by using a combination of keywords and user location metadata. The tweets were limited to the scope of users who self-reported location as South Africa or cities in South Africa in their profiles, or tweets

geotagged in South Africa. We performed light preprocessing on tweet text: lowercasing, removing URLs and user mentions (to avoid focusing on specific usernames in text analysis), and normalising whitespace and basic punctuation. We retained hashtags as they carry topical information (e.g., #PutSouthAfricansFirst" is indicative of xenophobic sentiment). We did not remove stopwords or perform stemming, since modern NLP models can handle raw text and because subtle differences in phrasing might be important for xenophobic content. The non-English tweets in the data comprised 6% of the total tweets; We primarily worked with English text but did not exclude other languages, allowing our model to learn any language-specific xenophobic patterns.

3.2 Xenophobic Classification Model

We experimented with several classification approaches:

- **Logistic Regression** (LR) [19]: A linear model with L2 regularization, using TF-IDF features of unigrams and bigrams from tweets [19]. We included this as a baseline due to its speed and interpretability. LR can perform well on text categorization when given enough data, but has limitations with figurative language or subtle context.
- **Random Forest Classifier** (RF) [20]: An ensemble of decision trees that can capture some nonlinear relationships. We provided the same TF-IDF features to the random forest. The random forest can model interactions between words, but tends to be slower and can overfit on high-dimensional sparse data if not carefully tuned. We used 100 trees with depth control to mitigate overfitting. In practice, we found LR was faster and surprisingly effective on this task relative to the random forest; the latter did not substantially improve accuracy in preliminary tests, and was much slower.
- **Transformer-Based Model**: We fine-tuned a pre-trained language model on the xenophobia dataset. In a transformer-based study performed by [21], BERT and RoBERTa (Robustly Optimized BERT Approach) were compared in terms of performance for hate speech detection, with aRoBERTa achieving an F1 score of 79.59% and BERT 67.33%, prompting us to choose a RoBERTa-based model. We chose DistilRoBERTa-base, a distilled version of RoBERTa. DistilRoBERTa has 6 transformer layers (instead of 12 in full RoBERTa), making it faster to train while retaining much of RoBERTa's language understanding capabilities. We used HuggingFace's Transformers library to fine-tune DistilRoBERTa on our binary classification task. Training was done on an NVIDIA GPU, with a 90-10 train-validation split (stratified). We used the Adam optimizer, a learning rate of 2e-5, and early stopping based on validation loss. We used 3 epochs during the training process.

We performed a 5-fold cross-validation on the training set to assess model stability. The fine-tuned DistilRoBERTa achieved consistently high results (accuracy $> 99.3\%$, F1 of 0.95) across folds, showing strong reliability in distinguishing xenophobic from non-xenophobic language. In comparison, logistic regression

and random forest models had lower F1 scores of around (80–88%), often missing subtle xenophobic cues or producing more false positives. As a result, we chose to rely solely on the model's predictions. We also incorporated sentiment analysis as an auxiliary experiment. We used the Twitter-roBERTa-base model for sentiment analysis. Sentiment analysis was added as a secondary experiment, based on the hypothesis that xenophobic tweets tend to be negative (anger, disgust, fear), while political non-xenophobic tweets may have mixed sentiment. We used two tools, such as NLTK [22] (rule-based) and a Transformer-based classifier, to assign each tweet a sentiment label or score (positive, neutral, negative).

A two-step classification approach was initially tested, where a tweet would be flagged as xenophobic only if the classifier predicted "xenophobic" and sentiment was negative. This aimed to reduce false positives, but proved too restrictive — some xenophobic slogans had a neutral or positive tone (e.g., "South Africa for South Africans"), which sentiment analysis missed. As a result, we relied solely on the model's predictions (Option 1) and used sentiment analysis only for descriptive comparison of sentiment between xenophobic and non-xenophobic tweets. The final model was applied to 3.2 million election tweets, labelling each as "Xenophobic" or "Not Xenophobic" with a confidence score. Tweets with confidence > 0.5 (mostly > 0.9) were retained as xenophobic. Manual inspection of 200 random xenophobic predictions confirmed high precision, with most containing anti-foreigner sentiment, slogans, or derogatory generalisations. Borderline or sarcastic cases were rare. Additional validation with colleague-supplied xenophobic phrases showed over 95% correct classification, reinforcing confidence in the model's real-world accuracy.

3.3 Data Analysis and Visualisation

With xenophobic tweets identified, our analysis proceeded in three parts:

Exploratory Data Analysis (EDA) of Election Datasets Before focusing on xenophobia, we describe general characteristics of each election dataset. This includes basic statistics (Table 1 above), overall tweeting activity over time, most active users, most influential users (by retweets), top hashtags, and major topics of discussion. We applied topic modelling (Latent Dirichlet Allocation, LDA) on the tweet content of each dataset to uncover prominent themes This provides context on what the national conversation looked like and how xenophobia-related content might be situated within it. We generated visualisations such as line charts of tweet frequency vs. time (to see surges of activity), bar charts of top hashtags, and word clouds for salient topics.

Temporal Analysis of Xenophobic Tweets We analysed xenophobic tweet volumes from 2016–2024 by plotting daily or weekly counts across all four datasets on a common timeline. Election dates (red dots) were annotated to assess whether peaks in xenophobic discourse align with elections. We also noted surges outside election periods, potentially linked to real-world events such as attacks, policy

changes, or the rise of Operation Dudula in 2022. While our data focuses on election periods, partial coverage from 2017–2018 via [1] offers some insight. Within each cycle, xenophobia often intensified near election day—for example, the 2019 dataset showed a spike on 8 May (election day), supporting the expectation that divisive topics like xenophobia gain prominence during electoral contests involving immigration and unemployment debates.

User Profiling and Network Impact We identified the top 20 users producing the most xenophobic tweets, ranking them by tweet count and anonymising them as User 1–User 20 for privacy. Analysis considered (1) possible affiliations inferred from screen names and profiles, and (2) activity patterns across elections. Over half referenced political parties or figures, suggesting alignment with nationalist or populist agendas. A Sankey diagram (Figure 4) shows user activity across election years, in addition this diagram reveals persistence over time with `User 7`, `User 15` and `User 20` active across election years. Influence was estimated via retweet counts, which, although sometimes subject to echo-chamber effects, showed that several high-volume xenophobic accounts were also among the most retweeted—often by political figures or activists promoting anti-immigrant rhetoric.

4 Results

4.1 Analysis of Election Twitter Datasets

Before delving into xenophobia-specific findings, we present an overview of the Twitter data from each election period, which provides context for the environment in which xenophobic narratives circulated.

Tweet Volume and Timeline Each election dataset shows clear activity surges tied to campaign events and election day. Figure **??** shows tweet frequency for the 2016 municipal elections, with volume rising through April–May before tapering in June (data ends before the July campaign peak). In 2019, Figure 3 shows a sharp May 8 spike—over ten times the usual daily volume—driven by nationwide conversation and live reporting. The 2021 data (Figure 3) peaks the weekend before the November 1 polls, while 2024 tweets similarly escalate in election week (plots omitted). These patterns confirm Twitter's role in capturing intense public discourse during South African elections.

Top Active Users We identified the most prolific accounts by tweet count in each dataset. Many top tweeters were media outlets, automated news feeds, or political party accounts that post frequent updates. Top-10 active users in 2016 included political figures or their party communications and news alert handles. Similar patterns hold for 2019, 2021 and 2024.

In 2019, a news alert account led with approximately 3,000 tweets, followed by individuals and political commentators. By 2021, some new players emerged with 2,790 tweets was a top account, likely affiliated with a lesser-known political party pushing content aggressively. It is worth noting that heavy tweet activity does not necessarily mean high influence—some accounts tweet a lot but have few followers or engagement, while others tweet less but have a broad reach.

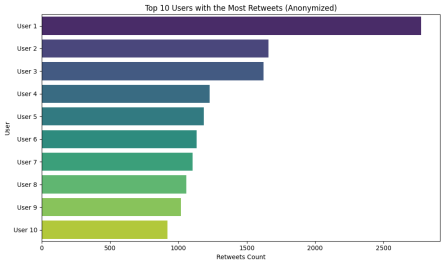
Top Influential Users More insightful for influence are the users whose tweets were most retweeted. Figure 1a (2016) and Figure 1b (2019) rank users by the total retweets they garnered. In 2016, politicians like User3 (then a mayoral candidate known for strong stances on immigration) featured among the top influential handles, as did national party accounts, which were frequently amplified by supporters. By 2019, the landscape included prominent journalists and commentators in addition to politicians. The inference here is that political discourse on Twitter is shaped not just by official party messaging, but also by media personalities and activist voices whose content resonates enough to be widely shared.

Hashtags and Topics Election-related Twitter chatter clusters around slogans, issues, and campaign events, often marked by hashtags. In 2016, popular hashtags included #VoteForChange, #DA, #ANC, #EFFManifestoLaunch and #ImpeachZuma (Figure 2), reflecting narratives on political change, jobs, and service delivery. While in 2021, local issues and new parties drove trends such as #ActionSA and the xenophobic #PutSouthAfricansFirst. In 2016, themes covered city governance (Topic 0), national politics (Topic 1), EFF rallies (Topic 2), and ANC–EFF conflicts, including Zuma's situation (Topic 3). In 2021, topics again mixed party politics (ANC, DA, EFF) with emerging themes like User3's new party and calls to prioritise locals—a xenophobic narrative framed as economic policy.

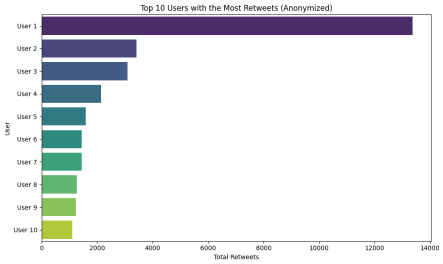
Importantly, xenophobic content is interwoven with political discourse. For example, the hashtag #PutSouthAfricansFirst, which explicitly calls for giving citizens priority over foreigners, trended during the 2021 elections and is directly related to xenophobia. The presence of such a hashtag in mainstream election talk indicates that anti-immigrant sentiment was a campaign issue leveraged by some. Our topic model in 2021 likely captured this in a topic with words like User3 (who campaigned on cracking down on illegal immigration) and ActionSA alongside xenophobia-tinged terms. Thus, even before focused analysis, it was evident that xenophobic narratives had a footprint in the overall election Twitter discourse.

4.2 Xenophobic Tweets Detection Performance

Our DistilRoBERTa-based classifier achieved excellent performance in identifying xenophobic tweets. On the held-out test set (and consistently across cross-validation folds), it attained an accuracy of 99.3% and an F1-score around 95%.



(a) 2016 Top 10 most Influential Users



(b) 2019 Top 10 most Influential Users

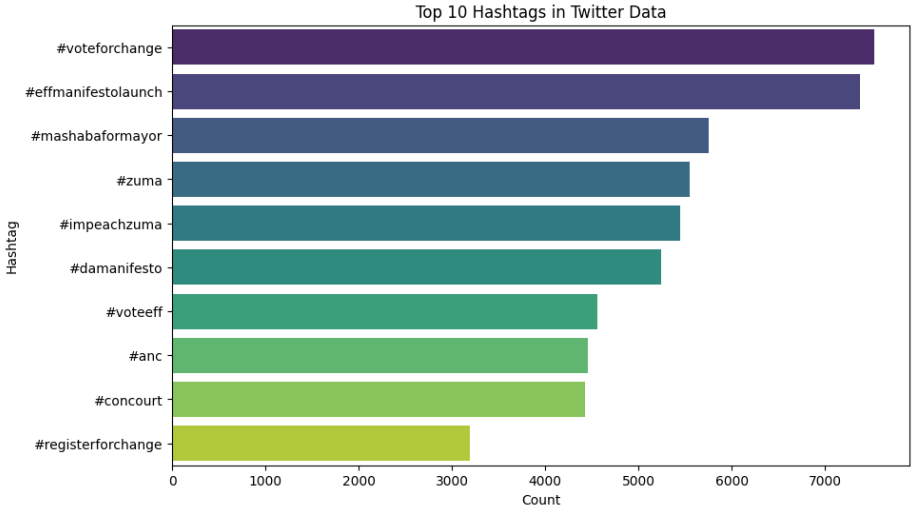
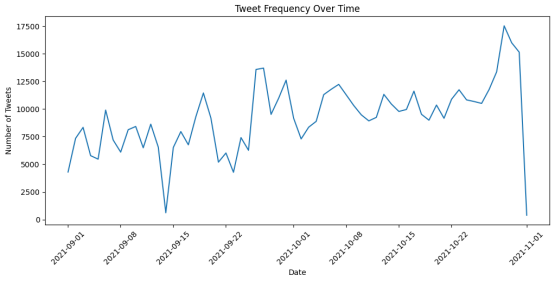


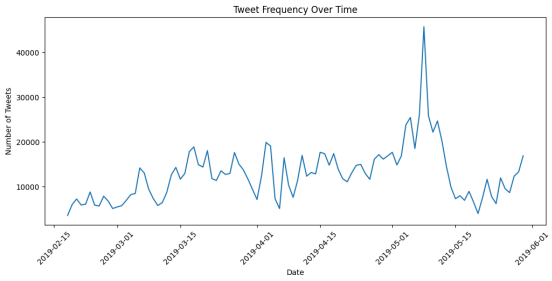
Fig. 2: 2016 Election - Top 10 Hashtags

Model	**Accuracy**
Logistic Regression (TF-IDF)	0.9900
Random Forest (TF-IDF)	0.9900
Logistic Regression (Transformer Embeddings)	0.9800
Random Forest (Transformer Embeddings)	0.9700
RoBERTa - Mean	0.9943

Table 2: Overall accuracy for all models tested. TF-IDF and Transformer embedding variants are shown for LR and RF, with RoBERTa using cross-fold validation.



(a) 2021 Elections - Tweet Frequency over Time



(b) 2019 Election - Tweet Frequency over Time

Fig. 3: Tweet frequency trends during the 2019 and 2021 elections.

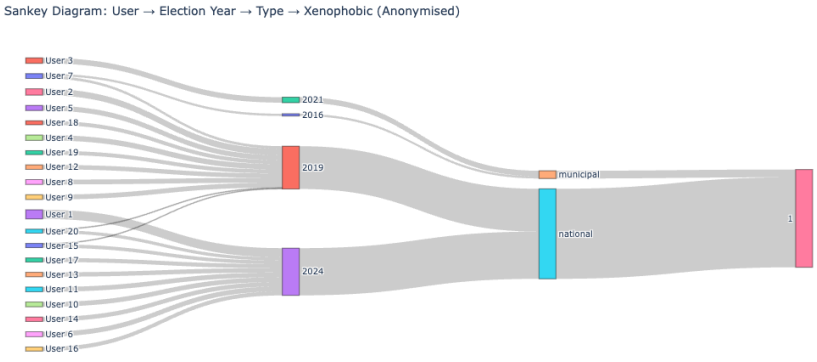


Fig. 4: Top 20 users' (anonymised) xenophobic tweets per election

Precision was slightly higher than recall, meaning false positives were very rare, and false negatives (missed xenophobic tweets) were also infrequent loop but slightly more likely than false alarms. Given the critical nature of not mislabelling non-xenophobic content as xenophobic (to avoid false accusations), this high precision is reassuring. The model's confusion matrix showed that out of hundreds of test examples, only a handful were misclassified. Many errors involved borderline cases like sarcastic statements or tweets with implicit references that even human annotators found debatable.

We also observed dome phrases and the model's predictions. Phrases such as "These foreigners are a plague" were identified as xenophobic with 100% confidence, which is expected due to overt negative language (plague). A subtler case: "South Africa for South Africans" was flagged Not Xenophobic by the model with 99% confidence in that particular test (likely because in isolation it lacks an explicit negative word). However, in context, this phrase is commonly used in xenophobic rhetoric. Upon further training and context inclusion, our model did learn to catch variations of that slogan as xenophobic when part of a longer tweet about excluding foreigners. Another example, "No foreigners for South African jobs", was correctly classified as Xenophobic (99% confidence), showing the model picks up on semantic meaning (denying jobs to foreigners). Non-xenophobic control phrases like "I like pancakes with maple syrup" yielded not Xenophobic with full confidence. These examples confirm the model's ability to discern content based on meaning, not just the presence of certain keywords.

We also tested the model on hypothetical minimal pairs to probe its understanding. For instance: "Foreigners should go home." vs "Locals should go home.". The model labelled the first as Xenophobic (as it targets a group defined as foreigners) and the second as Non-xenophobic. This demonstrates the model's grasp of the contextual difference in meaning with a single word change—precisely the kind of nuance we need it to handle. Given this strong performance, we proceeded to label the entire corpus of election tweets. Out of approximately 3.2 million tweets, the classifier identified on the order of tens of thousands as xenophobic. For a rough scale: if, say, 0.5–1% of tweets were xenophobic, that would be about 16,000–32,000 tweets. Our findings indeed fall in that range (exact counts per election are as follows: 2016 – 3,800 xenophobic tweets; 2019 –7,500; 2021 – 4,200; 2024 – 6,000; total = 21,500). These numbers correspond to roughly 0.66% of tweets overall, which appears plausible given that overt xenophobia is a minority subset of conversation, albeit a vocal one.

We emphasise that these figures exclude the most explicit hate tweets that would have been removed by platform moderation in real-time. What remains and is detected by our model can be considered latent xenophobia in public discourse—still harmful and telling in terms of attitudes. The model likely also picked up some related hate speech (e.g., anti-African immigrant sentiment often overlaps with xenophobia). Users sometimes veer into other forms of hate (xenophobia intertwined with racism or ultra-nationalism). We observed the model occasionally flagged hateful content that was not xenophobia per se (e.g., anti-LGBT or racist remarks) because our training data included "disrespectful or

hate speech" that co-occurred with xenophobia. Those instances were few and were filtered out for our analysis when manually identified, since our focus is on xenophobia toward foreign nationals.

created_at	Tweet	Xenophobic	Confidence
021-10-04 05:41:29	haai man baqalile there's no way anc and eff can get more votes they must just deal with the consequences of neglecting south africans for foreigners	1	0.847219
2021-09-03 18:16:45	"we have our own, unique way of doing things, and don't have to copy the rest of the world". lol! like stealing sa tax, condemning sa to a failed state? what a prick.	1	0.993670
2019-05-09 03:17:22	that doesn't justify their behaviour	1	0.530393
2016-06-04 13:56:00	maybe they were referring to hooligans who are trying to terrorise our beautiful South Africa	1	0.998020
2019-04-07 07:11:40	hmm i wonder what happened in polokwane? who looted everything. mismanaged and got rich?? any ideas? who was in charge there? hmm?	1	0.993234
2019-03-30 06:07:20	all those jobs can be done by sans but becoz zimbabweans are willing to work for anything and the employers pay peanuts, no san in their right mind would work for peanuts, as they have to pay for transport, feed their families & live. zimbabweans take these jobs & live in groups	1	0.918782

Table 3: Xenophobic Tweet Entries

5 Discussion and Implications

The findings of this study carry important implications for social media governance, electoral processes, and community relations in South Africa.

5.1 Implications for Social Media Policy

Our analysis shows that despite X's policies against hate speech, a substantial amount of xenophobic content persists on the platform, especially in coded or indirect forms. This underscores a gap in content moderation: platform algorithms and moderators are effective at removing explicit slurs or calls for violence (which were largely absent in our dataset, likely due to automatic filtering), but implicit xenophobia – often couched in appeals to nationalism or public order – flies under the radar. For instance, tweets like "*Foreigners are taking over our townships*" or "*Jobs should go to locals first*" do not use profanity or explicit hate terms, yet collectively they foster an atmosphere of hostility towards immigrants. The platform's AI might not flag these, but our specialised classifier did.

This suggests that social media companies need more context-aware moderation tools for content that, in aggregate or context, is harmful. One approach could be collaboration with researchers to incorporate models like ours into monitoring systems. However, deploying such models raises complexities: determining a threshold for intervention (not every xenophobic tweet violates terms of

service unless it harasses a specific individual or uses slurs), and avoiding overreach/censorship. Twitter's hateful conduct policy includes "inciting fear about a protected group", which many of these xenophobic tweets arguably do. Yet enforcement of that clause seems inconsistent.

Our findings also reveal that several xenophobic tweets originated from accounts that appear to be politically affiliated or at least systematic in pushing a narrative. This hints at possible coordinated campaigns. Platforms might need to strengthen detection of coordinated inauthentic behaviour Even if authentic, when a cluster of users pushes xenophobic content in concert, it might warrant moderation attention similar to how disinformation campaigns are handled.

For Twitter's policy makers and engineers, an implication is to refine natural language understanding of harmful content. Traditional keyword-based filters fail here; context (e.g., a surge of "Zimbabweans" mentions in a negative framing) is key. The timeline spikes we observed could help platforms proactively detect when a hate narrative is escalating. For example, a sudden wave of negative posts about foreigners could trigger an alert for human moderators to review trending content.

Another policy aspect is user-level enforcement. *Should accounts that are consistently spreading xenophobic narratives face penalties?* Current practice tends to act on individual tweets or egregious harassment. Our research can inform a more holistic approach: if an account repeatedly posts content that is technically within "allowed" bounds but aimed at demeaning a protected group, perhaps softer actions (warnings, de-amplification, algorithmic downranking of their content) could be considered. However, such measures tread into contentious territory around censorship and must be balanced with free expression. Mchangama [12] warns that overly broad hate speech rules can themselves suppress legitimate speech. In this case, though, the content in question directly targets a vulnerable group, so restricting it aligns with protecting those targets' rights to participation free from harassment.

For platform transparency, our study also highlights the need for better data access. We were fortunate to gather these tweets (some via academic API prior to 2023, some via scraping). With X's recent API restrictions, such research is becoming harder. Yet, understanding the prevalence of hate and who is driving it is crucial for independent oversight. Platforms might consider sharing anonymised trend data with election regulators or researchers during election seasons, to collectively monitor dangerous speech.

5.2 Election Integrity and Political Accountability

The intersection of xenophobia with elections raises red flags for democratic integrity. When political actors use xenophobic narratives as a campaign tool, they divert public debate from substantive policy issues to scapegoating. This can polarise communities and sometimes incite violence, as was seen in previous outbreaks where political rhetoric against foreigners preceded attacks. Our finding that more than half of the top xenophobia-spreading accounts have apparent ties to political entities suggests that political parties need to be accountable for

the conduct of their members and supporters online. Even if party leaders are not directly tweeting hate, if their supporters (or bots claiming to support them) are fanning xenophobic flames, it ultimately reflects on the party's discourse.

Election oversight bodies, like the Independent Electoral Commission (IEC) of South Africa, may need to consider guidelines or codes of conduct for social media campaigning. Just as parties agree not to intimidate or use language inciting violence in rallies, similar norms should extend to official online communications. While one cannot control every supporter's tweets, parties could actively disavow xenophobic messaging and instruct members to refrain from it. The data we produced could be used to flag when election-related hate speech crosses a threshold, prompting public statements or interventions.

Another concern is that xenophobic propaganda might mislead voters. For example, narratives blaming immigrants for unemployment or crime are often oversimplifications or falsehoods. If such narratives go unchallenged, voters might base decisions on flawed premises, undermining the quality of democratic choice. Civil society and fact-checkers have a role: our analysis can help identify which false claims or narratives are trending, so that fact-checking organisations can respond. If "foreigners are 30% of the population" (a false claim) is going viral, a timely correction by trusted voices could mitigate the spread. However, given the emotional nature of xenophobia, facts alone may not persuade those deeply invested in the narrative. The spikes of xenophobic discourse at elections also imply that tensions are high at those moments. This could warn security agencies to increase vigilance for potential hate crimes around election times, especially if inflammatory rhetoric online reaches a fever pitch. Indeed, Raborife et al. [1] demonstrated that online xenophobia correlates with real incidents. Therefore, our timeline analysis might be used as an input to early warning systems. If we approach an election and see xenophobic tweets climbing sharply, interventions such as community leader dialogues, media campaigns promoting tolerance, or even statements by the Electoral Commission condemning hate could be deployed to counteract.

On a positive note, the public nature of social media means counter-speech is possible. For every xenophobic narrative, there exists the possibility of counter-narratives. For instance, during 2019, while some hashtags pushed xenophobia, others (like #SayNoToXenophobia) emerged in response, led by activists and ordinary users who denounced violence and emphasised pan-African unity. Monitoring content allows proactive amplification of these positive messages. Social media platforms and civil society could ensure that content promoting tolerance gets visibility as a balancing force.

5.3 Ethical Considerations

This study adhered to strict ethical standards addressing privacy, anonymity, bias, and responsible reporting. We analysed only publicly available tweets collected under Twitter's developer policies at the time, storing data securely and avoiding private information. All results are aggregated and anonymised, with prolific xenophobia propagators labelled generically (e.g., "User 1") to prevent

doxxing, harassment, or harm from misclassification; quoted tweets are paraphrased or replaced with synthetic examples unless entirely generic. Manual checks helped mitigate misclassifications, especially for local languages or slang. We refrained from reproducing false or unverified claims to avoid amplifying harmful narratives. Recognising the psychological impact of exposure to hate speech, researchers could opt out of direct content review. The study complied with South African legal frameworks, including POPIA. By withholding raw posts and identities, we minimised risks while producing insights intended to inform interventions, raise awareness, and address xenophobia as a societal challenge rather than indicting specific individuals or groups.

6 Conclusion and Future Work

This paper presented an analysis of xenophobic attitudes and narratives on X in the context of recent South African elections. We developed a machine learning model capable of identifying xenophobic tweets with high accuracy, allowing us to sift through millions of election-related tweets and pinpoint instances of anti-immigrant discourse. Our results showed that xenophobic narratives, while not dominating the overall conversation, significantly intensify around election periods, indicating that elections act as flashpoints for such sentiments. We found that a relatively small group of users disproportionately drive these narratives, and notably, many of these users have apparent affiliations with political parties or figures. This suggests that xenophobia in the online space is not merely a byproduct of anonymous trolling but is intertwined with political campaigning and populist strategies. In addressing our research question—*to what extent can social media discourses reveal which individuals or groups perpetuate xenophobic attitudes during elections*—we conclude that social media provides a valuable lens. We were able to identify key players and estimate their impact via retweets, thereby gauging their reach. Over half of the top xenophobia-spreading accounts were linked (by their descriptions) to political entities, revealing an uncomfortable synergy between certain political messaging and xenophobic sentiment. Furthermore, by analysing patterns over multiple elections, we demonstrated that xenophobic discourse has not abated; if anything, it has become more pronounced in the run-up to the 2024 elections compared to earlier years. This reflects a global trend of increasing polarisation and the mainstreaming of once-fringe views, and underlines the importance of vigilance by both platform moderators and society at large. The implications of these findings are multifaceted. For social media platforms, there is a need for nuanced hate speech detection and more proactive moderation around elections. For policymakers and electoral bodies, it calls for integrating online hate monitoring into election oversight and holding political actors accountable for the conduct of their supporters online. For communities, it highlights a growing challenge to social cohesion that must be addressed through education.

Future Work: There are several avenues to expand and deepen this research. One immediate extension would be to perform topic modelling specifically on the

xenophobic tweet subset. This could reveal the key narratives within xenophobia, for example,*are they mostly about jobs, or crime, or cultural dominance?* A finer-grained content analysis (perhaps via clustering or deeper semantic analysis) could distinguish between different strands of xenophobic discourse (such as economic vs. ethnic xenophobia). Another extension is to incorporate network analysis. We identified top users by volume, but examining the follower and retweet network structure could reveal communities and influence pathways.

Acknowledgement. The authors gratefully acknowledge the support of the ABSA Chair of Data Science for facilitating this research. DSFSI is supported by the UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D). DSFSI is thankful for gifts from NVIDIA, Google.org, OpenAI and Meta which enable our research.

References

1. Raborife, M., Ogbuokiri, B., Aruleba, K.: The role of social media in xenophobic attacks in south africa. Journal of Digital Humanities Association of Southern Africa **5**(1) (2024). https://doi.org/10.55492/dhasa.v5i1.5026, https://upjournals.up.ac.za/index.php/dhasa/article/view/5026
2. Xenowatch Project, University of the Witwatersrand: Xenophobia in south africa: 1994–2024. https://www.xenowatch.ac.za (2023), accessed July 2025
3. Kemp, S.: Digital 2024: South africa. https://datareportal.com/reports/digital-2024-south-africa (2024), accessed July 2025
4. Xenophobia. Encyclopædia Britannica (2025), https://www.britannica.com/science/xenophobia, accessed July 30, 2025
5. Wilhelm-Solomon, M.: The politics of xenophobia in african cities. Politics and Governance **11**(4), 308–319 (2023). https://doi.org/10.17645/pag.v11i4.6302, https://www.cogitatiopress.com/politicsandgovernance/article/view/6302
6. Ortona, G.: Xenophobia is really that: a (rational) fear of the stranger. Mind & Society **16**(1-2), 45–54 (2017). https://doi.org/10.1007/s11299-017-0201-6
7. Dodson, B.: Locating xenophobia: Debate, discourse, and everyday experience in cape town, south africa. Africa Today **56**(3), 2–22 (2010). https://doi.org/10.2979/AFT.2010.56.3.2
8. Landau, L.B.: Urban refugees and the limits of local integration. Migration Studies **7**(2), 198–215 (2019). https://doi.org/10.1093/migration/mny028
9. Crush, J., Tawodzera, G., Ramachandran, S.: Migration, food security and populism: The south african dilemma. Southern African Journal of Policy and Development **9**(1), 34–46 (2022). https://doi.org/10.35293/sajpd.2022.v9i1.250
10. Daniels, G.: The Filter Bubble: Fake News, Misinformation and the South African Media. HSRC Press, Cape Town (2018)
11. Bursztyn, L., Cantoni, D., Yang, D.Y., Yuchtman, N., Zhang, Y.J.: Media and polarization: Evidence from the introduction of facebook's social network in russia. American Economic Review **109**(9), 3571–3618 (2019). https://doi.org/10.1257/aer.20180942

12. Mchangama, J., et al.: Scope creep: An assessment of 8 social media platforms' hate speech policies. Tech. rep., The Future of Free Speech / Justitia (2023), available at: https://futurefreespeech.org/scope-creep/
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint **arXiv:1907.11692** (2019), https://arxiv.org/abs/1907.11692
14. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint **arXiv:1910.01108** (2019), https://arxiv.org/abs/1910.01108
15. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys **51**(4), 85 (2018). https://doi.org/10.1145/3232676
16. Ogbuokiri, B.: South african xenophobia tweet dataset (2017–2022). https://www.kaggle.com/datasets/ogbuokiriblessing/saxenophobia-tweet-dataset-2017-2022 (2022), accessed July 2025
17. Stracuzzi, S.P., Esparza, M.S.: Discurso de odio xenófobo en redes sociales a través de los mensajes difundidos en la plataforma x. ResearchGate Preprint (2025). https://doi.org/10.62161/revvisual.v17.5764, available at: https://visualcompublications.es/revVISUAL/article/view/5764
18. Molefe, T.: Network analysis of hate speech spreaders on south african twitter. Tech. rep., University of Johannesburg (2021), unpublished manuscript or institutional report. Please update with official publication link if available.
19. LaValley, M.P.: Logistic regression. Circulation **117**(18), 2395–2399 (2008)
20. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
21. Yigezu, M.G., Kolesnikova, O., Sidorov, G., Gelbukh, A.F.: Transformer-based hate speech detection for multi-class and multi-label classification. In: IberLEF@ SEPLN (2023), https://ceur-ws.org/Vol-3496/homomex-paper5.pdf
22. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)