INSTITUTE FOR SECURITY STUDIES

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

South African
CRIME QUARTERLY
No. 73 | 2024

Protest events have a 'Twitter signature'

Evidence from South Africa's #FeesMustFall

Hraklis Papageorgiou, Joseph Baggott and Martin Bekker[1]

papageorgiou.hraklis7@gmail.com
joebaggott@gmail.com
martin.bekker@wits.ac.za

https://doi.org/10.17159/2413-3108/2024/vn73a 16815

Protest levels in South Africa remain notoriously high. However, Protest Event Analysis (PEA) within the South African context has predominantly relied on media or administrative data, with scant regard for the insights that social media records may provide. Using data gathered during the #FeesMustFall movement, we employ a machine learning model trained on historical South African Twitter data combined with event records from a protest database. Our findings indicate that such events establish a distinctive time-series pattern, suggesting a robust approach to modelling, identifying or predicting protests within South Africa. Moreover, this implies that protests can be characterised using social media sources in ways that can enhance the insights from traditional data sources, and that automated analysis can unearth and assemble PEA databases. Moreover, automated PEA may soon provide valuable information about the actors, level of turmoil, size, duration, grievances, and motivations of protest.

Protest has long been a significant aspect of South African society, both during and after the struggle against Apartheid. While South Africa's transition to democracy heralded as the road to 'a better life for all', the country continues to experience numerous community and labour-related protests daily, by some claims more than any other

country.[2] Contemporary protests tend to raise grievances associated with the neoliberal dispensation of the post-Apartheid state.[3] These protests are geographically diffuse[4] and despite their ubiquity and resolve, have thus far generally failed to link and coalesce into broad, sustained, emancipatory movements.[5]

Protests in South Africa represent a censure of the state, and arise from several factors. While these factors have not been formally modelled, they appear to include a community having crossed some threshold of grievances, alongside a community having access to sufficient organisational capital, and political opportunity. South African protests cover a broad spectrum of concerns, including labour disputes, community service delivery, and education-related grievances. The 'driving factors' behind protest appear to be economic (e.g., unemployment), social (e.g., inequality), political (e.g., electoral participation), and demographic (e.g., the dependency ratio).[6]

Globally, the frequency and volume of protests appear to be rising.[7] South Africa, despite its already extremely high levels of protest, has not bucked this trend. Estimates suggest that the number of South African protest events is rising annually, with the country exceeding an average of 14 protest events per day.[8]

Scholars of protest, including those focused on South Africa, tend to view and categorise protests into paradigmatically different 'kinds', including community protests (or service delivery protests, to evoke the South African incantation), labour protests, anti-police protests, and education-related protests, as the main kinds.[9] Notably, within the South African context, even among protests of the same kind – say community protests – these tend not to coalesce into sustained emancipatory movements but are generally disconnected events, showing little promise of bringing about real social change.[10]

There are, however, rare instances of a series of events transcending the tendency towards short-lived, disconnected protest events. One such example – one that became a widespread and sustained social movement – is the student-led #FeesMustFall (FMF) movement, regarded as part of South Africa's most significant social movements in recent times, the #MustFall movement.[11]

The #MustFall movement was a widespread phenomenon in South Africa, characterised as a 'meme event' by Frassinelli.[12] Beginning in March 2015 with the #RhodesMustFall campaign, the movement called for the removal of a centrally located statue of Cecil John Rhodes at the University of Cape Town. The term 'MustFall' was later adopted by what would become the FMF movement, which sought to reduce fee increases (initially to remove fees altogether) at public universities across the country. Other movements – even those with limited connection to the original Rhodes campaigns (and decolonisation of universities) – subsequently used the same term to gain similar levels of attention and support, giving rise to "Fallism", the collective term for such movements. Given the heightened role that various social media platforms played in assisting organisers and supporters of these events, Fallism in general, and FMF in particular, lend themselves to a Social Media Analysis of protest.

In parallel to these social developments, recent advancements in machine learning, (ML), have opened new possibilities for analysing and predicting social phenomena, including protests. By leveraging large datasets, ML algorithms can identify patterns and trends that may not be immediately apparent through traditional analytical methods. In the context of protest prediction, machine learning models can be trained on historical data to recognise the complex interplay of factors that typically

precede protest events, such as social grievances, economic stressors, or political tensions. These models can then be applied to real-time data sources, like social media, to detect early signals of unrest. For instance, spikes in specific keywords, hashtags, or changes in sentiment on platforms like Twitter (in this article, we use Twitter instead of 'X', the new, often confusing name of the platform) could serve as indicators of rising discontent. By accurately identifying these precursors, machine learning offers a powerful tool for forecasting when and where protests are likely to occur, offering a kind of civic barometer of discontent.

Given the centrality of social media to the FMF movement, this paper explores the potential of using Twitter data to analyse and predict protest events in South Africa. By utilising ML, data processing and examining the "Twitter signature" of protest events, this study aims to contribute to the growing body of research on PEA and to provide new insights into the dynamics of protest in the digital age.

Protest trend data sources

PEA is a widely used approach for studying protests and is typically based on content surveys of sources, such as newspaper reports or police records, to map, analyse, and interpret protests. Koopmans and Rucht provide a summary outline of PEA in *Methods of Social Movement Research*.[13]

News record databases

News record databases constitute the mainstay source for PEA. The value of news-based records for PEA lies in the combination of automation, which allows for the collection of large volumes of records, and careful curation, which involves cleaning and labelling data, often using human annotators. Massive news-aggregated datasets such as the Armed Conflict Location Event Database, ACLED,[14] are by far the most drawn-upon source of PEA. However, media-based datasets can be limited by the newsworthiness criteria of the publications they aggregate. They can imply an urban bias, omit small protests or favour reports on violent or spectacular events. This selective reporting has been criticised as 'riot porn,'[15] indicating that media-based datasets contain, by definition, fewer events than took place in reality, as media agencies do not report on protests deemed un-newsworthy.

Administrative data

The use of administrative data, instead of media-based data, constitutes a noteworthy alternative source for PEA. This data typically comes from governmental agencies, such as the state's security apparatus, labour departments or statistical agencies. Although this type of information is not widely disseminated due to confidentiality, it offers insights into protest events since data is collected without the constraints of newsworthiness criteria. The South African Police Service's highly-regarded Incident Registration Information System is the preeminent example here,[16] and has allowed for the development and testing of theories of protest.[17]

Social media data

Beyond media-based and administrative data, social media data represents a salient and distinct third source of protest event data. Social media platforms, particularly Twitter, provide an unfiltered, first-person perspective on protest events. Unlike traditional data sources, social media is not limited to text and may also include multimodal elements (images) and metadata (time and geolocation). Posts are often composed as real-time, participant-generated accounts of ongoing events and reveal specific grievances and on-the-ground perspectives, ostensibly providing a more comprehensive corpus of information than that

which is derived from traditional data sources.[18] However, it is important to note that not all social media content is from individuals – some posts are made by organisations (which include media outlets and government agencies), bots (automated accounts), and other entities, which can influence the data.

The use of social media data warrants broad consideration by PEA scholars in both developed and developing countries. Although South Africa boasts relatively high smartphone and internet usage, digital access is not uniformly distributed,[19] and only a fraction of the population – approximately 9.3 million, or 15% – were active Twitter users in 2019. Despite these limitations, the sheer volume and quality of the data, coupled with the youthful demographic that social media tends to attract (also reflecting the demographic of protesters in South Africa), suggest that social media data has untapped potential for analysis. Indeed, the study of protest events and of social movements more broadly teeters on the brink of a transformative moment, akin to that experienced by international relations and peace studies at the turn of the century, as large conflict datasets enabled the testing of specific theories (e.g., the Stockholm International Peace Research Institute and Polity IV databases). Social media data can be mined for sentiment analysis (where sentiment is not filtered via the writings of officials or journalists), used to track the real-time development of protest events, monitor fluctuations in public opinion from on-the-ground sources, and offer insights into the activities of agent provocateurs, all of which could revolutionise traditional PEA. Elsewhere, social media data has been used to analyse public interest in natural disaster-related events,[20] elections[21] and football matches.[22]

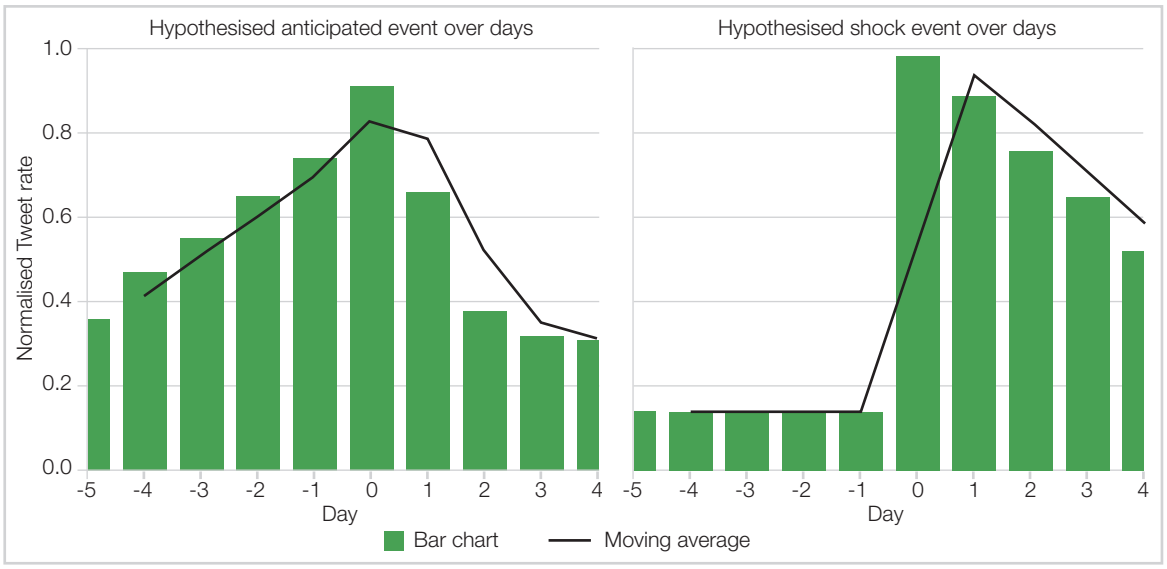
The 'Twitter signature' of an event

Foundational work has seen researchers employ various methods to analyse social media data for event detection and tracking, such as natural language processing and machine learning algorithms.[23] Together, these techniques can identify keywords and phrases related to an event, allowing researchers to track changes in the volume and sentiment of tweets relating to a specific protest event over time. The resulting graph, which shows patterns of public interest in a given protest subject over time, can be thought of as the 'shape of an event', the 'time-varying volume of subject-related tweets' or more poetically, an event's 'Twitter signature.'[24] Exploratory studies of various Twitter event signatures[25] have suggested the possibility of paradigmatically different types of events – ranging from public celebrations and sports matches to civil unrest – each exhibiting distinct patterns of public interest and engagement and that 'the shape of... [an event's] signature may convey information about the nature of [the event] itself.'[26]

Figure 1 offers a hypothetical, stylistic comparison of Twitter event signatures. One graph represents a hypothesised signature of an event with a steady build-up of interest (for example, a football match[27]), while the other graph shows a 'shock event' (for example, the death of a public figure or a natural disaster) where there would be no anticipatory build-up (apart from some noise or natural mention of a person/event). In the latter case the event occurs unexpectedly, and presents a 'tail' to the right, with interest waning over time.

These hypothetical Twitter event signatures raise the question of whether protests (or particular types of protests) exhibit identifiable signatures. If so, could these signatures be used to improve anticipation of protest events, support retroactive classification of event types, or study protest event duration? Such event anticipation, social media interest curve shape-informed classification, and real-time participant-data-based estimates of duration are underdeveloped

Figure 1: Comparison between the social media signatures of a hypothetical anticipated event and a hypothetical unanticipated shock event



applications in PEA: our technique, described below, presents one possible approach to studying these signatures.

Materials and methods

Twitter data, like other social media data, requires significant preprocessing to be rendered appropriate for PEA.[28] Given the volume of tweets – approximately 850 million per day – the first processing challenge lies in effectively isolating tweets related to specific protest events. There are several methods that can identify protest-related records, including keyword-based approaches,[29] geo-location metadata-based approaches,[30] and automated content analysis approaches.[31] By combining these approaches, using a set of protest-identifying keywords/phrases, location data, and content features (such as sentiment), we isolated tweets related to the FMF protests in 2016.

Tweet database assembly

To obtain the Twitter data, we utilised the Twitter API, which enabled us to download millions of tweets based on criteria such as dates, locations, and content (i.e., a tweet containing specified keywords or phrases). Note that the monitoring or predicting of 'sensitive events', including protests, rallies or community-organised meetings was not permitted by Twitter's terms of use at the time of data collection.[32] Instead, we focused on the FMF series of protest events some years after the fact, aiming to establish a proof of concept on historical protests.

We developed a query system using four categories of keywords to filter tweets related to the FMF protests:

1. University-related terms: e.g., campus, student, lecture
2. University abbreviations or popular names: e.g., UCT, UNISA, Wits
3. Civil unrest actions: e.g., barricade, strike, protest
4. FMF-specific terms: e.g., free education, mustfall

For a tweet to be included it had to contain both a university keyword or abbreviation and at least one civil unrest action or FMF-specific keyword.

The query system was case-insensitive and included only English-language tweets. Each tweet was accompanied by metadata such as the number of replies, likes, quotes, location, retweets (to identify duplicates), and the follower count of the tweet originator. Additionally, tweets were filtered to ensure they originated from South Africa.

Twitter provides tweet location data in the form of a boundary box (b-box), which approximates the actual location based on IP addresses. The size of the b-box varies, depending on the actual location range (smaller b-boxes in urban areas, larger b-boxes in rural areas). To reduce complexity, Twitter approximates the b-box to a single location at its centre. However, nearly 10% of SA-based tweets were tagged with only 'South Africa,' which provides insufficient location information. Tweets with insufficient geo-location information were excluded from the dataset.

Protest event calibration

To identify the dates of specific protest events, we used data from ACLED. ACLED's data is available through an API and can be filtered by location, date, event type, estimated size, actors, and keywords. South African protest data is available from 1994, with updates typically three months behind real-time. Using the API, we filtered the dataset based on protest criteria, limiting it to 2016 (the peak of the FMF movement). This yielded 147 candidate protest events, which we manually checked and reduced to 99 FMF-related events.

Data preparation

Social media data and Twitter data, in particular, are notoriously 'dirty' and difficult to 'wrangle', requiring significant preparation.[33] In contrast to conventional print media, the premise of social media is that anyone can tweet about almost anything, implying that truth and nuance are optional (as are spelling and grammar rules), and the fabrication of facts is common. This consideration implies that cleaning social media data is more arduous than cleaning administrative or news-media data, where a degree of editorial scrutiny casts a conforming lens, and outlets tend to prize their reputations for being truthful (and stylistically consistent).

Tweet preparation involved several standard data cleaning techniques, including:

1. Removing ambiguous and erroneous data: parameter manipulation (e.g., location) and removing unrelated posts (e.g., advertisements).
2. Deriving data: additional fields, such as tweet rate and sentiment analysis, were derived from existing metadata.
3. Feature creation: this included estimating tweet locations based on metadata, adjusting locations to better match protest events (especially where metadata and hashtags were misaligned), and conducting content and sentiment analysis to standardise the intent of each tweet.

While most tweets were correctly assigned using the automated b-box allotment, FMF-related protests were sometimes tweeted about from distant locations (e.g., solidarity from students at a university other than where the protest occurred). We assumed that tweets related to a specific protest event even if originating from elsewhere in the country, should be reassigned to the referenced university (since FMF protests were, for the most part, campus-based). For instance, a tweet from Johannesburg mentioning protests at The University of Cape Town was reassigned to the geographical coordinates of The University of Cape Town.

Content and sentiment analysis

Content analysis involves a qualitative breakdown of text to estimate meaning.

To uncover the context of each tweet, we employed a modified content vocabulary adapted from Bekker.[34] Bekker developed a vocabulary of keywords and phrases designed to capture different aspects of protests – for example, tactics used in a protest or the grievance of the protest. Building on this foundation, we tailored the vocabulary to align with the specific nuances of our dataset, enabling a focused exploration of FMF protests. We searched each tweet for keywords or phrases indicative of:

- Categories (8): grievances, triggers, tactics, actors, locations, weapons, eventualities, curiosities, and non-protests
- Sub-Categories (92): e.g., 'arrest' under 'triggers or 'crowd projectiles' under 'weapons'

Some subcategories required multiple or conditional keyword matching (distinguishing between 'firing a gun' and 'firing an employee'), while others only needed a single keyword. The content analysis also identified whether a tweet was associated with a protest by using the non-protest category and specific rules (e.g., dropping tweets belonging to the non-protest category and lacking a grievance, trigger, or tactic).

Sentiment analysis is the process of using natural language processing and machine learning techniques to identify subjective information in text. Through tokenisation (breaking sentences in words, and words into lemmas) and feature extraction (identifying emotive words or combination of words), and then applying a classifier (awarding the content an emotive score), we estimated each tweet's sentiment with a score from -1 to 1, ranging from negative to positive sentiment respectively.

Machine learning

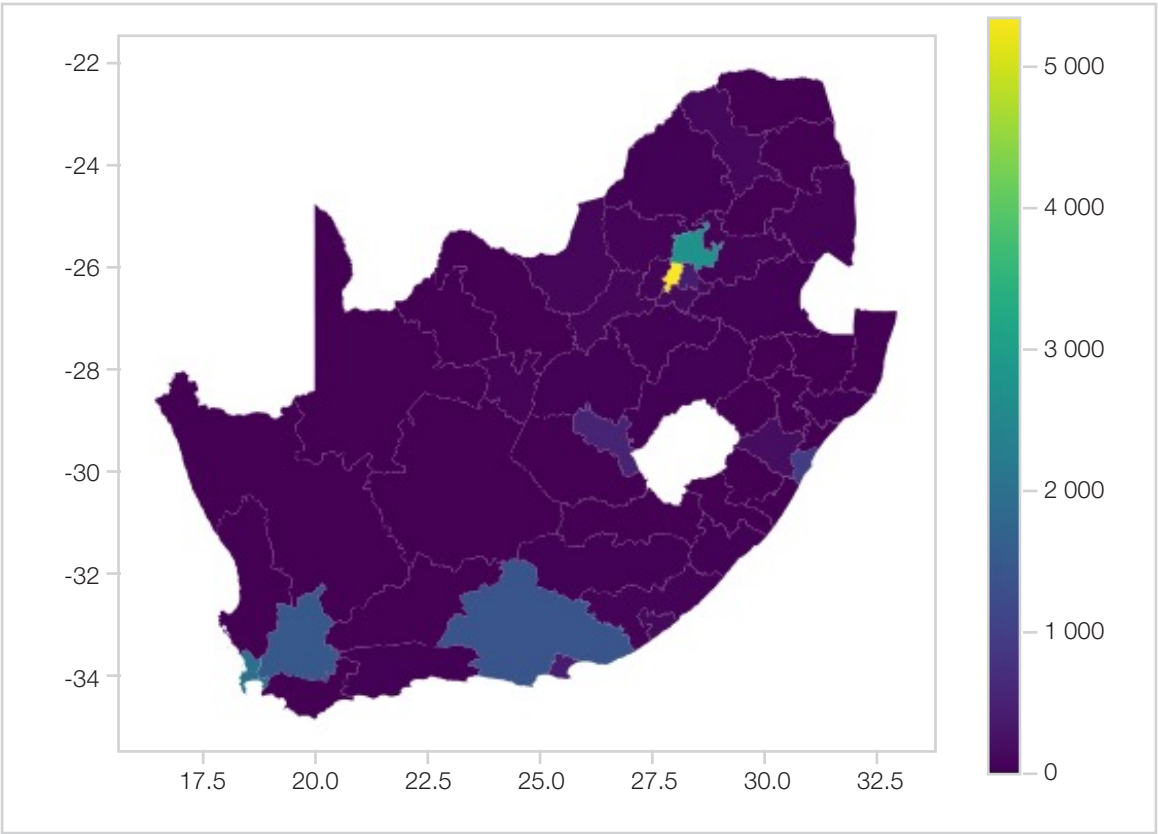
Machine learning (ML) involves training algorithms to learn patterns in data, allowing them to automatically identify these patterns and make predictions on new data. In our study, we used supervised learning, a stream of ML that trains a model on labelled data to map inputs to desired outputs.[35] Our task was a binary classification problem, where the model was tasked with predicting whether a protest would occur or not.

We constructed our dataset using known protests from the ACLED dataset, while the Twitter data served as the feature set. Data was labelled with '1' for actual known protest events and '0' for non-protest events. We grouped tweet and protest locations into standard coordinates, matching them to local municipalities (LMs) in South Africa. Date matching involved comparing the tweet's date to a ten-day interval around a protest event. We retained only one event for days with multiple protests in the same LM to avoid duplicate data.

Tweets were aggregated to show combined totals and averages of various features, such as the number of tweets, retweets, likes, followers, and replies. Sentiment analysis was employed to estimate the mean polarity and subjectivity of each tweet, while content analysis assigned the tweet to a content category (e.g., if the 'education-related' grievance category was the most common keyword identified for a series of tweets, then the event's protest grievance would be assigned as 'Education'). Lastly, the tweet rate was calculated for each day in the ten-day cycle and normalised to a value between 0 and 1.

Figure 2 depicts the LM boundaries of South Africa. The choropleth map shows the relative abundance of tweets in the different LMs across South Africa. It is important to note that a select few LMs have a higher relative number of tweets when compared to the rest of South Africa. Some of these higher tweet density LMs include, Johannesburg, Tshwane, Cape Town, Cape Winelands, Mangaung, Sarah

Figure 2: Choropleth showing location of Tweets used in training dataset



Bartman and eThekwini municipalities. These LMs are expected as they contain most of the major universities of South Africa; Wits and UJ, UP, UCT, Stellenbosch, UFS, Rhodes and UKZN respectively.

Model training and evaluation

Initial models struggled to distinguish between protest events and non-protest events largely due to outliers in the training data. Protest events consisting of a low volume of tweets were difficult to identify due to limited Twitter presence for said protests. To improve model performance, we excluded protests with fewer than 20 tweets over a ten-day period, as they were deemed insufficient in scope and likely to introduce noise into the model. From this, any prospective event that did not surpass the 20-tweet threshold would be directly labelled as a non-protest.

The dataset was divided into training and testing subsets to allow for evaluation. The training set was used to develop and fine-tune the model, while the testing set was reserved for assessing performance on unseen data. Candidate models tested include the most common classification algorithms, including logistic regression, decision tree classifiers, random forest classifiers, naive Bayes classifiers and various boosting algorithms. The best-performing model was a random forest classifier, drawing on features that included the tweet signature, sentiment and content analysis outcomes, and other tweet metadata. While the model was trained at the local municipal level, it is able to predict at other administrative levels (such as district municipalities) with similar prediction accuracy.

We conducted several rounds of hyperparameter tuning to further improve the

random forest model. Hyperparameter tuning involves adjusting the parameters that govern the training process of the model, such as the number of 'trees' in the 'forest', the maximum depth of each tree, and the minimum number of samples required to split a node – the optimum arrangement is determined by model performance. We assessed the model using metrics such as accuracy and F1 score to ensure it accurately predicts both classes. Additionally, we used a cross-validation procedure to prevent overfitting. The specifics of the results are highlighted below.

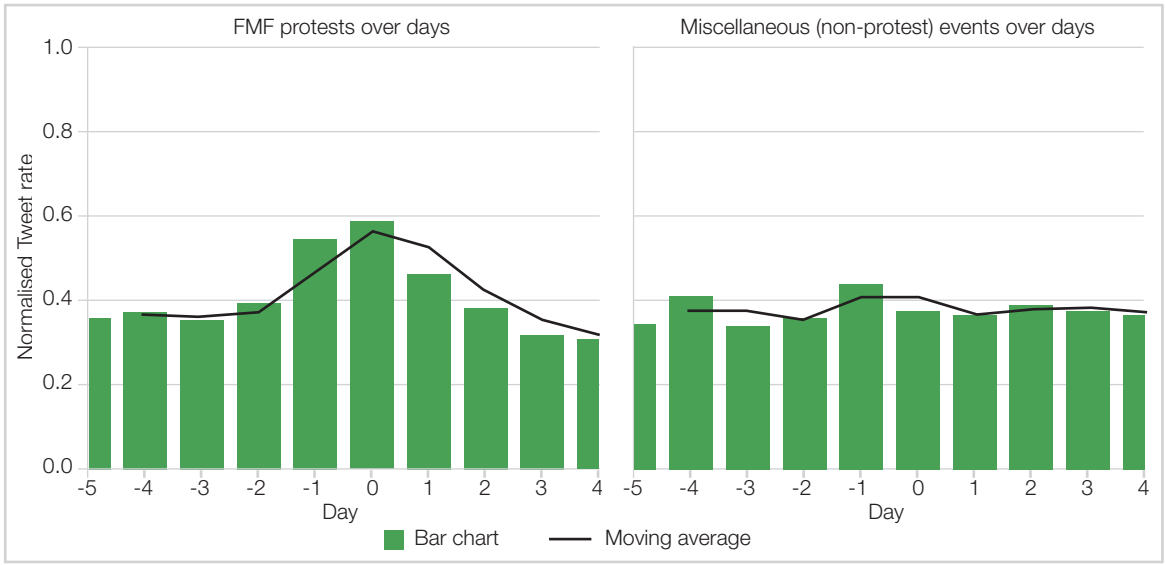
Results and discussion

The query system and the accompanying machine learning model identified several notable trends, based on the FMF-series of protest events. Notably, the system successfully identified a grievance in approximately one-third of tweets and identified a tactic in a similar proportion. Content analysis confirmed that the patterns behaved 'as expected', e.g. that the most frequent location subcategory for FMF-protests was 'tertiary education institution,' and the most common grievance subcategory was 'education.'

Significantly, a consistent shape was observed in the tweet rate curve associated with protests. On aggregate, the tweet rate peaked at the time of the protest event or shortly thereafter. As illustrated in Figure 3, the typical tweet rate curve for FMF-related protests shows a gradual increase in interest over the five days before the protest event, peaking on the day of the protest and a slow decrease in interest following the protest. This pattern contrasts with non-protest events, which exhibited either different 'signatures' or a noisy tweet rate distribution over the ten-day periods, also depicted in Figure 3.

A conceivable explanation for the distinct tweet rate curve is that it reflects the confluence of articulated grievances leading to widespread public reaction, combined with organisational efforts as activists mobilise for public intervention. The media attention and public discussion, including other digital phenomena, such as counter-arguments, trolling, and opportunistic posts, may account for the heightened interest immediately following the protest event.

Figure 3: Comparison between the typical FMF protest events' Twitter signature and SA-based Twitter activity for non-protest events, where day 0 is the day of the protest event



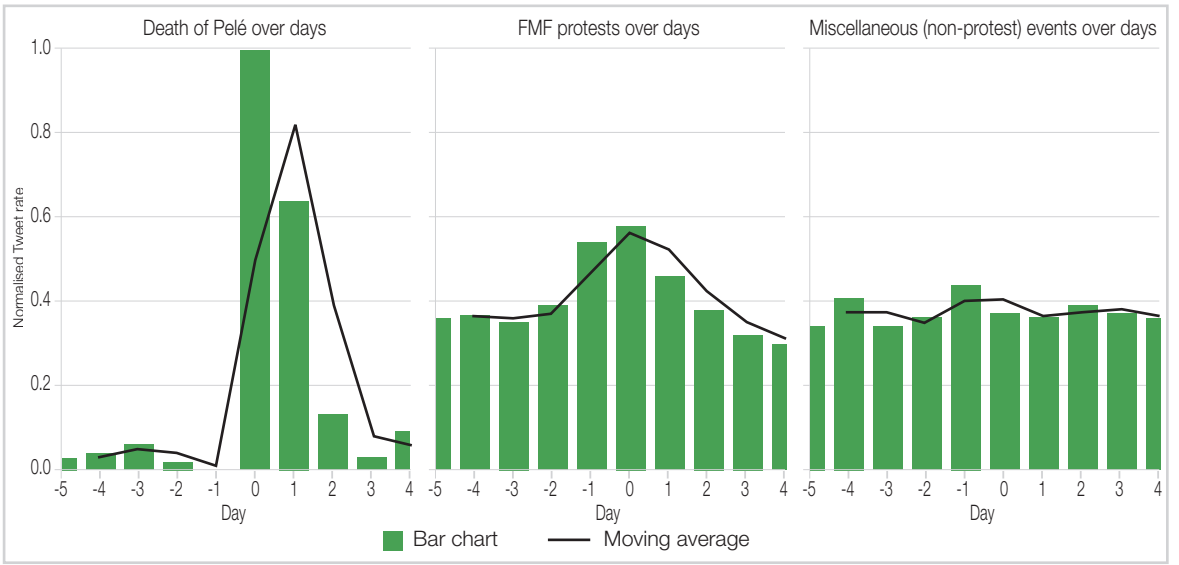
A prominent aspect of this pattern is the consistency of protest-related interest (albeit confined to FMF protests) regardless of variances in geography, timing and actors. This suggests that protest events exhibit a distinct pattern of interest over time, independent of the involvement of specific actors, grievances, sentiment, locations, and patterns of escalation and de-escalation. In particular, the shape of the event's Twitter signature emerged as a highly predictive factor in reconstructing when (and whether) a protest event took place.

The robustness of the Twitter signature model extends to its ability to differentiate between 'random events' (non-protest, non-high interest, as shown in Figure 3) and other social events that, despite generating similar levels of public interest, follow different patterns to the FMF-protest signature. Figure 4 demonstrates this by showing the signature of the death of Pelé (a largely unforeseen event) next to the FMF-protest signature. The Twitter signature model proved effective not only in retrospectively identifying protest events from social media data but also in potentially forecasting imminent protests. The Twitter signature model effectively differentiates between "random events" and other social events, as demonstrated by the stark contrast between the two signatures. The model's ability to identify these distinct patterns suggests its potential for real-world applications, such as monitoring public sentiment or predicting potential crises. Future research could explore the applicability of this model to other types of events or social media platforms.

Machine learning model on unseen data

Our random forest classifier demonstrated consistent and reliable results. Hyperparameter tuning increased the model's accuracy score from 81.1% to 89.9%. Furthermore, the F1 score of the model is notably high, with a score of 0.95 for protest events and 0.96 for non-protest events. Cross-validation ensured the model was not overfitted, resulting in a final cross-validation accuracy score of 88.1%. The difference in overall accuracy and cross-validation accuracy was approximately 2%, indicating a robust and well-generalisable model. These metrics affirm the model's capability to accurately identify and differentiate between protest and non-protest events, even in novel scenarios.

Figure 4: Comparison between anticipated and unanticipated events' social media signatures to FMF-protest events' signatures



The model's strong performance can be attributed mainly to our effective feature engineering approach. Robust content and sentiment analyses, in particular, played a crucial role in extracting meaningful information from the data. This, combined with the classifier's ability to handle complex relationships between features, contributed significantly to the model's high accuracy and generalisability.

The consistent performance across datasets, as indicated by the cross-validation, highlights the efficacy of our machine-learning approach and enhances confidence in its practical applicability for real-world protest prediction and analysis. Overall, the model's reliability and precision across various datasets underscore its potential for practical application in real-world protest event prediction.

Predictive ethics

While the potential of ML models to predict protest events based on social media data is promising, it is crucial to acknowledge the ethical implications of such technology. The ability to foresee protest activities, although valuable for understanding social dynamics, also raises concerns about its possible misuse. For instance, state authorities could leverage these predictive tools for undemocratic purposes, such as surveilling and suppressing dissent, rather than addressing the underlying causes of unrest. As Bekker notes,[36] protest is often a critical form of popular expression and a symptom of deep social distress. Therefore, researchers must approach the development and application of these models with caution, ensuring that their use aligns with ethical standards and promotes transparency, accountability, and the protection of civil liberties. For this consideration, we have focused purely on historical protests and data in this study.

Future improvements

Our method presents several opportunities for further development. The current approach is restricted to tweets with English text, this decision was largely due to the size limitation of the Twitter API token length. Another potential improvement involves protest cataloguing. This approach requires greater computational requirements, it involves locally downloading all tweets (negating the limit of the API token) emanating from a particular country or other geography, for a specific timespan, and running a model to identify and index historical protest incidents for further study. Such an approach could be largely 'language-neutral,' as the model would focus on event signatures (such as the timing and frequency of identical hashtags) rather than relying on a language-specific query system. Event catalogues based on social media data may offer new insights into protest dynamics, including the role of state violence and provocation, changes in protest size over time, and shifts in sentiment during protest events.

Yet another avenue for exploration is to study the signature shapes of other types of group phenomena, including social media activity during times of significant public phenomena, including emancipatory change, popular celebrations, communal trauma, elections, coups, and revolutions. It is our hope that this line of enquiry might enhance our understanding of social change, civil distress and beyond.

Potential replacement and supplementation of research

The advent of Large Language Models (LLMs) such as OpenAI's GPT series, Meta's LLaMA models, and Google's Gemini models has the potential to significantly influence the field of computational social science, including PEA. LLMs are designed to process and generate human-like text, enabling them to perform a wide range of natural language processing

tasks, such as sentiment analysis, content categorisation, and event detection, with high accuracy and minimal human intervention. This capability suggests that LLMs could either enhance or, in some cases, replace traditional methods of content and sentiment analysis used in protest research.

LLMs could streamline the process of content and sentiment analysis by automating the identification of grievances, tactics, and other relevant protest characteristics from social media data. With their ability to process vast amounts of text data efficiently, LLMs could potentially outperform traditional machine learning models by providing more nuanced interpretations of social media content. Additionally, LLMs can adapt to various languages and dialects, making them well-suited for analysing multilingual datasets, which is a common challenge in global protest research and further emphasised in our methodology.

LLMs offer a new method to data cleaning, sentiment analysis, grievance cataloguing and model simplification. For example, LLMs could be used to extract additional features from the text, such as underlying themes, complex sentiment shifts, or even the identification of less overt forms of discourse (e.g., sarcasm, slang or loanwords). These features could then be incorporated into our event signature model to improve its predictive accuracy and robustness. By combining the strengths of LLMs with our temporal pattern analysis, we can develop a more comprehensive understanding of protest dynamics and social media behaviour.

Other developments

Since completing this research, the field of computational social science has encountered a significant challenge. In early 2023, following Elon Musk's highly publicised acquisition of Twitter (which has since been rebranded as 'X' and tweets as 'X's'), the Twitter API, including the academic research access, was monetised. The introduction of charges has effectively ended the decade of academic research on a wide range of human behaviour, viewed through social media records. With such records now virtually inaccessible to researchers, many potential insights will remain unexplored, except through well-resourced entities.

Conclusion

We have demonstrated the presence of a 'Twitter signature' attending protest events on social media. By analysing historical protest events, particularly the FMF protests in South Africa, our approach shows promise in both retrospectively identifying and potentially predicting similar events. The utilisation of social media data provides a unique vantage point, offering insights into protest dynamics that may not be captured by traditional media or administrative records, especially in contexts where media coverage is limited or censored.

However, several caveats must be noted. The findings are context-specific, applying predominantly to the FMF protests within South Africa. Further research is needed to assess the generalisability of our methodology to other types of protests and geographic locations. Utilising historical datasets of tweets presents an opportunity for further replication and validation, and for exploring the model's applicability in diverse contexts.

Ethical considerations are paramount in this research. Protest tends to present 'an important popular censure on rulers and the institutions they represent' and generally represents 'a symptom of severe social distress.'[37] While our study illustrates the potential of machine learning to predict protests, this capability raises concerns about its use for state-sponsored repression of would-be protests – which, we have shown, can be predicted with some confidence.

Monitoring social media for signs of protest could infringe on individuals' privacy and freedom of expression, especially if such data is used to preemptively suppress or control dissent. The ability to anticipate protest events could be misused for undemocratic surveillance and control, highlighting the importance of using such technology responsibly. Moreover, any deployment of these technologies must be guided by ethical principles that prioritise transparency, accountability, and the protection of civil liberties. Our goal is to contribute to a deeper understanding of collective action events and their underlying dynamics rather than to facilitate real-time monitoring of social media.

Our research indicates that machine learning models can accurately predict protest events based on historical Twitter data. This is particularly effective within the specific context of FMF protests in South Africa. The distinctiveness of protest-related Twitter signatures suggests that social media can offer valuable insights into protest characteristics, potentially surpassing traditional methods, and could even be anticipated in advance to a degree of certainty.

Notes

1 Hraklis Papageorgiou is a full stack data scientist at Standard Bank in South Africa. Joseph Baggott is a AI developer and MSc candidate at the University of the Witwatersrand. Martin Bekker is a computational social scientist and AI ethicist at the University of the Witwatersrand.

2 Peter (Kate) Alexander, "Rebellion of the Poor: South Africa's Service Delivery Protests – A Preliminary analysis." *Review of African Political Economy* 37, no. 123 (2010): 25–40; Martin Bekker, "Language of the Unheard: Police-Recorded Protests in South Africa, 1997–2013," *Review of African Political Economy* 49, no. 172 (2022): 226–245.

3 Steven Friedman, “People are Demanding Public Service, Not Service Delivery,” *Business Day Live*, 29 July 2009; Trevor Ngwane, "Ideology and Agency in Protest Politics," (PhD dissertation, Faculty of Humanities, University of KwaZulu-Natal, Durban, 2011), https://ccs.ukzn.ac.za/files/Ngwane%202012%20masters%20thesis%20at%20UKZN%20CCS.pdf ; Luke Sinwell, “Is ‘Another World’ Really Possible? Re-Examining Counter-Hegemonic Forces in Post-Apartheid South Africa”. *Review of African Political Economy*, 38, no. 127 (2011): 61–76, DOI:10.1080/03056244.2011.552588; Karl Von Holdt, Langa Malose, Sepetla Molapo, Nomfundo Mogapi, Kindiza Ngubeni, Jacob Dlamini, and Adele Kirsten, *Insurgent Citizenship, Collective Violence and the Struggle for a Place in the New South Africa*, (Johannesburg: Centre for the Study of Violence and Reconciliation; Society, Work and Development Institute, Faculty of Humanities, University of the Witwatersrand, 2011).

4 Martin Bekker, "Depends on How You Count Them: The Value of General Propensity Choropleth Maps for Visualising Databases of Protest Incidents," *Journal of Maps* 19, no. 1 (2023): 2064778, DOI:10.1080/17445647.2022.2064778.

5 Patrick Bond and Shauna Mottiar, "Movements, Protests and a Massacre in South Africa," *Journal of Contemporary African Studies* 31, no. 2 (2013): 283–302, https://doi.org/10.1080/02589001.2013.789727.

6 Martin Bekker, *Rebellion with a Cause: An Enquiry into the Nature of South African Post-apartheid Protest, using Computational Social Science* Methods (PhD Dissertation, University of Johannesburg, 2020), https://search.proquest.com/openview/770b4f80c23e9858f5952c79eb6d703c/1?pq-origsite=gscholar&cbl=2026366&diss=y.

7 Samuel Brannen, Christian Haig, and Katherine Schmidt, *The Age of Mass Protest: Understanding an Escalating Global Trend* (Washington, DC: Center for Strategic and International Studies, 4 March 2020), http://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/200303_MassProtests_V2.pdf.

8 Bekker, "Language of the Unheard.”

9 Carin Runciman, Peter (Kate) Alexander, Mahlatse Rampedi, Boikanyo Moloto, Boitumelo Maruping, Eunice Khumalo, and Sehlaphi Sibanda, *Counting Police-Recorded Protests: Based on South African Police Service Data* (Johannesburg: Social Change Research Unit, University of Johannesburg, 2016).

10 Bond and Mottiar, "Movements, Protests and a Massacre in South Africa."

11 Glenda Daniels, “Scrutinizing Hashtag Activism in the #MustFall Protests in South Africa,” in *Digital Activism in the Social Media Era*, ed. Bruce Mutsvairo (Cham, Switzerland: Palgrave Macmillan, 2016), DOI:10.1007/978-3-319-40949-8_9.

12 Pier Paolo Frassinelli, “Hashtags: #RhodesMustFall, #FeesMustFall and the Temporalities of a Meme Event,” in *Perspectives on Political Communication in Africa*, ed. Bruce Mutsvairo and Beschara Karam (Cham, Switzerland: Palgrave Macmillan, 2018) DOI: 10.1007/978-3-319-62057-2_4.

13 Ruud Koopmans and Dieter Rucht, “Protest Event Analysis,” in *Methods of Social Movement Research*, ed. Bert Klandermans and Suzanne Staggenborg, (Minneapolis: University of Minnesota Press, 2002), 321.

14 Clionadh Raleigh, Rew Linke, Håvard Hegre and Joakim Karlsen, "Introducing ACLED: An Armed Conflict Location and Event Dataset," *Journal of Peace Research* 47, no. 5 (2010): 651–660, DOI:10.1177/0022343310378914.

15 Jane Duncan, *Protest Nation: The Right to Protest in South Africa*, (Durban: University of KwaZulu Natal Press, 2016), 183.

16 Lizette Lancaster, "Unpacking Discontent: Where and Why Protest Happens in South Africa," *South African Crime Quarterly* 64 (2018): 29–43, DOI:10.17159/2413-3108/2018/v0n64a3031.

17 For example, see Bekker, "Rebellion with a Cause."

18 Michael Goodchild, "Citizens as Sensors: The World of Volunteered Geography," *GeoJournal* 69 (2007): 211–221, DOI: 10.1007/s10708-007-9111-y; Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proceedings of the 19th International Conference on World Wide Web* (2010) ,pp. 851–860.

19 Tanja Bosch, *Social Media and Everyday Life in South Africa*, (London: Routledge, 2020), DOI: 10.4324/9780429316524.

20 See, for example, Son Doan, Bao-Khanh Ho Vo and Nigel Collier, "An Analysis of Twitter Messages in the 2011 Tohoku Earthquake," in *Electronic Healthcare: 4th International Conference, eHealth 2011, Málaga, Spain, November 21–23*, 2011, Revised Selected Papers 4, (Berlin: Springer, 2012), pp. 58–66; Sakaki, Okazaki and Matsuo, "Earthquake Shakes Twitter Users."

21 See, for example, Yu-Ru Lin, Brian Keegan, Drew Margolin and David Lazer, "Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events." *PloS One* 9, no. 5 (2014): e94093, DOI: 10.1371/journal.pone.0094093.

22 See, for example, Hamed Abdelhaq, Christian Sengstock and Michael Gertz, "Eventweet: Online Localized Event Detection from Twitter," *Proceedings of the VLDB Endowment* 6, no. 12 (2013): 1326–1329, DOI: 10.14778/2536274.2536.

23 Hassan Sayyadi, Matthew Hurst and Alexey Maykov, "Event Detection and Tracking in Social Streams," in *Proceedings of the International AAAI Conference on Web and Social Media* vol. 3, no. 1 (2019), pp. 311–314, DOI: 10.1609/icwsm.v3i1.13970; Jianshu Weng and Bu-Sung Lee, "Event Detection in Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media* vol. 5, no. 1 (2011), pp. 401–408, DOI: 10.1609/icwsm.v5i1.14102.

24 See, for example, Haji Mohammad Saleem, Yishi Xu and Derek Ruths, "Effects of Disaster Characteristics on Twitter Event Signature," *Procedia Engineering* 78 (2014): 165–172, DOI: 10.1016/j.proeng.2014.07.053.

25 Guandan Chen, Qingchao Kong and Wenji Mao, "Online Event Detection and Tracking in Social Media based on Neural Similarity Metric Learning," in *2017 IEEE International Conference on Intelligence and Security Informatics* (ISI) (Beijing, China: Institute of Electrical and Electronics Engineers, 2017), pp. 182–184; Taiwo Kolajo, Olawande Daramola and Ayodele Adebiyi, "Real-time Event Detection in Social Media Streams through Semantic Analysis of Noisy Terms," *Journal of Big Data* 9, no. 1 (2022): 90, DOI: 10.1186/s40537-022-00642-y.

26 Saleem et al, "Effects of Disaster Characteristics on Twitter Event Signature."

27 Hamed et al, "Eventweet."

28 Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim and Doheon Lee, "A Taxonomy of Dirty Data," *Data Mining and Knowledge Discovery*, 7 (2003): 81–99.

29 Sathappan Muthiah, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz and Naren Ramakrishnan, "Planned Protest Modeling in News and Social Media." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 2 (2015), pp. 3920–3927, DOI: 10.1609/aaai.v29i2.19048; Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wan, Jose Cadena, Anil Vullikanti and Gizem Korkmatz, "'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2014) pp. 1799–1808, DOI: 10.1145/2623330.2623373; Ryan Compton, Craig Lee, Tsai-Ching Lu, Lalindra De Silva and Michael Macy, "Detecting Future Social Unrest in Unprocessed Twitter Data: Emerging Phenomena and Big Data," In *2013 IEEE International Conference on Intelligence and Security Informatics*, pp. 56–60, (IEEE, 2013), DOI: 10.1109/ISI.2013.6578786.

30 Mohsen Bahrami, Yasin Findik, Burcin Bozkaya and Selim Balcisoy, "Twitter Reveals: Using Twitter Analytics to Predict Public Protests," *arXiv preprint arXiv:1805.00358*, DOI: 10.48550/arXiv.1805.00358.

31 Kartikeya Bajpai and Anuj Jaiswal, "A Framework For Analyzing Collective Action Events on Twitter," In *Proceedings of the 8th International ISCRAM Conference* (2011); Zuoming Wang and Kara Caskey, "#Occupywallstreet: An Analysis of Twitter Usage During a Protest Movement," *Social Networking* 5, no. 04 (2016): 101, DOI: 10.4236/sn.2016.54011.

32 Twitter, Inc. 2023, "Developer Agreement and Policy," accessed at https://developer.x.com/en/developer-terms/agreement-and-policy.

33 Kim et al, "A Taxonomy of Dirty Data."

34 Martin Bekker, "Better, Faster, Stronger: Using Machine Learning to Analyse South African Police-Recorded Protest Data," *South African Review of Sociology* 52, no. 1 (2022): 4–23, DOI: 10.1080/21528586.2021.1982762

35 Tom Mitchell, *Machine Learning* (New York: McGraw-Hill, 1997).

36 Bekker, "Depends On How You Count Them."

37 Ibid.