Uncovering Influential Factors of Civil Unrest in South Africa: A Machine Learning and OSINT Approach

Lungisani Ndlovu[1], Anton De Kock[2], Nenekazi N. P. Mkuzangwe[3], Ntombizodwa Thwala[4], Chantel J.M. Mokoena[5] and Rethabile Matimatjatji[6]

[*] Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
{ lndlovu3, adkock2, nmkuzangwe, nthwala1, cmokoena, rmatimatjatji}@csir.co.za

Abstract. Continuous monitoring of the risk of civil unrest events and predicting their occurrence is of paramount importance. This task involves identifying and understanding the primary factors that contribute to such events, especially in regions with unique dynamics, such as South Africa. Although many global and South African-specific studies have conducted research on predicting the frequency or probability of these events, there is a notable gap in identifying the influential factors behind them. This study unveiled several contributing factors, including demanding behaviour, power outages, service delivery, wage disputes, acts of violence, gender-based conflicts, and unemployment rates. These factors, individually or collectively, contribute to the complexity of civil unrest in the region. The 2021 South African unrest, also known as the July 2021 riots, the Zuma unrest, or Zuma riots, serves as an example. This event was triggered by the imprisonment of former president Jacob Zuma for contempt of court, inciting his followers to demand his release, a situation aligning with the 'demanding behaviour' influential factor identified in our study. We used advanced data analysis and machine learning techniques to explore these factors. Specifically, the Logit model was used to determine the coefficients that optimally fit the data, establishing significant relationships between these factors and incidents of civil unrest. Our research not only offers insights on influential factors, but also presents a predictive framework. We evaluated logistic regression, support vector clustering, decision tree classifier, and random forest classifier models to predict civil unrest. The results showed that the decision tree and the random forest classifiers perform better, achieving an accuracy of 98%, compared to logistic regression and support vector clustering, which have an accuracy of 97%.

Keywords: Civil Unrest, South Africa, Influential Factors, OSINT, Machine Learning.

1 Introduction

With the advent of the information age since the 1990s, open source intelligence (OSINT) has gained a special place in the intelligence community (IC) [1]. Consequently, we have seen a significant contribution to harnessing the vast amount of OSINT data to provide insights and predict civil unrest. Civil unrest has profound

implications for the stability, development, and well-being of nations [2]. In the context of South Africa, a country with a complex history of political, economic, and social dynamics, understanding and predicting the occurrence of such events is of paramount importance. Over the years, the country has made significant improvements towards building a more inclusive and equitable society, but it continues to face a range of challenges that can spark civil and civil unrest. As a result, it has become crucial to continuously monitor the risk of civil unrest and predict the probability of such events.

According to Ungureanu [3], the research community has developed numerous systems to provide valuable insights using OSINT data collected from various sources, with social media platforms playing an important role. Such systems enable users to better understand unfolding events, public opinions, and potential risks. In addition, these systems provide accurate and timely predictions of these events, which are essential for policymakers, law enforcement agencies, and civil society organisations to proactively address the underlying issues, improve public safety, and promote social cohesion.

However, most studies have been conducted in international contexts, which may not be completely applicable to South Africa given the unique dynamics of civil unrest. In addition to that, most studies, including South African-specific studies, focused on predicting the frequency or probability of these events. Although the ability to predict future occurrences of these events is important to help governments, organisations, and individuals anticipate them and allow proactive measures to be put in place to maintain peace, public safety, and economic stability, it is also important to understand the factors that cause or influence these events. Knowledge of these influential factors in society will enable corrective measures to be put in place to reduce the occurrence of these influential societal factors and hence reduce the occurrence of civil unrest events in the future.

Therefore, this study aims to uncover and understand the factors that have a significant influence on the occurrence of civil unrest in the South African context and presents a predictive tool. This study uses data from Twitter and news media platforms, including SABC News, Eyewitness News (EWN) and News24. We also employ data from the Armed Conflict Location & Event Data Project (ACLED). We use natural language processing (NLP) to process and explore the data to gain insight. Subsequently, we train and evaluate supervised learning models that include support vector clustering, decision tree classifier, and random forest classifier with label encoder to normalise the data where necessary. From the model evaluations, the decision tree and random forest classifiers perform better than the others, achieving an accuracy of 98%, compared to logistic regression and support vector clustering, which have an accuracy of 97%. This approach is designed as a valuable tool for early warning and proactive intervention to reduce social instability and promote sustainable development in the region. However, we propose future work to address the problem of identifying misinformation in data before predicting the probability of civil unrest.

2 Related Work

In recent years, much work has been done to use OSINT data to provide insights and predict the occurrence of civil unrest.

Xu et al. [4] focused on detecting emerging civil unrest events by analysing massive microblogging streams. Specifically, they proposed an early detection system consisting of a novel cascade of text-based filters to identify posts of civil unrest events based on topics, time, and location. In contrast to model-based prediction approaches, their method was purely extractive as it detected relevant posts from massive volumes of data directly. They designed and implemented such a system in a distributed framework for scalable processing of real-world data streams. Subsequently, they carried out a large-scale experiment with their system on the entire dataset from Tumblr for three consecutive months. Experimental results indicated that the simple filter-based method provided an efficient and effective way to identify posts related to real-world civil unrest events. Although similar tasks have been investigated on different social media platforms (e.g., Twitter), little work has been done on Tumblr despite its popularity. Their analysis showed the collective microblogging patterns of Tumblr. Future work included a more advanced geocoding mechanism for Tumblr posts. The preliminary analysis showed that it was possible to align users on different social media platforms and then propagate their geoinformation. The authors also planned to analyse different types of social movements on a global scale.

McMinn and Jose [5] proposed a novel, efficient and real-time event detection approach for Twitter, using the role that named entities play in events. They used a clustering technique that partitioned tweets according to the entities they contained, along with burst detection and cluster selection techniques, to extract clusters related to ongoing real-world events. Their approach demonstrated superior performance compared to state-of-the-art methods while maintaining very low computational complexity and ensuring real-time performance. The authors observed that nouns and verbs play a significant role in determining recall and precision, respectively, and that Twitter-specific features either have no effect or may have a detrimental impact on detection performance when using their entity-based approach.

Fengcai, Jinsheng, and Li [6] present a hidden semi-Markov model-based framework to take advantage of large-scale digital history-coded events captured from GDELT to use frequent subgraph patterns extracted from GDELT event streams to uncover the underlying event evolution mechanics and formulate the prediction of civil unrest events as a sequence classification problem. Extensive empirical tests with data from Thailand in Southeast Asia demonstrated the effectiveness of this framework by comparing it with the traditional HMM, the logistic regression model, and the baseline model. It shows that the GDELT dataset does reflect some useful precursor indicators that reveal the causes or evolution of future events. The authors plan to conduct their future work in the following three aspects: a) introduce a multilevel prediction mechanism to their frameworks, such as city-level or province-level; b) use GDELT 2.0 as the event-mentioned details and global knowledge graphs are also provided in real-time, thus providing detailed insights into the events; c) improve the framework by distinguishing widespread news coverage from localised coverage.

Jiang et al. [7] highlighted the challenging nature of predicting events of civil unrest, emphasising its crucial importance for government departments in managing public opinion and preserving social stability. They introduce a temporal dual graph convolutional network model as a means to predict civil unrest events, employing GDELT's unrest events as the ground truth and utilising social media data as historical texts. Their future research directions include the exploration of novel semantic association methods to improve the accuracy of labelling social networks and news media data, along with the construction of an event evolution graph aimed at illustrating the developmental trajectory of unrest events, thereby providing a more comprehensive interpretation of prediction outcomes.

Wu and Gerber [8] conducted a study on predicting the onset of protests in Cairo, Egypt, during the early Arab Spring. They used social media data and found that an increase in expressions about future protests on Twitter was correlated with upcoming protests. However, they did not find support for other related hypotheses. They built two sets of predictive models: protest advertisement models and GDELT models, with the former performing better for short-term predictions and the latter for longer horizons. The study also suggested that an increase in military presence could predict the start of protests. This research aimed to bridge social science theory with civil unrest forecasting and provided valuable information for decision-makers. Future research could focus on real-time model testing, exploring grievances' role in the unrest, and understanding factors motivating online political participation.

3 Methodology

3.1 Data Collection

This study uses data collected from various sources, including Twitter, news media platforms, and ACLED. The Twitter data and news media data span from November 2022 to August 2023, with ACLED data included for June to August. The selected news media platforms comprise SABC News, News24, and EWN (Eyewitness News), renowned for providing reliable news, opinions, business, sports, politics, interactive media, and lifestyle content in South Africa [9]. News data is collected hourly through news feeds. On the other hand, Twitter data is collected with the help of Twint, an advanced Twitter intelligence scraping tool. This Python-written tool enables the scraping of tweets from Twitter profiles, bypassing the need for Twitter's Application Programming Interface (API) [10].

3.2 Data Cleaning and Pre-processing

Data cleaning, also known as text cleaning, as defined in [11], involved cleaning data, handling missing values, deleting duplicates, and altering variables within a dataset as needed. Pre-processing refers to the steps undertaken to make data more manageable and efficient for data mining, machine learning, and other data science tasks [12]. In this study, this process included removing URLs, removing stop words, and performing lemmatisation. Lemmatisation refers to the process of grouping different inflected

forms of the same word [13]. These steps are essential to perform before analysing the data to ensure its quality and suitability for the intended purposes [11]. In addition, data transformations, including standardisation or normalisation, were performed as needed to ensure variable compatibility.

4 Data Analysis and Results

4.1 Feature Selection

Feature selection refers to the process of selecting relevant information features or subsets of variables from a large number of data attributes to improve the efficiency and effectiveness of machine learning or statistical models [14]. This selection helps reduce dimensionality, improve model performance, and simplify data analysis [14]. Table 1 shows the features that were extracted or selected from the news articles.

Table 1. Key features extracted from articles.

#	Feature	Data Type
1.	Subjectivity	float64
2.	word_count	int64
3.	sentence_count	int64
4.	Location	int32
5.	top_pos	int32
6.	Demanding	int64
7.	Electricity	int64
8.	power outage	int64
9.	service delivery	int64
10.	Wage	int64
11.	Salary	int64
12.	load shedding	int64
13.	Loadshedding	int64
14.	Unemployment	int64
15.	Violence	int64
16.	gender-based	int64
17.	Gbv	int64
18.	Employment	int64

4.2 Correlation Heatmap

We used a heat map and plotted it to visualise the correlation between different features in the dataset. This helped us identify the relationships among features, providing early insights into latent associations among the defined features or variables. Understanding

these associations is crucial to grasping the possible correlations or patterns amongst the chosen features and between the chosen features and the target variable. Furthermore, this understanding of the association amongst features assists with identifying the existence of multicollinearity amongst the chosen features and addressing it since it tends to make statistically significant features insignificant [15]. Multicollinearity was examined using the pairwise correlation coefficients values where if the absolute value of the correlation coefficient between two features is close to 0.8 indicates the possibility of collinearity [16], [17]. Fig. 1 shows the correlation heatmap/matrix that provides the pairwise correlation coefficient values amongst the independent features.

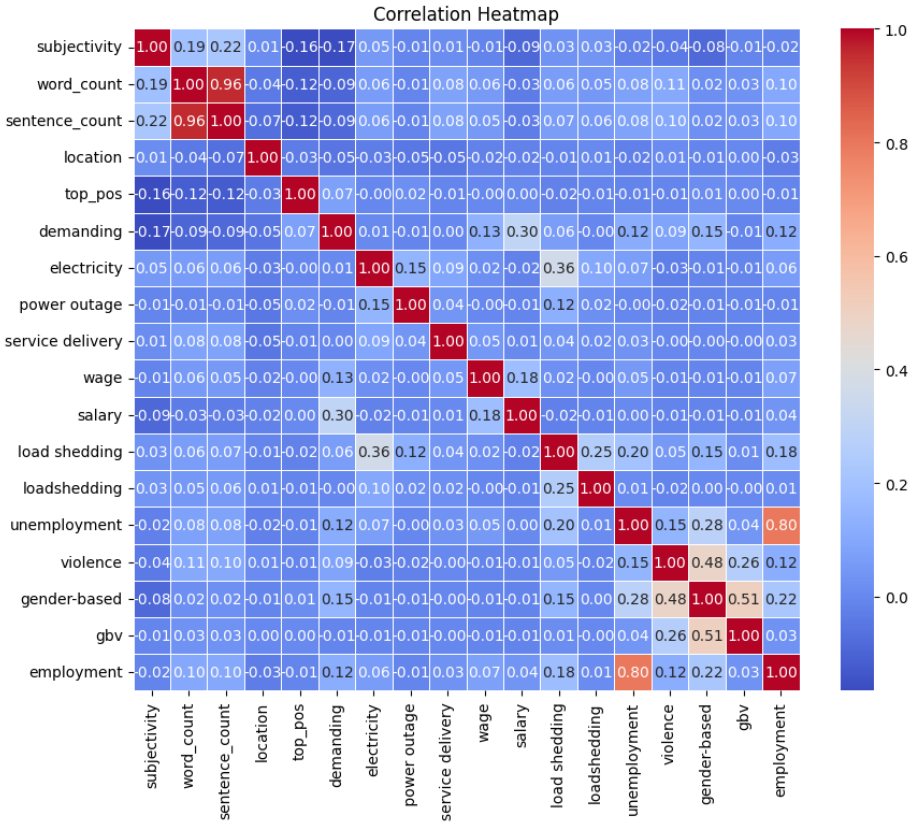


Fig. 1. Correlation heatmap.

In this study, multicollinearity is reduced by removing one of the two highly correlated independent variables (pairwise correlation coefficient of 0.8 or more). That is, of the two highly correlated independent variables, only one will be used in the fitting of the model. From Fig. 1 it can be seen that most of these variables have pairwise correlation coefficient values of less than 0.8, indicating that they are suitable for inclusion in the model. However, the pairs word_count and sentence_count, as well as employment and unemployment features, have correlation coefficient values greater than 0.8. Consequently, only sentence_count and employment were selected for further analysis using the logit regression model.

4.3 Modelling and Evaluation

We used logistic regression to evaluate the association between the features and the target variable. With logistic regression, it is possible to estimate the probability of the occurrence of an event (e.g., voting or not voting) based on a dataset of independent variables [18]. We used the fitted Logit to determine the model coefficients and their associated p-values to establish/identify the features that have a significant relationship with the target variable, civil unrest.

To determine if the fitted features had a significant relationship with civil unrest, the corresponding model coefficients' p-values were checked if they were less than 0.05 level of significance. Of the 18 features that were used to build the model, 11 features were found to have a significant relationship with civil unrest and are listed in Table 2. This means these features are the influential factors of civil unrest in South Africa.

Table 2. Logit regression results.

Feature	P-value (P>|z|)
Subjectivity	0.000
sentence_count	0.000
Location	0.000
top_pos	0.004
Demanding	0.000
power outage	0.035
service delivery	0.009
Wage	0.000
Violence	0.000
gender-based	0.015
Employment	0.000

To evaluate the model's fit to the data, this study employs statistical accuracy measurements, including log-loss (cross-entropy) and Brier score. Log-loss is a common metric to assess the quality of predicted probabilities in binary classification problems [19], [20]. The Brier score measures the accuracy of probabilistic predictions [19], [20].

Both the log-loss and Brier score performance metrics quantify the alignment between the model's predictions and actual outcomes. These methods provide insight into the effectiveness of the model in capturing the underlying patterns and trends.

Table 3. Logistic regression results.

Metric	Value
Log-Loss	0.0202
Brier Score	0.0028

Table 3 presents the evaluation metrics, providing information on the performance of the logistic regression model. The calculated log-loss is approximately 0.0202, representing the average difference between the actual and predicted values, expressed in the same units as the target variable. A lower log-loss means a model that better fits the data, indicating that the predicted values align closely with the actual values. The Brier score is approximately 0.0028, representing the mean absolute difference between the actual and predicted values. Similarly, a lower Brier score indicates a more accurate model. In the context of classification problems where the quality of predicted probabilities is critical, the Brier score is often favoured, especially when there are concerns about outliers in the predicted probabilities. It offers a robust assessment of the model's calibration and prediction quality under such circumstances. In conclusion, these metrics collectively provide a comprehensive insight into the model's predictive accuracy.

Table 4. Models' evaluation results.

Model	Accuracy (%)
Logistic regression	97
Support vector classification	97
Decision tree classifier	98
Random forest classifier	98

Subsequently, we compared logistic regression with other well-known classification algorithms. As shown in Table 4, the model evaluations have shown that the decision tree and random forest classifiers performed better, achieving 98% accuracy, compared to logistic regression and support vector clustering, which achieved 97% accuracy.

However, the potential limitations inherent in this study arise from the nature of the dataset used. Although the dataset offers valuable information, it is important to recognise that the news media may have biases in the areas they cover and in their reporting styles. This can result in an incomplete or distorted representation of civil unrest events. The over-reporting of some incidents and under-reporting of others can affect the accuracy of predictions. Furthermore, future research is recommended to address the problem of identifying misinformation in data before predicting the probability of civil disturbance.

5 Conclusions

This study aimed to identify and understand the primary factors that contribute to civil unrest, especially in regions with unique dynamics, such as South Africa. The heatmap visualisation technique was used to explore the correlations between various features defined in the data set. It provided a visual representation of the potential relationships among the features, offering initial insights into hidden associations. Also, the Logit model was used to determine the coefficients that optimally fit the data, establishing significant relationships between these factors and incidents of civil unrest.

The objective was to uncover potential patterns and connections between the selected features and civil unrest. Subsequently, the focus shifted to model development and evaluation. A logistic regression approach was chosen for analysis, given its suitability for exploring the relationships between features and the target variable.

The metrics and coefficients offered valuable insights into the performance of the model. The Log-Loss and Brier score functions facilitated a comprehensive understanding of the model's predictive alignment with actual outcomes such as how well the model fits the data. Furthermore, the coefficients and p-values calculated for each feature revealed the features that have a significant relationship with civil unrest.

Furthermore, we compared logistic regression with other well-known classification algorithms. From the experimental evaluations, the results showed that the decision tree and random forest classifiers performed better, achieving 98% accuracy, compared to logistic regression and support vector clustering, which achieved 97% accuracy.

Future research will focus on investigating biases within news media platforms, particularly in terms of areas covered and reporting methods. Addressing the challenge of identifying misinformation in the data remains a crucial area for exploration.

References

[1] B. Senekal and E. Kotzé, "Open source intelligence (OSINT) for conflict monitoring in contemporary South Africa: Challenges and opportunities in a big data context," *African Security Review*, vol. 28, no. 1, pp. 19–37, 2019.

[2] H. Prime, M. Wade, and D. T. Browne, "Risk and resilience in family well-being during the COVID-19 pandemic.," *American Psychologist*, vol. 75, no. 5, p. 631, 2020.

[3] G. T. Ungureanu, "Open-source intelligence (OSINT). The way ahead," *Journal of Defense Resources Management (JoDRM)*, vol. 12, no. 1, pp. 177–200, 2021.

[4] J. Xu, T.-C. Lu, R. Compton, and D. Allen, "Civil unrest prediction: A tumblr-based exploration," in *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings 7*, Springer, 2014, pp. 403–411.

[5] A. J. McMinn and J. M. Jose, "Real-time entity-based event detection for twitter," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction:*

6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings 6, Springer, 2015, pp. 65–77.

[6] Q. Fengcai, D. Jinsheng, and W. Li, "An online framework for temporal social unrest event prediction using news stream," in *2020 international conference on cyber-enabled distributed computing and knowledge discovery (CyberC)*, IEEE, 2020, pp. 176–182.

[7] Y. Jiang, B. Zhou, H. Tu, and L. Gao, "A temporal dual graph convolutional network for social unrest prediction," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 12005.

[8] C. Wu and M. S. Gerber, "Forecasting civil unrest using social media and protest participation theory," *IEEE Trans Comput Soc Syst*, vol. 5, no. 1, pp. 82–94, 2017.

[9] S. Banjac and F. Hanusch, "A question of perspective: Exploring audiences' views of journalistic boundaries," *New Media Soc*, vol. 24, no. 3, pp. 705–723, 2022.

[10] A. K. Buvanasri, R. Meenakshi, and S. Karthika, "Applications of Open Source Intelligence in Crisis Analysis—A COVID-19 Case Study," in *ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Volume 1*, Springer, 2022, pp. 313–325.

[11] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "Big data preprocessing," *Cham: Springer*, 2020.

[12] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[13] R. Rosa and Z. Žabokrtský, "Unsupervised lemmatization as embeddings-based word clustering," *arXiv preprint arXiv:1908.08528*, 2019.

[14] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[15] E. Pedhazur, "Multiple regression in behavioral research: Explanation and prediction. 3rd Harcourt Brace," *Fort Worth, TX*, 1997.

[16] D. S. Young, *Handbook of regression methods*. CRC Press, 2018.

[17] B. Barton and J. Peat, *Medical statistics: A guide to SPSS, data analysis and critical appraisal*. John Wiley & Sons, 2014.

[18] A. A. T. Fernandes, D. B. Figueiredo Filho, E. C. da Rocha, and W. da S. Nascimento, "Read this paper if you want to learn logistic regression," *Revista de sociologia e politica*, vol. 28, p. 006, 2021.

[19] P. Tabacof and L. Costabello, "Probability calibration for knowledge graph embedding models," *arXiv preprint arXiv:1912.10000*, 2019.

[20] Ž. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021.