ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL
Volume 7 (Issue 1) (2023) Pages 47-51

Towards Election Forecasting Using Sentiment Analysis: The Zambia General Elections 2021

Yasin Musa Ayami[a] and Mayumbo Nyirenda[b]

Department of Computer Science, University of Zambia, Lusaka, Zambia

a. yasin.ayami@cs.unza.zm b. mayumbo.nyirenda@cs.unz.zm

Abstract

Forecasting of election results is one of the key activities prior to elections. In Zambia, like many other countries, opinion polls have been used to predict the outcome of elections since 1999. During the run up to the 2021 general elections, two opinion polls were conducted. One poll suggested that HH would emerge victorious whilst the other predicted that ECL would emerge victorious. Actual results announced on the 16th of August 2021 by the Electoral Commission of Zambia (ECZ) had HH obtaining 59.02% of the votes. The variance in the two opinion polls leaves room for alternative approaches to predicting election results. This study proposes sentiment analysis as part of the initial stage to building an alternative solution to predicting the outcome of an election. The study analysed sentiments shared on social media during the build up to the August 2021 general elections. A total of 3,519 tweets were scrapped from Twitter and sentiment analysis was performed on the tweets. The findings of the study reveal that as election day drew closer, there was an exponential increase in the number of tweets that were posted on a daily basis. Further, our analysis of the tweets revealed that the majority of the tweets were neither positive nor negative (they were neutral) in line with the Afrobarometer opinion poll. Topic modelling was subsequently also performed on the tweets using BERTopic. Some of the topics learnt include voter engagement, the shutdown of the internet and the election day. Initial findings are promising to drive towards election forecasting using sentiment analysis.

Index Terms - Forecasting, Natural Language Processing, Sentiment Analysis, Social Media, Elections, Opinion Polls.

I INTRODUCTION

In Zambia, opinion polls have been used to predict the outcome of elections since 1999. During the run up to the 2021 general elections, two opinion polls were conducted. One was by the Political Science Association of Zambia (PSAZ) and the other by Afrobarometer. The opinion poll by PSAZ suggested that Edgar Chagwa Lungu (ECL), would get 40.4% and the opposition leader, Hakainde Hichilema (HH), would get 30.33%. Meanwhile, an opinion poll conducted by Afrobarometer had a representative of 1,200 Zambians drawn across all the ten provinces. The results of this opinion poll showed that support for ECL declined to 22.9%; 25.2% said that they would vote for HH and 45.6% refused to answer [1]. However, HH was declared president elect of Zambia on the 16th of August, 2021 by the Electoral Commission of Zambia (ECZ) after obtaining 59.02% of the while ECL obtained 38.71% of the votes [2]. The variance in the opinion polls and the actual results indicate there is room for more accurate predictions.

Artificial intelligence (AI) has been used by scholars elsewhere to predict rainfall [3], diseases [4,5], financial fraud [6] and many phenomena. This study aims to use AI to analyse the sentiments that were expressed online in relation to the 12th August Zambia general elections. To do this we first extracted selective data from Twitter before subjecting it to data cleansing. We then analysed the frequency of tweets as the election date got closer. After this, analysed the polarity of each tweet and finally analysed the sentiments of the tweets. Using the analysed tweets, we extracted topics that were of relevance to the Zambia 2021 Presidential elections. We believe that the work presented in this study is a fundamental initial step towards performing a trend analysis which can ultimately lead to an alternative approach to election results forecasting. Therefore, our analysis sought to answer the following research questions:

- What were the tweeting patterns of Zambians in relation to the elections?
- What was the polarity of sentiments shared in relation to the elections?

- What were the significant topics during the election period?

We present the rest of the study as follows: section 2 presents literature on the use of sentiment analysis to forecast elections; section 3, presents methodology taken and techniques used to conduct this study; section 4 provides the findings while section 5 discusses the findings; and finally section 6 concludes the study and makes recommendations for future work.

II. Literature Review

According to [7], social media, particularly Twitter can be used to forecast elections. However, a study by [8] on the usage of Twitter to predict the outcome of elections suggested that tweets are more reactive rather than predictive. The authors further asserted that Twitter can be used to generate 'buzz', but this 'buzz' cannot translate into a victory. However, research by other scholars like [7] that use sentiment analysis show promise of election forecasting using tweets. Two key areas are important in this process. These are sentiment analysis and topic modelling.

A. Sentiment Analysis

Sentiment analysis measures the mood of online conversations, and provides insight into the emotion behind the words by categorizing tweets into positive, neutral or negative categories[14, 15]. Sentiment analysis can be considered as a classification process involving three main classification levels which include: document level, sentence level, and aspect level. Presently, sentiment analysis is being applied in various domains including social media, health care, management and many other cases [14].

B. Topic Modelling

Topic modelling is an unsupervised learning technique that aims to group similar documents based on the tokens that are present in the document [16]. The authors [16] explain that it is particularly well suited for use with text data; however, it has also been used for analysing bioinformatics data, social data and environmental data. The Latent Dirichlet Allocation (LDA) is the most used technique for topic modelling [17]. LDA classifies text into a document and the words per topic, these are modelled based on the Dirichlet distributions and processes. However, topic modelling has several shortcomings which include noise sensitivity, and instability which can result in data which is unreliable; some techniques are also not representative of real-world data relationships [16].

In the last few years, there has been a huge surge in using neural networks to deal with complex, unstructured data. Language is inherently complex and unstructured. Therefore, there is a need to use models with better representation and learning capability to understand and solve language tasks [9]. It is on the basis of such studies that focus on how AI can be used to predict the future taking into consideration natural language processing that we build this study.

III. Materials and Methods

A. Data Source

The dataset for this study consisted only of a corpus of tweets that were extracted from Twitter. These tweets were embedded with multimedia data such as emojis, audio, images and videos. However, this multimedia data was not considered in this study. This is because audio and visual processing require different forms of analysis which are beyond the scope of this study.

A total of 3,519 were extracted from Twitter using six hashtags which were frequently used during the election build-up. These hashtags included: '*#ZambiaDecides2021', '#ZambiaVotes2021', '#ZambiaElections2021', '#ZambiaDecides', '#ZambiaVotes'* and *'#ZambiaElections'*. After this data cleansing and filtering was performed on the extracted tweets.

B. Data Cleansing and Filtering

Data cleansing is an operation performed on data to remove anomalies in the data. After extracting the data from Twitter, the data underwent various cleansing processes which included the removal of Uniform Resource Locators (URLs), emojis and special characters, retweet (RT), user mention (@), and unwanted punctuation. The study further converted tweets to lowercase and removed stopwords (i.e., words that did not express any meaning, such as is, a, the, he, them, etc). The data collected for this study was between the day the mandate for ECL began, the 13th of September, 2016, and the eve of the elections of 2021 that is the 11th of August, 2021.

C. Sentiment Analysis

The Valence Aware Dictionary and sEntiment Reasoner, also known as VADER was used to analyse the polarity of sentiments of the extracted tweets. VADER is a lexicon and rule-based model for the analysis of textual data. It measures the sentiment and polarity of the text by examining a list of lexical features (e.g. words) which are labelled according to their semantic orientation as either positive, negative or neutral. This is because VADER not only gives information about the positivity and negativity score but it also informs on the polarity or degree of positivity or negativity of a sentiment [10].

D. Topic Modelling using BERTopic

To extract topics, this study made use of BERTopic. According to [11], BERTopic is a topic model that extracts coherent topic representation through the development of a class-based variation of term frequency–inverse document frequency (TF-IDF). BERTopic generates topic representations through three steps. Firstly, a pre-trained language model is used to convert a document to its corresponding embedding. Secondly, the size of the embeddings is reduced through a process called dimensionality reduction to optimize the clustering process.

Lastly, from the clusters of documents, topic representations are extracted using a custom class-based variation of TF-IDF.

IV. FINDINGS

A. *Tweeting trends*

The findings of the study reveal that as the election day drew closer, the number of tweets exponentially increased. For example, over 75% of the total sample were tweets that were posted 11 days before the elections. A similar pick is observed during the 2016 elections. This is illustrated in Fig 1.

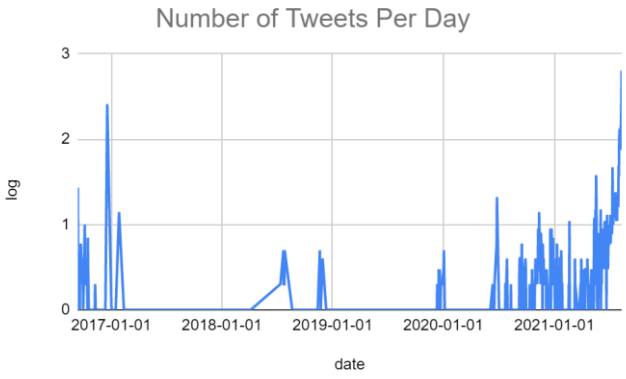

Fig. 1. Summary of the number of tweets from 2016 to 2021 using the log function

B. *Sentiment Analysis*

The study made use of VADER to analyse the polarity of the sentiments. As shown in Fig 2, 64.5% of the tweets were classified as neutral, 24.7% were positive and 10.6% were negative.

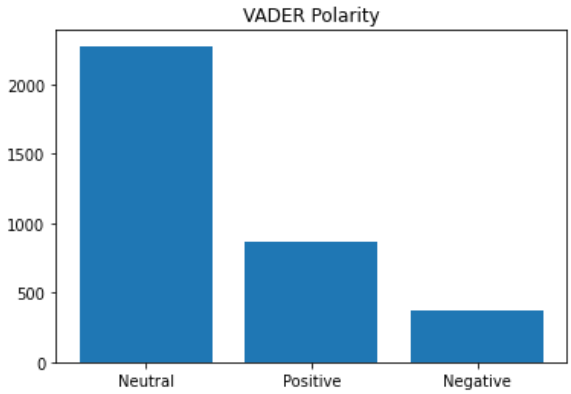

Fig. 2. Polarity of tweets using VADER

After classifying the tweets into positive, negative and neutral, the study made a deeper analysis of each of these polarities using BERTopic. At this point the analysis was focused on classifying the tweets into topics and sentiments.

1) Negative Sentiment Analysis: The keywords in Table 1 were the top words that appeared after doing a sentiment analysis on the negative tweets. Interestingly, the following keywords: internet, online, shutdown and elections appeared both in the TF-IDF and BERTopic keywords.

TABLE 1. Negative keywords using TF-IDF and their corresponding TF-IDF score

Keyword	TF-IDF Score
internet	0.35
Online	0.22
Election	0.18
shutdown	0.14
Threat	0.10

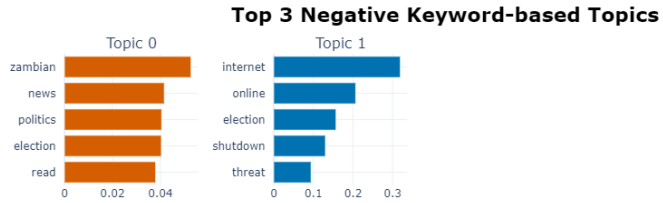

Fig. 3 Top 3 Negative Keywords using BERTopic

2) Neutral Sentiment Analysis: Based on the analysis of the neutral sentiments as shown in Table 2 and Fig 4, it was observed that as the election day drew closer, there was anxiety and excitement among the electorate, the majority of whom were voting for the first time. This can be evidenced from the number of tweets that were tweeted everyday as the election day drew closer.

TABLE 2. Neutral keywords using TF-IDF and their corresponding TF-IDF score

Keyword	TF-IDF Score
Sleep	0.73
Live	0.57
Decides	0.48
Week	0.43
Cry	0.35

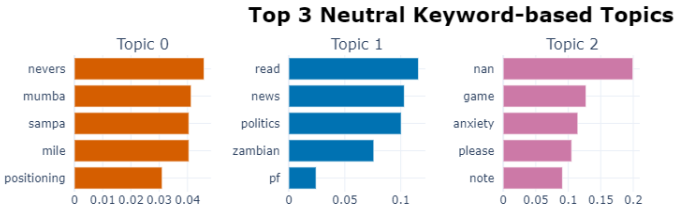

Fig. 4. Top 3 Neutral Keywords using BERTopic

3) Positive Sentiment Analysis: Tweets that were classified as positive were mostly about voter engagement encouraging electorates to vote wisely.

TABLE 3. Positive keywords using TF-IDF and their corresponding TF-IDF score

Keyword	TF-IDF Score
mwamba	0.26
bishop	0.26
politics	0.20
news	0.19
life	0.18

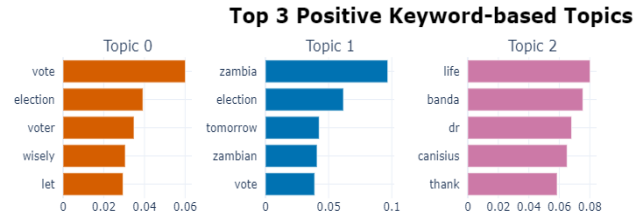

Fig. 4. Top 3 Positive Keywords using BERTopic

V. DISCUSSION

From the findings a particular trend of importance is that tweets increase towards and just after elections. This can be explained by the moment towards elections as well as anxiety to get results just after voting. This implies that the electorate tend to use social media more towards the elections and thus social media is a good source of data for sentiment analysis. It can be further noticed from these preliminary findings that the majority of tweets were of neutral sentiment in line with the opinion poll conducted by Afrobarometer in which most users declined to say who their preferred candidate was. During the build up to the elections, various media houses such as Lusaka Times[1] reported that the Zambian government had resolved to restrict access to the internet by completely shutting down the internet beginning the voting day [12] which came to pass on the election day. This is captured as one of the topics in the topic modelling. Sentiment analysis also shows that there was negativity in the sentiments related to the internet shutdown. According to [13], the government mismanaged the economy, citizens' freedom of expression was suppressed and police brutality was on the increase. Resulting from this, electorates were looking forward to the polls hence the huge voter turnout. This is also evident in topics related to voting and elections which had highly positive sentiments. The findings of this study are in line with the report that was published by [1,2], where it was reported that the voter turnout was high and the majority were youths. A study by [1] further reported that ECL would lose the elections.

VI CONCLUSION AND RECOMMENDATIONS

The study sought to analyse the sentiments that were expressed online in relation to the 12th August Zambia general elections. A total of 3519 tweets were extracted from Twitter and sentiment analysis was performed on them. Our analysis of the data revealed that the majority of the tweets were neither positive nor negative (they were neutral) in line with the Afrobarometer opinion poll. The findings further reveal that as the election day drew closer, there was excitement among electorates. Sentiment analysis with regards to mined topics clearly is in line with the election results. This shows that further study can culminate into an alternative form of election forecasting. Arising from this work, this study recommends the following as future work:

1) Explore the possibility of using keywords instead of hashtags. During the extraction of data, we noticed that some tweets did not contain any hashtags.

2) The model used in this study to analyse the polarity of tweets (VADER) is rule based, future work can explore using pretrained models that are trained using actual tweets.

[1]https://www.lusakatimes.com/2021/08/05/zambia-to-shut-down-the-internet-on-voting-day-as-facebook-urges-government-to-keep-internet-open/

3) Explore alternative sources of text to be used for sentiment analysis.

REFERENCES

[1] O. Magasu, "Credibility of an Opinion Poll: The Case of the 2021 General Elections in Zambia," 2022.

[2] E. C. o. Z. (ECZ), "General Election Results and Statistics – Electoral Commission of Zambia," 2021

[3] Mzyece, L., Nyirenda, M., Kabemba, M. K., & Chibawe, G. Forecasting Seasonal Rainfall in Zambia: An Artificial Neural Network Approach. Zambia ICT Journal, 2(1), 16–24, 2018

[4] Nyirenda M, Omori R, Tessmer HL, Arimura H, Ito K, Estimating the Lineage Dynamics of Human Influenza B Viruses. PLoS ONE 11(11): e0166107. https://doi.org/10.1371/journal.pone.0166107, 2016

[5] SH Chiwamba, J Phiri, POY Nkunika, M Nyirenda, An Application of Machine Learning Algorithms in Automated Identification and Capturing of Fall Armyworm (FAW) Moths in the Field, ICICT2018, Lusaka, Zambia, 2018

[6] K. Kamusweke, M. Nyirenda and M. Kabemba, "Data mining for fraud detection in large scale financial transactions", EasyChair, no. 1729, Oct. 2019, [online] Available: https://mail.easychair.org/publications/preprint_download/G5sK.

[7] A. Sharma and U. Ghose, "Sentimental analysis of twitter data with respect to general elections in India," *Procedia Computer Science,* vol. 173, pp. 325-334, 2020.

[8] D. Murthy, "Twitter and elections: are tweets, predictive, reactive, or a form of buzz?," *Information, Communication & Society,* vol. 18, no. 7, pp. 816-831, 2015.

[9] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli, "Exploring chemical space using natural language processing methodologies for drug discovery," *Drug Discovery Today,* vol. 25, no. 4, pp. 689-705, 2020.

[10] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014, vol. 8, no. 1, pp. 216-225.

[11] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794,* 2022.

[12] L. Times. "Zambia to shut down the internet on voting day as Facebook urges government to keep internet open." (accessed. 12th November 2022)

[13] H. Siachiwena, "A SILENT REVOLUTION Zambia's 2021 General Election," 2021.

[14] P. P. Patil, S. Phansalkar, and V. V. Kryssanov, "Topic modelling for aspect-level sentiment analysis," in Proceedings of the 2nd International Conference on Data Engineering and Communication Technology, 2019: Springer, pp. 221-229.

[15] T. A. Rana, Y.-N. Cheah, and S. Letchmunan, "Topic Modeling in Sentiment Analysis: A Systematic Review," Journal of ICT Research & Applications, vol. 10, no. 1, 2016.

[16] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," Information Systems, vol. 94, p. 101582, 2020.

[17] H. Jelodar et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78, no. 11, pp. 15169-15211, 2019.