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

Yasin Musa Ayami
Department of Computer Science
University of Zambia
Lusaka, Zambia
yasin.ayami@cs.unza.zm

Mayumbo Nyirenda
Department of Computer Science
University of Zambia
Lusaka, Zambia
mayumbo.nyirenda@cs.unz.zm

Abstract

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. 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. The findings of the study reveal that as the 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.

1 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. The opinion poll by Political Science Association of Zambia (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. This study aims to use AI to analyse the sentiments that were expressed online in relation to the 12th

August Zambia general 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.

We present the rest of the study as follows: section 2 presents the literature review; section 3, presents the methodology; section 4 provides the findings while section 5 discusses the findings; and finally, section 6 concludes the study and makes recommendations for future work.

2 Literature Review

According to [3], 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 [3] 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.

2.1 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 [4, 5]. 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 [4].

2.2 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 [6]. The authors [6] 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 [7].

3 Materials and Methods

3.1 Data Source

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.

3.2 Data Cleansing and Filtering

The subsequently 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. The data collected for this study was between the 13th of September, 2016, and the 11th of August, 2021.

3.3 Sentiment Analysis and Topic Modelling

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 [8].

To extract topics, this study made use of BERTopic. According to [9], 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).

4 Findings

4.1 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.
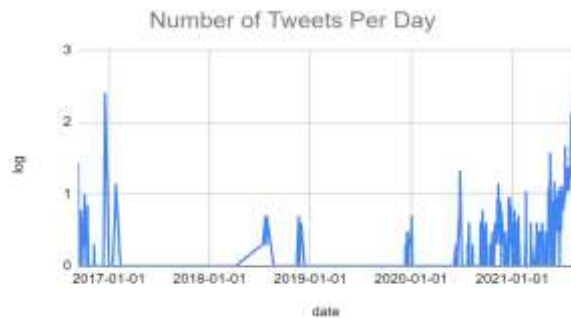

Figure 1: Summary of the number of tweets from 2016 to 2021 using the log function

4.2 Tweeting Trends

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.
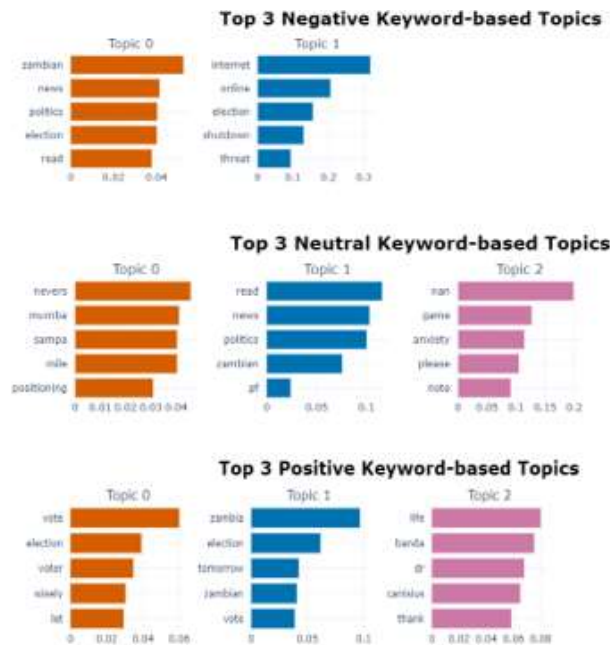

Figure 2: Summary of the positive, negative and neutral keywords from the sentiments.

3

5 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 reported that the Zambian government had resolved to restrict access to the internet by completely shutting down the internet beginning the voting day [10] 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 [11], 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.

6 Conclusion and Recommendations

The study sought to analyse the sentiments that were expressed online in relation to the 12th August Zambia general elections. 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.
3. Explore alternative sources of text to be used for sentiment analysis.

References

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

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

[3] Sharma, A. & Ghose, U. 2020. Sentimental analysis of twitter data with respect to general elections in india. Procedia Computer Science, 173, 325-334.

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

[5] Rana, T. A., Cheah, Y.-N. & Letchmunan, S. 2016. Topic Modeling in Sentiment Analysis: A Systematic Review. *Journal of ICT Research & Applications*, 10.

[6] Vayansky, I. & Kumar, S. A. 2020. A review of topic modeling methods. *Information Systems*, 94, 101582.

[7] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.

[8] Hutto, C. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, 2014. 216-225.

[9] GrootendorsT, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

[10] TIMES, L. 2021. Zambia to shut down the internet on voting day as Facebook urges government to keep internet open [Online]. [Accessed].

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