Topic Modelling Of News Articles For Two Consecutive Elections In South Africa

1st Avashlin Moodley
CSIR
University Of Pretoria
Pretoria, South Africa
amoodley1@csir.co.za

2nd Dr. Vukosi Marivate
University of Pretoria
CSIR
Pretoria, South Africa
vukosi.marivate@cs.up.ac.za

***Abstract*—In election cycles, the political-themed articles published by news providers present a rich source of information about election discourse. Extracting useful themes from a large article corpus manually is infeasible, text mining techniques such as topic modelling provide a mechanism to automatically infer themes from a corpus of text. Exploring the coverage of a single election period uncovers topical discourse that is relevant to current affairs in that election period. Analysing two consecutive election periods allows one to analyse the evolution of discourse from one period to another. Articles published by News24 were sourced to conduct the analysis and answer the research questions set forth. The articles were cleaned and topic models were built to identify 20 latent topics. The articles are classified with their topic before a pairwise cosine similarity comparison is applied on topic corpora to identify similar topics between election periods. The results of this study provide important insights relating to the two election periods, some of these include: coverage of corruption-related content is consistent between the two election periods and most political-themed articles in this corpus address problematic themes.**

***Index Terms*—natural language processing, elections, topic modelling, cosine similarity**

I. Introduction

What can we learn from the media coverage of an election? Piecing together an election using media coverage has its benefits and shortcomings. The media, as part of the democratic process, provides information to the public about political parties and the election. Even when trying to be objective, the news providers do choose what to cover to engage with their audience. Covering everything is impossible, especially when you have a large multi-party system like South Africa. In this paper, we present an analysis of 2 election cycles in South Africa using Natural Language Processing (NLP) on news from one of South Africa's largest online news providers. We identify common and standout themes across both election periods. Furthermore, from personalities being mentioned, we look at the correlations between election results.

Election cycles present an interesting text mining problem. The text from news articles can be explored to identify latent topics in the data that highlight the coverage associated with parties and provide a window into important themes during the election cycle. Studies to understand the landscape of an election period is common. [1], [2] focused on analysing the 2016 United States (US) presidential election, both studies focused on news and were interested in understanding the themes of discussion, [1] opted for a text mining approach to infer topics whereas [2] chose to employ trained personnel to annotate topics manually. [3] performed a text mining study on the 2012 Korean presidential elections using tweets to understand the topics and networks of interactions present during that election period. Whilst the approach used in these studies may overlap, the value of each study resides in the analysis of the landscape in the context of the country and the election period being observed. Our study analyses the context of the election in South Africa and instead of focusing on a solitary election, we analyse two consecutive election periods.

The contribution made by this study is the application of topic modelling techniques, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorisation (NMF), to uncover latent themes within articles published by News24[1] preceding the 2014 and 2019 South African general elections. In South Africa, the three major political parties are the African National Congress (ANC), Democratic Alliance (DA) and the Economic Freedom Fighters (EFF). These three parties account for 90% of the votes in both elections. Therefore, the analysis is limited to these three parties. Our analysis aimed to answer the following questions:

- What were the prominently covered political topics in the 2014 and 2019 election cycle?
- How has the coverage evolved between two consecutive election periods?
- Does the news coverage of a political party correlate to the proportion of votes they receive?

The remainder of the paper is organised as follows: Section II provides a brief description of the data. In Section III the analysis process is discussed. This is followed by a discussion on the observations uncovered in Section IV. Lastly, the findings are summarised and a conclusion is presented in Section V.

II. Article Data

The News24 dataset consists of 53897 unique articles. The dataset contains articles from 2014 and 2019 for the first five months of the year (01 January - 31 May). News24 provided

[1]https://www.news24.com/

the corpus of articles for both periods to the team for use in this study. The data was received in comma separated value (CSV) format. There are 26087 articles from 2014 and 27810 articles from 2019. The analysis focuses on the election, therefore the full dataset was segmented by selecting articles that contained the name of the political parties, their acronyms or their leaders. This resulted in a dataset containing 7253 articles (3477 and 3776 published in 2014 and 2019 respectively).

III. Mining News Article Data for Latent Topics/Themes

When one looks back on an election cycle, we could have individuals read news articles and then extract the main themes discussed during an election, this is costly and time consuming. Applying text mining to automatically infer themes is more time efficient and results in objective results. The remainder of this section discusses the approach taken to analyse the News24 articles.

A. Data Preprocessing

Raw text data has a high-dimensional nature and is uninterpretable by machine learning algorithms [1]. The dimension of a corpus of text is determined by the number of unique tokens present in the vocabulary of the corpus [1]. A token is defined as a character sequence containing no whitespace. In order to be processed by machine learning algorithms, the text within an article needs to be represented by a sparse vector. The length of the vector is equal to the size of the vocabulary of the training corpus. The remainder of this section describes the dimensionality reduction, corpus enrichment and vector representation procedures undertaken to transform the raw text present in the articles into a format that is applicable for use with the machine learning techniques being explored.

1) Dimensionality Reduction: The first step in processing the text was to standardise the case of all the words to lowercase to avoid duplicate tokens. Stop words, common words within a language that are semantically meaningless, were removed from the corpus to reduce the vocabulary size. Language is complex and many words have extensions that provide little semantic difference to the root form of the word (for example, "evident" and "evidently" are used in different situations but possess the same meaning). Stemming was applied to convert words into their root form to reduce the vocabulary size. Some words in a text corpus can commonly occur across the corpus and add noise to the formulation of topics where the aim is to create heterogeneous topics. Similarly, having words that rarely occur can also negatively affect topics if the rare words are treated as important to the topic even though they provide little semantic value. To alleviate this, thresholds are applied to filter out words that occur less than 20 times and in more than 50% of the corpus. This reduces the size of the vocabulary and makes the distribution more dense as it removes the outlier words.

2) Corpus Enrichment: A combination of words close to each other may be more insightful than the individual words itself (for example, the bigram "ice cream" provides more semantic meaning to a sentence in comparison to "ice" and "cream" individually). Bigrams and trigrams were generated to enrich the tokens present in the corpus.

3) Vector Representation: A term frequency, inverse document frequency (TF-IDF) vectoriser is used to weight each token in the corpus. a TF-IDF weight can be broken down into two components: the term frequency (TF) and the inverse document frequency (IDF). Equation 1 depicts the term weight calculation for a token, $w_{x,y}$.

$$w_{x,y} = tf_{x,y} \times log(\frac{n}{df_x}) \tag{1}$$

where $w_{x,y}$ represents the weight for token x in document y, $tf_{x,y}$ represents the frequency of token x in document y (the TF component of TF-IDF). $log(\frac{n}{df_x})$ is the IDF component of TF-IDF where df_x represents the document frequency of token x in the corpus and n is the number of documents in the corpus [4].

The TF weighting in isolation will indicate the importance of a token to a document but is guilty of assigning importance to frequently occurring words that don't add any value to the semantic meaning of the document [1]. To counter this, the IDF component of the weighting provides a normalisation of the weight to favour terms that occur less frequently in the corpus to better distinguish tokens that may be of semantic significance to a document [1]. The TF-IDF weighting is applied to all tokens in all documents to create a document-term matrix that consists of rows that represent documents and columns that represent each term in the vocabulary.

B. Topic Modelling To Extract Latent Themes

Topic modelling is an unsupervised learning technique that aims to group similar documents based on the tokens that are present in the documents. The application of topic modelling techniques to a text corpus is done with two goals in mind, grouping similar documents together and reducing the dimensionality of the text corpus from a document-term matrix to a document-topic matrix [1]. A latent topic is defined as a set of words with weighted importance that define a theme that occurs across multiple documents in a corpus. To classify an unseen document, the document is transformed into a document-term vector and evaluated by a built topic model that produces a distribution of topics to represent the constitution of the document.

Latent Dirichlet Allocation (LDA) [5] is a commonly used algorithm for topic modelling. The premise behind LDA is to learn the relationship that exists between tokens, documents and a fixed set of topics by assuming a specific probabilistic model is responsible for generating documents [6]. Non-negative Matrix Factorization (NMF) is a matrix decomposition technique that has proven to be a useful for working with high dimensional data [7], [8]. The defining constraint of using NMF is that the data needs to be non-negative [7], which is the case when working with a document-term matrix built from a text corpus. Text mining studies have used NMF to perform document clustering and topic analysis [8], [9].

The technical details of LDA and NMF have been omitted for brevity, refer to [5], [6], [9] for more information on how LDA and NMF identifies latent topics from a text corpus. Both techniques have strengths and weaknesses, [6] found that LDA is better at learning descriptive topics whereas NMF is better at creating compact representations of topics for a corpus. This study explores both techniques to identify which performs better on the news articles.

Whilst topic modelling provides a quick way to understand the latent themes present in a text corpus, there are three caveats that require manual intervention. The number of topics is not automatically inferred and requires tuning to determine the optimal number of topics for a given corpus. The coherence score is commonly employed to gauge model performance of a topic model [6]. When evaluating a document with a topic model, a distribution is produced and not a single value to indicate the topic of the document. Human intervention is required to determine a threshold value applicable to assign a document to a specific topic. Other methods include selecting the maximum value from the distribution. A topic is represented as a weighted sum of important tokens, a topic label requires human intervention to interpret the weighted sum of tokens. In a coherent topic, it might be easy to identify the theme present in that topic however this can become difficult to decipher with weak topics.

Preliminary experiments conducted on the articles revealed that the optimal number of topics for the politically-themed articles was 20 topics for LDA and NMF for both election periods. The dominant topic was selected as the label for a document.

C. Topic Similarity

An important experiment for this study is to analyse the evolution of topics between election periods. Documents with a dominant topic that has a probability of less than 50% were excluded for this experiment. Filtering articles with a weak membership to a topic ensures that the corpus representing each topic consists of articles that strongly represent the topic. A pairwise comparison occurs between topic corpora by creating a TF-IDF vocabulary of the union of the two corpora, vectorising each corpus and performing a cosine similarity calculation [10] on the two resulting vectors. The result is a similarity score that can be used determine whether two text corpora are similar or dissimilar.

D. Election Coverage

The election coverage of a political party is calculated by dividing the sum of party keywords by the sum of keywords for all parties for an election period. The keywords consist of the party name, the acronym and the name of their leader.

IV. Observations & Discussion

This section highlights the observations that resulted from text mining. The rest of this section discusses the popular trigrams identified, the highlight topics, the topics that were similar to each other between election periods and the correlation between coverage and election results.

A. Popular Trigrams

Inspecting words that frequently occur together is one way to get insight into recurring themes in the dataset. In Figure 1 and Figure 2 the frequencies of the top 10 occurring trigrams in the data is illustrated for 2014 and 2019 respectively.

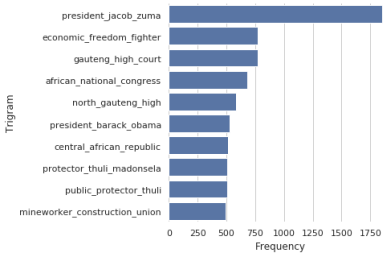

Fig. 1. 10 Most Frequently Occurring Trigrams for 2014 Articles

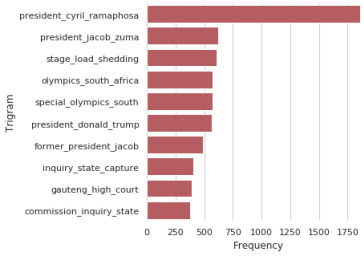

Fig. 2. 10 Most Frequently Occurring Trigrams For 2019 Articles

1) 2014: Many articles referenced the president at the time, Jacob Zuma. The trigrams in 2014 also reference two high courts, the public protector and the mining sector. The references to the public protector (Thuli Madonsela) and the president could be in relation to the release of the Nkandla report in 2014 which was controversially set to be released just before the election. Furthermore, some of the references relating to the president could be related to the government or the upcoming elections at the time. The mining references could indicate discussions about the Marikana massacre or a strike that may have occurred in 2014.

2) 2019: The 2019 articles indicated a similar trend where the current president, Cyril Ramaphosa, was the most frequently occurring trigram in the data. The trigrams are also well represented by the former president, Jacob Zuma. These references were expected due to his implication in the state capture enquiry and links to his reign over the country in discussions relating to the ANC. Load shedding and the special Olympics are also represented in 2019.

B. Important Topics

It is infeasible to discuss all the topics produced by the four models, intelligible and important topics are presented in Table I and II for articles published in 2014 and 2019 respectively.

1) 2014: The controversial Nkandla report published by the Public Protector features in both models. The NMF model appears to separate discussions on the report and the former

TABLE I
2014 TOPICS WITH THEIR DESCRIPTION AND A HUMAN GENERATED LABEL

Model	Topic	Keywords	Label
NMF14	1	report, madonsela, release, nkandla, public_protector, public_protector_thuli_madonsela, public, finding, upgrade, find	Nkandla Report
NMF14	2	da, march, zille, johannesburg, councillor, democratic_alliance, party, job, luthuli_house, supporter	DA
NMF14	6	malema, sars, tax, order, sequestration, trust, owe, sa_revenue_service_sars, provisional_sequestration, julius_malema	Malema/ SARS
NMF14	10	ramphele, agang_sa, zille, agang, announce, leader, presidential_candidate, mamphela_ramphele, agang_sa_leader_mamphela, merger	DA/ Agang
NMF14	11	eff, member, party, economic_freedom_fighter_eff, red, economic_freedom_fighter, stadium,house, mpofu, supporter	EFF
NMF14	7	zuma, president, president_jacob_zuma, nkandla, house, home, million, address, build, supporter	Nkandla / Zuma
NMF14	13	party, election, nfp, ifp, support, campaign, leader, member, cope, political_party	Small Parties/ Election
LDA14	0	malema, sars, tax, owe, debt, sequestration, order, sa_revenue_service_sars, provisional_sequestration, trust	Malema/ SARS
LDA14	10	madonsela, report, nkandla, zuma, public_protector_thuli_madonsela, upgrade, president, public_protector, release, committee	Nkandla Report
LDA14	11	party, da, vote, people, zuma, election, member, johannesburg, eff, government	Election
LDA14	17	ramphele, agang_sa, agang, zille, mamphela_ramphele, da, merger, agang_sa_leader_mamphela, presidential_candidate, da_presidential_candidate	DA/ Agang

TABLE II
2019 TOPICS WITH THEIR DESCRIPTION AND A HUMAN GENERATED LABEL

Model	Topic	Keywords	Label
NMF19	0	south_africa, government, country, investment, ramaphosa, people, south_african, economy, new, business	Economy
NMF19	1	eskom, power_utility, load_shed, power, unbundling, gordhan, announce, address, electricity, newsletter	Eskom
NMF19	4	ramaphosa, bosasa, corruption, mp, accuse, act, clean, malatsi, bench, someone	Bosasa
NMF19	6	mentor, former, commission, capture, saa, anc_mp_vytjie, flight, claim, zuma, evidence	State Capture
NMF19	9	da, party, vote, election, eff, people, committee, free_state, black, city	Election
NMF19	13	lekota, eff, shout, ramaphosa_write_special, sonadebate, jan_gerber_gerbjan_february, billion_health_minister, lekota_thing_smear, people, house	SONA
NMF19	18	malema, matter_move_high, court, case, challenge, matter, refer_gordhan_dog, white_monopoly_capital, gordhan_daughter, deputy_floyd_shivambu	EFF/ Gordhan
LDA19	2	event, gordhan, dam, load_shed, apple, stadium, cell_phone, last, signal, technical	Eskom
LDA19	6	south_africa, government, ramaphosa, country, investment, people, president, us, south_african, give	Economy
LDA19	8	party, da, eff, zuma, people, bosasa, leader, election, land, manifesto	EFF/ Election
LDA19	13	member_parliament, mantashe, mentor, ramaphosa_reply_debate, shark, debate_president_cyril, contact, ramaphosa_state_nation, club, former	SONA
LDA19	17	eskom, ramaphosa, country, pec, electricity, directorate, announce, load_shed, new, mtn	Eskom

president into two topics. Julius Malema, the leader of the EFF, and his battle with the South African Revenue Service (SARS) about tax in 2014 also features in both models. The proposed merger of the DA and Agang features in both models. The highlight themes uncovered from the 2014 articles relate to the Nkandla report, Julius Malema, the election and the proposed merger of the DA and Agang.

2) 2019: Both, NMF and LDA, models for 2019 produced Eskom related topics with keywords referencing load shedding and the proposed splitting of the state owned enterprise into multiple units. Load shedding was a major problem in the months preceding the election. The need for economic growth in the country feature in both models, Cyril Ramaphosa emphasised this in many of his speeches. The NMF model isolated a topic about the battles between the EFF and Pravin Gordhan whereas the LDA model was able to isolate a topic about the EFF relating to the election, their stance on land expropriation and discussions relating to Zuma and Bosasa controversies. The NMF model was able to extract intelligible topics relating to Bosasa and state capture enquiry discussions surrounding former member of parliament, Vytjie Mentor. Both models also captured topics relating to the state of the nation address. The themes uncovered in 2019 related to current affairs happening in the country, with the most notable coverage focusing on corruption, the need for economic growth and Eskom frailties.

C. Topic Distributions

The topic distributions of the 2014 and 2019 LDA and NMF models are illustrated in Figure 3. The LDA models appear to produce large dominant topics that cluster large proportions of the articles into a topic whereas the NMF models appear to have more compact topics. Manual inspection of the topics produced by both models indicate that both models produce a set of intelligible topics. Both of the dominant topics produced by LDA relate to the election and contains keywords about political parties. The segmentation of the data on political party keywords could be accountable for this, further experiments with these keywords removed might result in more compact topics for LDA.

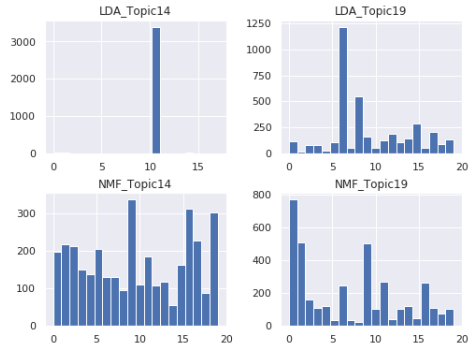

Fig. 3. Topic Distribution for the 4 models

D. Topic Similarities

Heat maps were created to illustrate the similarity between topics created for both election periods. Figure 4 and Figure 5 illustrate the evolution of topics for the LDA and NMF models respectively. The LDA heat map shows the dominant topic 11 in 2014 exhibiting similarities to many topics from 2019. The strongest of these similarities occur with topic 6 and 8. Inspecting Table I and Table II reveals that all three topics reference the election or government. The LDA model for 2014 is only represented by 11 topics in Figure 4. The 9 missing topics were excluded because none of the documents labelled with these topics were dominated by the topic (i.e. the probability was less than 50%). The dominant topic in the 2014 LDA model most likely accounts for a significant share of the probability for those articles.

The high similarity seen in Figure 5 between topic 2 of 2014 and topic 9 of 2019 for NMF can be attributed to the

presence of keywords relating to the DA, both topics reference the DA but one is in relation to the election and the other in relation to the DA's march to Luthuli House in 2014.

These observations indicate that the topic similarity technique applied can be used to measure similarities between two text corpora, however, more targeted experiments are required to validate this claim.

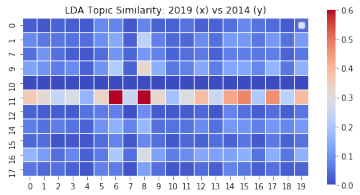

Fig. 4. LDA Topic Similarity: 2019 (x) vs 2014 (y)

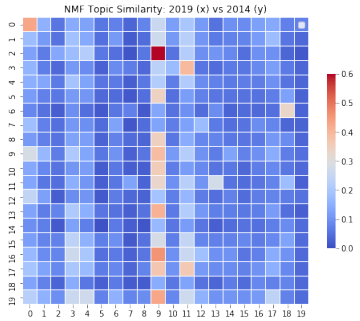

Fig. 5. NMF Topic Similarity: 2019 (x) vs 2014 (y)

E. Election Coverage vs Election Result

News coverage can either help a party gain or lose votes. Table III lists the coverage and vote percentages received by each party. The increase in coverage for the ANC from 2014 to 2019 has an inverse relationship to the votes the party received. The EFF experienced the opposite situation (their coverage decreased whilst their votes increased) and the DA experienced a decrease in both coverage and votes. Manual inspection of the topics in Table I and Table II describe problems rather than progress. In light of this, more coverage of negative events could be counterproductive to a party's ambitions in an election. Whilst this can be used as a rationale to explain the correlation between coverage and results for the ANC and EFF, it does not explain the relationship seen for the DA. Furthermore, the coverage only includes the top three parties whereas the results include the lesser parties that contested the election.

V. Conclusion

In this study, NMF and LDA topic models were created to uncover latent themes in the News24 articles published during the 2014 and 2019 election period. The analysis uncovered themes relating to corruption, the need for economic growth and Eskom frailties in the 2019 corpus. The models applied to the 2014 corpus uncovered themes relating to the Nkandla report, Julius Malema's tax battle and the proposed merger of the DA and Agang. The topics produced in the analysis indicate that NMF produce more compact topics than LDA. A pairwise cosine similarity calculation was done between the corpora of two topics which indicated that the topics identified to be similar contained a similar theme and keywords. The coverage of the ANC and the EFF showed an inverse relationship between coverage and votes, whereas the DA saw a decrease in both coverage and votes.

TABLE III
Political party coverage vs election result for the 2014 and 2019 election periods

	2014		2019	
Party	**Coverage (%)**	**Result (%)**	**Coverage (%)**	**Result (%)**
ANC	53.85	62.15	64.29	57.50
DA	28.01	22.23	18.76	20.77
EFF	18.13	6.35	16.94	10.79

A. Future Work

We plan on processing and analysing tweets to understand the discussions on Twitter during the election period. Thereafter, we plan on contrasting the article topics with the tweet topics using NMF to determine whether a relationship exists between them. We also plan on analysing the interactions on Twitter to the articles to understand the propagation of politically-themed information.

References

[1] G. C. Calafiore, L. E. Ghaoui, A. Preziosi, and L. Russo, "Topic analysis in news via sparse learning: a case study on the 2016 us presidential elections," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 13 593 – 13 598, 2017, 20th IFAC World Congress.

[2] T. E. Patterson, "News coverage of the 2016 general election: How the press failed the voters," 2016.

[3] M. Song, M. C. Kim, and Y. K. Jeong, "Analyzing the political landscape of 2012 korean presidential election in twitter," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 18–26, 2014.

[4] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf* idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[6] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 952–961.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[8] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, "Text mining using non-negative matrix factorizations," in *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 2004, pp. 452–456.

[9] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[10] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, 2008, pp. 9–56.