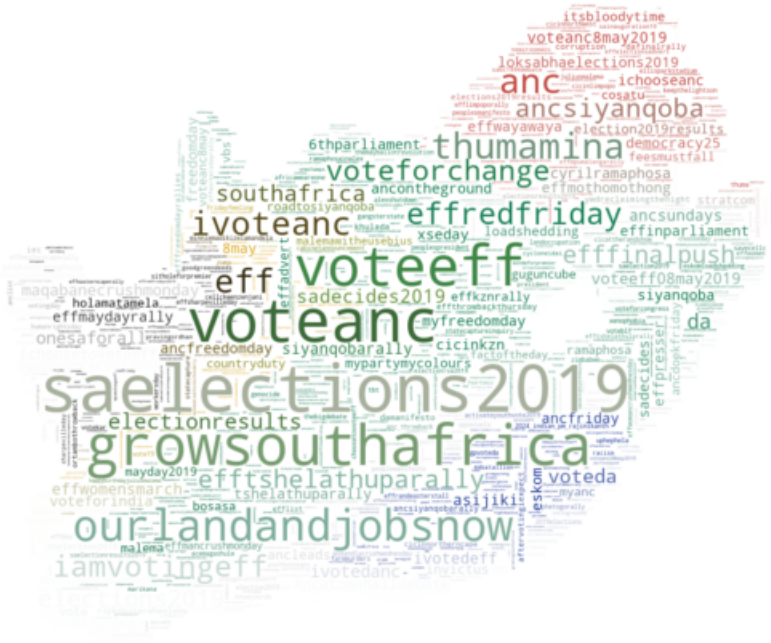
Ukhetho: A Text Mining Study Of The South African General Elections

by

Avashlin Moodley



Submitted in partial fulfillment of the requirements for the degree
Master In Information Technology (Big Data Science)
in the Faculty of Engineering, Built Environment and Information Technology
University of Pretoria, Pretoria

December 2019

Publication data:

Avashlin Moodley. Ukhetho: A Text Mining Study Of The South African General Elections. Masters Dissertation, University of Pretoria, Department of Computer Science, Pretoria, South Africa, December 2019.

Ukhetho: A Text Mining Study Of The South African General Elections

by

Avashlin Moodley
E-mail: avashlin@gmail.com

Abstract

The elections in South Africa are contested by multiple political parties appealing to a diverse population that comes from a variety of socioeconomic backgrounds. As a result, a rich source of discourse is created to inform voters about election-related content. Two common sources of information to help voters with their decision are news articles and tweets, this study aims to understand the discourse in these two sources using natural language processing. Topic modelling techniques, Latent Dirichlet Allocation and Non-negative Matrix Factorization, are applied to digest the breadth of information collected about the elections into topics. The topics produced are subjected to further analysis that uncovers similarities between topics, links topics to dates and events and provides a summary of the discourse that existed prior to the South African general elections. The primary focus is on the 2019 elections, however election-related articles from 2014 and 2019 were also compared to understand how the discourse has changed.

Keywords: Election analysis, text mining, natural language processing, latent dirichlet allocation, non-negative matrix factorization

Supervisors : Dr. Vukosi Marivate
Department : Department of Computer Science
Degree : Master of Information Technology (Big Data Science)

Acknowledgements

I would like to express my appreciation to the following people and organisations for their continued support throughout this degree:

- CSIR, for their financial support and time allowances that helped me to complete this degree.
- My supervisor, Dr. Vukosi Marivate, for opening my eyes to the possibilities and opportunities available. I am also grateful for the encouragement, guidance and support he has afforded to me throughout this endeavour.
- My girlfriend, Tiffany Mari, for her understanding and support throughout this stressful period of my life. I am truly grateful for everything you did to help me cope with this endeavour.
- My family, for encouraging me and motivating me to overcome obstacles I encountered in this process.
- My colleagues and friends who listened to me speak about this project continuously and provided valuable feedback to help me complete it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The discourse of an election is both noisy and informative. In a multi-party context like the South African election where the population expresses diversity in wealth, race, culture and socio-economic backgrounds, the election discourse presents an interesting mix of campaigns and rhetorics. The abundance of text generated about the election presents an interesting text mining task to uncover latent information that describes the events and campaigns that occurred. Election-related text is broadcast via many information streams. Different information streams contribute different perspectives to the election discourse. It would be insightful to summarise different streams of information to better understand the themes present in the discourse of an election from multiple perspectives.

In a democratic election, every eligible and willing individual needs to decide on who they would prefer to vote for. Political parties canvas voters by spreading their rhetoric on what they will do if elected to office. In a small town, perhaps its possible for candidates to engage with voters individually, however on a larger scale like a country's presidential election, it is not possible for political parties or their candidates to engage with every potential voter individually. Voters receive their election-related information from many sources: media agencies publishing news articles and stories about the election, public

figures expressing their opinions about the election, political discussions between friends or directly from political parties through distribution channels such as rallies, social media and email.

Media agencies provide coverage for a wide range of information relating to many genres. In an election period, media agencies provide a source of information about the election for potential voters to consume. The election coverage in news articles is a curated perspective of events and campaigns that are deemed newsworthy by the media agency who publishes it. With finite resources, media agencies are selective in what they cover, with many factors contributing to their election coverage strategy. Political parties don't have control over the rhetoric propagated about them in news articles.

Since Barrack Obama's successful use of social media in the 2008 United States presidential election, the use of social media campaigns by political parties have increased [17]. On social media platforms political parties can propagate their rhetoric directly to potential voters and engage in discussions with voters. The voting population can use platforms such as Twitter to voice their opinions on campaign policies, current affairs, events and many other aspects relating to the election and the country. Political parties have full control over the messages they propagate but they also become open to criticism from the public for the content they publish.

An election in a diverse country like South Africa presents a rich discourse environment with content appealing to the different demographics present in the country. A multi-party election environment further adds to the richness of the discourse since different parties adopt different policies based on their rhetoric. This presents a source of text data that is diverse, rich and noisy. Summarising and indexing the discourse of an election manually can be time consuming, costly and subject to human bias. Natural language processing techniques provide an automated and deterministic approach to uncover valuable information amidst noisy and irrelevant text data. Text data contains characters that form words, or more generally tokens. Tokens are sequences of characters separated by whitespace. Text is high dimensional with each unique token considered to be a dimension, thus a text corpus with 500 unique words in it's vocabulary is considered

to inhabit 500-dimensional space. Natural language processing techniques such as topic modelling aim to reduce the dimensionality of a text corpus by finding latent groupings that cluster documents together based on the similarity of the vocabulary contained in the documents. Once the documents have been clustered, it becomes easier to understand the themes that occur in a text corpus. If we examine the example of 1000 documents being grouped into 10 topics, understanding the 10 topics are easier than understanding the 1000 documents.

Text mining studies that aim to understand the context of an event are common, more specifically many studies have been conducted to understand the discourse of elections across the world [12]. News articles and tweets have commonly been analysed in the context of elections however not many studies perform contrasting analysis between the two information streams. Furthermore, studies of this nature occur frequently in other countries but there are limited prior studies [2] aiming to understand the text contained in South African election discourse.

In this study, election-related news articles and tweets are analysed to uncover latent information that describes the stories, campaigns and events that are relevant in the election discourse. Election-related articles from 2014 and 2019 are analysed using topic modelling techniques to understand the discourse and the relationship between election periods. Tweets curated by election-related keywords are collected and analysed to understand the themes present in the 2019 election discourse on Twitter and the campaign strategies of political parties for the 2019 election. Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) are the topic modelling techniques used in this study to identify themes that are present in these corpora.

The analysis is the focal point of this study, however, the tweet corpus collected represents a record of the 2019 election discourse that can be used for numerous other research endeavours. 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.

1.1 Motivation

The South African general elections creates a rich and diverse election discourse environment due to the diversity of the population, the political parties and the contrasting socioeconomic backgrounds of voters. The insights derived from the election discourse supports the understanding of the content and themes deemed relevant in an election period.

Text mining studies analysing election discourse in South Africa are rare, the only other study found in the literature [2] analysed the 2014 election from the perspective of news articles and tweets. Analysing the discourse of an election provides a window into the themes of discussion that was deemed important by the media and users on Twitter. This study aims to analyse the information present in the 2019 election discourse in both articles and tweets. Furthermore, a comparison between articles from 2014 and 2019 provides a novel insight into how election discourse has evolved in news articles between consecutive elections.

The development and evaluation of suitable election discourse analysis techniques can give rise to: a more informed voter, if deployed proactively; a historic summary of election discourse surrounding a particular election and insight into the evolution of election discourse over multiple elections.

1.2 Objectives

This study aims to uncover latent themes of discussion that emanate from the discourse related to the South African election in articles and tweets. Topic modelling is applied to the text corpora to uncover themes that are present in the corpora. Furthermore, a contrast between the topics uncovered from the tweets and articles is conducted to understand the relationship between two different information streams contributing to the election discourse. The objective of this study is to provide insights to help answer

the following questions:

1. What insights can be extracted from the news articles?
   (a) What were the prominently covered topics in the 2014 and 2019 election-related articles?
   (b) How has the discourse in articles changed from 2014 to 2019?
2. What contrasts about LDA and NMF can be derived from the experiments?
3. What insights can be extracted from the tweets collected?
   (a) Can descriptive topics be produced from the short text present in tweets?
   (b) What were the prominent topics in the Twitter election discourse?
4. What insights can be extracted from political party campaigns on Twitter?
   (a) What themes are central to the Twitter campaigns of the ANC, DA and EFF?
   (b) What events can be uncovered from the party tweets?
   (c) What similarities exist between the tweets of the ANC, DA and EFF?
5. Does a relationship exist between the articles and tweets?

1.3 Contributions

This study makes the following novel contributions:

1. The collection of a tweet corpus with curated content related to the 2019 South African general elections.
2. An analysis of themes uncovered from the 2014 and 2019 election-related news articles.

3. An analysis of the Twitter discourse landscape of the 2019 election.
4. An analysis of Twitter campaigns conducted by political parties for the 2019 South African elections.
5. An analysis of the relationship that exists between articles and political party tweets in an election period.
6. A discussion on the discourse surrounding the 2019 election from the perspective of tweets and articles.

1.4 Derived Publications

The following publication was derived from the work done in this study:

- Avashlin Moodley and Vukosi Marivate. Topic Modelling Of News Articles For Two Consecutive Elections In South Africa. *In proceedings of the 6th International Conference on Soft Computing Machine Intelligence*, pages 131-136, 2019.

1.5 Dissertation Outline

This study contains the following structure:

- **Chapter 2** provides a review of election-related text mining studies that have been conducted. The techniques relevant to this type of analysis are discussed briefly to justify the decision to perform the analysis done in this study.
- **Chapter 3** discusses the technical details of the algorithms used in this study. This provides the reader with a working understanding of how the techniques work.

- **Chapter 4** introduces the data used in this study. This chapter discusses the collection process and privacy protocol employed to collect election-related tweets. A summary of the articles are also presented in this chapter.
- **Chapter 5** explores the articles and tweets to provide baseline insights about the data. These baseline insights provide a reference point for other observations to build upon.
- **Chapter 6** provides an account of the experimental process employed in this study. The experiments are briefly introduces to set the tone for the chapters to follow.
- **Chapter 7** discusses the experiment that focused on uncovering themes contained within the 2014 and 2019 news article election discourse. The experiment aims to meet the objectives of the study by addressing a subset of the questions that relate to the articles.
- **Chapter 8** provides an account of the analysis performed on the tweets. This experiment focused on the breadth of the Twitter data collected to understand to overarching themes present in the Twitter election discourse.
- **Chapter 9** contrasts the Twitter election campaign of political parties to understand the rhetorics propagated for the 2019 election. Furthermore, this chapter provides insights into the events and campaigns related to parties during the observed period.
- **Chapter 10** builds topic models on a combined corpus containing articles and party tweets. The experiment analyses issue-centric topics to identify patterns between articles and tweets that may explain the relationship that exists between the two data sources.
- **Chapter 11** combines the insights from all the experiments to present a discussion on the insights uncovered from this study. The discussion addresses insights about each data source and provides an analysis of the relationship that exists between the articles and tweets in the 2019 election discourse.

- **Chapter 12** summarises the findings and discusses the limitations of this study. A brief discussion is also presented on future work that can extend on the work in this study.

Chapter 2

Literature Review

Election cycles present an interesting text mining problem. An election provides a rich information environment with different actors playing different roles in the environment [2]. Political parties develop policies and campaigns to appeal to the interests of voters in an effort to obtain more votes. In a large scale election like the national election, it is not possible for parties to reach large proportions of voters directly. Since it is more effective and easier to reach the voting population through the media, political parties rely on media coverage, social media and advertisements to make the voting population aware of their policies and campaigns.

News articles are one such information stream that inform the voting population prior to them casting their vote. The information found in news articles are professionally written, curated and deemed newsworthy by the media agency that publishes it. Whilst news articles provide information to the voting population, the information is filtered and it is not possible to cover everything. Social media has become increasingly popular for election campaigns since Barrack Obama's successful 2008 election campaign in the United States of America (USA) [17]. Social media allows political parties to engage directly with their target population. Furthermore, social media allows the voting population to engage with political parties and share their views with others. The text from news articles and tweets can be explored to identify latent topics in the data that

highlight the coverage associated with parties. The news articles and tweets present two contrasting perspectives to view the election discourse and better understand the context of the election.

This chapter discusses the election-related studies that were done globally in Section 2.1, describing the election under review, the data that was used and the techniques that were applied. Thereafter, prior work for South African elections are discussed and contrasted with the objectives of this study. This study focused on using topic modelling techniques to understand the latent themes present in a text corpus, therefore in Section 2.3 topic modelling approaches used in election analysis are briefly discussed. Lastly, a summary of this chapter is presented in Section 2.4.

2.1 Political Discourse Analysis Globally

Political discourse, more specifically election discourse has been the subject of interest for many studies. [6, 21] focused on analysing the 2016 USA presidential election, both studies focused on news and were interested in understanding the themes of discussion, [6] opted for a text mining approach to infer topics whereas [21] chose to employ trained personnel to annotate topics manually. Whilst the approach used in these studies may overlap, the value of each study resides in the analysis of the information in the context of the country and the election period being observed.

In [23], the 2012 Korean presidential election was analysed from the perspective of tweets and news articles. LDA was applied to the tweets and articles. Furthermore, network analysis was done on the tweets. The tweets were collected using candidate names as keywords. [23] used the latent topics identified to understand the temporal properties that exist in the data by contrasting article and tweet timelines and tying topics to events. The network analysis aspect is not discussed since it is outside the scope of this study.

[1] studied the 2016 USA presidential election from the perspective of Facebook, specifi-

cally fake news stories shared on Facebook. [1] analysed 150 fake news stories and found that social media is a significant contributor to fake news spread in the context studied. Facebook data was not an option for this study since ethical clearance was only obtained for the collection of tweets. Furthermore, [1] had an objective to understand the spread of fake news on social media whereas this study focuses on understanding the discourse.

In [16], the 2010 USA midterm election is studied from the perspective of tweets. The tweets were subjected to graph mining, text mining and user profiling. [16] performed naive NLP tasks such as visualisations of top terms and top hashtags. Network analysis of the networks associated with candidates were studied to understand the overlap between parties [16]. [16] also applied LDA to the tweets but this resulted in poor quality topics due to the short length of the tweets.

European Union (EU) parliamentary speeches were studied in [10]. [10] contrasted LDA and NMF by measuring the associated coherence scores (more information on the coherence score is provided in Section 3.2.1), their results found that NMF produced more coherent topics than LDA. [10] also applied a two-layer NMF technique called dynamic-NMF which modelled the evolution of topics over multiple periods. Dynamic-NMF was considered for this study but due to time constraints and the need for modifications to apply the technique to this study, it has been excluded from this study and reserved for future work.

[8] analysed tweets from the 2018 USA midterm election to understand political ideologies of Twitter users and to detect bots in the discourse. The bot detection in [8] had the objective of detecting political manipulation using social media. [13] sought to understand political popularity of Bernie Sanders (a politician from the USA) by applying LDA to tweets. [13] measured the importance of topics in discourse related to Bernie Sanders. [26] aimed to understand the sentiment in tweets related to the 2009 German federal election. [26] observed that Twitter is a platform for political discussion. The application of sentiment analysis in [26] aimed to understand the nuances related to the election campaign.

2.2 South African Election Discourse Analysis

Globally there has been many data mining studies that aim to understand certain aspects of an election that is relevant in the context of those studies. In the South African context, studies of a social science nature analyse the election from a social, economic and political perspective. [19] analysed comments made on politically-themed articles in 2014 produced by multiple media agencies in South Africa. [19] aimed to understand the issues discussed in the comments and the concerns of the public in a qualitative manner. [14] provided a reflection on the 2019 South African election in a qualitative manner, addressing the election from a social science perspective. [14] discussed focal points of political party campaigns. The context of [14, 19] are not directly relevant to this study, however the discussions contained within provide perspectives that can be contrasted with the observations made in this study as future work.

[2] was a text mining study that focused on analysing the election discourse related to the 2014 South African general elections. [2] focused on analysing news articles and tweets for the 2014 election. A correlated topic model [3] was applied in [2] to uncover latent topics from the article and tweet corpora. This study resonates with [2] because both studies analyse articles and tweets, focus on the South African election and make use of topic modelling. Both studies aim to analyse election-related discourse, however they differ in the following ways:

- [2] aims to understand emerging democracies using the South African election as a use case whereas this study aims to understand the information that can be uncovered from text to explain the events that unfolded in the discourse of a South African election.
- The analysis in this study examines tweets from a broader context as well as political party tweets whereas [2] only analysed political party tweets.
- The tweets in this study were from 2019 whereas the tweets from [2] where from 2014.

- This study looks at news articles from 2014 and 2019 whereas [2] analysed 2014 articles only.
- [2] had a variety of news article sources whereas this study focused on articles published by News24.
- This study employs multiple topic modelling techniques whereas [2] applies a solitary approach.

2.3 Topic Modelling Techniques

Studies that focus on uncovering information from text using text mining techniques [2, 6, 10, 13, 16] tend to make use of topic modelling in their studies. This study operates on two text corpora in the form of articles and tweets therefore topic modelling is a relevant approach for this study.

LDA [4] is a popular choice for topic modelling, LDA is a probabilistic model that identifies topics in a corpus of text. Another popular technique for topic modelling is NMF [15], NMF is a matrix decomposition technique that finds two matrices, a document-topic matrix and a topic-term matrix, which represents the model. [24] contrasted LDA and NMF and found NMF to produce more compact topics. This study aims to apply both LDA and NMF to the article and tweet corpora to uncover latent information and provide an auxiliary comparison of these techniques.

2.4 Summary

Since Barrack Obama's successful use of social media in his election campaign in 2008 [17], social media has become increasingly popular in election campaigns around the world. Existing studies covering the South African election aim to analyse the election from a social science perspective. [2] shares similarities with this study however, the aim

of [2] was to understand emerging democracies using South Africa as a use case whereas this study aims to understand the election-related discourse and the contrasts between information streams to understand the information received by voters from news articles and tweets.

The use of text mining techniques to understand election-related discourse in a South African context is limited thus highlighting a gap in the literature to analyse the election-related discourse from a data science perspective for the 2019 South African election. The analysis of the 2019 South African election-related content provides a novel window into the information surrounding the election which is lacking in the literature. Furthermore, contrasting news articles between two election periods provides a novel perspective on the evolution of news articles published by News24.

This chapter provided a review of the literature related to text-based election analysis on articles and tweets. The topic modelling techniques applied in this study are introduced in greater detail in Chapter 3 to provide the reader with background on the semantics of the algorithms prior to introducing the data in Chapter 4.

Chapter 3

Technical Background

This chapter provides background information to help the reader better understand the chapters that follow. The topic modelling and model evaluation techniques that were used in this study are discussed in this chapter. No modifications were made to the respective algorithms for this study. The contents of this chapter aim to provide the reader with a brief understanding of how the algorithms work. The remainder of this chapter is organised as follows:

- Section 3.1 discusses the topic modelling techniques used in this study.
- Section 3.2 provides insight into the topic modelling evaluation metric used to determine the best model.
- Section 3.3 summarises the contents of this chapter.

3.1 Topic Modelling

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 [6]. A topic is a set of tokens with weighted importance that define a theme that occurs across multiple documents in a corpus [3].

3.1.1 Data Representation

Given a text corpus containing multiple documents, transform the corpus to a document-term matrix with size $n \times m$ where n represents the number of documents in the corpus and m represents the number of tokens in the vocabulary of the entire corpus. This matrix is called the document-term matrix, for simplicity it is referred to as A in this discussion.

3.1.2 Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is a matrix decomposition technique that has proven to be useful for working with high dimensional data [15, 22]. The defining constraint of using NMF is that the data needs to be non-negative [15], 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 [22, 28].

Algorithm

The application of NMF to matrix A produces two matrices, W and H, which approximates A, i.e. $A \approx WH$. The objective of applying NMF to matrix A is to minimize the reconstruction error between A and WH by minimizing the Frobenius norm of $A - WH$ [22, 24]. This becomes a classical optimisation problem in which W and H are iteratively updated until the error is minimised below a threshold value or a iteration limit is reached [10, 22]. Algorithms aimed at solving this minimization problem are often ran-

domly initialised [10]. This study follows the strategy of [10] by setting the initial values in W and H using the Non-negative Double Singular Value Decomposition (NNDSVD) initialisation approach presented in [5].

Interpretation

W represents a matrix that has dimension $n \times k$ where n maintains its earlier definition and k is a hyperparameter that represents the number of topics. Thus, W can be viewed as a topic membership matrix. For each document in the corpus, the corresponding row in W indicates the membership weights of each topic. H represents a matrix that has dimension $k \times m$. For each of the k rows in H, the columns contain the importance of each token in vocabulary, m, to the topic. Ordering a row in descending order will highlight the important words for the topic [10].

3.1.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [4] is a generative probabilistic model commonly used for topic modelling. LDA posits that every document in a text corpus is generated by a distribution of topics containing words in the vocabulary of the corpus, m [3]. LDA aims to learn the relationship that exists between tokens, documents and topics under the assumption that a specific probabilistic model is responsible for generating documents in a corpus [24]. LDA operates on an assumption that there is a fixed number of topics, k, that generate all documents in the corpus under inspection [3].

Algorithm

The model estimated by LDA consists of two matrices. The first matrix, Φ, represents the probability of selecting a particular token from sampling a topic for all tokens in the vocabulary of the corpus. The second matrix, Θ, represents the probability of selecting a particular topic when sampling a document from the corpus.

Φ is built by drawing samples for each topic from a Dirichlet distribution. Each topic is a multinomial distribution of tokens from the vocabulary, m [3, 24]. Since each topic in Φ contains m tokens, many tokens will have zero probabilities thus Φ is a sparsely distributed matrix.

Θ is built by sampling a Dirichlet distribution for each document. Each row representing a document contains probabilities that it is generated by the topic represented in the column [3, 24]. The topics in a document in Θ is a multinomial distribution over the k topics.

The values in Φ and Θ are iteratively updates in an optimization task until the model converges to an appropriate error or a maximum number of iterations have concluded.

Interpretation

The document-topic distributions, Θ, represents the proportion of each document that belongs to a topic, thus providing an indication of the topics present in each document. This is analogous to the W matrix from NMF. The topic-token distributions, Φ, represents the probability that a token in the vocabulary belongs to a certain topic. This distribution is analogous to the H matrix from NMF, providing an indication of the tokens that are important to a topic.

3.2 Model Evaluation

Topic modelling is an unsupervised learning technique. Unlike supervised learning, there is no source of truth to measure performance. In a knowledge discovery study like this, the objective is to uncover latent information from large text corpora. There is still a need for the unsupervised models to be evaluated to determine whether the model is a good representation of the corpus it is modelled on. Evaluation metrics such as perplexity and coherence are commonly used for evaluating unsupervised topic models

[24, 27]. [7] uncovered that intrinsic techniques such as perplexity are not always good measures of the semantic interpretability of topics. Topic coherence is the evaluation metric used in this study. There are many variants of topic coherence [20, 24], for this study the Topic Coherence Word2Vec (TC-W2V) [20] measure is used.

3.2.1 Topic Coherence Word2Vec (TC-W2V) Score

A Word2Vec (W2V) [18] model is built on a corpus of text. A W2V model can be built using the continuous bag of words (CBOW) or skipgram method [20]. The skipgram method is used in TC-W2V [20]. A W2V model is a word embeddings model that captures the semantic relatedness of tokens. This property allows for arithmetic and similarity calculations to be performed on vectors representing tokens. Before calculating the TC-W2V measure, the x top-ranked tokens for each topic in a topic model are extracted, where x is a positive integer. A pairwise similarity score is calculated for each pair of top-ranked tokens in a topic by calculating the cosine similarity [11] between the pair. The average of these scores represent the TC-W2V score for a particular topic. The TC-W2V score of the entire topic model is the average TC-W2V score of all the individual topic's TC-W2V score [20].

Application To Topic Models

The topic modelling techniques discussed in Section 3.1 requires that the value of k, the number of topics, is provided when building the model. The value of k is problem dependant [10] and requires tuning to identify the most coherent model for a corpus. A TC-W2V score is calculated for different values of k to determine the optimal value for k for a corpus. The most coherent model is used to perform further analysis of the topics within the model.

3.3 Summary

The topic modelling and model evaluation techniques were discussed in this chapter. The topic modelling techniques discussed were LDA and NMF. LDA is a probabilistic topic modelling technique whereas NMF is a matrix decomposition technique [24]. The evaluation method for topic models in this study was also discussed, the TC-W2V score is used to identify the most coherent model from a range of k values. The contents of this chapter provides the reader with background information on the algorithms to allow for the chapters that follow to focus more on the application domain of interest.

Chapter 4

Data

An election in a diverse country like South Africa presents a rich discourse environment with content appealing to the different demographics present in the country. The objective of this study is to inspect text corpora to uncover the themes present in the election discourse. The text corpora of interest are tweets and news articles. Both data sources provide contrasting perspectives, the articles reveal themes deemed newsworthy whereas the tweets provide a window into discussions held by political figures and the general population of the country. The articles can be classified as monologues that are broadcast to people to consume. Conversely, tweets can contain monologues and conversations between multiple parties. Articles are professionally written, contain formal language and are not restricted in length. The tweets on the other hand can contain formal or informal language and anyone can produce and broadcast a tweet. Tweets are restricted to 280 characters and thus are usually significantly shorter than articles.

This chapter aims to provide the reader with an understanding of the properties of each corpus. A summary of the fields and text contained in each corpus is discussed. A discussion is also presented on the tweet data collection process and privacy preservation strategy employed.

The rest of the chapter is organised as follows:

- Section 4.1 presents an overview of the articles and the curation process applied to select election-related articles.
- Section 4.2 discusses the data collection process and privacy preservation strategy for collecting tweets. This is followed by a summary of the tweet corpus.
- Section 4.3 provides a summary of this chapter, highlighting notable properties of the articles and tweets.

4.1 Articles

The articles used in this study were published by News24[1]. News24 provided the corpus in comma separated value (CSV) format. The corpus contained news articles published between 1 January and 31 May for 2014 and 2019. Every article published in this period was provided. Since News24 provides coverage for a broad spectrum of content covering many genres, election-related content needed to be curated from the corpus.

4.1.1 Raw Data Summary

The initial corpus contained 53 897 articles. There were 26 087 articles published in 2014 and 27 810 articles published in 2019. Each article contains 13 fields, the fields and their associated data types are listed in Table 4.1.

The Permatitle field is useful for filtering out duplicate articles. The publication date allows for temporal analysis to be performed. The title, synopsis and body fields provide rich text sources for topic modelling and other natural language processing (NLP) tasks. The category and keywords fields can be used as validation for clustering tasks or to perform supervised learning in future work. The article URL can prove useful in linking articles to tweets. The Boolean fields do not provide much value for this study.

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

Publish Date (Date)	Synopsis (String)
Permatitle (String)	Article URL (String)
Site Name (String)	Has Gallery (Boolean)
Category (String)	Has Video (Boolean)
Keywords (String)	Has Images (Boolean)
Title (String)	Has Audio (Boolean)
Body (String)	

Table 4.1: Data fields in the article corpus

4.1.2 Curation Process

The corpus provided by News24 contains articles covering a multitude of themes. This study is only interested in the election-related articles thus a curation strategy was employed to select relevant articles. The curation was done by selecting articles that contained a keyword from a predefined list of keywords. The name of a political party, their associated acronym and the name of the party leader were used as keywords. The study is limited to content relating to the ANC, DA and EFF thus the keyword list was short and contained nine entries. Duplicate articles were filtering out using the Permatitle field. Articles that were composed solely of media elements and contained no text were also removed. The curation resulted in a corpus containing 5 483 articles (2 202 and 3 282 published in 2014 and 2019 respectively).

4.2 Tweets

There are millions of tweets being published every second. Collecting tweets pertaining to an event needs to either be ring-fenced around a location, curated by a set of keywords or gathered by collecting tweets from accounts. In this study, a keyword list is defined that contains accounts and hashtags that are relevant to the South African election. The tweets were curated according to the defined list of keywords. Tweets were collected from

15 February to 15 May 2019. This section provides a discussion on the data collection process, the raw data extracted, the curation process and the privacy protection protocol being observed.

4.2.1 Privacy Preservation

The collection of information that can contain personal information for research purposes usually requires informed consent. In the case of collecting tweets from Twitter, it is unfeasible and impractical to provide each tweet publisher with an opt-out option for this research. A user with a public Twitter account understands the terms and conditions of Twitter, by making the decision to leave your account public, anyone can consume what you tweet at any time. This study has operated within the bounds of the Twitter privacy policy[2].

The following steps are taken in the collection of information to assure the confidentiality of Twitter users that contribute to the data set:

- Only publicly available tweets are collected and analysed.
- No individual information of the general population will be revealed. In accordance with the guidelines presented in [25], this study will only reveal information relating to political parties and election candidates as these are public entities and their publications are done with the objective of reaching a large audience and thus do not have privacy expectations for their content.
- Tweets from the general population will be analysed in an aggregated manner to uncover latent themes and not isolate specific accounts.
- Obfuscation of personal information like the scrambling or hashing of usernames and names will be applied before sharing content from the general population, in the event that this needs to occur.

[2]https://twitter.com/privacy?lang=en

- Results will not reveal personal information of individuals that are not political parties or public figures affiliated with political parties.

This privacy protection protocol together with the data collection discussed in the next section was submitted to the University of Pretoria ethics committee (Reference: EBIT/90/2019) for consideration and was granted approval.

4.2.2 Data Collection Process

The data collection involved implementing a Python module that encapsulated the Twarc[3] tool to extract tweets that contained relevant keywords. The keywords relevant in this study are depicted in Table 4.2. The keywords contain the ANC, DA and EFF party and leader accounts, hashtags linked to political party campaigns and election related hashtags. Multiple data collection modules were run in parallel to maximise the amount of tweets collected.

Effsouthafrica	#ANC	VoteEFF8May
Myanc	#DA	ElectionResults
Our_da	#EFF	SAElections
Julius_S_Malema	DAmanifesto	SAElections2019
CyrilRamaphosa	ThumaMina	Xse2019
MmusiMaimane	DAManifesto	Elections2019
Voteeff	VoteForChange	Ivotedanc
Voteanc	PeoplesManifesto	Ivotedda
Voteda	EFFManifesto	Ivotedeff
SAElections2019	VoteANC8May	ivoted
Xse2019	VoteDA8May	

Table 4.2: Twitter data collection keywords

Tweets that met the criteria were stripped of media related fields and stored in an

[3]https://www.docnow.io/

Elasticsearch[4] index. An Elasticsearch index is a non-relational database that stores information in key-value pairs. The next subsection discusses the raw data that was collected and stored.

4.2.3 Raw Data

There were approximately 10 million tweets collected in the 3 month period. Each tweet contained multiple fields, most of the fields were irrelevant for this study and were removed. Table 4.3 lists the fields that were kept. The *Full Text* attribute contains the text that is relevant to topic modelling.

User Creation Date (Date)	Mention Screen Name (String)
User ID (long)	Retweet Count (Numeric)
Screen Name (String)	ID (long)
Full Text (String)	In Reply To Screen Name (String)
Is Quote Status (Boolean)	In Reply To Status ID (String)
Tweet Creation Date (Date)	In Reply To User ID (String)
Retweeted (Boolean)	URL (String)
Hashtag Text (String)	Expanded URL (String)
Symbols Text (String)	Display URL (String)

Table 4.3: Data fields in the tweet corpus

Some of the tweets collected were not relevant to the election, there were tweets relating to the keyword 'DA' that were a mix of languages that contain 'DA' in their vocabulary (these include Russian and Portuguese). The next section discusses the process applied to isolate and remove irrelevant tweets.

Curation Process

The corpus contained irrelevant tweets that added noise to the data. These tweets contained text from languages outside of South Africa. These tweets were isolated to a

[4]https://www.elastic.co/

two-week period in April and resulted from a minor change in keywords that resulted in 'DA' being collected instead of '#DA'. The two week segment that contained noise was removed from the corpus, it contained 5.5 million tweets. The segment was subjected to a language identification module and a manually curated list of words seen in a wordcloud to remove irrelevant tweets. Approximately 4 million tweets were removed from the tweet corpus. Thereafter, the cleaned segment was merged with the rest of the corpus.

The decision to isolate the problem segment, remove it, clean it and reattach it was made because the removal method is imperfect. Applying this to the entire corpus would result in unnecessary data loss. Isolating the problem segment and applying the cleaning approach to that segment minimises the scope of data that can be lost. There is undoubtedly noise remaining in the corpus but the extent of the noise has been minimised. The resulting corpus contained 6.1 million tweets.

4.3 Summary

This chapter introduced the corpora used in this study. The articles were obtained from News24 and curated to select election-related content. The election-related content accounted for 10% of the total articles received from News24. The tweets were collected from Twitter by ingesting tweets that contained keywords of interest. The tweet corpus contained noise for a two-week period in April, the corpus was curated to remove the noise by isolating the problem period and applying language identification to remove tweets in foreign languages like Russian. This approach is not perfect and can potentially leak data thus it was only applied to the problem segment to reduce data loss.

The privacy preservation strategy employed with the tweets provided an overview of the ethical boundaries for this study. The strategy highlighted that this study focused on aggregate analysis and account-specific analysis is reserved for political parties and their affiliates. The curated corpora are used for the experiments and analysis in this study. The next chapter provides an account of the exploratory analysis conducted on the curated article and tweet corpora.

Chapter 5

Exploratory Data Analysis

In Chapter 4, the two corpora of interest were introduced. The curation protocol employed for the articles were discussed along with a summary of the contents of the corpus. The collection process, curation protocol and a summary of the tweets were also presented. This chapter focuses on uncovering what is contained within the corpora. Exploratory analysis provides a naive window into the corpora which contributes to the overall analysis of the discourse as well as providing baseline insights that the analysis in later chapters can leverage to explain observations.

The rest of the chapter is organised as follows:

- Section 5.1 presents exploratory insights derived from the article corpus.
- Section 5.2 discusses the exploratory insights derived from the tweet corpus.
- Section 5.3 provides a summary of this chapter, highlighting notable insights derived from the article and tweet corpora.

5.1 Exploring The Article Corpus

The exploratory analysis was done on the curated corpus discussed in Section 4.1. The exploratory analysis done on the articles aim to provide:

- Descriptive statistics about the article corpus.
- Insight into the temporal properties of the articles.
- Insight into prominent words and concepts within the corpus by applying naive NLP techniques.

5.1.1 Descriptive Statistics

This study focuses on text mining thus the corpus statistics in Table 5.1 focuses on the properties of the text. The vocabulary consists of all unique words that exist in the corpus. The vocabulary used in 2019 has decreased in size in comparison to 2014, this is perhaps an indication of denser themes in 2019. There were significantly more articles published in 2019 in comparison to 2014. The articles in 2019 have also become longer with the average number of words increasing. The political party mentions in this curated corpus paints a contrasting picture. The mentions are calculated by counting the number of occurrences of a party, it's acronym and it's leader has within an article. The ANC was mentioned in 25% of the articles in 2014, this increased to 66% of articles in 2019. The DA and EFF both saw declines in mentions in News24 articles from 2014 to 2019. The EFF received the most coverage in 2014, with the party being mentioned in 40% of the articles. The ANC received the most coverage in 2019.

	2014	**2019**	**2014+2019**
Number of Articles	2 202	3 282	5 484
Vocabulary size	78 266	47 318	106 988
Average number of words	392	556	491
Average number of characters	2 489	3 545	3 121
Average number of ANC mentions	0.25	0.66	0.50
Average number of DA mentions	0.29	0.09	0.17
Average number of EFF mentions	0.40	0.14	0.24

Table 5.1: Summary of the curated article corpus

5.1.2 Temporal Properties

Understanding the temporal properties of the article corpus can inform insights made with more complex analysis later on in this study. Figure 5.1 illustrates the distribution of articles published in 2014 and 2019. Election-related content started to appear from March in 2014, whereas in 2019 the election-related content started appearing much earlier with the first articles appearing in the middle of January. The moment of highest traffic in 2014 is at the start of May which coincides with the election date in that year. In 2019, a peak is seen at the start of February, which can most likely be attributed to the state of the nation address (SONA). Both election periods have a large mass of articles being published around the election, however 2014 has significantly more published in that period.

Another interesting aspect is the coverage that political parties receive. Figure 5.2 illustrates the mentions received by the parties in both election periods. In 2014, all parties experience a spike in coverage in the week of the election. The ANC is prominently mentioned in 2019 consistently. The DA and EFF are covered less in 2019 in comparison

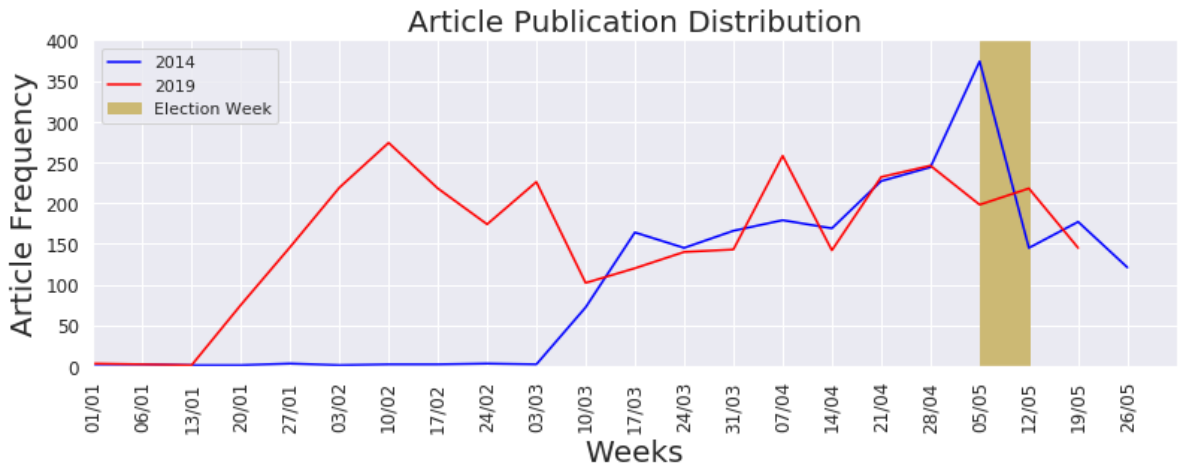


Figure 5.1: Distribution of article publications

to 2014.

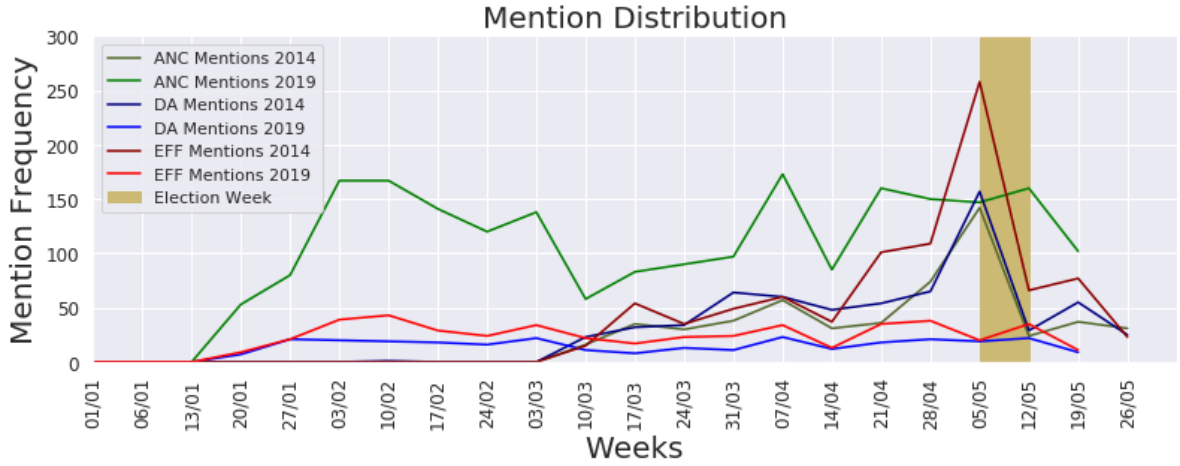


Figure 5.2: Distribution of party mentions in the articles

5.1.3 Naive NLP

Inspecting words that frequently occur together is a naive way to gauge the recurring themes occurring in a corpus. In Figure 5.3 and Figure 5.4 the frequencies of the top 10

occurring trigrams in the data is illustrated for 2014 and 2019 articles respectively.

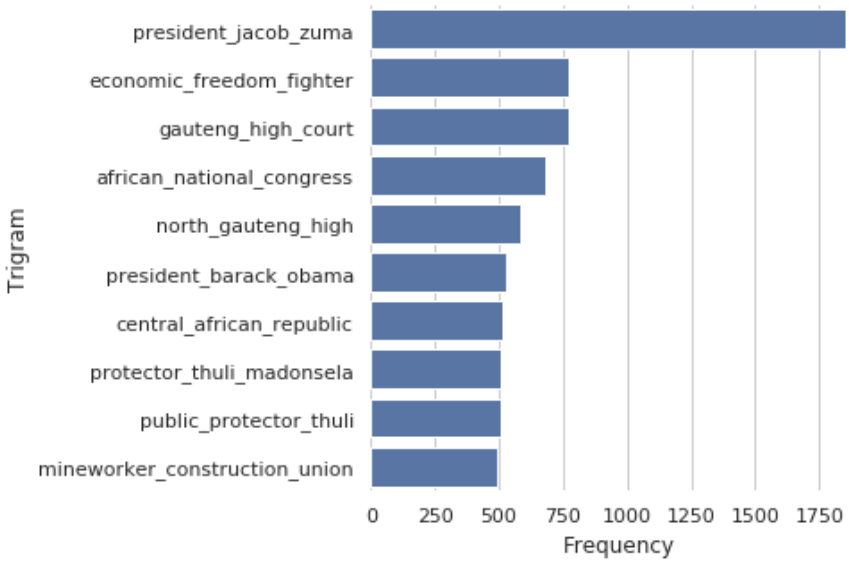


Figure 5.3: 10 most frequent trigrams for 2014 articles

In 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.

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 were also represented in 2019.

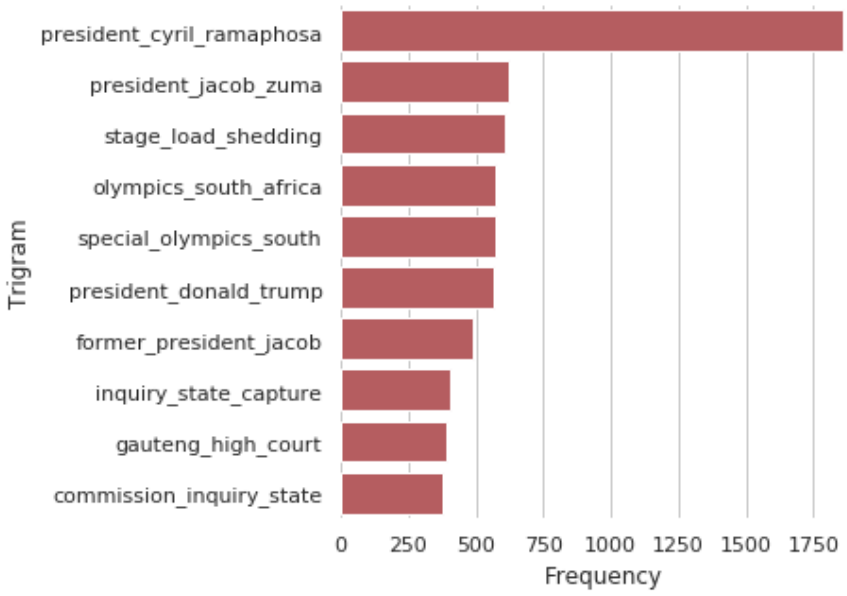


Figure 5.4: 10 most frequent trigrams for 2019 articles

5.2 Exploring The Tweet Corpus

The exploratory analysis is done on the curated tweet corpus discussed in Section 4.2. The exploratory analysis done on the tweets aim to provide:

- Descriptive statistics about the tweet corpus.
- An analysis of the temporal properties of the corpus.
- Preliminary NLP analysis done using naive techniques.

5.2.1 Descriptive Statistics

Statistics were generated for the entire corpus as well as sub-corpora for the political party tweets. These statistics are depicted in Table 5.2. The corpora containing tweets from political parties include tweets from the official party accounts as well as the party

	Corpus	ANC Tweets	DA Tweets	EFF Tweets
Number Of Tweets	6 131 747	8 331	6 105	11 646
Average Number Of Tokens	19	21	28	18
Average Number Of Characters	130	158	185	137
Vocabulary Size	2 395 944	18 477	17 522	25 738
Number Of Hashtags	1 720 850	4 979	3 143	6 990
Number Of Unique Hashtags	139 466	922	596	1 139
Number Of Mentions	5 614 404	3809	3 951	7 984
Number Of Unique Mentions	616 797	553	636	2 170

Table 5.2: Summary of the curated tweet corpus

leader's account. The corpus contains 6.1 million tweets with the parties accounting for a very small percentage of the tweets. This indicates that the majority of the discourse on Twitter originates from tweets from other sources, most likely composed of a large proportion of general population tweets. The average tweet consists of 19 tokens which is much smaller than the article corpus. The average tweet also consists of 130 characters which indicates that tweets being published don't contain much text and are most likely augmented by other media such as web links and images to deliver the intended message. The EFF were more active on Twitter in comparison to the ANC and DA. The ANC and DA published tweets that were longer than the average tweet. The vocabulary of the tweet corpus consisted of 2.4 million tokens and contained 1.7 million hashtags and 5.6 million mentions. The ratio between the number of mentions and the number of tweets in the corpus is 0.9. This means that on average every 10 tweets will contain 9 mentions which supports the argument that Twitter is a platform for political discussion in election periods.

5.2.2 Party Following

The collection process of the tweets implicitly provided information on the following of political parties, these values are depicted in Table 5.3. The following of a party provides insights into the extent of their public reach on Twitter. The information in Table 5.3

are representative of values at the end of May 2019.

	Followers	**# Tweets**
Cyril Ramaphosa	539 734	3 864
ANC	663 525	56 630
ANC + Ramaphosa	1 204 259	60 494
Mmusi Maimane	1 120 887	15 575
DA	556 001	96 168
DA + Maimane	1 676 8881	111 743
Julius Malema	2 421 143	29 574
EFF	761 950	45 192
EFF + Malema	3 183 093	74 766

Table 5.3: Political party follower and tweet count

The EFF had the largest Twitter following in comparison to their opponents, with the party account having 750 000 followers and their leader having almost 2.5 million followers, collectively reaching 3.1 million followers. The DA has the next largest following with the party account amassing just over 500 000 followers and their leader having slightly more than 1 million followers, collectively the DA had 1.6 million followers. The ANC has the smallest following with both the party and leader accounts amassing roughly half a million followers each, giving the ANC a collective following of 1.2 million. The party accounts produce more tweets than their leaders, this indicates that parties in SA approach public engagement on Twitter by communicating as a collective entity rather than leaders campaigning as individuals. Whilst leaders do support the campaigns of the parties, the campaigns are primarily shared via the party accounts.

The variation in the Twitter activity of leaders provides an indication of their usage of the platform. This is not limited to the election, but a general indication of their interactions with the public on Twitter. The DA and EFF leaders are significantly more active than the ANC leader on Twitter. Maimane is five times more active than Ramaphosa and Malema is ten times more active than Ramaphosa and twice as active as Maimane. The correlation between the level of activity and number of followers of

the party leaders indicate that their following is directly proportional with the volume of tweets they produce. Malema and Maimane have a higher following than their respective parties, this could be indicative of the public being more interested in engaging with party leaders. They could also use the platform for sharing aspects that are not of a political nature, it is social media after all. The contrast between the usage of Twitter between the leaders could be related to their ages, Maimane and Malema are under 40 and are probably more inclined to using social media whereas Ramaphosa, who is 67, is not as accustomed to the platform as his counterparts.

There are many contributing factors to the following of political parties and their leaders. Observations made on the data in Table 5.3 is an analysis of correlations and trends observed and are not causal explanations. The collection process did not track how the following of parties change during the election period, this could be a useful insight to derive in future studies of a similar nature.

5.2.3 Temporal Properties

Exploring the temporal properties of the tweets can lead to insights being uncovered about moments of high traffic that occurred in the election period. Figure 5.5 illustrates the distribution of tweet publications in the corpus. The volume of tweets increased throughout the period with the largest peak in volume occurring in the week of the election. After the election concludes, the volume of tweets around the election decreased significantly. Other peaks occur in the middle of March and towards the end of April.

Political parties and their leaders are central figures in the election discourse on Twitter as many of the engagements by the general population are targeted at these figures. The contrast between the tweets published by these figures and the mentions they receive provide an indication of the direction of the conversation in the discourse involving these figures. Figure 5.6 illustrates the tweet publishing behaviour of these figures. The illustration contains the tweets from the parties and the leaders individually as well as aggregated to show the contribution of the leaders to the discourse. Julius Malema

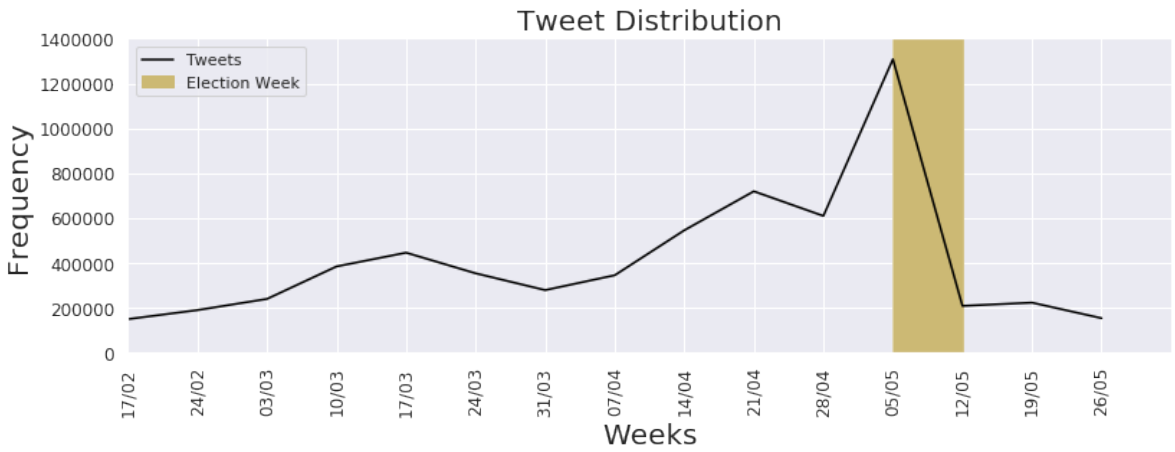


Figure 5.5: Distribution of tweet publications in the corpus

was the most active leader on Twitter with his Twitter activity in this period being significantly higher than his opponents. The ANC and DA leaders do not contribute much to the discourse and these parties appear to rely on the official account to spread their rhetoric. The EFF maintained the largest volume of tweet publication until two weeks before the election where the ANC's tweet volume significantly increased. The ANC's activity more than doubled from 14 April to 28 April, producing more tweets than the EFF from 21 April until the election. The DA was less active on Twitter in comparison to the ANC and EFF. Throughout the observer period, tweets from the political parties did not exceed 2000 in any given week.

The tweet publication behavior of political parties indicates the volume of their engagement on the platform. In contrast, the volume of mentions a party receives is an indication of the engagement directed towards a party. The illustration in Figure 5.7 depicts the volume of engagement received by political parties.

The EFF received the most engagement on Twitter for the majority of the period observed. The engagement received by the DA was minimal in comparison to their opponents. The party accounts were engaged with more than the leader accounts consistently throughout the period. All parties experience peak volume of mentions in the week of the election. The contrast between the mentions received and tweets published shows

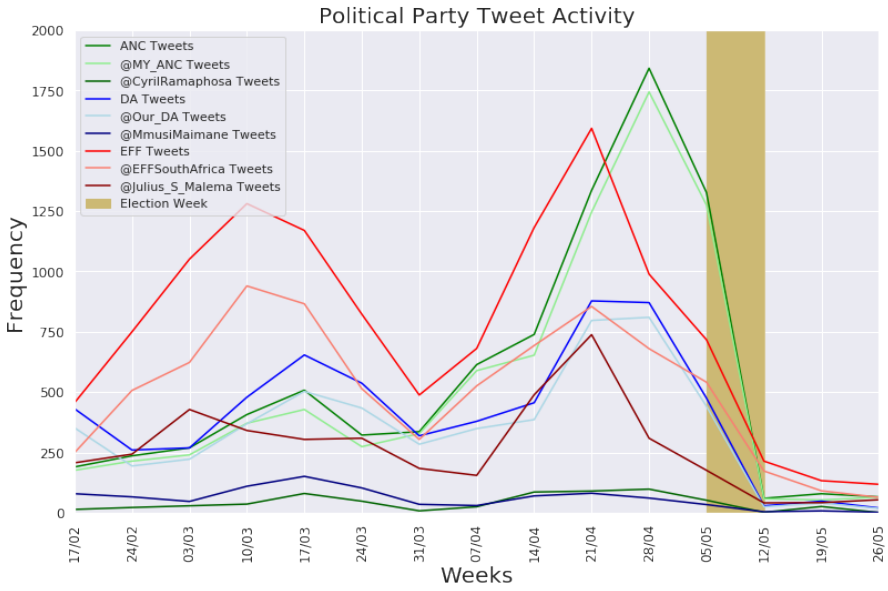


Figure 5.6: Twitter activity of political parties in the election period

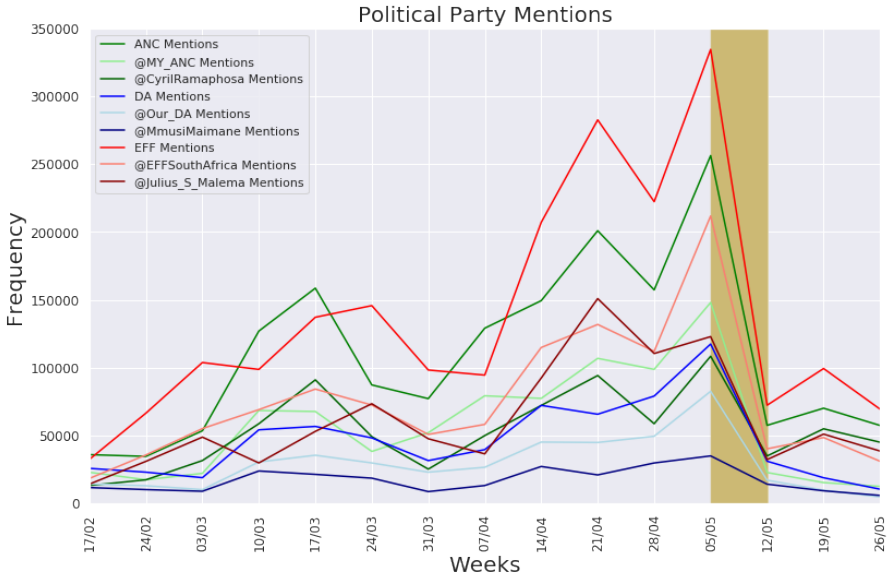


Figure 5.7: Political party mentions in the election period

that engagements on Twitter are directed towards the political parties.

5.2.4 Naive NLP

Wordclouds provide an indication of frequently used terms in a text corpus. The size of the terms in the wordcloud is determined by the frequency of the term occurring in the corpus. The wordclouds were restricted to contain a maximum of 2000 words. Building wordclouds on hashtags provide a good initial view into themes present in the tweet corpus. The use of hashtags in tweets is done to indicate that the tweet belongs to a user-defined or existing theme. A tweet publisher can make use of hashtags to express their opinion relating to a topic or to start a discussion with others on the matter. Evaluating the hashtags contained in the tweet corpus provides a window into high-level discussion points that exist in the discourse. Furthermore, the hashtags contained in tweets produced by political parties provides an indication of themes central to their Twitter campaigns. The density of the wordclouds for the political party tweets is indicative of the breadth of each party's Twitter campaign.

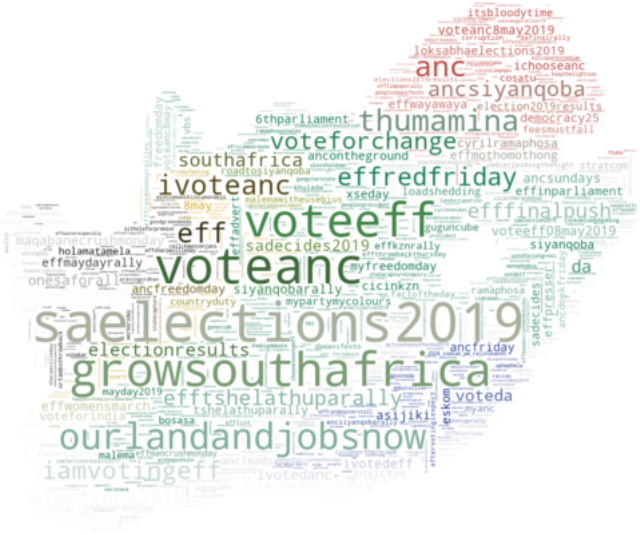


Figure 5.8: Corpus-wide hashtag wordcloud

The wordcloud in Figure 5.8 highlighted the set of frequently used hashtags in the corpus. The election hashtag *SAElections2019* is the most used hashtag in the corpus. ANC and

EFF related hashtags are prominent in the corpus whereas the DA related hashtags are minimally represented in the wordcloud. The tweet and mention activity related to the ANC and EFF that were illustrated in Figure 5.6 and Figure 5.7 show that the volume of activity surrounding the ANC and EFF was more than that of the DA thus explaining why ANC and EFF related hashtags are more prominent in the wordcloud.

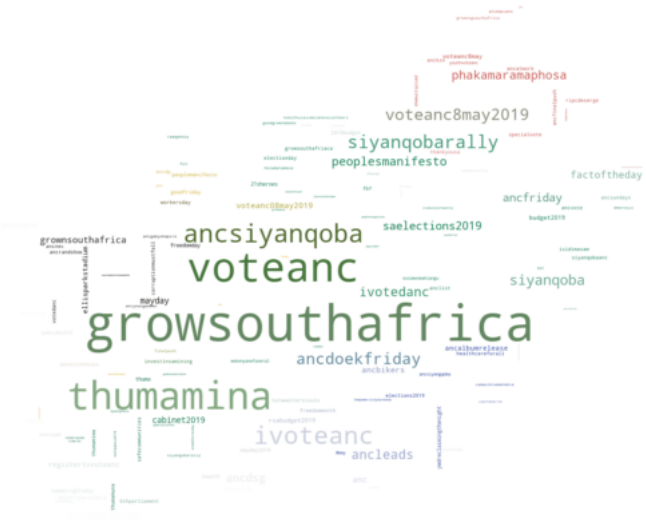


Figure 5.9: ANC hashtag wordcloud

Analysing the hashtags at a corpus level provides a window into the global themes present in the corpus. Analysing the tweets produced by political parties provide a window into the campaign strategy of the party. The wordcloud in Figure 5.9 highlights the hashtags used in tweets by the ANC. Campaign slogans like *ThumaMina*, *VoteANC* and *GrowSouthAfrica* are the most frequently used hashtags by the ANC. The ANC wordcloud is sparse, containing few hashtags which indicates that the ANC Twitter campaign was narrow in scope.

The hashtags used in the DA's Twitter campaign is illustrated in Figure 5.10. The DA made use of more hashtags in comparison to the ANC. The DA made use of slogan hashtags such as *OneSAForAll*,*KeepTheLightsOn* and *VoteDA* in their campaign. Hashtags

Figure 5.10: DA hashtag wordcloud

related to *Eskom* and *Bosasa* are prominent which indicates that the DA were expressive of their views on these matters as part of their campaign. The DA's campaign had a broader scope than the ANC campaign but the volume of tweets produced by the DA was small in comparison to their opponents.

The EFF hashtag wordcloud is illustrated in Figure 5.11. The hashtags central to the EFF's campaign were *OurLandAndJobsNow*, *EFFRedFriday* and *VoteEFF*. The EFF wordcloud also contains hashtags related to EFF related rallies and events. The EFF's hashtag usage was wider than their opponents. The EFF was also the most active party on Twitter, which also explains why their wordcloud is more dense and covers a wide range of hashtags dealing with slogans and events.

The corpus level wordcloud provided a high-level view of hashtags that were central themes in this corpus whereas the party wordclouds provided an indication of the hashtags that were used as central themes in the campaigns of the parties. Analysing party hashtags also provided an indication of the breadth of a campaign employed on Twitter. The overlap in hashtag usage by the top 3 parties is minimal. All parties have a

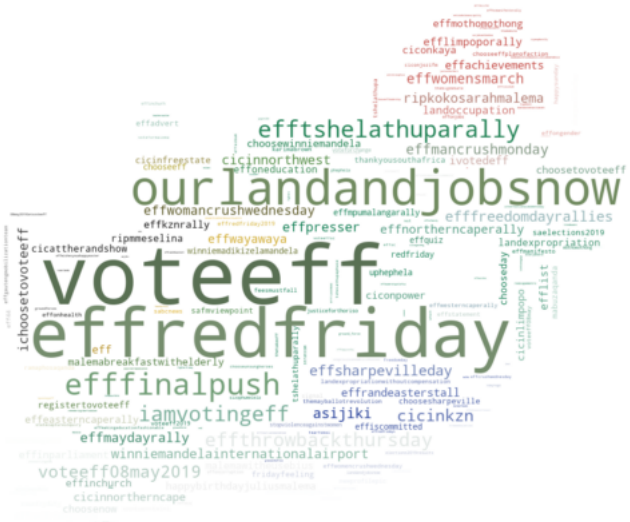


Figure 5.11: EFF hashtag wordcloud

few hashtags that are central to their campaign, which are propagated in many of their tweets. The DA hashtags indicate discussions around national problems and scandals whereas the ANC and EFF hashtags are more focused on the election campaign, with no problem-related hashtags being highlighted in their respective wordclouds.

5.3 Summary

The exploratory analysis conducted on the articles focused on providing descriptive statistics on the text contained in the corpus, analysing temporal factors relating to publication behavior and party coverage. Thereafter, trigram frequencies were analysed to identify high-level concepts and terms that are important in the article corpus in both 2014 and 2019. The exploratory analysis uncovered that the ANC received the most coverage in 2019, the ANC's coverage increased from 2014 whereas the coverage received by the DA and EFF had decreased. The election-related articles in 2014 start appearing in the discourse from March whereas in 2019, the election-related articles start appear-

ing from January. The trigrams highlight themes relating to the Public Protector, state capture and Eskom issues.

The exploratory analysis performed on the tweets was similar to the analysis performed on the articles. Descriptive statistics on the tweet corpus found that tweets contain on average 19 tokens which is significantly shorter than the average length of the articles which was about 500 tokens. The temporal analysis applied uncovered that the EFF and ANC were the most active parties on Twitter with the EFF consistently publishing tweets in the observed period and the ANC producing a last minute push on Twitter, surpassing the EFF's tweet frequency in the last two weeks prior to the election. The DA was the worst performing party in terms of article and tweet coverage. Their agenda was not propagated on these platforms to the extent of their competitors.

Wordclouds were generated to illustrate the frequently used hashtags in the entire corpus, the themes of the ANC and EFF are prominent in the wordcloud whereas the DA related hashtags are minimally represented. Wordclouds were also created for the tweets published by each party and provided a window into the hashtags that were central to each campaign. The ANC's hashtag vocabulary was narrow and indicates a campaign focused on attracting votes. The DA's hashtag vocabulary contained campaign slogans as well as hashtags relating to problems in the country. The EFF's Twitter campaign focused on campaign slogans as well as unique hashtags for rallies and events held by the EFF in the observed period.

The insights derived in the exploratory analysis provides a basis for further analysis and discussions about the election discourse present in both corpora. The next chapter provides insights into the experimental setup used for the topic modelling experiments.

Chapter 6

Experimental Setup

Chapter 3 provided background on the techniques that are used in this study and Chapter 4 discussed the article and tweet corpora that are explored. This study aims to answer many knowledge discovery questions about the articles and tweets that make up the political discourse space being explored. The content of this chapter provides an account of each experiment performed. Since there are two data sets explored in this study, the methodology employed for experiments can overlap. Therefore, the different modelling and analysis processes are discussed in this chapter to avoid redundancy.

The remainder of this chapter is organised as follows:

- Section 6.1 provides an account of the modelling processes applied in this study.
- Section 6.2 provides an overview of the different analysis techniques used in this study.
- Section 6.3 provides an account of the experiments conducted in this study.
- Section 6.4 summarises the contents of this chapter.

6.1 Modelling Processes

This section's objective is to provide an understanding of the modelling processes that are common between experiments. The contents of this section cover the data preparation and modelling processes needed to transform a raw text corpus into topic models.

6.1.1 Data Preparation

Raw text data has a high-dimensional nature and is uninterpretable by machine learning algorithms [6]. The dimension of a corpus of text is determined by the number of unique tokens present in the vocabulary of the corpus [6]. 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.

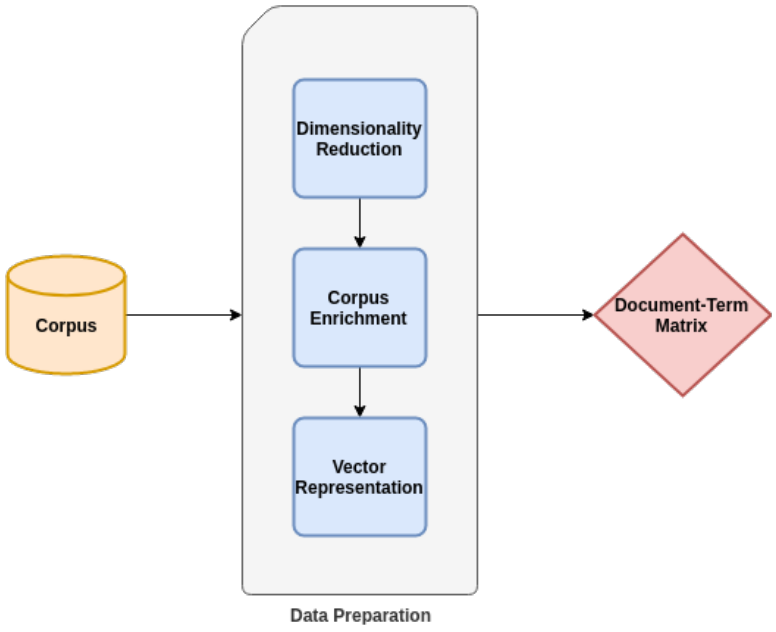


Figure 6.1: The data preparation process

Figure 6.1 illustrates the data preparation process that was followed. 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.

Dimensionality Reduction

The first step in processing the text was to convert all words to lowercase to avoid duplicate words that only differ in case. Punctuation and special characters were also removed. 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 corpus can commonly occur across the corpus and add noise to the created topics. Similarly, having words that rarely occur can 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 of words more dense as it removes the outlier tokens.

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 a different meaning to a sentence in comparison to "ice" and "cream" individually). Bigrams and trigrams were generated to enrich the tokens present in the corpus to capture concepts or identifiers that exist in the corpus.

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 6.1 depicts the term weight calculation for a token, $w_{x,y}$.

$$w_{x,y} = tf_{x,y} \times log(\frac{n}{df_x}) \tag{6.1}$$

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 [29].

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 [6]. 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 [6]. 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.

6.1.2 Word2Vec Modelling

The raw corpus is used to train a W2V model that can be used for the TC-W2V coherence score calculations. The raw data is used to allow the W2V model to capture the semantic relatedness of the tokens in context of the sentence containing it. The W2V model is created and provides a 200-dimensional vector to represent each token in the data provided. The model undergoes 5 epochs during training. The resulting model can be

used in TC-W2V coherence calculations when evaluating the produced topic models.

6.1.3 Topic Modelling

The process followed to build a topic model is illustrated in Figure 6.2. This process is applicable to building both, LDA and NMF, models. The data preparation process discussed in Section 6.1.1 is applied to the corpus of interest as a preprocessing step. Topic models are built on the prepared data. Models are built for k values ranging from a to b where a and b are positive integers and $a < b$. For these experiments, the values of k range from 5 to 50. This is done to identify the model with the highest coherence score. The optimal model is the one used for further analysis that addresses the objectives of the experiment.

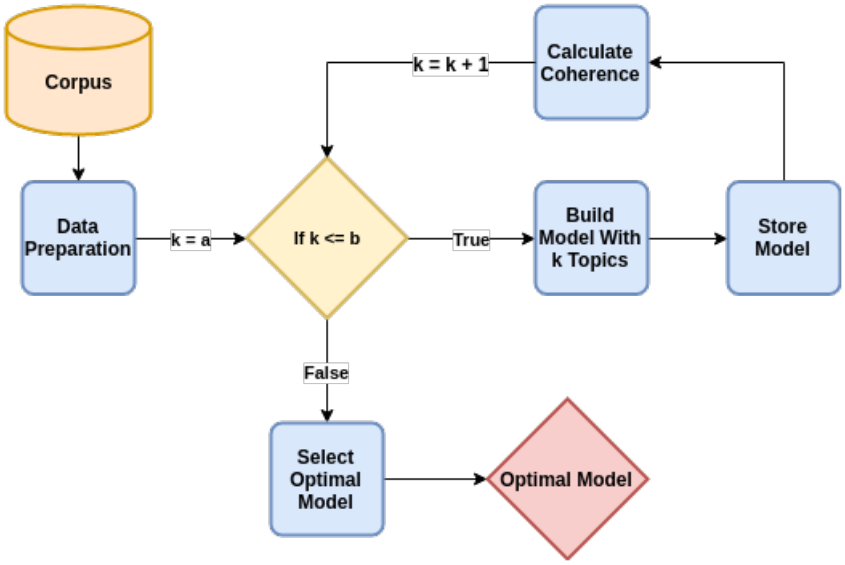


Figure 6.2: The topic model build process

6.2 Analysis Processes

6.2.1 Topic Distributions

The topic models are used to label documents in a text corpus by giving a document the label corresponding to the dominant topic in the topic distribution of that document. Labelling the documents with the dominant topic allows for the visualisation of topics contained in a corpus. This provides insights into a model's ability to cluster text documents. Furthermore, this provides a reference point to understand the importance of a topic based on the volume of documents attributed to it.

6.2.2 Highlight Topics

A topic model produces topics in an unsupervised manner, the quality of topics produced cannot be predetermined. Text data can contain noise therefore analysing the topics produced is a qualitative task that requires manual inspection. Some of the topics can be unintelligible and are excluded from further analysis. The annotation of a topic is subjective and requires human intervention. Topic keywords are analysed and a human annotated label is attached to that topic.

6.2.3 Topic Keyword Wordclouds

Wordclouds provide insight into important terms contained in a text corpus. The size of terms within a wordcloud is directly proportional to the number of times the term appears in a corpus. The documents contained in the corpus are labelled according to the topic number of the dominant topic. The topic keywords remain constant for documents that are members of the same topic, thus simulating importance in a wordcloud because the keywords of topics with large volumes will contain higher frequencies. The wordclouds generated from topic keywords provide insights into tokens that are important in the

discourse. Larger tokens either belong to topics that represent a high volume of the documents or occur in a high volume of documents across multiple topics. This represents a distribution of topics with a focus on the keywords that make up topics instead of the topic clusters.

6.2.4 Topic Timelines

When a text corpus has documents that have temporal information then topic timelines can be inspected. Topic timelines provide insights about the life-cycle of a topic. Documents in a corpus are grouped by topic labels and plotted over time. The timeline of a topic provides insights into when a topic is receiving attention. In some instances, it can provide insights into the dates of events.

6.2.5 Topic Similarity

Comparing two topic corpora to determine the similarity that exists in their vocabulary provides insights into whether the topics cover similar things. The topic similarity calculation starts with creating sub-corpora corresponding to the documents from each topic. Thereafter, creating vocabularies from the corpora of topic pairs. The vocabulary of each topic is represented as a bag of words (BOW) vector. The terms in the BOW vector are scored with TF-IDF. Pairs of topic vectors are subjected to a cosine similarity [11] calculation. The cosine similarity score ranges from zero to one. The closer the score is to one, the closer the vectors are to each other.

An efficient way to compare topics is to select a set of topics from two models, calculate the similarity between each pair and visualise the associated similarity in a heatmap. A heatmap provides an easy way to identify topics that share a similar vocabulary so that further investigations can be done to draw insights from the similarity.

6.3 Experiments

In Section 6.1 the modelling processes that are relevant to multiple experiments were discussed. The analysis techniques employed to understand the topics produced by the models were discussed in Section 6.2. This section provides an overview of the experiments performed in this study. The experiments exhibit a similar structure to each other but each experiment has unique objectives.

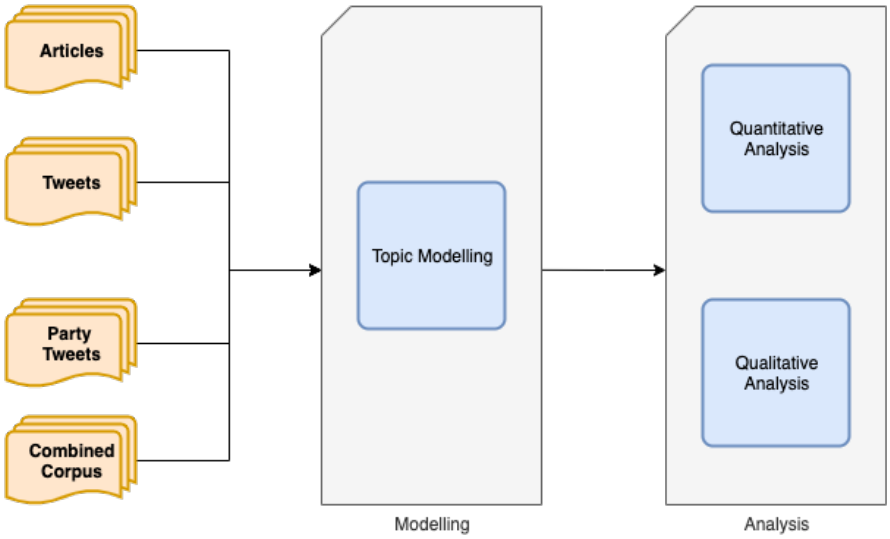


Figure 6.3: High level overview of the experiments

Figure 6.3 illustrates an overview of the experiments conducted. The aim of this study is to uncover insights by applying topic modelling and analysing the clusters of documents produced by the topic models. The experiments are structured as follows:

- The first experiment (Chapter 7) applies topic modelling to the articles to uncover insights about topics and relationships that exist between the 2014 and 2019 articles.
- The second experiment (Chapter 8) focuses on applying topic modelling to the tweets with the aim of understanding the themes of discussion in the election-related discourse on Twitter.

- The third experiment (Chapter 9) applies topic modelling to the tweets produced by political parties to uncover insights about their Twitter election campaigns.
- The fourth experiment (Chapter 10) applies topic modelling to a combined corpus consisting of 2019 articles and political party tweets. The experiment aims to identify a relationship that exists between the articles and tweets in the combined corpus.

6.4 Summary

The modelling and analysis processes employed in the experiments were discussed in this chapter. The modelling processes included data preparation, model building and model evaluation. The different analysis approaches used and the types of insights that can result from them were briefly covered. These included using heatmaps and wordclouds to view topic similarities and topic keyword importance. The modelling and analysis processes are used for multiple experiments. This study contains four experiments: an analysis of election-related news articles; an analysis of election-related tweets; a dissection of political party campaigns on Twitter and an analysis of a combined corpus to identify relationships that exist between the articles and political party tweets. The experiments were introduced to set the scene for the experiments in the chapters that ensue.

Chapter 7

Topic Modelling Of News Articles for Two Consecutive Elections

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, events 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 in a country like South Africa with a diverse population and multi-party election system.

In this experiment, the focus is on the election discourse found in news articles. The insights extracted from the articles provide a window into newsworthy stories and events discussed prior to the election. Since the article data is from a solitary source (News24), this analysis focuses on the insights from the data available and is not a generalisation for the overall article discourse surrounding the election. Both, LDA and NMF, are applied to the articles to gauge the quality of topics created from both types of models. The comparison between LDA and NMF is not a primary objective of this study, but the evaluation of these techniques allows for the selection of a preferred method for the experiments that follow in later chapters.

7.1 Objectives

This experiment aims to uncover insights that will assist in answering the following research questions:

- What were the prominently covered topics in the 2014 and 2019 election-related articles?
- How has the article discourse changed from 2014 to 2019?
- What contrasts between LDA and NMF can be derived from the experiment?

7.2 Experimental Setup

The process that was used to build a topic model from the article corpus is illustrated in Figure 7.1. Articles containing election-related content were selected from the larger article corpus.

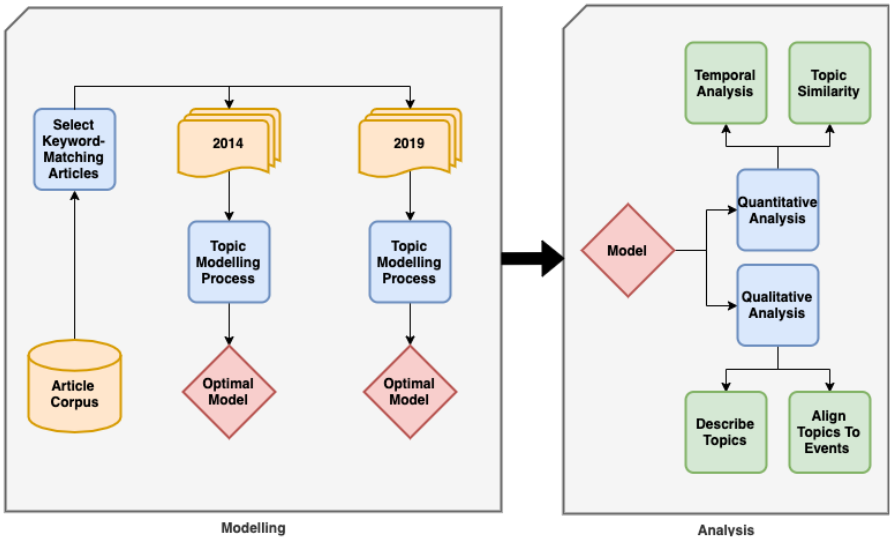


Figure 7.1: The process applied in the article experiment

The selected articles were segmented into two mutually exclusive subsets split by the publication year. The topic modelling process illustrated in Figure 6.2 is applied to build LDA and NMF models for the two subsets of articles. The optimal model is used to perform analysis. The analysis consists of qualitative and quantitative analysis. The qualitative analysis entails the manual inspection of topics to make observations about the topics and the correlation to events that exists. The quantitative analysis involves performing temporal analysis to understand the life-cycle of a topic and performing similarity comparisons on the corpora created from articles belonging to topics.

7.3 Finding The Optimal Model

The articles contained two sources of text, the article body and the article summary. Topic models were built in preliminary experiments to contrast the topics created from both sources. There were no notable differences in the topics produced by these models, therefore the article body was used as the text source for this experiment since it contained more text pertaining to the contents of the articles.

The topic models built in this experiment conform to the topic modelling process described in Figure 6.2. Models were built for k values ranging from 5 to 50. There were 46 models built for each variant, resulting in 184 models built for this experiment. Four models are selected for further analysis, one LDA and one NMF model per period. Figure 7.2 illustrates the coherence scores for the models built. The coherence of models are calculated using the top terms in the topic descriptor using the TC-W2V approach discussed in Section 3.2.1.

The 2014 models achieve similar coherence scores at the optimal k values. The optimal k values occur at k=*21* and k=*29* for the 2014 NMF and LDA models respectively. The coherence scores of the 2019 models depicts NMF producing more coherent topics in comparison to LDA. For the 2019 articles, the optimal NMF model occurs at k=*16* whereas the LDA counterpart is achieved at k=*44*. The 2019 models achieved lower coherence scores than the 2014 models. The LDA models achieved lower coherence

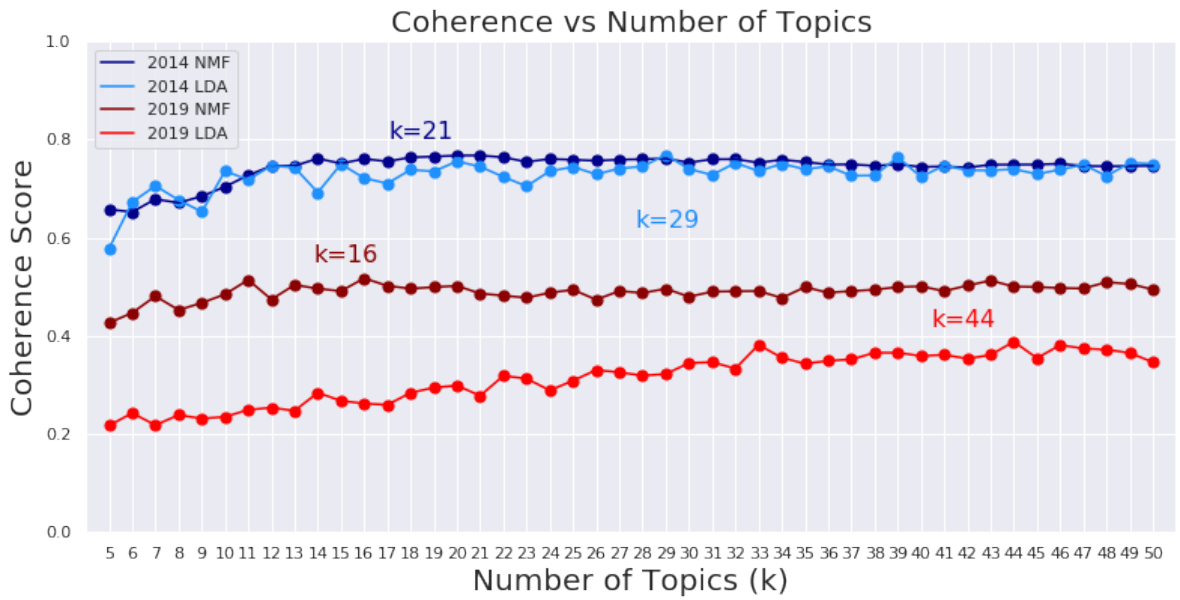


Figure 7.2: Coherence scores for article topic models

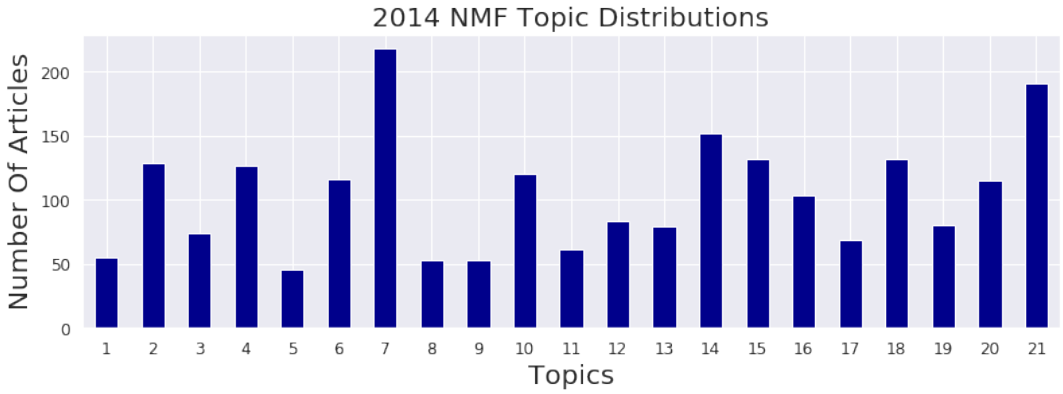
scores than their NMF counterparts in both periods.

7.4 Topic Distributions

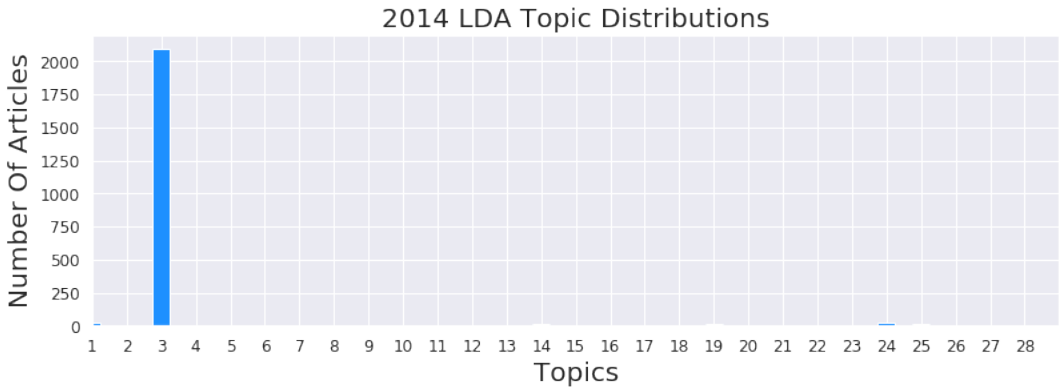
Analysing the distribution of topics produced by topic models provide an indication of the quality of clusters produced by the model. The topic distributions of the 2014 and 2019 models are illustrated in Figure 7.3 and Figure 7.4 respectively.

The illustrations in Figure 7.3 depict a contrasting observation about the two techniques. The NMF model (Figure 7.3a) created topics that average 100 documents per topic with all topics containing a proportion of the documents. In contrast to the NMF model, the LDA model (Figure 7.3b) does not segment the 2014 articles well. There is a single dominant topic that claims majority of the documents, 14 topics from the 29 identified are attributed no documents whilst the remaining topics averaged less than 20 articles.

The distributions of topics for the 2019 articles are illustrated in Figure 7.4. A similar



(a) NMF topic distribution for the 2014 articles

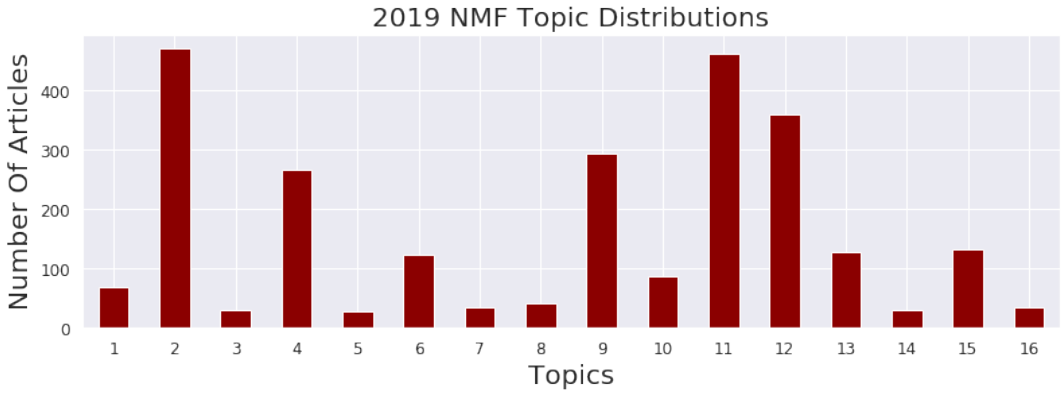


(b) LDA topic distribution for the 2014 articles

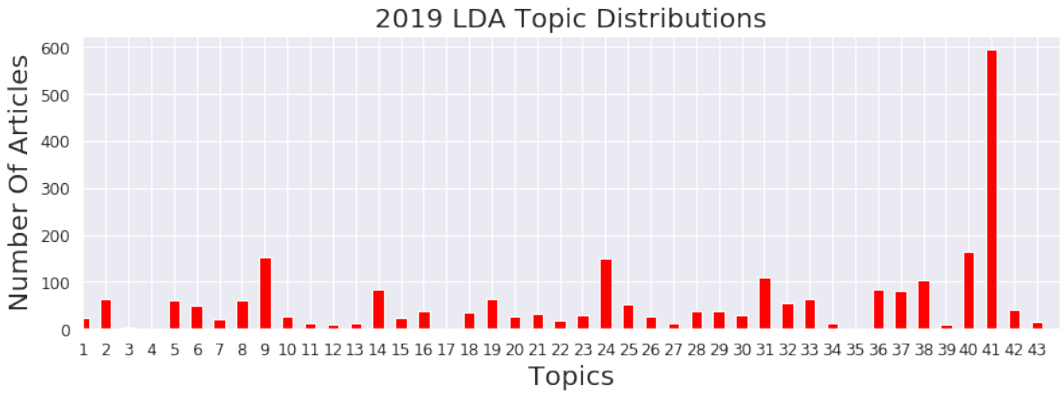
Figure 7.3: LDA & NMF topic distributions for the 2014 articles

trend to Figure 7.3 is seen in Figure 7.4, the topics from the NMF model contains denser topics than the LDA model. The 2019 LDA model (Figure 7.4b) produces a large dominant topic, however it is not as dominant as the one produced on the 2014 articles, the 2014 dominant topic contained almost all the articles whereas the 2019 one contains about 5 times the amount of articles in comparison to other topics. The NMF model (Figure 7.4a produced a representative distribution of topics where some topics are well populated and others contain around 50 articles. All topics were represented by some articles, unlike the LDA model.

The LDA models appear to produce large dominant topics that cluster large proportions



(a) NMF topic distribution for the 2019 articles



(b) LDA topic distribution for the 2019 articles

Figure 7.4: LDA & NMF topic distributions for the 2019 articles

of the articles into a topic whereas the NMF models appear to have more compact topics. Both of the dominant topics produced by LDA contain keywords about political parties. The segmentation of the data on political party keywords could be accountable for this or perhaps LDA requires more data to build topics with better clusters. Since both techniques are subjected to identical data, the NMF models cluster documents in a more compact and representative manner in comparison to LDA in this experiment.

7.5 Highlight Topics

The topic distributions of the models showed that NMF clustered the topics better than LDA. LDA created dominant topics that claim membership from most of the articles in both time periods. Nonetheless, the topic descriptors produced can still provide a summary of the events that occurred in the respective periods. The topic models produced 110 topics in total, it is infeasible to discuss all the topics produced by the four models, therefore the highlight topics are presented in Table 7.1 and 7.2 for articles published in 2014 and 2019 respectively.

7.5.1 2014

The highlight topics produced by the 2014 models are depicted in Table 7.1. Both models featured the Nkandla report, the NMF model highlighted a topic about the Public Protector's findings and one about a court application by the DA which was most likely done to get the report released. The NMF model captures a topic on DA affairs, with mentions of Helen Zille (the leader at the time) and Mmusi Maimane. An EFF-related topic in the NMF model mentions two members of the EFF leadership, Dali Mpofu and Mbuyiseni Ndlozi, no other context can be derived from the description. The LDA model's dominant topic (topic 3) contained references to parties and voting in the election. The NMF model also captured a topic related to comments made by Jacob Zuma about not requesting the security upgrades to Nkandla. Julius Malema and his battle with the South African Revenue Service (SARS) about tax in 2014 also features in both models. Both models produced a topic about an advert that was reported to the Independent Communication Authority of South Africa (ICASA) which was then banned. The reason for the ban could not be inferred from the topics. The NMF model produced a topic relating to a mine-worker strike which occurred at the Lonmin platinum mine.

The highlight themes uncovered from the 2014 articles provides a summarised view of news reported in this period. The themes related to the Nkandla report, a mine-worker

Model	Topic	Keywords	Label
NMF14	**1**	anc, campaign, member, election, support, former, people, leader, political, kasrils	ANC Related
NMF14	**2**	report, madonsela, release, nkandla, public_protector, upgrade, find, findings, public_protector_thuli_madonsela, public	Nkandla Report
NMF14	**8**	malema, sars, tax, trust, order, owe, fail, sa_revenue_service, return, court	Malema & SARS
NMF14	**10**	da, zille, maimane, da_leader_helen, leader, western_cape, mazibuko, job, democratic_alliance, party	DA Related
NMF14	**11**	strike, amcu, platinum, worker, wage, union, offer, lonmin, mine, association_mineworker_construction_union	Mine-worker Strike
NMF14	**12**	sm, court, application, da, fair, send, hellens, nkandla_report_show, dismiss, judge_mike	DA Court Application / Nkandla Report
NMF14	**15**	zuma, president, president_jacob_zuma, nkandla, ask, upgrade, pay, home, security_upgrade, family	Zuma & Nklandla
NMF14	**16**	eff, mpofu, party, economic_freedom_fighter_eff, economic_freedom_fighter, dali_mpofu, city, spokesperson_mbuyiseni_ndlozi, member, red	EFF Related
LDA14	**1**	sabc, advert, icasa, ban, advertisement, air, public_broadcaster, da, complaint, independent_communication_authority_sa,	SABC Banned Advert
LDA14	**3**	anc, party, vote, da, people, election, zuma, member, eff, government	Elections / Parties
LDA14	**6**	sm, da, court, fair, send, hellens, application, nkandla_report_show, zuma_steal_money, judge_mike	DA Court Application / Nkandla Report
LDA14	**14**	sars, malema, tax, bill, award, infrastructure, contract, agriculture, amount, prevent	Malema & SARS

Table 7.1: Highlight topics from the 2014 articles

strike and Julius Malema's tax battles with SARS. Some party-related themes were also revealed, which were most likely related to party activity in the observed period.

7.5.2 2019

The highlight topics produced by the 2019 models are described in Table 7.2. Both models 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. Both models also captured topics relating to the SONA and the SONA debate that occurred afterwards. Both models captured topics related to state capture and the testimony of Vytjie Mentor (a former member of parliament) at the commission of enquiry on state capture.

The NMF model isolated a topic about the court case between Julius Malema and Pravin Gordhan (the minister of public enterprises) where Malema referred to Gordhan as a dog. The NMF model also uncovered a topic related to the motorist demerit system that rolled out in March 2019. The LDA model was able to uncover a topic relating to land expropriation without compensation. The LDA model's dominant topic (topic 41) referenced the ANC and Eskom but it is unclear what the topic is about.

The themes uncovered in 2019 related to current affairs happening in the country at the time. The most notable coverage focused on state capture, Eskom issues, SONA and land expropriation without compensation. There were no intelligible topics found that referenced party-related content however party keywords did appear in topics that could not be deciphered.

7.6 Topic Keyword Wordclouds

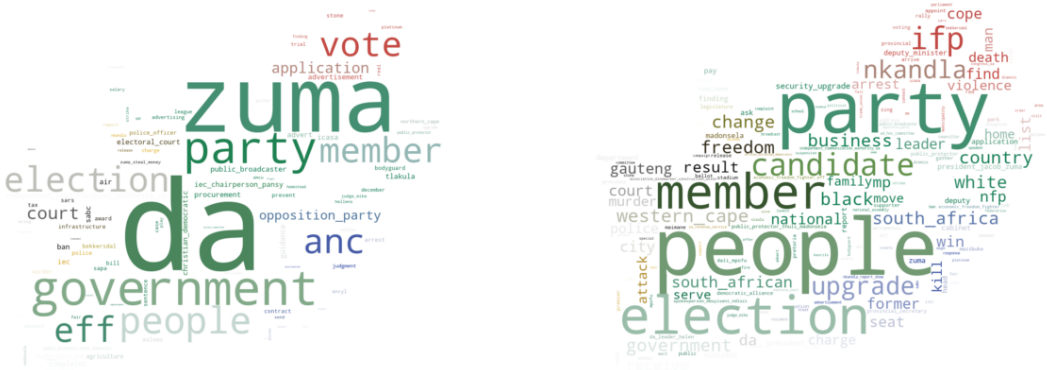
Topic keyword wordclouds were generated from the top terms of the 2014 and 2019 models. The wordclouds generated for 2014 are illustrated in Figure 7.5.

The LDA wordcloud (Figure 7.5a) indicates importance to keywords of topic 3 of the 2014 LDA model which is expected since that is the dominant topic in the model that contains almost all the documents. The NMF wordcloud (Figure 7.5b) contains a variety

Model	Topic	Keywords	Label
NMF19	**1**	free_state, pay, free_state_government, company, da, reinstate, charge, bank_statement_reveal, evidence, money	Free State / Corruption
NMF19	**2**	eskom, load_shed, power_utility, power, gordhan, unbundling, announce, address, electricity, continue	Eskom / Load Shedding
NMF19	**4**	ramaphosa, president, debate, state_nation_address, anc, president_cyril_ramaphosa, mp, member_parliament, accuse, reply_debate,	SONA Debate
NMF19	**6**	mentor, commission, former, anc_mp_vytjie, flight, claim, capture, saa, evidence, zuma	State Capture
NMF19	**8**	malema, matter_move_high, refer_gordhan_dog, white_monopoly_capital, government_seemingly_result, magistrate_court, gordhan_lodge_complaint, speech_november_last, ian_levitt_represent, allegedly	Malema & Gordhan Court Case
NMF19	**11**	investment, government, south_africa, country, water, plan, people, world, company, economic	Government Plans
NMF19	**15**	committee, da, election, propose, mean, chairperson, motorist, process, demerit, parliament	Motorist Demerit System
LDA19	**1**	eskom, solution, problem, free_state_government, unbundling, increase, privatise, labour, government, generation	Eskom Unbundling
LDA19	**15**	free_state, da, black, gupta_family, mentor, evidence, anc, commission, gupta, charge	State Capture
LDA19	**25**	eff, ramaphosa, anc, da, mp, party, bank, doctor, sona, debate	SONA Debate
LDA19	**29**	mentor, flight, saa, former, claim, capture, anc_mp_vytjie, offer_position_public_enterprise, october, reflect	State Capture
LDA19	**30**	committee, african, election, trade, anc, http, lengthy, agreement, public_consultation, expropriation_without_compensation,	Land Expropriation
LDA19	**40**	load_shed, eskom, regard, effort, ongoing, truth, tuesday, hint, monday, bold	Eskom Load Shedding
LDA19	**41**	eskom, anc, friday, man, time, province, state, deliver, give, plan	Eskom / ANC

Table 7.2: Highlight topics from the 2019 articles

of keywords related to the highlight topics of the 2014 NMF model discussed in Table 7.1 as well as other topics in the model. The most notable terms are "Nkandla", "election" and "party". Inspecting Table 7.1 shows that "Nkandla" features often in the highlight



(a) 2014 LDA topic keyword wordcloud (b) 2014 NMF topic keyword wordcloud

Figure 7.5: Topic keyword wordclouds for the 2014 articles

topics. Furthermore, more than 800 articles are members of topics with Nkandla as keywords (split between topic 2 and 12). It is expected that "election" and "party" appear prominently in the wordcloud since the articles are curated with keywords related to the parties and there are topics that reference the political parties and the election.

The 2019 LDA model produced more distributed clusters resulting in more topics being relevant in comparison to the 2014 LDA model. Both models captured descriptive topics. Figure 7.6 illustrates the wordclouds for the 2019 articles. The LDA wordcloud (Figure 7.6a) has keywords relating to the ANC, government, corruption, Eskom and Bosasa which is in line with the topics identified for the 2019 LDA model in Table 7.2. A similar trend can be seen in the NMF wordcloud (Figure 7.6b) with keywords referencing the ANC, Eskom and Pravin Gordhan that match the highlight topics for the 2019 NMF model depicted in Table 7.2.

Both 2019 wordclouds provide insights into the content reported by News24. The discourse contains many references to problems faced in the country, especially Eskom frailties and corruption. This is in contrast to the 2014 wordclouds, which contained many references to the elections and political parties. The contrast between election periods indicates that a lot more emphasis was put on corruption and other problems in 2019. The 2014 topics covered the Nkandla report, which has evolved into the focus

(a) 2019 LDA topic keyword wordcloud (b) 2019 NMF topic keyword wordcloud

Figure 7.6: Topic keyword wordclouds for the 2019 articles

on state capture and corruption in the 2019 topics. These issues were linked to former president, Jacob Zuma, and his reign over the country.

7.7 Topic Timelines

Topic timelines provide insight into the life-cycle of a topic. The LDA models created dominant topics which made the timelines noisy. Timelines were created for the LDA models but the dominant topic overwhelmed the other topic and no real insights could be derived from the timelines. Upon removing the dominant topic, the other topics were short-lived and did not contain any notable peaks in volume. Therefore, only the NMF topic timelines are illustrated and discussed.

Figure 7.7 illustrates the timelines of the 2014 highlight topics. Both Nkandla related topics experience peaks in volume in the week of 17 March 2014. This is indicative of discussions relating to Nkandla occurring in March 2014. The DA court application topic peaks at the end of March which indicates that the application was most likely an effort to have the Nkandla report released before the election.

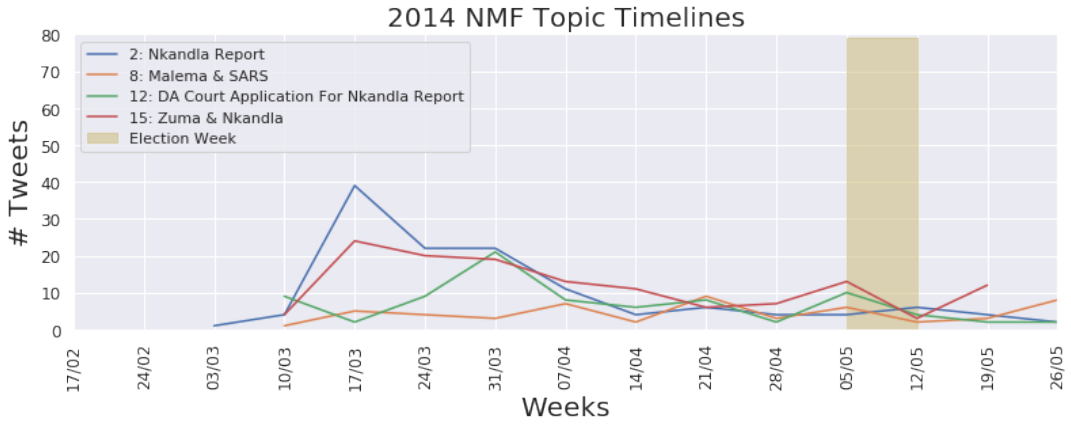


Figure 7.7: NMF topic timelines for the 2014 articles

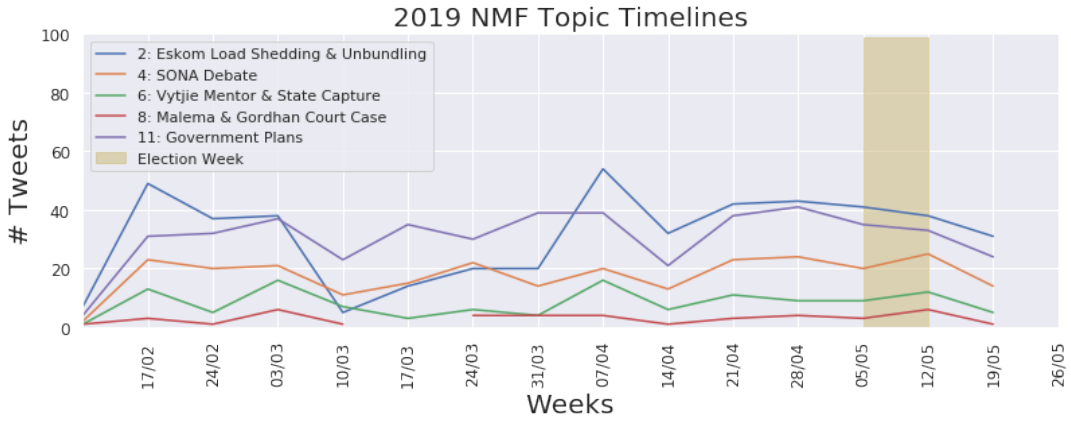


Figure 7.8: NMF topic timelines for the 2019 articles

The topic timelines of some highlight topics from 2019 are illustrated in Figure 7.8. The Eskom topics receive peaks in volume in the week of 17 February and 7 April. The state capture topic receives minor peaks in volume throughout the observed period. The SONA debate topic experiences an almost constant volume throughout the observed period indicating that the articles associated may deal with other aspects as well. The topic interpreted as government plans could be in relation to election campaigns, it receives attention throughout the observed period and the keywords are indicative of campaign terminology. The Malema and Gordhan court case receives minor attention, there is a notable gap in coverage of this topic between 10 and 24 March. In this two-week

period, the Eskom topic also experienced a decline in volume.

7.8 Topic Similarity Heatmaps

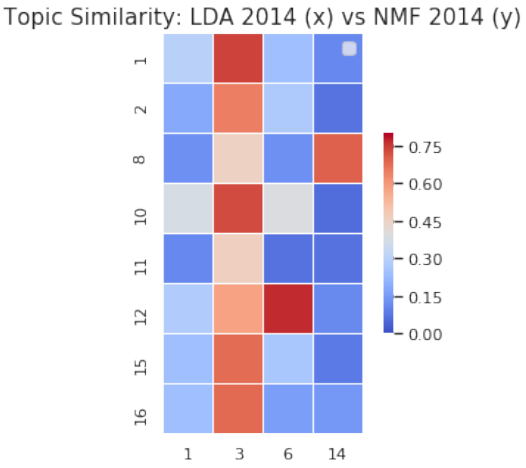
The topic similarity heatmaps illustrate the common vocabulary shared by topics. This can be used to inspect the topics produced by a pair of topic models. The topic similarity heatmaps contrasting the topics created by LDA and NMF for the 2014 and 2019 articles are illustrated in Figure 7.9.

The 2014 heatmap (Figure 7.9a) illustrates the effect of the dominant topic produced by the 2014 LDA model, The dominant topic is shown to be similar in vocabulary to all of the topics produced by the NMF model. Nonetheless, there are two other LDA topics that express a similarity to the NMF topics. Topic 14 from the 2014 LDA model and topic 8 from the 2014 NMF model share a similar vocabulary, inspecting Table 7.1 indicates that both topics describe Malema and his tax battles with SARS. Furthermore, topic 12 of the LDA model and topic 6 of the NMF model share a similar vocabulary. Both these topics address the DA's court application to release the Nkandla report.

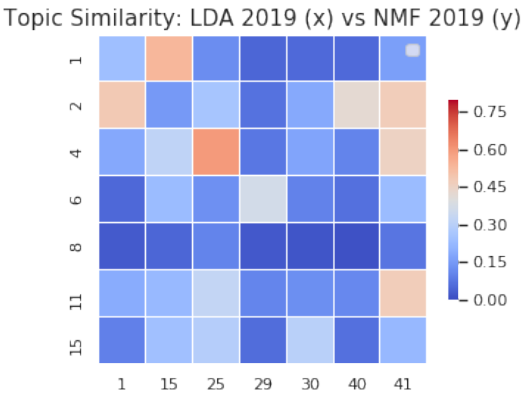
The 2019 heatmap (Figure 7.9b) also highlights topics that share a similar vocabulary. The dominant topic in the 2019 LDA model does not interfere as much as the one from the 2014 LDA model. Topic 1 from the LDA model and topic 2 from the NMF model are both related to Eskom. Topic 15 from the LDA model and topic 1 from the NMF model are both related to state capture. The similarity that exists between topic 40 from the LDA model and topic 2 from the NMF model is due to the common theme of load shedding and Eskom.

Contrasting articles between both election periods provides insights into common themes that are persistent between election periods. The heatmap measuring the similarity between 2014 and 2019 NMF topics is illustrated in Figure 7.10.

Topic 1 from the 2014 NMF model was ANC related whereas topic 4 from the 2019



(a) 2014 LDA & NMF heatmap



(b) 2019 LDA & NMF heatmap

Figure 7.9: Comparison of LDA & NMF topic similarity heatmaps

NMF model referenced the SONA debate, the similarity in vocabulary is likely related to common terminology related to the ANC. Topic 10 from the 2014 model contained keywords related to the DA and their affiliates and topic 15 from the 2019 model related to the motorist demerit system. Upon inspection of Table 7.2, references to the DA was seen in the motorist demerit system topic. This could be an indication that the DA featured in articles related to the motorist demerit system.

These observations indicate that the topic similarity technique applied can be used to measure similarities between two text corpora to identify common elements between

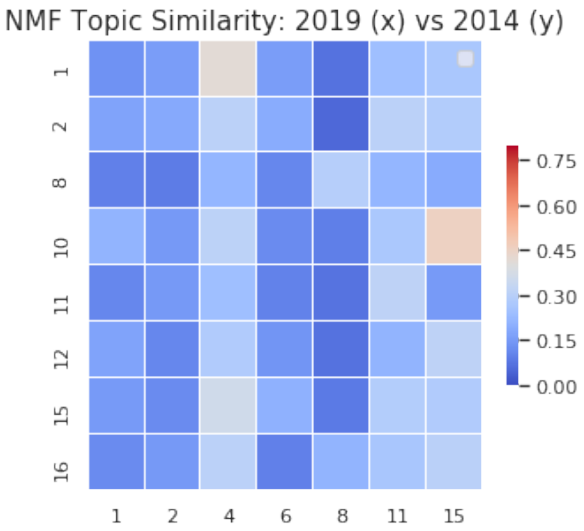


Figure 7.10: NMF topic similarity heatmap for the 2014 & 2019 articles

different topic models.

7.9 Summary

In this experiment, NMF and LDA topic models were created to uncover latent themes in the articles published during the 2014 and 2019 election period. The articles contained two sources of text, the article body and the article summary. Preliminary experiments showed that there were no notable differences between the topics created, therefore the article body was used for this experiment because it contained more text.

The article clusters produced by the models indicated that LDA produced clusters that inadequately split the corpus. In both 2014 and 2019, the LDA models produced dominant topics that prevented notable events from appearing in the topic timelines. The NMF models produced more clusters compared to LDA, this allowed for further analysis to be conducted. The analysis uncovered themes relating to corruption, state capture 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 with SARS and the

election.

The topic keyword wordclouds provided an insightful way of uncovering important terms in the corpus by using the topic keywords to generate wordclouds. The 2014 LDA wordcloud was more sparsely populated than the NMF counterpart, this was due to the dominant LDA topic accounting for majority of the documents thus limiting the vocabulary of the wordcloud and increasing the frequency of terms within the dominant topic. The 2019 LDA topic keyword wordcloud was more insightful than the 2014 one, providing a contrast with the NMF wordcloud. The 2019 LDA wordcloud uncovered terms relating to corruption, the ANC, Bosasa and Eskom. The 2019 NMF wordcloud uncovered terms relating to Eskom and ANC-related content. The contrast between 2014 and 2019 as viewed through these wordclouds indicate that the article discourse of the 2019 election focused more on corruption and problems whereas the 2014 discourse contained more references to political parties and election-related aspects.

The topic timelines were only illustrated for the NMF models as the dominant topics of the LDA models affected the timelines of other topics making these uninterpretable for identifying temporal properties. The 2014 NMF topic timelines illustrated peaks in volume in the middle of March for topics related to Nkandla and the Nkandla report, another notable spike in volume occurred at the end March when the DA filed a court application to release the Nkandla report. The 2019 topic timelines experienced spikes in volume in the middle of February and at the start of April. The volume in February was mostly due to SONA whereas the spike in the Eskom topic in early April was most likely due to load shedding.

Topic similarity heatmaps were created to highlight the similarity in articles belonging to topics from different models. The comparison of LDA and NMF topics for 2014 articles showed that the dominant LDA topic contained a similar vocabulary within its articles to majority of the NMF topics. This was due to the dominant topic claiming membership of most of the articles from 2014. The other notable similarities were between topics related to the DA's court application to release the Nkandla report and Julius Malema's tax battles with SARS. The notable similarities in the 2019 heatmap comparison of LDA

and NMF topics were between topics containing reference to Eskom and state capture. The dominant LDA topic in 2019 contained a similar vocabulary to 3 NMF topics. Contrasts between the NMF models for 2014 and 2019 revealed similarities in topics relating to the ANC, the SONA and the DA.

The next chapter analyses the tweets collected in this study. The analysis focuses on understanding the election discourse present on Twitter prior to the 2019 election.

Chapter 8

Understanding The Twitter Election Discourse In 2019

Twitter provides a platform for engagements in the form of micro-blogs that are limited to 280 characters. In election periods, political parties and the general population use Twitter to voice their opinions on campaign policies, current affairs and many other aspects relating to the election and the country. There are many contributors to the election discourse: political parties, politicians, celebrities, new agencies and the general population. Tweets are informally written and short in length (averaging 20 tokens per tweet in this corpus).

Political parties and their affiliates may directly engage with their target audience on Twitter without a third-party moderating the content, which is the case in news articles. Potential voters can use Twitter as an outlet to voice their views and contribute to the discussion on important topics. Furthermore, stories or events sometimes first break on Twitter prior to reporters becoming aware of the story. Twitter presents a rich, unfiltered source of discourse about the election that provides a different perspective into the important themes in this election period.

This experiment focuses on analysing the tweets collected to understand the contribution

it makes to the election discourse of the 2019 South African general election. The experiment in Chapter 7 focused on analysing an article corpus by applying LDA and NMF to the corpus to find topic models that separate and explain the themes in the articles. The experiment found that NMF is better at creating compact topics. The contrast in topic distributions for LDA and NMF indicated that LDA clumps documents into dominant topics and results in the remainder of topics being sparsely populated by documents. This does not split documents into representative clusters that support further analysis. Preliminary experiments were conducted on the tweets with LDA which presented a similar outcome. The main objectives of this experiment is to discover knowledge from the tweets and not to evaluate the topic modelling techniques thus this experiment excludes LDA and focuses on the analysis of the topics created by NMF.

8.1 Objectives

This experiment aims to uncover insights that will assist in answering the following research questions:

- Can descriptive topics be produced from the short text present in tweets?
- What were the prominent topics in the Twitter election discourse?
- Can any observations be made about spikes in the volume of topics?

8.2 Experimental Setup

Figure 8.1 illustrates the process followed to build topic models for the tweet corpus. NMF models are built by applying the topic modelling process illustrated in Figure 6.2 to the tweets.

The optimal model that results from this process was used to conduct further analysis.

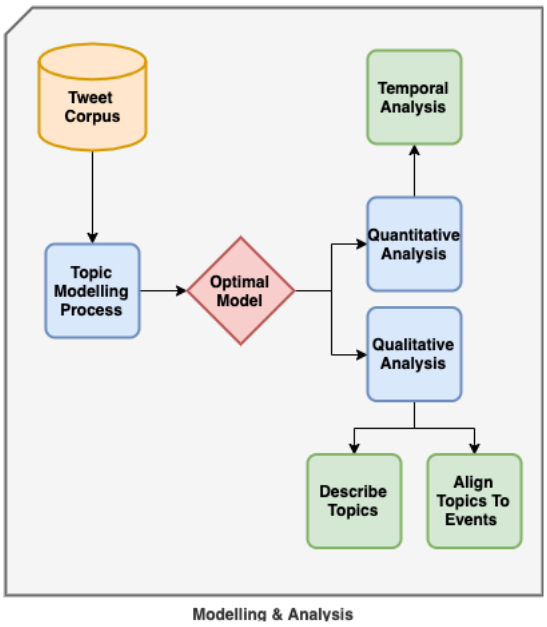


Figure 8.1: The process applied in the tweet experiment

The analysis consists of a qualitative and quantitative component. The qualitative analysis entails making observations about themes and events from the manual inspection of topics. The qualitative component consists of temporal analysis to understand the life-cycle of the topics.

8.3 Finding The Optimal Model

In accordance with the topic modelling process described in Figure 6.2, models were built on the entire tweet corpus for k values ranging from 5 to 50. Figure 8.2 illustrates the different coherence scores achieved for the different values of k. The optimal model for the tweet corpus was achieved at k=*26*.

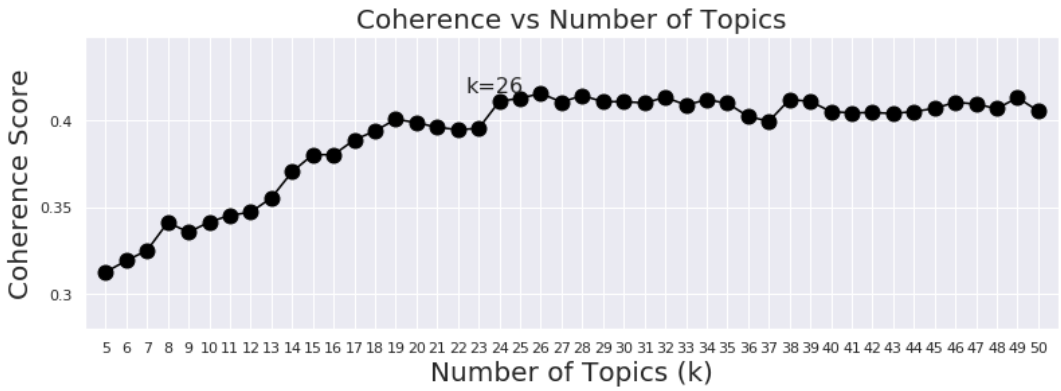


Figure 8.2: Coherence scores for tweet topic models

8.4 Topic Distributions

Each tweet in the corpus was tagged with a label corresponding to the dominant topic. Figure 8.3 illustrates the distribution of topics that resulted from tagging the tweets with topic labels. There were 11 topics that contained more than 200 000 tweets. The topics with the small distributions range from 14 000 to 30 000 tweets.

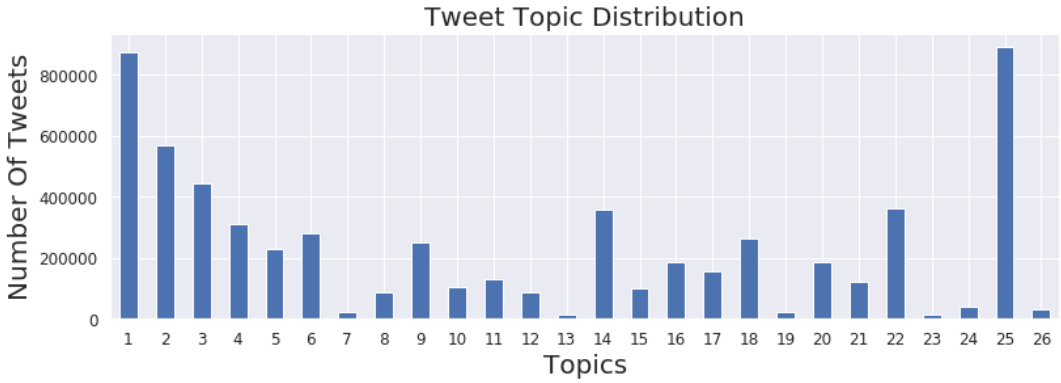


Figure 8.3: Distribution of topics in the tweet corpus

8.5 Highlight Topics

The optimal model selected for analysis contained 26 topics, the distributions of these topics were illustrated in Figure 8.3. Upon inspection of the topics created, most of the smaller topics contained unintelligible topic descriptions. Furthermore, topic 25 which was rather large also contained an unintelligible topic description. The highlight topics from the model are depicted in Table 8.1. The topics contain many usernames providing insight into the participants in these topics. Tweets also contain fairly short text and the trend seen in Table 8.1 indicates that tweets generally contain many mentions to other accounts in the text. Perhaps other media (images and videos) was used to convey majority of the intent with the text aspect of a tweet being used to convey a summarised view and tag others to participate in the conversation.

The topics provide a summarised view of the discourse that occurred. The discourse contained a few ANC and EFF topics and a solitary DA related topic, this trend was in accordance to the frequency of party tweets and mentions illustrated in Figure 5.6 and Figure 5.7 respectively, where the ANC and EFF were more active and engaged with on Twitter compared to the the DA.

The ANC topics mostly referenced their election campaign, the keywords for these topics contained campaign-related hashtags such as *ThumaMina* and *GrowSouthAfrica*. Another ANC-related topic contained a mixture of keywords. It contained references to Cyril Ramaphosa, the global citizen festival and congratulatory tweets to Professor Mashudu Tshifularo for leading the first team to use 3D-printed bones for middle ear implants at Steve Biko hospital. A peculiar insight from topic 3 was the reference to Herman Mashaba (who is affiliated with the DA), this is most likely due to Mashaba contributing to the discourse by responding to ANC related material. The DA topic (topic 5) contained keywords indicative of views important to their campaign. The topic contained references to DA affiliates and their rhetoric about bringing a job to every home in South Africa. The EFF possessed many topics in the discourse, the themes present in their topics mostly relate to their election campaign. The topics contain many references to EFF affiliates and references to the party's rhetoric on land and jobs.

Topic	Keywords	Label
1	effsouthafrica, must_watch_cic, effredfriday, voteeff_ourlandandjobsnow, effkzn, fighter, fighters, tumisole, rasta_efftshelathuparally, vngalwana	EFF Campaign/ Rally
2	cyrilramaphosa, presidencyza, hi, mashudu_tshifularo_amp, send_heartfelt_congratulation_professor, team_steve_biko_academic, thank_commit_billion_glblctzn, festival_mandela, matamela_cyril_ramaphosa_swear, faithful_republic_south_africa	Ramaphosa / Global Citizen
3	myanc, hermanmashaba, comrade, samkelemaseko, presjgzuma, mbindwane, growsouthafrica_thumamina, growsouthafrica_voteanc_ancsundays, enca, lesufi	ANC Campaign
4	vote, cast, change, tomorrow, station_sa, count, station, iecsouthafrica, first_time, register	IEC Voting / Registration
5	mmusimaimane, helenzille, leader, alettaha, hermanmashaba, mmusi, zilevandamme, bring_job, every_home, actually_south_africans	DA Campaign
6	anc, comrade, government, watch, lead, campaign, member, corruption, gauteng, voter	Gauteng Voting Corruption
7	iaaf_protest_ruling_caster, professor_steve_cornelius_resign, south_africans_appreciate, south_africans, cornelius, lwazberry, nt, dis, votingday, soon	IAAF Ruling / Caster Semenya
8	voteanc, thumamina, growsouthafrica_thumamina, siyanqobarally, growsouthafrica_ancsiyanqoba, thumamina_growsouthafrica, growsouthafrica_anconthegrounds, today_mark_days_general, elections, ivoteanc	ANC Campaign
14	people, black, white, young, work, country, government, fight, understand, land	Land / Fighting Government / People of SA
16	cic, address, arrive_parliament_registration, ahead_tomorrow_upon_arrival, today, watch, leadership, koko_sarah_malema_grandmother, floydshivambu_ladies_gentlemen_incoming, floydshivambu_sadden_pass_away	EFF / Malema
17	voteeff, effredfriday, floydshivambu_effelectionsadvert_land_jobs, iamvotingeff, efffinalpush, effredfriday_ourlandandjobsnow, voted_ourlandandjobsnow, mbuyisenindlozi_future_smile, thank_commissars_ground_forces, kzn_indeed_home	EFF Campaign
18	election, win, time, today, campaign, result, understand, willowcudi_electionresults_many, national, south_africans	Election Results
20	mbuyisenindlozi, floydshivambu, advbarryroux, happy_revolutionary_birthday_commissar, ourlandandjobsnow, tumisole, happen, deputy_president, gardeegodrich, visual_social_medium	EFF Related / Birthday Celebration
21	growsouthafrica, voteanc_anconthegrounds, thumamina, ancsiyanqoba, siyanqoba, ancsundays, voteanc_ichooseanc, encatownhalldebate, gautenganc, anckzn	ANC Campaign
22	amp, iecsouthafrica, today, register, least, secrecy_vote, photograph_marked_ballot_protect, remember_criminal_offence, station_sa, government	IEC Warning For Photographing Ballot
24	thanks, governmentza_commit_free_sanitary, product_school_girl_low, income_household_count_keep, continue_amp_expand, investment_menstrual_health_supply, amp_education_itsbloodytime, product_school, cyrilramaphosa, product_school_girl	Government / Sanitary pads

Table 8.1: Highlight topics from the tweet topic model

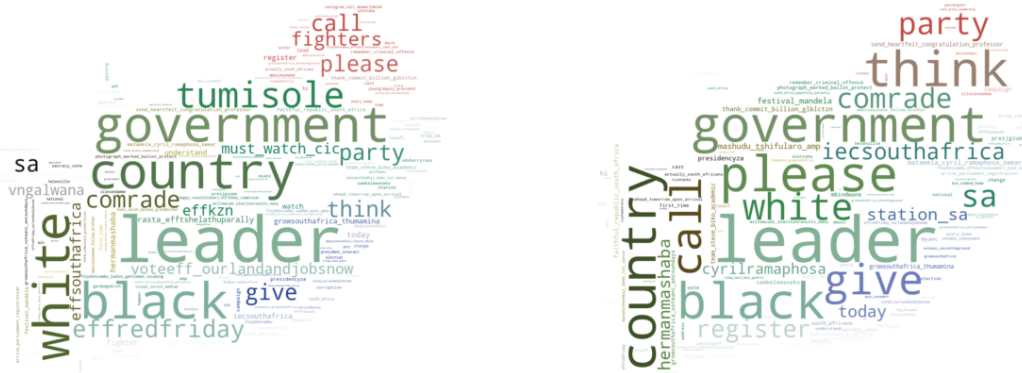
Other EFF-related topics were related to Julius Malema, one expressing birthday wishes and another expressing sadness at the passing of his grandmother.

There were a few topics that are not directly related to any of the parties. There were two IEC topics related to voting stations, first time voters and a warning that taking a photograph of a ballot paper is a criminal offence. There was a topic related to government providing free sanitary pads to school girls from low income households.

There was another semi-intelligible topic referencing land and fighting with government, this topic was most likely related to a land expropriation discussion.

8.6 Topic Keyword Wordcloud

The tweet corpus was collected by selecting tweets that contained predefined keywords. Anyone can publish tweets that contain these keywords, thus the corpus can contain noise. The highlight topics depicted in Table 8.1 were topics that were intelligible and deemed relevant to the election. The wordclouds illustrated in Figure 8.4 contain topic keywords from the entire corpus and keywords from the highlight topics. The wordcloud built on all the topics (Figure 8.4a) indicates that the EFF-related discourse floods the wordcloud, with references to the other parties barely noticeable. This was expected since the party produced the most tweets and were also mentioned the most from all the parties. The EFF made good use of Twitter and the wordcloud indicates how their campaign overshadowed their opponents in terms of volume. References to white and black in this wordcloud could be indicative of race-related discussions.



(a) Wordcloud for all topics

(b) Wordcloud for highlight topics

Figure 8.4: Tweet topic keyword wordclouds

The highlight topic wordcloud (Figure 8.4b) provides an indication of terms that were

important in the sub-corpus related to those topics. Government and the country still remain focal terms of the discourse. The references to black and white were persistent in the highlight wordcloud as well. The highlight wordcloud did not contain as many references to the EFF as the full corpus because there were many EFF-related topics that did not feature in the highlight topics either due to unintelligible topic descriptions or the concepts covered in the topic were covered in another topic.

8.7 Topic Timelines

The timeline of a topic indicates the moments in the observed period at which a topic was receiving attention. It also provides an indication of long lasting topics that are relevant throughout the observed period. Table 8.1 contains the topics, the keywords for the topic and a manually annotated label.

The timelines of ANC related topics are illustrated in Figure 8.5. Topic 3 experiences spikes in volume in the middle of March, the end of April and in the week of the election. All the other topics experience spikes in volume around the election week, which is expected as this is the final push to convince voters to cast their vote. Topic 6 was interesting because it contained references to voting corruption and had ANC keywords, this topic experiences a spike in the middle of March and another in election week.

The DA timeline illustrated in Figure 8.6 contains a solitary topic. The timeline indicates that a spike in volume occurs in the middle of March, middle of April and in election week.

The EFF topic timelines illustrated in Figure 8.7 contains timelines for topic 16,17 and 20. The timeline for the other EFF topic, topic 1, was excluded as it contained much more tweets than the other topics and diminishes the peaks in the visualisation. Topic 16 which relates to the passing of Julius Malema's grandmother experiences spikes in February, March and April which is odd since she passed away at the start of May. This indicated that other aspects of the topic which were relevant in the months preceding her

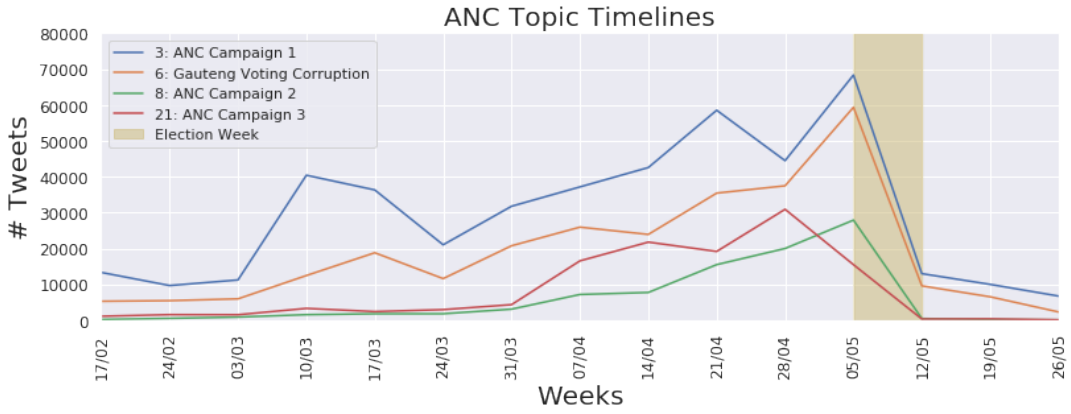


Figure 8.5: Timeline of ANC topics in the tweet model

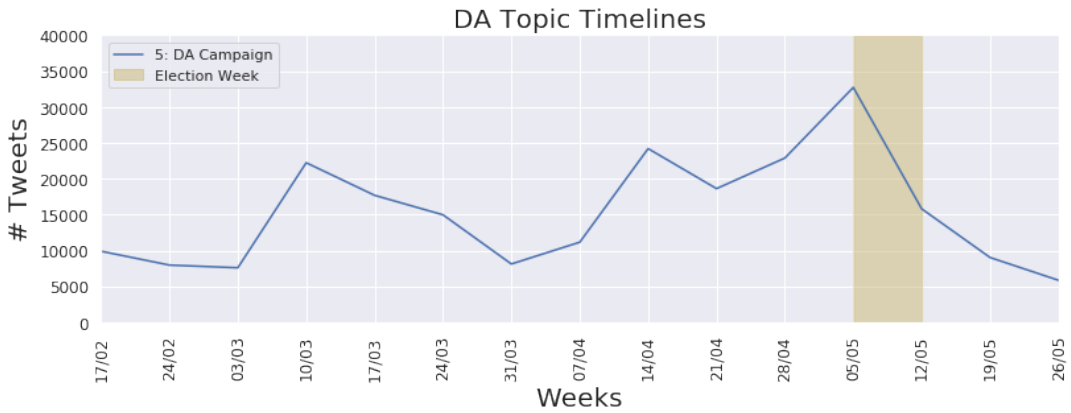


Figure 8.6: Timeline of DA topics in the tweet model

death have been overshadowed in this topic. A similar trend can be observed for topic 20, where the keywords indicate a theme about Malema's birthday which is in March but the timeline contains more than 10 000 tweets consistently from March until after the election. All 3 of these topics experience spikes in volume around election week.

The timelines of other topics not directly related to a party was illustrated in Figure 8.8. The sanitary pads topic experienced a spike in the middle of March, this was in accordance with the launch of the sanitary dignity initiative at the end of February. The IAAF and Caster Semenya topic has a short lifespan in the observed period as it is only relevant at the start of May.

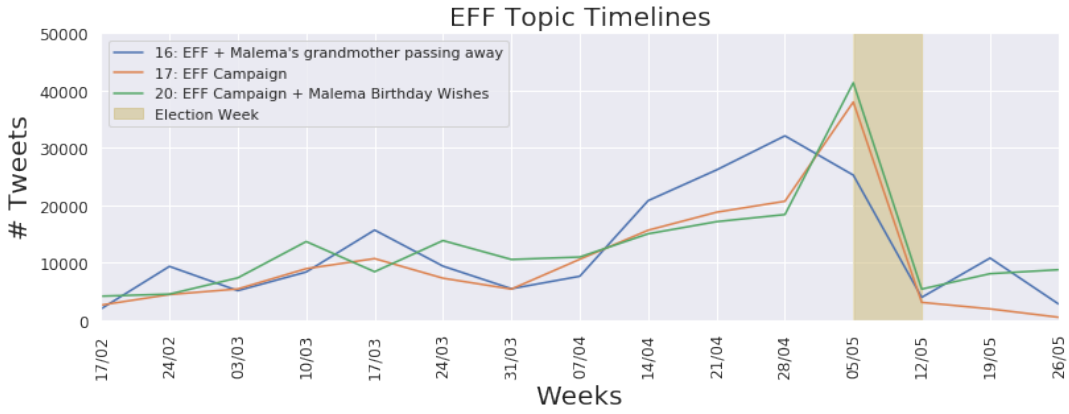


Figure 8.7: Timeline of EFF topics in the tweet model

Figure 8.8: Timeline of other topics in the tweet model

The topic models were trained on the entire corpus and clusters tweets into themes that may lose some of the localised context relating to the spikes in volume that occur throughout the period of interest. The timelines can only serve to highlight spikes which require further investigation to understand the localised context. The EFF topic timelines were insightful for the reason that two of these topics were related to localised events but contained volume in other periods. This indicates that the topics may relate to other aspects that are overshadowed by the main keywords representing the topic.

8.8 Summary

Preliminary experiments using LDA showed similar results to the article experiment in Chapter 7 where LDA did not produce adequate clusters. Therefore, only NMF models were built and analysed for the tweets. The optimal model produced 26 topics, with most of the smaller topics containing unintelligible topic keywords. The topic distribution illustrated in Figure 8.3 indicated that the highlight topics discussed were well represented. The topics uncovered from the tweets were more difficult to interpret than the ones uncovered from the articles in Chapter 7, however the topics were descriptive enough to infer participants and the broad theme described by the keywords. The quantity of topics for the ANC and EFF in relation to the DA follows the trend seen in the tweet frequency and mention frequency illustrations, where the DA is less active and less engaged with in comparison to their counterparts.

The highlight topics produced by the model described party-related topics. There were references to Ramaphosa and the global citizen festival, IEC-related topics and a topic about government providing sanitary pads to underprivileged girls. The EFF topics showed contrasting themes with references to their campaign, Julius Malema's birthday and the passing away of his grandmother. The ANC topics mostly related to their election campaign. The DA topic in the highlight topics table (Table 8.1) referenced many of their affiliates and campaign rhetoric.

Two topic keyword wordclouds were generated, for the highlight topics and the full set of topics. The wordcloud for the full set of topics highlighted the EFF's dominance of Twitter with references to them flooding the wordcloud. The highlight topics were selected based on intelligibility and provided a more diverse view of the Twitter discourse with some of the noise and uninterpretable topics excluded. The focus in the highlight topics were on government and the country. Both wordclouds contained references to black and white indicating that a race-related theme might have been present in the discourse.

The topic timelines illustrated the life-cycle of a topic during the election period with

some topics localised to a short period and others existing throughout the observed period. The topics were not descriptive enough to isolate events, however the topic timelines highlighted high volumes of tweets in the middle of March and middle to end of April. All party-related topics experienced their highest volumes in the week of the election.

The next chapter analyses the tweets produced by political parties. The analysis focuses on a smaller sample of tweets which could potentially shed some light on the reasons for high volumes in tweets in the middle of March and second half of April.

Chapter 9

Understanding The Twitter Election Campaigns Of Political Parties

Political parties design strategies and campaigns to appeal to voters. Twitter provides a platform that political parties can use to directly interact with their target population and spread their rhetoric without third party moderation. The content propagated by political parties provides a window into the themes that are pillars of their campaigns. These tweets contain information about events and rallies held by the parties. Furthermore, the tweets of political parties in an election period is a record of their respective campaigns which can be contrasted against future elections to study the evolution of the campaigns.

The experiment in Chapter 8 performed NMF topic modelling on the entire tweet corpus, the corpus contained 6 million tweets and the optimal model produced 26 topics. The topics provided insight into themes propagated on Twitter in the observed period of February to May 2019. These themes captured the high level discourse which included many participants. This chapter focuses on the discourse contribution of political parties to understand the themes political parties chose to emphasise. Furthermore, the temporal properties of themes propagated by parties could provide insights into the spikes in volume seen in Chapter 8.

9.1 Objectives

- What themes were central to the Twitter campaigns of the ANC, DA and EFF?
- What events can be uncovered from the party tweets?
- Can the insights drawn from this experiment explain spikes in volume in the larger tweet corpus?

9.2 Experimental Setup

Figure 9.1 illustrates the process followed to build topic models on the tweets produced by political parties.

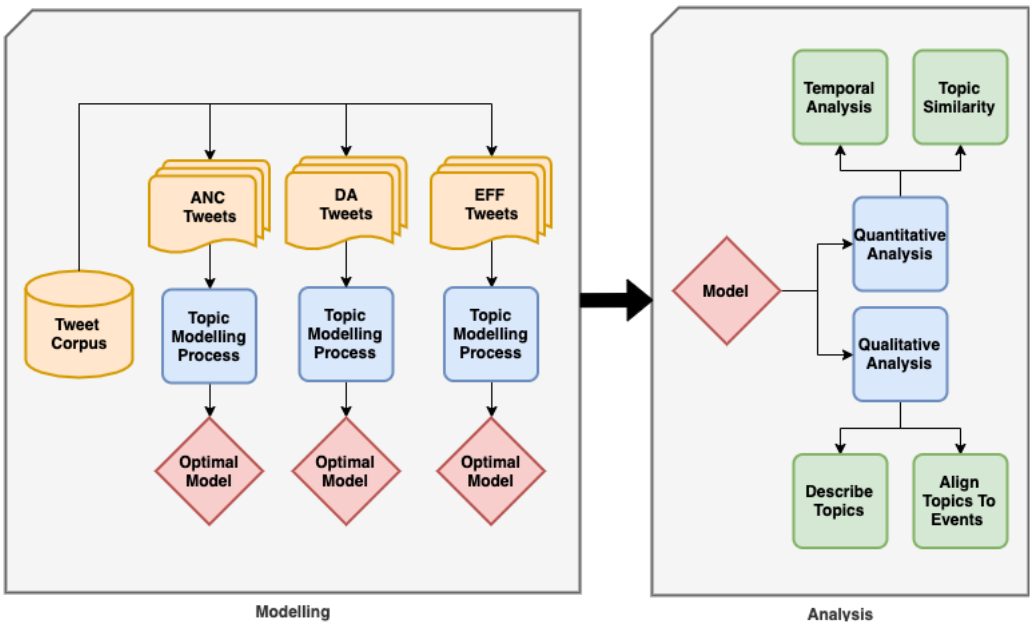


Figure 9.1: The process applied in the party tweet experiment

In Chapter 8 topic models were built on the entire tweet corpus, in this experiment party-specific corpora are segmented from the larger corpus. The corpus for each party

is made up of tweets from the party's official Twitter account and the Twitter account of the party's leader. NMF models are built for a range of k values and the model with the highest coherence is selected for further analysis.

The insights derived from the experiment is documented in the sections to follow. The analysis discussed include:

- The selection of the optimal model for each party corpus.
- Reviewing the highlight topics in each corpus by examining topic descriptions.
- Inspecting topic timelines to uncover information about the Twitter campaign of the ANC, DA and EFF.

9.3 Finding The Optimal Models

The coherence plot for the party models are illustrated in Figure 9.2. The coherence scores in Figure 9.2 are higher than the scores seen in Figure 8.2. The party models are built on smaller sets of tweets in comparison to the corpus model. The tweets published by parties serve a purpose of propagating the rhetoric of a party. The optimal values for k were achieved at 47, 13 and 48 for the ANC, DA and EFF respectively.

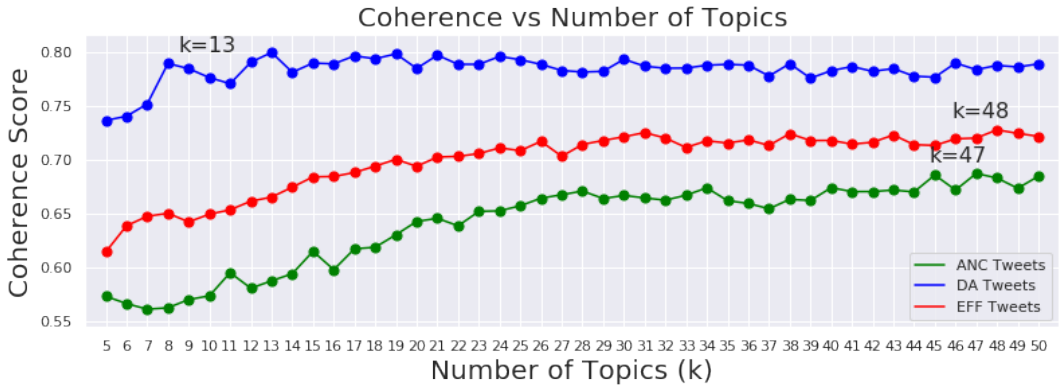


Figure 9.2: Coherence scores of party topic models

9.4 ANC Topic Analysis

The optimal ANC topic model produced 47 topics. This model had the lowest coherence amongst all the party models. According to Figure 5.6, the ANC increased the volume of tweets produced in the weeks prior to the election which is indicative of a last minute push. This section analyses the highlight ANC topics and the timelines associated with them to understand the context and temporal properties of the ANC's Twitter campaign.

9.4.1 Highlight Topics

The highlight topics produced by the ANC model are depicted in Table 9.1. There are two rally topics in the ANC model, these relate to the Siyanqoba Rally and a ANC Women's League Rally. The Siyanqoba Rally topic contains keywords that indicate that the president addressed this rally whereas the ANC Women's League Rally contains reference to Bathabile Dlamini who was most likely in attendance as the leader of the organisation. Another rally-related topic is topic 35 which describes the deputy secretary general engaging with the public in Khayelitsha, Western Cape. There are topics related to the ANC's election campaign containing hashtags that are central to their Twitter campaign. A topic also materialised regarding the results, which was expected because the ANC won the election.

Topic	Keywords	Label
5	siyanqoba, rally, ancsiyanqoba, evening, speaks, growsouthafrica, pledge, president, era, bikers	Siyanqoba Rally
20	amp, peoplesmanifesto, share, minister, provincial, growsouthafrica, corruption, find, education, reason	Manifesto
28	ancleads, phakamaramaphosa, iec, count, website, check, result, results, declaration, national	Election Results
34	job, create, economy, investment, decent, opportunity, number, black, invest, jobs	Economic Investment / Jobs
35	cape, western, town, eastern, resident, northern, ancdsg, delft, khayelitsha, engage	Engagement In Western Cape
38	government, build, house, work, africans, country, factoftheday, continue, commit, provide	Housing Commitment
40	voteanc, growsouthafrica, ancsiyanqoba, ivotedanc, reasons, eligible, voter, region, mamelodi, mayday	Election Campaign Final Push
44	question, land, respond, reform, ask, sustainable, programme, radical, provide, answer	Land Reform
45	freedom, ancwl, rally, hill, constitutional, women, night, right, league, bathabile	Women's League Rally

Table 9.1: Highlight topics from the ANC tweets

The remaining topics in Table 9.1 related to the rhetoric pushed by the ANC. There is a manifesto related topic which indicates that their manifesto deals with corruption, education and growing South Africa. Furthermore, their rhetoric contains topics that address economic growth in the country, job creation, land reform and the provision of housing.

The topics highlighted in Table 9.1 provide a summarised view of the ANC's Twitter campaign that breaks up into two categories: the party's events and rallies as well as their rhetoric for addressing critical issues in the country.

9.4.2 Topic Keyword Wordclouds

The topic keyword wordcloud for the ANC topics is illustrated in Figure 9.3. The terms *ANCSiyanqoba* and *GrowSouthAfrica* are prominent in the wordcloud, which were ANC election campaign slogans. The wordcloud indicates that the ANC tweets were focused on their pledge to the country, with references to education and their *GrowSouthAfrica* campaign. The ANC also urged the public to cast their votes.

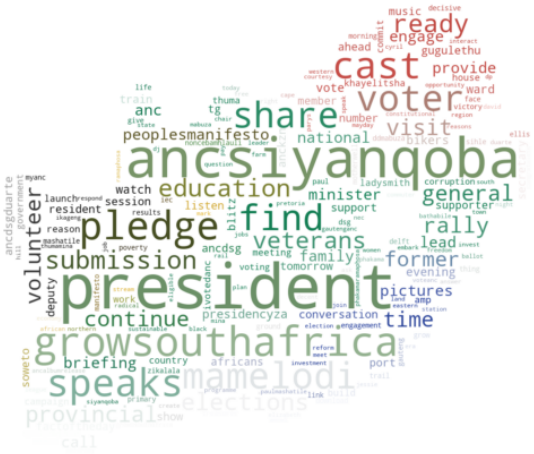


Figure 9.3: ANC topic keyword wordcloud

9.4.3 Topic Timelines

The ANC timelines illustrated in Figure 9.4 provide insight into the temporal aspects of their rallies, campaigns and public engagement. The Siyanqoba Rally is relevant only in the last week of April indicating that the event was announced or occurred close to the election. This was their final rally held on 5 May so it makes sense that it was publicised the week prior to the event. The ANC Women's League Rally and the engagements in the Western Cape saw spikes in volume in the week of 21 April, this was an indication of when these events took place. The ANC Women's League Rally occurred on 26 April thus explaining the spike in volume at the end of April. These topics also received a slightly smaller spike in volume in the week of 17 March.

The election campaign topic received increased volume in the two weeks prior to the election which is indicative of a last minute push to rally voters, the increase in volume for this topic coincided with the ANC Women's League Rally and Siyanqoba Rally topics. As the party that won the election, it was expected that an election results topic would materialise. This topic is relevant in the week of the election, most likely receiving volumes of tweets as the results were being counted and announced.

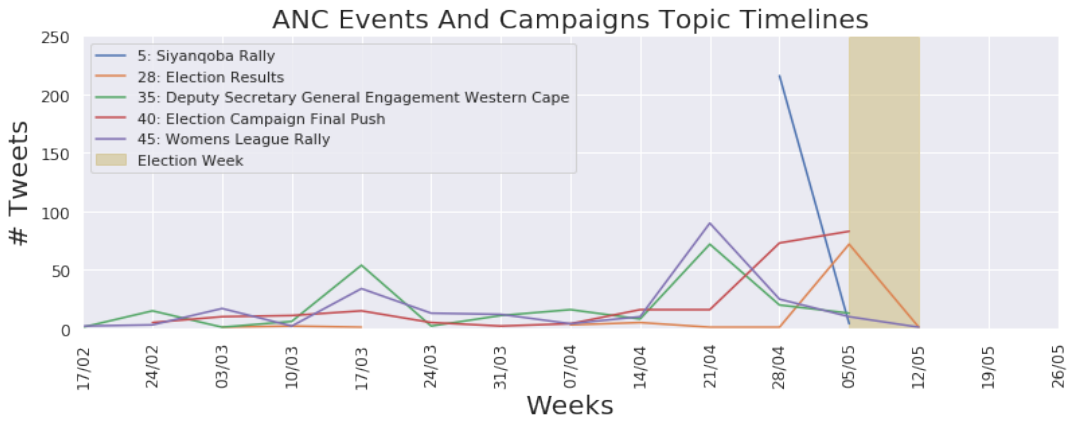


Figure 9.4: ANC events and campaigns topic timelines

The timelines in Figure 9.5 illustrates the life-cycle of topics relating to the ANC's rhetoric. The topic related to their manifesto is consistently propagated throughout the observed period with increased volume from the periods when the rallies occurred. The land reform topic received attention throughout the observed period with a spike in volume being observed in the same time periods as the two rallies. The economic growth and housing topics experienced similar trends with peaks in volume coinciding with the rallies.

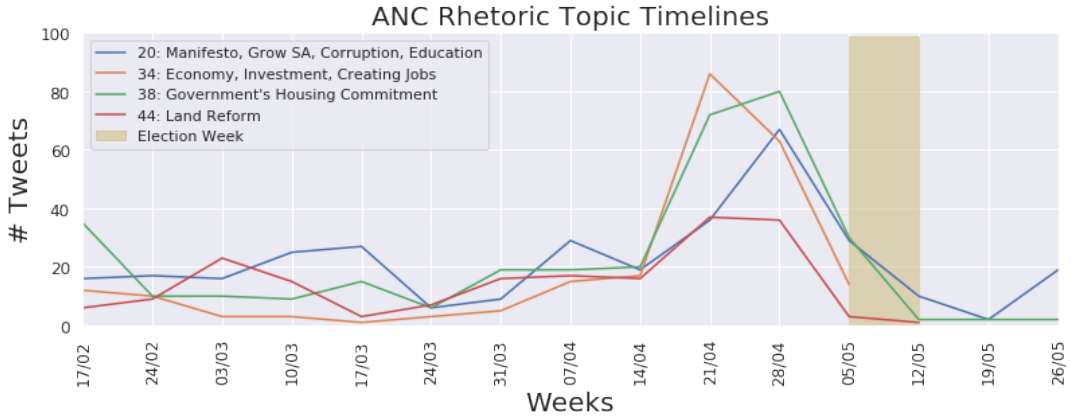


Figure 9.5: ANC rhetoric topic timelines

9.5 DA Topic Analysis

The DA topic model achieved the highest coherence amongst all the party models. The optimal model contains 13 topics, which is significantly lower than the models built on their opponents tweets. However, the DA was the least active party on Twitter. This section analyses the highlight topics from the DA topic model to understand the context of their Twitter campaign. Furthermore, the timelines associated with these topics are examined to understand the temporal properties of the DA's Twitter campaign.

9.5.1 Highlight Topics

Topic	Keywords	Label
1	mmusimaimane, leader, watch, address, explain, arrive, today, unpack, church, community	Maimane Addresses Church Community
3	anc, corruption, government, fail, election, corrupt, bosasa, burn, party, sollymalatsi	ANC Corruption & Failure
4	cape, western, alanwinde, northern, windeforpremier, province, eastern, town, premier, candidate	Alan Winde Premier Campaign
5	damanifesto, manifesto, full, launch, stadium, rand, live, saturday, johannesburg, offer	Manifesto Launch In Rand Stadium
6	vote, cast, voteda, khulada, special, election, important, time, voter, please	Election Campaign
7	dafinalrally, phetogorally, dobsonville, stadium, tomorrow, blue, soweto, join, lead, today	Final Rally In Dobsonville Stadium
8	president, ramaphosa, bosasa, cyril, son, andile, cyrilramaphosa, zuma, ask, deputy	Ramaphosa and Bosasa
9	amp, eskom, keepthelightson, crisis, power, sa, energy, sans, time, plan	Eskom
10	change, build, onesaforall, bring, ready, chance, voteda, message, choose, give	One SA For All
11	government, job, create, national, business, ensure, home, plan, put, work	Put A Job In Every Home
12	premier, gauteng, candidate, sollymsimanga, today, lead, msimangaforpremier, office, limpopo, zwakelem	Solly Msimanga Premier Campaign

Table 9.2: Highlight topics from the DA tweets

The highlight topics from the model built on the DA tweets are illustrated in Table 9.2. The DA model produced much fewer topics than their opponents but achieved a higher coherence score indicating that their Twitter campaign was narrow and focused on a specific agenda. The DA publicised campaigns for Alan Winde and Solly Msimanga for premiership of Western Cape and Gauteng respectively. The model contained a topic

related to Mmusi Maimane addressing a church community. There are two rally related topics present in the model, the first relates to the launch of the DA's manifesto in Rand Stadium whereas the other refers to their final rally in Dobsonville Stadium.

The topics relating to the DA's rhetoric focused on problems and solutions. They highlight failures and corruption in the current administration with references to the ANC, Eskom, Bosasa and Cyril Ramaphosa. Their campaign also involves the proposition of solutions that they would implement to put a job in every home and create one SA for all. Their topics highlight failures and problems in the current administration whilst promising jobs and unity if they are elected.

9.5.2 Topic Keyword Wordclouds

The DA topic keyword wordcloud is illustrated in Figure 9.6. The wordcloud indicates importance to terms relating to Bosasa, their premiership campaigns, their policies and Eskom. There are also references to the election and choice which was indicative of them trying to persuade the public to vote for them.

Figure 9.6: DA topic keyword wordcloud

9.5.3 Topic Timelines

The timelines of the DA topics are split into two categories. The first category, illustrated in Figure 9.7, contains tweets related to campaigns and events. Solly Msimanga's

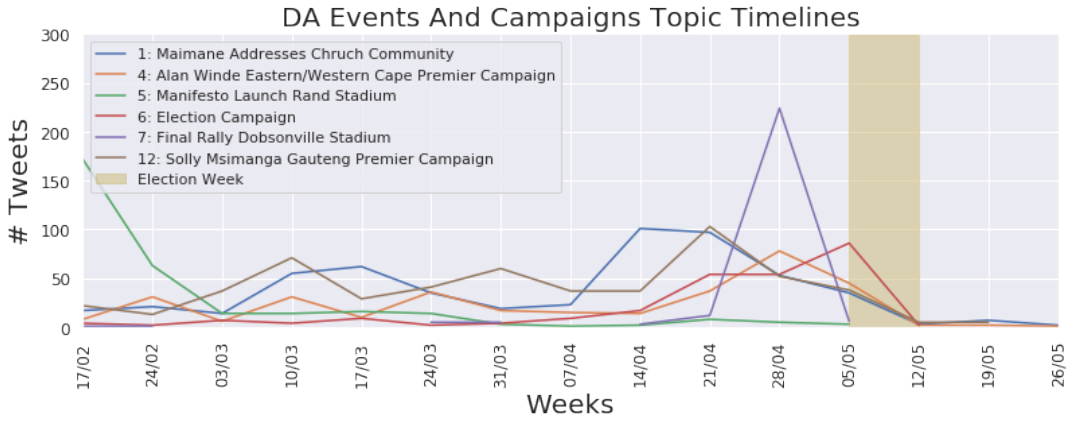


Figure 9.7: DA events and campaigns topic timelines

campaign to become premier of Gauteng consistently received more attention from the DA than Alan Winde's campaign to become premier of the Western Cape. This could be an indication of the DA feeling confident in securing the Western Cape premier race and giving more attention to Gauteng. The DA's final rally in Dobsonville happened at the end of April, the timeline for topic 7 experienced a large spike in volume in the week of 28 April. The election campaign topic received almost no activity until April, the volume gradually increases to around 100 tweets in the week of the election.

The topic timelines related to the DA's rhetoric are illustrated in Figure 9.8. The DA's "one SA for all" rhetoric received volume from the their final rally until election week. The DA's "putting a job in every home" rhetoric received volume in the middle of March and just before their final rally. The volume prior to their rally was expected as a strategy to draw attention to the party in their final push before the election. Eskom and load shedding was high on the DA's agenda. Their Eskom-related tweets reached peaks in March. This aligns with Eskom implementing stage 4 load shedding in the week of 17 March. This timeline also spiked just before their final rally, which most likely had the

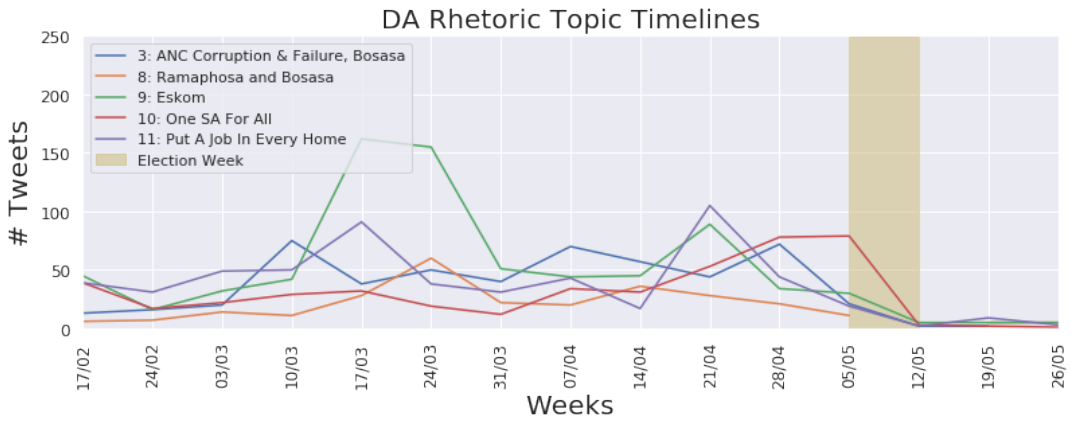


Figure 9.8: DA rhetoric topic timelines

intention of drawing attention to their party in their final push. ANC and Bosasa related corruption consistently received attention from the DA throughout the observed period, indicating that these were frequently brought up by the DA. This was probably done to draw attention to the problems in the current government and contrast it with the solutions proposed by them.

9.6 EFF Topic Analysis

The optimal topic model built on the EFF tweet corpus produced 48 topics. The EFF was the most active party on Twitter according to Figure 5.6. Their Twitter campaign was consistent and contained high volumes of tweets for most of the observed period. This section analyses the topics produced by the EFF model to understand the context of their tweets. Furthermore, the timelines of these topics are analysed to understand the temporal nature of the topics present in the EFF's campaign.

9.6.1 Highlight Topics

Topic	Keywords	Label
2	ourlandandjobsnow, voteeff, thing, occupy, future, efffinalpush, ieff, cicinkzn, situation, right	Election Campaign
5	cic, address, arrive, cicinkzn, conversation, cicinnorthwest, conclude, ask, walk, province	Malema address in KZN & NW
12	stadium, orlando, fill, ahead, efffreedomdayrallies, road, address, invite, join, nkowankowa	Shela Thupa Rally
17	meeting, community, address, cicinnortherncape, galeshewe, bushbuckridge, tswaing, underway, worker, hold	Malema address in NC
21	land, jobs, job, uphephela, landexpropriation, compensation, march, expropriation, news, shivambu	Land Expropriation
24	cape, western, northern, town, eastern, airport, international, metro, winnie, renaming	Renaming Cape Town Airport
30	government, ensure, free, education, settlement, local, house, commitment, state, current	Free Education \ Housing Settlements
36	efffinalpush, iamvotingeff, ichoosetovoteeff, ivotedeff, effmaydayrally, cast, choose, efffreedomdayrallies, voting, cicinlimpopo	Election Campaign Final Push
39	limpopo, province, rallies, mogalakwena, seshego, hold, freedom, efflimpoporally, spend, ntokozo	Limpopo Rally
45	leadership, lead, march, service, delivery, municipality, kzn, road, underway, demand	Service Delivery
46	sentletse, black, write, criminal, karima, pravin, ramaphosa, eskom, thing, campaign	Eskom / ANC

Table 9.3: Highlight topics from the EFF tweets

The highlight topics produced by the EFF topic model are depicted in Table 9.3. There are two rallies described in the topics, the Shela Thupa Rally and the Limpopo Rally. The Limpopo Rally took place in Seshego and the Shela Thupa Rally was held at Orlando Stadium. There are two topics related to addresses made by the EFF leader, Julius Malema. He addressed potential voters in Kwa-Zulu Natal (KZN), North West (NW) and Northern Cape (NC). The EFF election campaign topics focused on slogans that are central to their campaign and evident in the hashtag wordcloud of the party illustrated in Figure 5.11. There was a separation between their long-term election campaign and their final push before the election.

The topics that relate to the EFF's rhetoric include tweets about land expropriation, renaming Cape Town Airport, service delivery, free education, housing, Eskom and the ANC. The land expropriation topic was expected to be highlighted since this is central to the EFF's rhetoric with their campaign hashtag *OurLandAndJobsNow*. The party tabled a motion to change the name of Cape Town Airport to honor Winnie Mandela. The Eskom and ANC topic contained keywords referring to criminals, Pravin Gordhan,

Cyril Ramaphosa and Eskom.

9.6.2 Topic Keyword Wordclouds

The EFF topic keyword wordcloud is illustrated in Figure 9.9. The terms "statement" and "tomorrow" in the wordcloud indicate that the EFF made announcements to create anticipation of appearances and statements being made by the party. There are also references to the election, the EFF's *RedFriday* campaign and party members.

Figure 9.9: EFF topic keyword wordcloud

9.6.3 Topic Timelines

Similarly to the other parties, the EFF's timelines are split into two categories: topics related to their rhetoric and topics related to events and campaigns. The timelines for events and campaigns are illustrated in Figure 9.10. The topics related to Malema addressing voters in KZN and NW experienced a spike in volume in the weeks of 24 February, 17 March and the middle of April. These spikes could be indicative of when these addresses occurred. The KZN address occurred on 15 April whereas the NW

address occurred on 25 April, this explains the spike in volume that occur in April but does not explain the other spikes. This was likely due to other themes with a similar context getting consumed in this topic.

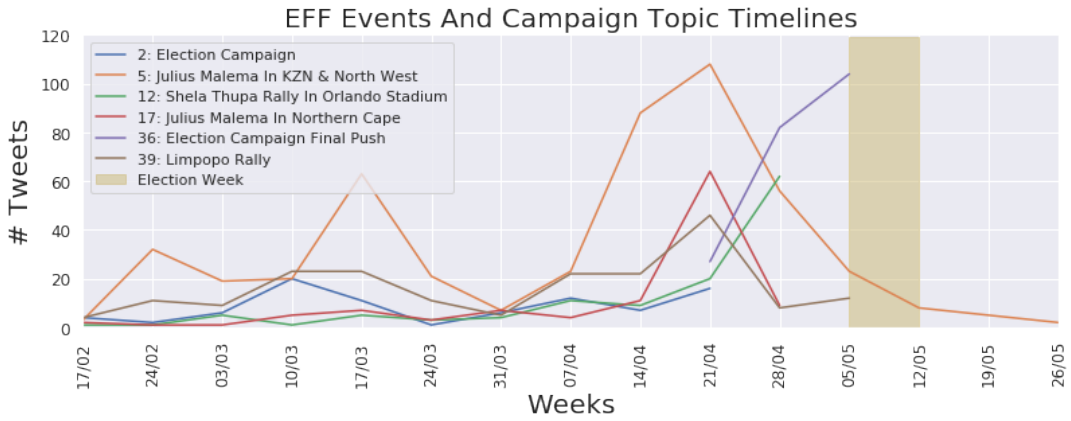


Figure 9.10: EFF events and campaigns topic timelines

The topic related to Malema's address in Northern Cape spiked in volume in the week of 25 April, this event occurred on 24 April thus explaing the spike in volume. The Limpopo Rally experienced a spike in volume in the week of 25 April, the event occurred on 27 April. The volume experienced elsewhere could be indicative of other discourse getting consumed in this topic. The Shela Thupa Rally occurred on 5 May, the timeline experienced a spike in volume in the week of 28 April which was most likely related to an announcement or reminder about the rally in the final weeks of the campaign. The election campaign final push topic receives a spike in volume that coincides with the rallies that occurred in the final weeks of April.

The topic timelines related to the EFF rhetoric are illustrated in Figure 9.11. The topic related to free education and housing consistently receives attention from the EFF throughout the observed period with spikes in volume occurring in the week of 10 March and 21 April. The 21 April spike was likely in relation to all the rallies held by the party at the end of April. Renaming Cape Town Airport to honor Winnie Mandela also experienced spikes in volume at similar periods.

The service delivery topic was consistently discussed by the EFF with peaks in volume

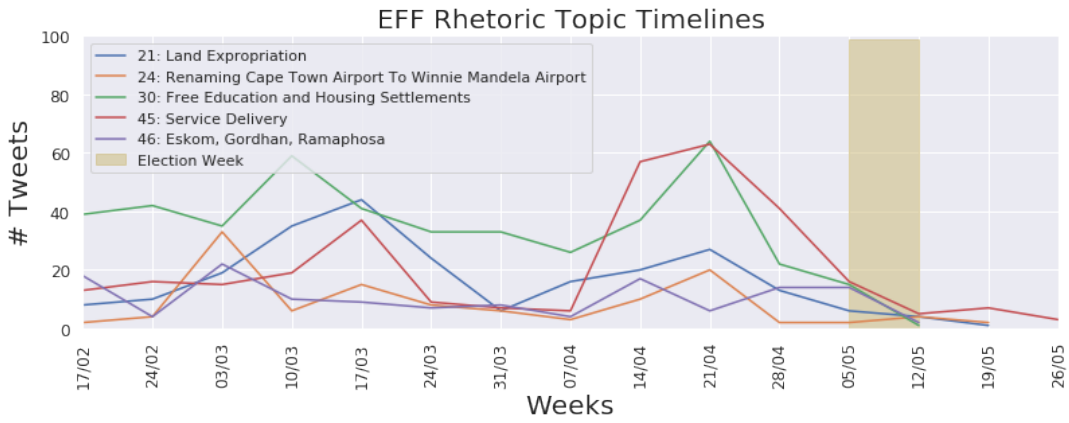


Figure 9.11: EFF rhetoric topic timelines

occurring in the week of 17 March and most of April. The spike in volume on 17 March coincides with Eskom implementing stage 4 load shedding. This aligns the topic to an event that explains the spike in volume. The land expropriation topic is consistently discussed throughout the observed period with peaks in volume coinciding with the service delivery topic. The topic related to Eskom and the ANC received some attention but appears to be a minimal part of the EFF's rhetoric in their election campaign.

9.7 Similarities Between Party Tweets

The topic similarity method described in Section 6.2.5 was applied to the topics produced by the party models. This was done to identify if any topics contain a similar vocabulary. The heatmap illustrated in Figure 9.12 shows the similarities that exist between ANC and DA topics. topic 9 of the DA model and topic 20 of the ANC model shared a similar vocabulary, the ANC topic addressed their manifesto and plan to grow SA and get rid of corruption whereas the DA's topic referenced Eskom and it's frailties.

The similarity most likely exists due to references to Eskom, mentions of the ANC and the government. The DA's topic 4 and the ANC's topic 35 share similarities, the similarity is the location, the Western Cape. The DA's topic referenced Alan Winde's premier

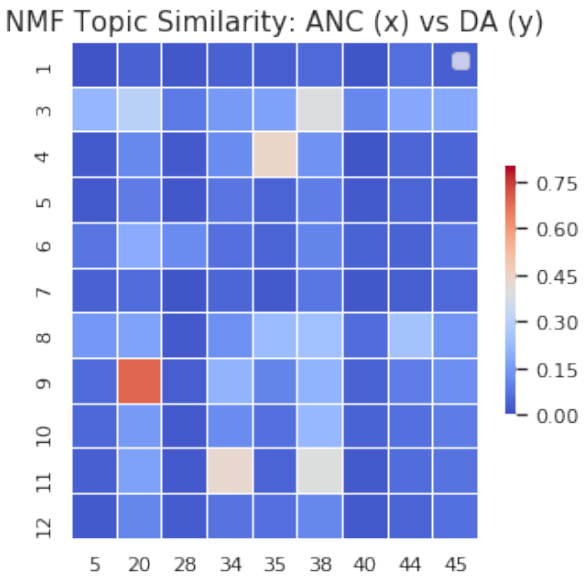


Figure 9.12: Topic similarity heatmap of ANC & DA topics

campaign and the ANC's topic referenced the deputy secretary general engaging with the public in Khayelitsha. The parties shared a similar vocabulary in their plans if elected to government. The DA's topic 11 shared similarities with topic 34 and 38 of the ANC model. The DA's topic dealt with their rhetoric to put a job in every home whereas the ANC's topics referenced their rhetoric around investments in the economy, creating jobs and continuing to build houses for the public. These topics contain campaign promises and the similarities can be explained by the intention of the topics.

The heatmap in Figure 9.13 illustrates the similarity that exists between ANC and EFF topics. The parties shared similarities in topics related to land reform in ANC topic 44 and EFF topic 21. Another theme that was shared was the commitment to provide housing within ANC topic 38 and EFF topic 30. Similarity occurred between ANC topic 35 and EFF topic 24 which referenced the EFF wanting to rename Cape Town airport and the ANC engaging with the public in Khayelitsha, the common element being the location, the Western Cape.

The heatmap illustrating the similarities between DA and EFF topics is shown in Figure 9.14. The comparison between the DA and EFF produced two similar pairs. DA

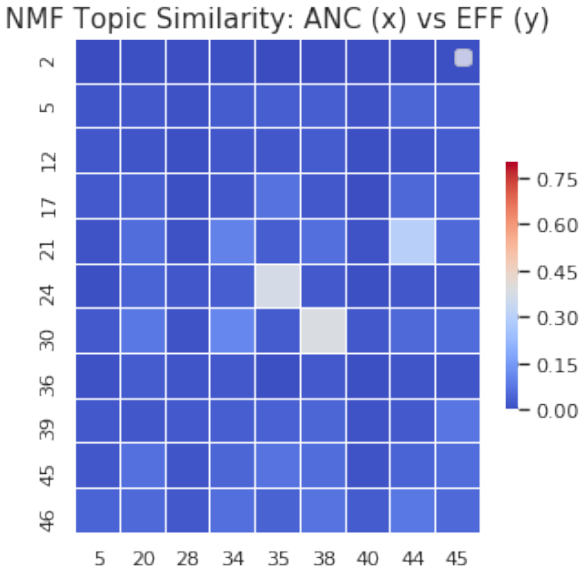


Figure 9.13: Topic similarity heatmap of ANC & EFF topics

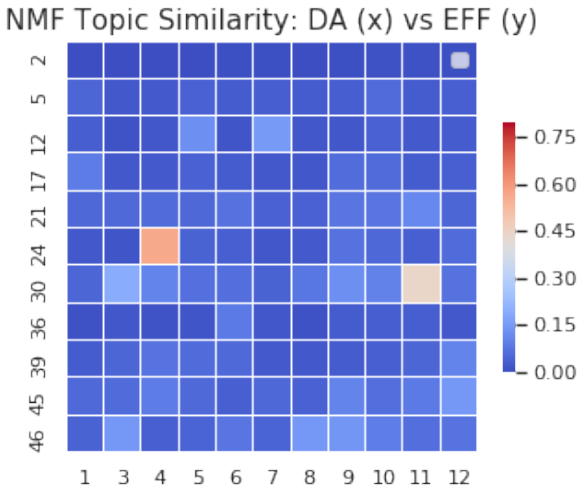


Figure 9.14: Topic similarity heatmap of DA & EFF topics

topic 4 and EFF topic 24 shared similarities, both reference the Western Cape. The EFF's topic referenced their motion to rename Cape Town airport whereas the DA's topic referenced Alan Winde's premiership campaign. The other notable pair was DA topic 11 and EFF topic 30, both these topics referenced the rhetoric of the party con-

cerned and their campaign to attract voters with election promises.

9.8 Summary

This chapter analysed the tweets produced by political parties. The ANC, DA and EFF tweets resulted in models that contained 47, 13 and 48 topics respectively. Inspecting these topics uncovered two categories of topics across all parties. Topics relating to events, campaigns and rallies and topics that were related to the rhetoric of the parties.

The ANC rhetoric focused on stopping corruption, housing, job creation, economic growth and land reform. The DA focused on highlighting corruption and failures in the ANC's current administration. They also focused on job creation, Eskom and proposing that they will create one SA for all. The EFF's focus was on free education, housing, land expropriation, service delivery and highlighting frailties in Eskom and the ANC. In contrast to the DA, the failures of the ANC was not central to the EFF's campaign.

The topic keyword wordclouds emphasised the important terms in each party's corpus according to topic keywords. The ANC wordcloud referenced their pledge to the country. The party urged voters to cast their votes. The DA wordcloud associated importance with their policies and their premiership campaigns. Furthermore, it highlighted corruption and frailties in relation to Eskom, Bosasa and the ANC government. The EFF wordcloud associated importance to the EFF announcing their statements either to create anticipation or attract attention.

The topic timelines explored for the parties shed some light on the relevance of the spike in volume in the larger corpus in the week of 17 March. This was due to stage 4 load shedding being implemented by Eskom. This received a spike in volume in the DA and EFF timelines as well as the larger tweet corpus, The ANC timeline indicated minimal spikes in volume in this period, however the context of their standpoint regarding Eskom could have gotten lost in an unintelligible topic. The spike seen in the larger corpus towards the end of April can also be explained by the numerous rallies and last minute

pushes by all 3 parties.

Furthermore, the topic timelines highlighted the campaign, rallies and events held by the parties. The EFF timelines were particularly descriptive of their numerous rallies held before the election, they visited multiple provinces in a 2 week period to rally voters. This busy period coincided with a decrease in tweets from the EFF in Figure 5.6. The dates for the rallies held by the ANC and DA were also easily identifiable from the timelines.

The topic similarity heatmaps explored for the parties uncovered relationships between the tweets of political parties. The similarities highlighted were either related to the intent of a topic or a location. Topics relating to activities in the Western Cape for all parties indicated similarities. Topics relating to the rhetoric of parties also indicated a similarity in the intent. These similar topics shared commonalities in promises to provide jobs, housing, free education and economic growth.

The next chapter analyses a combined corpus of 2019 articles and the political party tweets discussed in this chapter. The analysis focuses on identifying a relationship between the articles and party tweets in the context of the 2019 election discourse.

Chapter 10

Identifying A Relationship Between The Articles And Tweets

In the previous experiments the focus was on understanding the information present in either the articles (Chapter 7), the tweets (Chapter 8) or the party tweets (Chapter 9). The articles and the tweets describe the election from different perspectives. In Chapter 7 the 2019 articles were found to be issue-related. The tweets produced by political parties contained campaign-related content, however it also contained some issue-related content. Since political parties and news agencies are producers of information regarding the election, it would be interesting to understand the relationship that exists between the two information sources. The tweet corpus would not be useful for this experiment since the topics produced from those were broad and did not highlight specific issue-related content.

In this experiment the articles and party tweets are combined to uncover the relationship present between the two sources of information. Topic timelines and wordclouds are employed to examine topics that contain links between the two sources.

10.1 Objectives

- Can a relationship between articles and political party tweets be uncovered for issue-related topics?
- Which data source leads in terms of publishing content on issue-related topics?

10.2 Experimental Setup

Figure 10.1 illustrates the process followed to build topic models on the combined corpus of articles and party tweets.

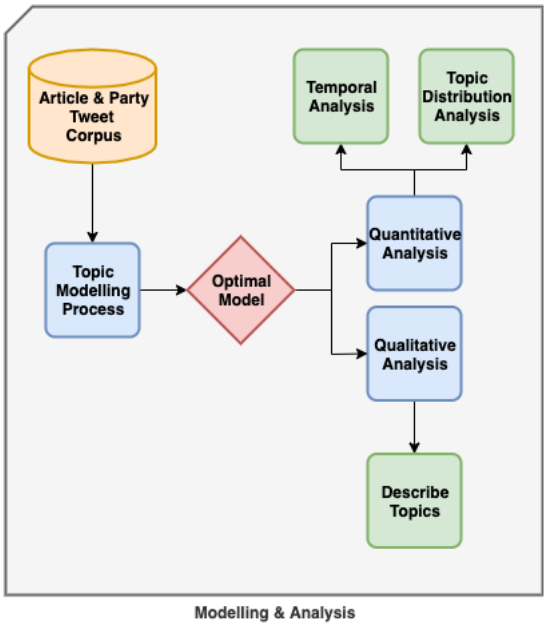


Figure 10.1: The process applied in the party tweet experiment

Previous experiments focused on a single corpus, whereas this experiment focused on a combined corpus. The party tweet corpus created in Chapter 9 was reused in this experiment. The party tweets were merged with the 2019 election-related articles. The

resulting corpus contained 28 000 documents. The articles accounted for 2 500 documents and the remainder consisted of party tweets. The skewed proportions between the two data sources were unavoidable since limited articles were available for this study. Nonetheless, the experiment focused on finding issue-related topics that contained articles and used those topics to determine the relationship between articles and party tweets. In this experiment, NMF models are built for a range of k values and the model with the highest coherence is selected for further analysis.

The insights derived from the experiment is documented in the sections to follow. The analysis discussed include:

- The selection of the optimal model.
- Reviewing the highlight topics that are relevant to this experiment.
- Inspecting topic timelines to uncover the relationship that exists between news articles and political party tweets.
- Analysing the content in high volume periods using wordclouds.

10.3 Finding The Optimal Model

The topic modelling process described in Figure 6.2 was used to build models on the combined corpus for k values ranging from 20 to 50. 20 topics were chosen as a lower bound for this experiment because the optimal model produced on a range of 5 to 50 was 10 topics. Those 10 topics were overwhelmed by tweet content that did not highlight issues thus the lower bound of the model building process was increased to 20.

Figure 10.2 illustrates the different coherence scores achieved for the different values of k. The optimal model for the combined corpus was achieved at k=37.

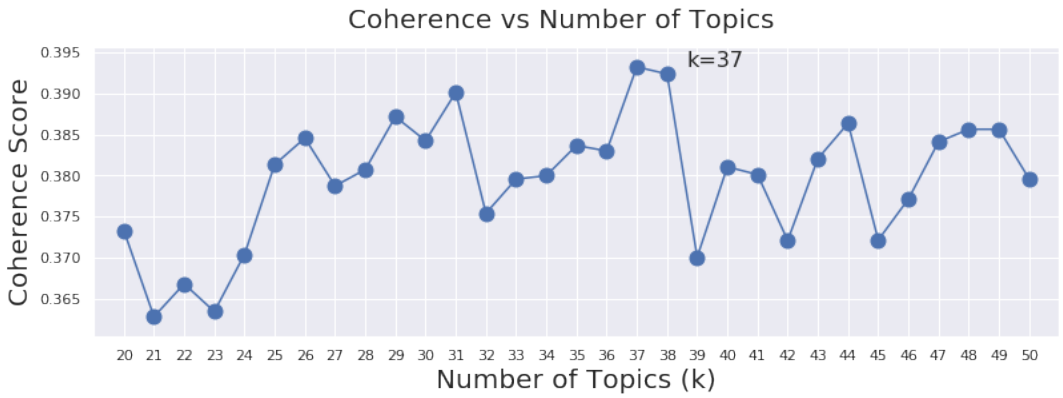
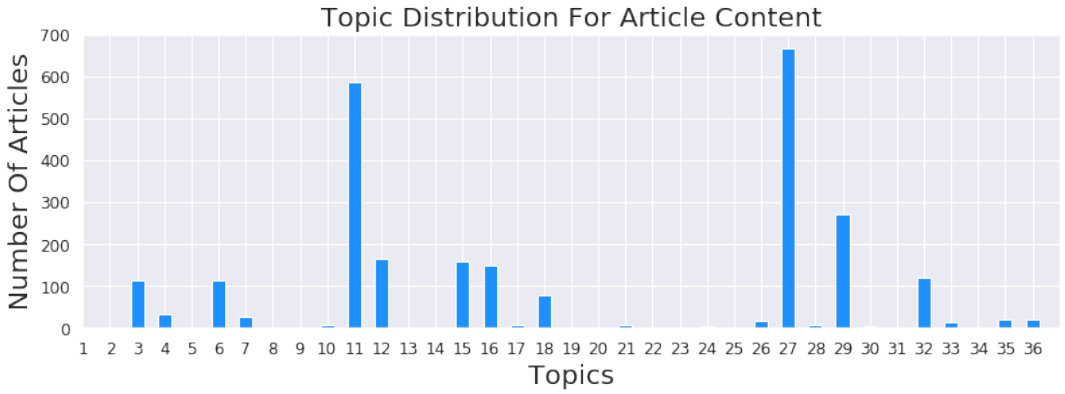


Figure 10.2: Coherence scores for the article and party tweet topic models

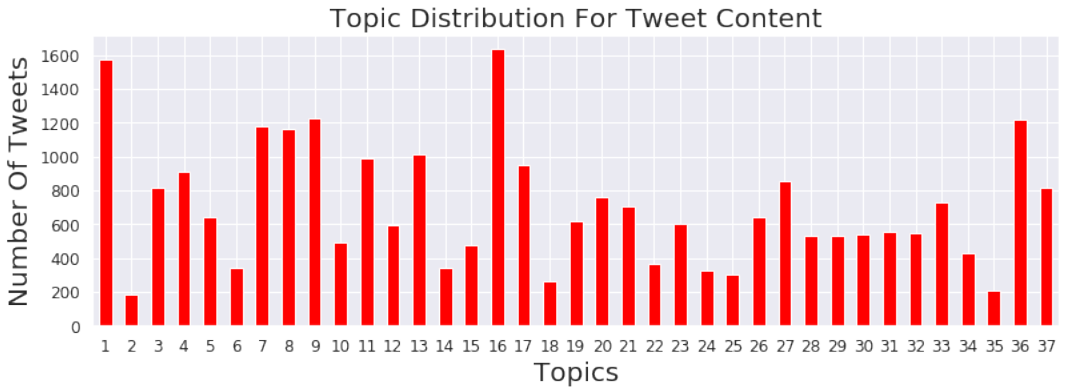
10.4 Topic Distributions

Each document in the corpus was tagged with a label corresponding to the dominant topic. Since the corpus contains two data sources, the distributions are discussed separately due to the disproportion between data sources. Figure 10.3 illustrates the distribution of topics within the optimal model. The article distribution (Figure 10.3a) illustrates that the articles are not represented in all the topics, this was expected since the articles account for 15% of the corpus. The tweet distribution (Figure 10.3b) illustrates that the tweets are strongly represented in most of the topics. Topics that are representative of articles include topic 11, 12, 15, 16, 27 and 29. These topics will be explored in the sections that follow to understand if a relationship between the articles and tweets exist.

The topic distributions in Figure 10.3 illustrate that a skewed corpus (where one data source is represented more than the other by a large proportion) does not produce topics that are representative of both data sources. Since these were the only election-related articles available for this study, it was not possible to overcome this problem. Future work should aim to collect more election-related articles to understand if balancing the corpus between data sources has an effect on the resulting model.



(a) Article Topic Distribution



(b) Tweet Topic Distribution

Figure 10.3: Topic Distribution For Combined Corpus Model

10.5 Highlight Topics

Since the articles are not represented in all topics, topics that contain articles were the focus of the experiment. Topics represented by articles were annotated to identify issue-related topics. Many of the topics uncovered in this experiment resonated with the topics described in Chapter 9. This was expected since majority of the corpus in this experiment contains party tweets. The highlight topics produced by the combined corpus model is depicted in Table 10.1. The topics contained references to Eskom and the political parties. Two Eskom topics were produced, topic 11 appeared to be associated with the DA campaign slogan *KeepTheLightsOn* that was central to their rhetoric regarding

Eskom whereas topic 12 contained keywords related to Eskom and the load shedding crisis. Topic 15 referenced the ANC Twitter campaign. Topic 16 contained references to Malema and race-related keywords. The DA and EFF final rallies featured as keywords in topic 29. Topic 27 referenced multiple aspects, including Limpopo, a premier candidate campaign and voting.

Topic	Keywords	Label
11	mmusimaimane, leader, arrive, explain, keepthelightson, damanifesto, kasitokasi, address, unpack, premier	DA / Eskom
12	eskom, power, load, shed, stage, utility, time, crisis, minister, energy	Eskom
15	voteanc, ivoteanc, ivotedanc, growsouthafrica, siyanqobarally, left, recap, voting, ready, reasons	ANC
16	people, young, black, malema, white, life, country, continue, call, fight	Malema / Race Related
27	today, freedom, mark, voting, premier, remember, limpopo, days, candidate, march	Voting / Premier / Limpopo
29	stadium, efftshelathuparally, orlando, efffinalpush, tomorrow, fill, dobsonville, dafinalrally, lead, road,	EFF Rally / DA Rally
33	land, reform, jobs, expropriation, malema, compensation, question, constitution, ownership, uphephela	Land Expropriation

Table 10.1: Highlight topics from the article & party tweet model

The article experiment in Chapter 7 found that the 2019 articles were issue-centric. The land expropriation topic, topic 33, did not contain a large volume of articles but it was a topic centred around an issue therefore it was included in Table 10.1 since it can provide an understanding of the relationship between the articles and the political party tweets when the topic timelines are explored in the next section.

10.6 Topic Timelines

The main objective of this experiment was to identify any relationship that exists between articles and party tweets. The topics listed in Table 10.1 were illustrated on topic timelines to identify if the articles and tweets displayed any noticeable relationship. Majority of the topics contained no identifiable links between the articles and tweets except for two issue-centric topics. These topics were related to Eskom (topic 12) and land expropriation (topic 33).

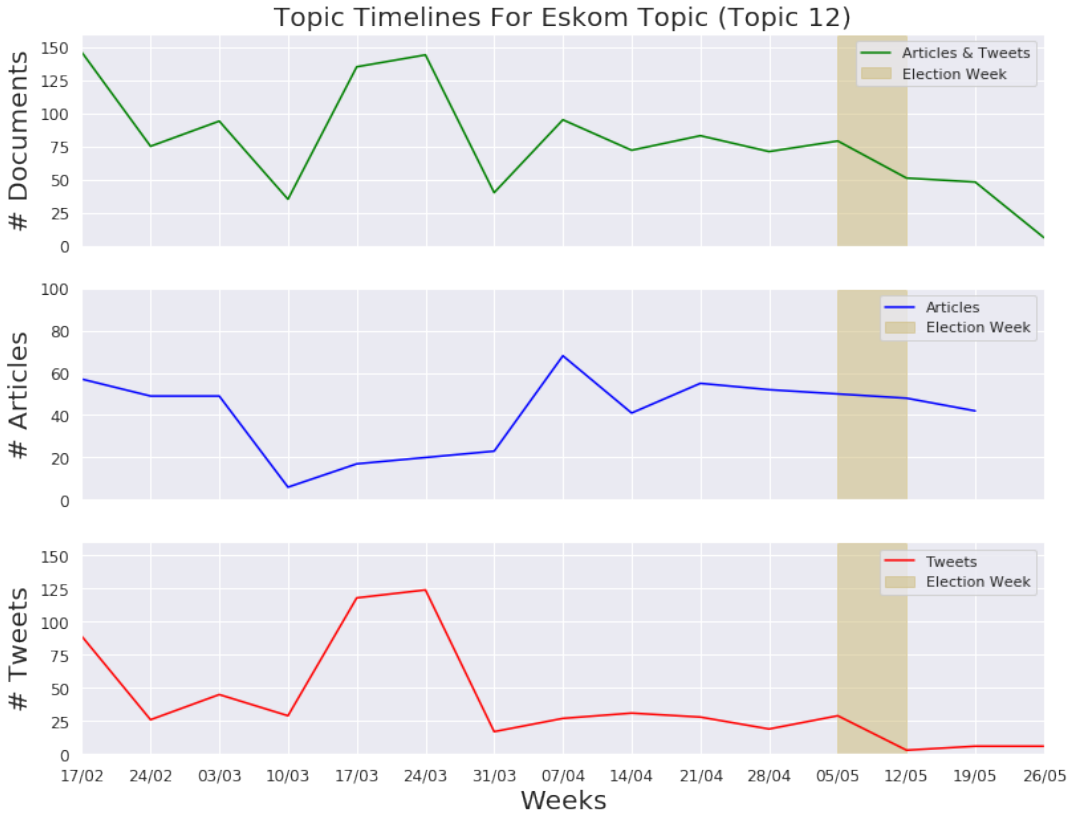


Figure 10.4: Eskom Topic Timelines

The timelines related to the Eskom topic are illustrated in Figure 10.4. There was a large spike in volume in the combined corpus that occurs in the second half of March. Contrasting the timeline of the article and tweet component of the topic indicated that the topic gains traction on Twitter in the second half of March followed by an increase

in volume in articles related to the topic at the start of April. This indicated that the party tweets created discourse around this topic with a two week lead on the articles.

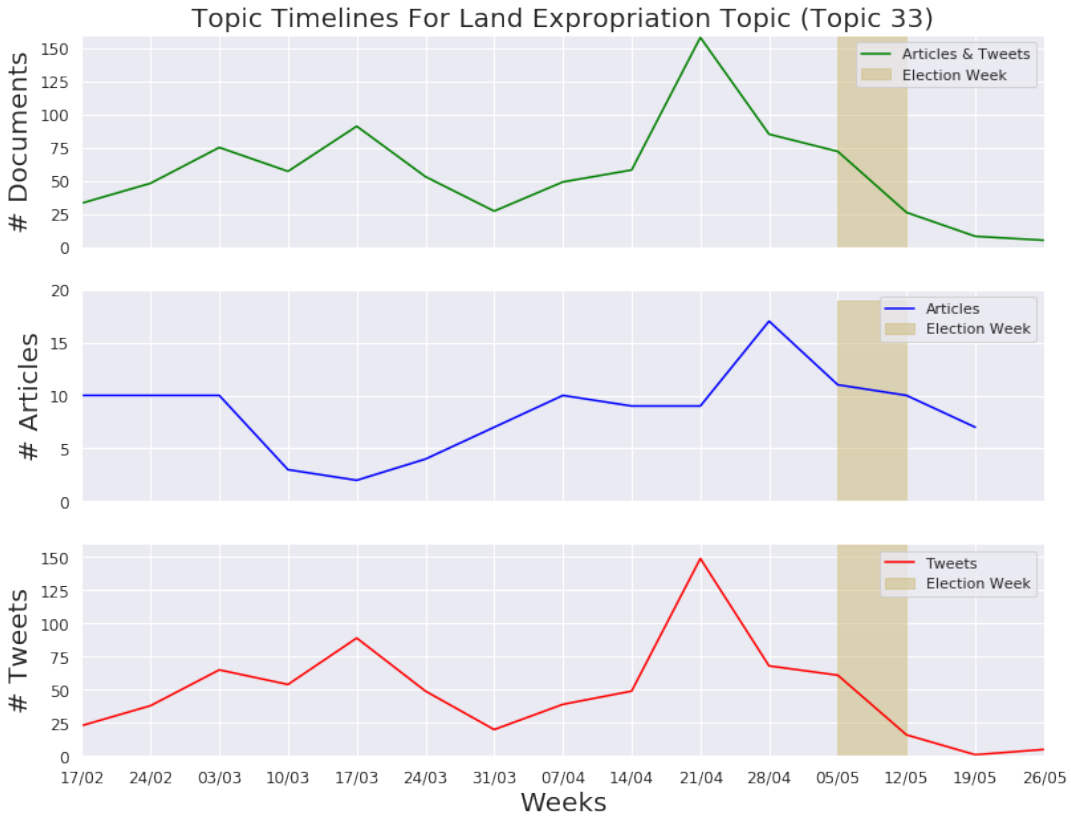


Figure 10.5: Land Expropriation Topic Timelines

Similarly, the land expropriation topic illustrated in Figure 10.5 indicated a similar trend to Figure 10.4. The land expropriation topic experienced a spike in volume towards the end of April. The constituent article and party tweet spikes indicated that attention is received on Twitter prior to the news articles. The article spike lagged behind the tweet spike by a week.

Both data sources contained different publishing trends and cannot be matched consistently throughout the observed period. However, in moments of high volume, the trend indicated that content relating to the two topics discussed received attention on Twitter prior to receiving attention in the news.

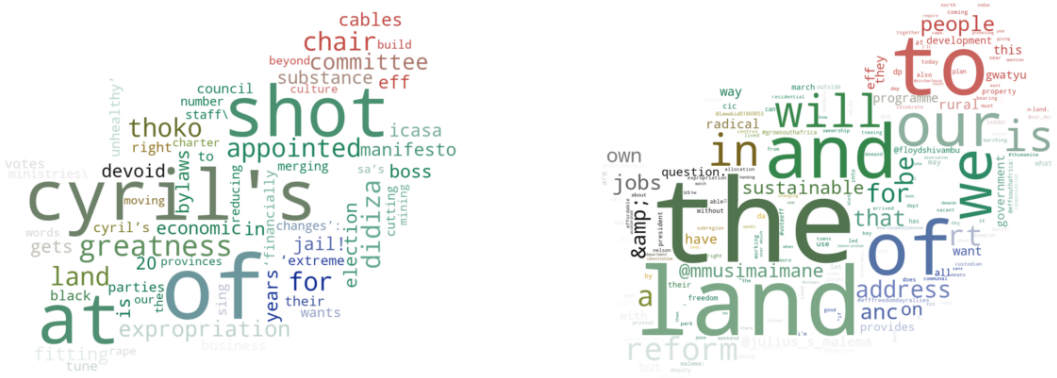
10.7 Topic Wordclouds

The topic timelines highlighted the relationship that exists between the articles and tweets within two issue-centric topics related to Eskom and land expropriation. The topic keyword wordclouds used to describe topics in previous experiments did not add value to this experiment since the article and tweet timelines were for the same topic and had the same keywords. Highlighting the content related to a spike in volume in the topics is achieved by building a wordcloud on the content of the articles and tweets in the week that the spikes occurred.

(a) Article wordcloud (b) Tweet wordcloud

Figure 10.6: Wordclouds for the Eskom topic volume spikes

Figure 10.6 illustrates the wordclouds for the Eskom topic. The wordclouds were built with the content from the documents in the weeks when spikes in volume occurred in the articles (week of 7 April) and tweets (weeks of 17 March and 24 March). The spike in volume seen in the tweets led the spike in volume seen in the articles by a two week period. The tweet wordcloud (Figure 10.6b) contained references to Eskom and the DA's *KeepTheLightsOn* campaign associated with load shedding. Load shedding happened in the middle of March coinciding with the spikes in volume seen in the tweets in the same period. The article wordcloud (Figure 10.6a also contained references to the load shedding crisis and Eskom. The context of both wordclouds were similar, indicating that both address the same event at different times.



(a) Article wordcloud (b) Tweet wordcloud

Figure 10.7: Wordclouds for the land expropriation topic volume spikes

The wordclouds in Figure 10.7 contains the content related to the land expropriation topic. The spikes in volume occurred in the week of 21 April and 28 April for the tweets and articles respectively. The tweets experienced a spike in volume for this topic a week prior to the articles. The tweet wordcloud (Figure 10.7b) contained references to land reform and political parties. The article wordcloud (Figure 10.7a) used different terminology to the tweets, referring to land expropriation instead of land reform. The articles mentioned the parties and the president. Both data sources referenced the same concept. The spike in the tweets led the articles by a week, most likely due to the conversation about land reform starting on Twitter by political parties prior to News24 giving attention to the topic.

Analysing both issue-centric topics indicated that information related to Eskom and land expropriation received attention on Twitter prior to it being reported in the news. However, the article sample used was from a solitary source. Examining more articles from different news agencies could provide a more complete analysis of which data source propagates information on issue-centric topics first.

10.8 Summary

In this chapter, articles and party tweets were used to created a combined corpus of 28 000 documents. The corpus contained mostly party tweets with the articles representing 15% of the corpus. Majority of the topics produced by the combined model contained keywords that were similar to the political party models discussed in Chapter 9. This was expected since most of the documents in the combined corpus were political party tweets. The articles were not represented in all the topics produced by the model therefore topics that contained articles were annotated in an effort to identify issue-related topics. Two issue-related topics were identified, Eskom and land expropriation.

The topics were inspected on topic timelines to understand the temporal relationship that existed between the articles and party tweets that relate to these two topics. In both topics, spikes in volume occurred on Twitter prior to the news articles. In the case of the Eskom topic, the articles lagged behind the tweets by two weeks. A similar trend was observed in the land expropriation topic where the articles lagged a week behind the tweets.

Topic wordclouds were generated for the periods in which spikes in volume occurred. The wordclouds indicated the the high volume periods for both the Eskom and the land expropriation topics addressed the same concept in both the articles and tweets. Therefore, based on the sample of data used in this study, it can be concluded that the issue-centric topics inspected are discussed on Twitter prior to being published in news articles. Increasing the diversity and size of the article corpus would provide a more complete analysis of this finding, however that is reserved for future work.

The next chapter provides a discussion on the analysis conducted in this study. Linking the insights found with the objectives of this study.

Chapter 11

Discussion

This study focused on discovering knowledge from articles and tweets by applying exploratory analysis and topic modelling techniques to uncover latent information about the election period. Chapter 5 performed exploratory analysis on the articles and tweets to uncover preliminary findings that provided a baseline for deeper analysis to build upon.

Chapter 7 analysed the election-related news articles from 2014 and 2019, providing insights into the newsworthy themes published by News24 in both these periods. LDA and NMF were contrasted in this experiment. LDA produced dominant topics which claimed majority of the documents. This hindered further analysis using topic timelines and topic similarity heatmaps therefore LDA was excluded from the tweet experiments because the technique inadequately grouped documents into clusters.

Chapter 8 aimed to draw insights from the tweet corpus by applying NMF to the tweets. The topics produced in this experiment provided insights into broad themes and participants however context of localised themes were consumed into broader themes.

In Chapter 9, a subset of tweets that were published by political parties and their leaders were analysed to understand the election campaign ran by political parties on Twitter. The topics produced in this experiment provided insight into the campaign strategies and

rhetorics of political parties. In this experiment, a latent separation between rhetoric-related topics and campaign and event related topics was found to exist for all three parties.

Chapter 10 analysed a combined corpus of 2019 election-related articles and political party tweets. The aim was to identify a relationship that existed between the articles and party tweets. Two issue-centric topics related to Eskom and land expropriation displayed a relationship where the party tweets that discussed these topics experienced a spike in volume prior to a spike in volume occurring in the articles.

This chapter focuses on answering the questions posed in Chapter 1 using the insights drawn from the experiments. The discussion focuses on the contrasts between the topic modelling techniques, the insights uncovered from the news articles and tweets, the Twitter election campaign ran by political parties and the relationship that exists between articles and tweets. Furthermore, brief discussions are presented on the experimental variations in the experiments that did not result in improvements and a contrast between this study and [2].

11.1 Contrasts Between LDA & NMF

The two topic modelling techniques of interest were applied to articles and tweets. NMF consistently produced models that were more coherent than LDA models. The article experiment showed that NMF produces more evenly distributed clusters from topic assignments in comparison to LDA. LDA produced dominant topics for 2014 and 2019 articles which negatively influenced the analysis on the clusters created by applying topic modelling. Analysis such as topic timelines, topic keyword wordclouds and topic similarity comparisons were hindered by the presence of the dominant topic. The analysis in this study relied on the topic modelling techniques producing representative clusters of documents to support further analysis into the topics. Preliminary experiments on the tweets indicated a similar outcome which resulted in LDA being excluded from the tweet experiments.

The LDA article topic keyword wordclouds were sparsely populated in comparison to their NMF counterparts, this was due to the dominant LDA topics accounting for large proportions of documents thus limiting the vocabulary of the wordcloud and increasing the frequency of terms within the dominant topic. Topic timelines could only be produced for the NMF models. The dominant topics of the LDA models affected the timelines of other topics making these uninterpretable for identifying their temporal properties. The comparison of LDA and NMF topics for 2014 articles showed that the dominant LDA topic contained a similar vocabulary within its articles to majority of the NMF topics, this is due to this topic claiming membership of most of the articles from 2014 thus making the comparison very noisy.

LDA produced inadequate clusters for these corpora thus justifying the exclusion of the technique. This study focused on discovering knowledge from text corpora and the contrast between techniques was an auxiliary objective therefore the exclusion of LDA for experiments on the tweets did not affect the objectives of the study.

11.2 The Insights Extracted From The Articles

The exploratory analysis uncovered that the ANC received the most coverage in 2019, the ANC's coverage increased from 2014 whereas the coverage received by the DA and EFF had decreased. The election-related articles in 2014 start appearing in the discourse from March whereas in 2019 the election-related articles start appearing from January. The popular trigrams highlighted themes relating to the president, public protector, state capture and Eskom issues.

The article experiment focused on analysing the topics created from the 2014 and 2019 articles. Preliminary experiments showed that there were no notable differences between the topics created on the body or the summary of the articles, therefore the article body was used for this experiment because it contained more text. The analysis uncovered themes relating to corruption, state capture 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 with SARS and the upcoming election.

The 2019 LDA topic keyword wordcloud provided a contrast with the NMF wordcloud where LDA covered terms relating to corruption, the ANC, Bosasa and Eskom and the NMF model covered mostly Eskom and ANC related content. The contrast between 2014 and 2019 as viewed through these wordclouds indicated that the article discourse of the 2019 election focused more on corruption and problems whereas the 2014 discourse contained more references to political parties and election-related aspects.

The 2014 NMF topic timelines illustrate peaks in volume in the middle of March for topics related to Nkandla and the Nkandla report, another notable spike in volume occurred at the end March when the DA filed a court application to release the Nkandla report. The 2019 topic timelines experienced spikes in volume in the middle of February and at the start of April. The volume in February was mostly due to the SONA whereas the spike in the Eskom topic in early April was most likely due to load shedding.

Notable topic similarities were discovered between LDA and NMF in 2014 for topics related to the DA's court application to release the Nkandla report and Julius Malema and his tax battles with SARS. The notable similarities for the 2019 comparison of LDA and NMF topics were between topics containing reference to Eskom and state capture. Contrasts between the NMF models for 2014 and 2019 revealed similarities in topics relating to the ANC and SONA as well as similarities between topics relating to the DA.

11.3 The Insights Extracted From The Tweets

Descriptive statistics on the tweet corpus found that tweets contain on average 18 tokens which is significantly shorter than the average length of the articles which was 500 tokens. Wordclouds were generated to illustrate the frequently used hashtags in the tweet corpus, the themes of the ANC and EFF were prominent in the wordcloud whereas the DA related hashtags were minimally represented.

The analysis of the tweets performed in Chapter 8 focused on the topic models built using NMF. There were 26 topics produced by the optimal model, many of the smaller topics contained unintelligible keywords. The topics uncovered from the tweets were more difficult to interpret than the ones uncovered from the articles. However, the topics were descriptive enough to infer some high profile participants and the broad themes described by the keywords. The model contained more topics for the ANC and EFF in comparison to the DA which followed the trend seen in Figure 5.5 and Figure 5.7 where the DA is less active and less engaged with in comparison to their counterparts.

The highlight topics produced by the model describes party-related topics which contained references to Cyril Ramaphosa and the global citizen festival, IEC-related topics and a topic about government providing sanitary pads to underprivileged girls. The EFF topics showed contrasting themes with references to their campaign and the passing away of Julius Malema's grandmother. The ANC and DA topics focused mainly on their campaigns. The topics uncovered from the tweets did not provide much insight into localised discussions and topics but rather provided a broad view of the context. The topics contained many references to participants in the discussions. This is most likely due to information being conveyed in other media (images, videos and links) and the text of the tweet serving the purpose of bringing participants into the discussion.

There were two topic keyword wordclouds generated for the tweets, one that contained all the topics and another that contained the highlight topics. The wordcloud for all the tweets highlighted the EFF's dominance of Twitter with references to their campaigns and personnel flooding the wordcloud. The highlight topics were selected based on intelligibility and provided a more diverse view of the Twitter discourse with some of the noise and uninterpretable topics excluded. The focus in the highlight topics were on government and the country. Both wordclouds contained references to black and white indicating that a race-related theme might have been present in the discourse. The contrast between both wordclouds indicates that much of the discourse related to the EFF was assigned to unintelligible topics, this is likely a result of the short context present in tweets.

The tweet topic timelines illustrated the life-cycle of topics during the election period with some topics localised to a short period and others existing throughout the observed period. The topics were not descriptive enough to isolate events, however the topic timelines highlighted high volumes of tweets in the middle of March and second half of April. All party-related topics experienced their highest volumes in the week of the election.

The analysis of the entire tweet corpus did not provide very informative results, it did group together topics related to parties but content related to parties also got lost in uninterpretable topics. Majority of the topics could be understood but only on a broad level. Party-related topics could only be classified as belonging to a party and addressing their campaign, specific insights about events could not be drawn from the topics. The topic timelines could not uncover localised events but did uncover two periods of high volume that warranted further investigation.

11.4 The Insights Extracted From The Twitter Campaigns Of Political Parties

The ANC's hashtag vocabulary was narrow and indicates a campaign focused on attracting votes. The DA's hashtag vocabulary contained campaign slogans as well as hashtags relating to problems in the country at the time. The EFF's Twitter campaign focused on campaign slogans as well as unique hashtags for rallies and events held by the EFF in the observed period.

The temporal analysis on party tweet behaviour uncovered that the EFF and ANC were the most active parties on Twitter with the EFF consistently publishing tweets in the observed period and the ANC producing a last minute rally on Twitter, surpassing the EFF's tweet frequency in the last two weeks prior to the election. The DA was the worst performing party in terms of tweet coverage. Their agenda was not propagated on Twitter to the same extent as their competitors.

Analysing the topics produced by political parties uncovered two underlying categories in the topics: topics related to the policies of the parties and topics related to events, campaigns and rallies. The ANC rhetoric focused on stopping corruption, providing housing, creating job, economic growth and land reform. The DA focused on highlighting corruption and failures in the ANC's current administration. They also focused on job creation, Eskom and proposing that they will create one SA for all. The EFF's focus was on free education, housing, land expropriation, service delivery and highlighting frailties in Eskom and the ANC. In contrast to the DA, the failures of the ANC was not central to the EFF's campaign.

The ANC topic keyword wordcloud referenced their pledge to the country and the party urging voters to cast their votes for them. The DA wordcloud uncovered the importance associated with: their plan if elected, their premiership campaigns and highlighting corruption and frailties in relation to Eskom, Bosasa and the ANC government. The EFF wordcloud associated importance to the EFF releasing frequent statements, either to create anticipation or attract attention to the party.

The topic timelines explored for the parties shed some light on the relevance of the spike in volume in the larger corpus in the week of 17 March, this was due to stage 4 load shedding being implemented by Eskom. This received a spike in volume in the DA and EFF timelines as well as the larger tweet corpus, The ANC timeline indicated minimal spikes in volume in this period, however the context of their discussion regarding Eskom could have gotten lost in an unintelligible topic. The spike seen in the larger corpus towards the end of April can also be explained by the numerous rallies and last minute pushes by all three parties. The topic timelines also highlighted the campaign rallies and events held by the parties. The EFF timelines were particularly descriptive of their numerous rallies held before the election, they visited multiple provinces in a 2 week period to rally voters. This busy period coincided with a decrease in tweets from the EFF in Figure 5.6. The dates for the rallies held by the ANC and DA were also easily identifiable from the timelines.

The topic similarity heatmaps explored for the parties uncovered relationships between

the tweets of political parties where the similarities highlighted were either related to the intent of a topic or a location. Topics relating to activities in the Western Cape for all parties indicated similarities. Topics relating to the rhetoric of parties also indicated a similarity in the intent. These similar topics shared commonalities in promises to provide jobs, housing, free education and economic growth.

The analysis related to the political parties produced high quality topics that highlighted events and rhetorics and provided a good summary of their Twitter campaigns.

11.5 The Relationship Between Articles And Political Party Tweets

The 2019 articles and party tweets were used to created a combined corpus of 28 000 documents. The party tweets accounted for 85% of the combined corpus. Due to the disproportion between the articles and party tweets, the articles were not represented in all the topics produced by the model. Topics that contained articles were annotated in an effort to identify issue-related topics for analysing the relationship between both data sources. Two issue-related topics were identified, Eskom and land expropriation.

Topic timelines were inspected to understand the temporal relationship between the articles and party tweets. In both topics, spikes in volume occurred on Twitter prior to the news articles. In the case of the Eskom topic, the articles lagged behind the tweets by two weeks. A similar trend was observed in the land expropriation topic with the articles lagging a week behind the tweets.

Topic wordclouds were created for the periods in which spikes in volume occurred. The wordclouds indicated that the high volume periods for both the Eskom and land expropriation topics addressed the same concept in both the articles and tweets. Therefore, it was concluded that the issue-centric topics inspected received mass attention on Twitter prior to being published in news articles. Increasing the diversity and size of the article corpus would provide a more complete analysis of this finding. The extension of the

article corpus was outside the scope of the project and is reserved for future work.

11.6 Excluded Experimental Variations

The article corpus contained a body and synopsis field. The experiment in Chapter 7 built models on the body field. Models were built with the synopsis field as well. The synopsis models did not produce topics that contained a notable difference between the keywords thus the experiment proceeded with the body field. Another variable factor for the article experiment was the degree of n-gram to include. Models were built for unigrams, bigrams and trigrams. The trigram models produced the most descriptive topics therefore the experiment utilised trigram tokens.

The tweet corpus underwent several variations. The experiment in Chapter 8 built models on trigrams and contained mentions, the resulting model was broad and did not provide much context to localised topics. The failed variations included unigram and bigram models which were not as descriptive as the trigram model. The tweets were short, averaging 19 tokens. The short context of text available would, in most cases, indicate that unigrams would be better than trigrams but the volume of tweets available resulted in the unigram model emphasising tokens that were not descriptive. In contrast, the trigram model grouped together triplets of tokens that provided more context. Even though it only described a broad context, the topics could be interpreted. A variation was also built to remove mentions and hashtags from the tweets but this resulted in uninterpretable topics therefore this variation was excluded. The removal of mentions and hashtags resulted in very short context for topic modelling since majority of the tweets predominantly contained mentions and hashtags.

There were a limited amount of political party tweets, ranging from 6 000 to 11 000 per party. The experiment in Chapter 9 built topic models on unigrams. Variations with bigrams and trigrams were tried. Upon inspection, the unigram topics were more descriptive that the bigram and trigram topics which was attributed to the limited number of tweets present in the party corpus. The tweet corpus contained 6 million tweets

whereas the party corpus contained approximately 26 000 tweets. A similar variation to the tweet experiment was tried where the mentions and hashtags were removed, the topics produced did not show an improvement on the unigram model including mentions and hashtags therefore this variation was excluded.

The experiment in Chapter 10 built models on a combined corpus of articles and political party tweets. The lessons learnt from the previous experiments were carried over to this experiment. The model was built on unigrams and contained mentions. Variations with the article part of the corpus included attempts to build models on the synopsis and title field of the article instead of the body field. These did not produce better topics since the combined corpus was skewed towards the party tweets and the keywords from the models contained keywords that were mostly found in the party tweets. Due to this observation, the experiment proceeded with using the article body so that the longer context could offset the low proportion of articles in the corpus.

11.7 South African Election Discourse Analysis

This study built upon the ideas of [2]. Both studies analysed the South African election from the perspective of articles and tweets. This study focused more on the modelling and analysis techniques that would uncover insights from the data whereas [2] aimed to understand how emerging democracies work using South Africa as a use case.

The approach that [2] took to collect articles by scraping content from websites will be a beneficial task to include for future work to get a more representative article corpus for this study. Applying the modelling and analysis approaches from this study to the data collected by [2] will allow for a more complete comparison between articles and tweets across two election periods. This could benefit the literature in understanding the evolution of discourse in both data sources. Uncovering the relationship between articles and tweets in both election periods could provide an indication of the rise of Twitter as an election discourse platform. Furthermore, the prospect of a future collaboration between this study and [2] can result in the data science aspects of this study and the

understanding of the social aspects of elections from [2] complementing each other. This would provide a more complete view of the South African election discourse and its implications for the election.

11.8 Summary

This chapter discussed the insights that answer the research questions posed in Chapter 1. NMF was found to produce better quality topics in comparison to LDA for the corpora being studied. The insights drawn from the analysis for the four experiments were addressed to highlight the information that was uncovered from the corpora of interest. This study uncovered numerous insights from two sources of information that characterise the 2019 election period.

The contrast between articles from 2014 and 2019 showed that News24 changed its focus from producing more election and party-related content to producing more issue-centric content. The analysis of the tweet corpus did not produce high quality topics, however, the context of the topics could be deciphered broadly. The analysis of topics dervied from political party tweets shed light on the campaigns, events and rhetorics of the political parties in granular detail. The analysis of the relationship between articles and party tweets found that the party tweets experienced spikes in volume prior to the articles for topics related to Eskom and land expropriation. The party tweet spikes led the article spikes by a week and two weeks for the land expropriation and Eskom topics respectively.

A discussion was presented on the experiment variations that were tried and excluded for the four experiments. The reasoning behind the exclusions could assist in informing future work on these corpora. Lastly, contrasts between this study and [2] were discussed highlighting that a future collaboration between both teams can result in a more complete study of the South African election discourse.

Chapter 12

Conclusion

This study aimed to understand the discourse surrounding the South African election. Two corpora were used to understand the discourse, a news article corpus and a tweet corpus. The tweet corpus was collected using a predefined list of keywords whereas the news articles were supplied by News24. The tweet collection process was subjected to an ethical clearance process and received approval from the University of Pretoria (Reference: EBIT/90/2019).

The analysis in this study was broken up into four experiments. The article experiment (Chapter 7) analysed the news articles from 2014 and 2019 to uncover themes and contrasts between the content from two different periods. Furthermore, the article experiment contrasted LDA and NMF and found that LDA produced inadequate clusters from the data. The tweet experiment (Chapter 8) analysed the collected tweets to identify broad themes and pivotal figures in the Twitter discourse. The party experiment (Chapter 9) analysed the tweets produced by political parties to uncover themes and events that were central to the campaigns of the three parties under scrutiny. The combined corpus experiment (Chapter 10) analysed a corpus made up of 2019 articles and political party tweets to understand the relationship that exists between the two data sources in specific topics. Chapter 11 provided a detailed discussion of the insights derived from the experiments. Furthermore, Chapter 11 links the insights to the objectives set forth

in Chapter 1.

Understanding an event such as an election from the perspective of news articles and tweets using text mining techniques provides an automated, unbiased way to summarise a large quantity of text. Understanding a large proportion of the discourse surrounding an election provides a record of a time period as well as a window into the information that is newsworthy and central to the campaigns of political parties. The modelling and analysis approach detailed in Chapter 6 can be applied to understand any event described by a corpus of text.

The novel contributions of this study includes a corpus of 2019 tweets related to the South African election, an analysis of election-related news articles from two election periods, an analysis of the 2019 election discourse from the perspective of Twitter, an analysis of the Twitter campaigns of political parties for the 2019 South African election and an analysis of the relationship that exists between articles and political party tweets in the 2019 South African election discourse.

12.1 Summary Of Conclusions

This section provides an account of highlight findings from this study, a more detailed discussion is provided in Chapter 11.

12.1.1 Comparison Of LDA & NMF

NMF produced more represented topic clusters compared to LDA. The NMF topics promoted further analysis whereas the LDA models produced dominant topics that added noise to the topics that hindered further analysis. The NMF and LDA comparison occurred in the first experiment (Chapter 7), which analysed articles. LDA produced inadequate clusters therefore it was excluded from the remainder of the experiments.

12.1.2 Notable Insights From The Articles

The analysis uncovered themes related to corruption, state capture and Eskom frailties in the 2019 corpus. The models applied to the 2014 corpus uncovered themes related to the Nkandla report, Julius Malema's tax battle with SARS and the upcoming election. The ANC received the most coverage in 2019, the ANC's coverage increased from 2014 whereas the coverage received by the DA and EFF had decreased. The contrast between 2014 and 2019 as viewed through topic keyword wordclouds indicated that the article discourse of the 2019 election focused more on corruption and problems whereas the 2014 discourse contained more references to political parties and election-related aspects.

12.1.3 Notable Insights From The Tweets

The topics uncovered from the tweets were more difficult to interpret than the ones uncovered from the articles. However, the topics were descriptive enough to infer some high profile participants and the broad themes described by the keywords. The topic keyword wordcloud for all the tweets highlighted the EFF's dominance of Twitter with references to their campaigns and personnel flooding the wordcloud. Party-related topics could only be classified as belonging to a party and addressing their campaign, specific insights about events could not be drawn from the topics. The topic timelines could not uncover localised events but did uncover two periods of high volume that warranted further investigation.

12.1.4 Notable Insights From Party Twitter Campaigns

The ANC's hashtag vocabulary was narrow and indicated a campaign focused on attracting votes. The DA's hashtag vocabulary contained campaign slogans as well as hashtags relating to problems in the country at the time. The EFF's Twitter campaign focused on campaign slogans as well as unique hashtags for rallies and events held by the EFF in the observed period. Analysing the topics produced by political parties uncovered

two underlying categories in the topics: topics related to the policies of the parties and topics related to events, campaigns and rallies. The ANC rhetoric focused on stopping corruption, providing housing, creating job, economic growth and land reform. The DA focused on highlighting corruption and failures in the ANC's current administration. They also focused on job creation, Eskom and proposing that they will create one SA for all. The EFF's focus was on free education, housing, land expropriation, service delivery and highlighting frailties in Eskom and the ANC.

12.1.5 The Relationship Between Articles And Party Tweets

Relationships between the articles and political party tweets were identified for two issue-centric topics related to Eskom and land expropriation. The topic timelines related to these two topics indicated that the spikes in volume seen in the party tweets led the spikes in volume in the articles. In the case of the Eskom topic, the articles lagged behind the tweets by two weeks. A similar trend was observed in the land expropriation topic with the articles lagging a week behind the tweets.

The experiment was conducted on a combined corpus. The corpus was disproportionate, with party tweets making up 85% of the documents. The inclusion of more articles would balance this corpus and could produce topics that are a better representation of the articles.

12.2 Future Work

1. In this study, it was not possible to capture the evolution of topics over the observed period. A time-sensitive technique such as the dynamic-NMF topic modelling [10] could provide a mechanism to understand how the context of a topic changes over time. An extension to this study would be to use dynamic-NMF to understand how topics evolve over time.

2. One of the limitations with the topic modelling techniques used in this study was that the semantic co-occurrence of tokens were not captured. Word embedding models such as W2V captures the relationship between words. An extension to this study would be to employ topic modelling techniques that operate in embedded space. The application of the embedded topic model [9] could be an option to explore the outcome of topics created from a model that captures the semantic co-occurrence of tokens.

3. The news articles used in this study were from a single source which provides a limited window into the news article discourse surrounding the election. The acquisition and inclusion of articles from more news agencies will enrich the corpus and enable a broader view of the discourse present in news articles.

4. The application of the modelling and analysis approaches described in Chapter 6 for future elections. This will provide a comparative analysis that can be contrasted with the discourse present in the 2019 election.

Bibliography

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[2] Jeffrey Arnold, Aaron Erlich, Danielle Jung, and James Long. Covering the campaign: News, elections, and the information environment in emerging democracies, 2018.

[3] David M Blei and John D Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

[6] Giuseppe C Calafiore, Laurent El Ghaoui, Alessandro Preziosi, and Luigi Russo. Topic analysis in news via sparse learning: a case study on the 2016 us presidential elections. *IFAC-PapersOnLine*, 50(1):13593–13598, 2017.

[7] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. pages 288–296, 2009.

[8] Ashok Deb, Luca Luceri, Adam Badaway, and Emilio Ferrara. Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. In

Companion Proceedings of The 2019 World Wide Web Conference, pages 237–247, 2019.

[9] Adji B Dieng, Francisco J R Ruiz, and David M Blei. Topic modeling in embedding spaces, 2019.

[10] Derek Greene and James P Cross. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1):77–94, 2017.

[11] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, pages 9–56, 2008.

[12] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1):72–91, 2016.

[13] Amir Karami and Aida Elkouri. Political popularity analysis in social media. In *International Conference on Information*, pages 456–465, 2019.

[14] Joleen Steyn Kotze and Narnia Bohler-Muller. Editorial: Quo vadis? reflections on the 2019 south african general elections. *Politikon*, 46(4):365–370, 2019.

[15] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[16] Avishay Livne, Matthew Simmons, Eytan Adar, and Lada Adamic. The party is over here: Structure and content in the 2010 election. In *Fifth international AAAI conference on weblogs and social media*, pages 201–208, 2011.

[17] Onyedikachi Madueke, Celestine Nwosu, Chibuzo Ogbonnaya, and Adaeze Anumadu. The role of social media in enhancing political participation in nigeria. *International Digital Organization for Scientific Research (IDOSR): Journal of Arts and Management*, 2(3):44–54, 2017.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[19] Lungisani Moyo and Osunkunle O Oluyinka. Economic and socio-political views of online participants in online political commentary during the 2014 south african general elections. *Journal of Human Ecology*, 68(1-3):42–58, 2019.

[20] Derek Ocallaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

[21] Thomas E Patterson. News coverage of the 2016 general election: How the press failed the voters, 2016.

[22] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 452–456, 2004.

[23] Min Song, Meen Chul Kim, and Yoo Kyung Jeong. Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intelligent Systems*, 29(2):18–26, 2014.

[24] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David 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*, pages 952–961, 2012.

[25] Leanne Townsend and Claire Wallace. Social media research: A guide to ethics, 2016.

[26] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, pages 178–185, 2010.

[27] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

[28] Wei Xu, Xin Liu, and Yihong 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*, pages 267–273, 2003.

[29] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.