Social Media Analytics and the Role of Twitter in the 2014 South African General Election: A Case Study

Asheen Singh (450623)

Supervisor: Michael Mchunu

A Dissertation submitted to the Faculty of Science,
University of the Witwatersrand, Johannesburg,
in fulfilment of the requirements for the degree of Master of Science.

February 21, 2018

Declaration

I declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

_____

Signature of candidate

_____ day of _____ 2018_____ in _____

Please Note

The views and opinions expressed in this document are those of the author and do not necessarily reflect the official policy or position of the University of the Witwatersrand. The data used in this research was publicly available and collected through the Twitter API. The analysis of the data and the interpretation of the results were carried out to the best of the authors ability, without any intention of favouring or being biased towards any political party.

Abstract

Social network sites such as Twitter have created vibrant and diverse communities in which users express their opinions and views on a variety of topics such as politics. Extensive research has been conducted in countries such as Ireland, Germany and the United States, in which text mining techniques have been used to obtain information from politically oriented tweets. The purpose of this research was to determine if text mining techniques can be used to uncover meaningful information from a corpus of political tweets collected during the 2014 South African General Election. The Twitter Application Programming Interface was used to collect tweets that were related to the three major political parties in South Africa, namely: the African National Congress (ANC), the Democratic Alliance (DA) and the Economic Freedom Fighters (EFF). The text mining techniques used in this research are: sentiment analysis, clustering, association rule mining and word cloud analysis. In addition, a correlation analysis was performed to determine if there exists a relationship between the total number of tweets mentioning a political party and the total number of votes obtained by that party. The VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment classifier was used to determine the public's sentiment towards the three main political parties. This revealed an overwhelming neutral sentiment of the public towards the ANC, DA and EFF. The result produced by the VADER sentiment classifier was significantly greater than any of the baselines in this research. The K-Means cluster algorithm was used to successfully cluster the corpus of political tweets into political-party clusters. Clusters containing tweets relating to the ANC and EFF were formed. However, tweets relating to the DA were scattered across multiple clusters. A fairly strong relationship was discovered between the number of positive tweets that mention the ANC and the number of votes the ANC received in election. Due to the lack of data, no conclusions could be made for the DA or the EFF. The *apriori* algorithm uncovered numerous association rules, some of which were found to be interest-

ing. The results have also demonstrated the usefulness of word cloud analysis in providing easy-to-understand information from the tweet corpus used in this study. This research has highlighted the many ways in which text mining techniques can be used to obtain meaningful information from a corpus of political tweets. This case study can be seen as a contribution to a research effort that seeks to unlock the information contained in textual data from social network sites.

Keywords. Data Mining, Text Mining, Sentiment Analysis, Correlation Analysis, Cluster Analysis, Association Rule Mining, Word Cloud Analysis, Social Network Sites, South African Election, Votes, Tweet, Twitter, Case Study

Acknowledgements

First and foremost, I would like to thank my supervisor, Michael Mchunu, for the tremendous effort and time he has put into this research. Without his guidance this would not have been possible. I would also like to thank my family and friends for their support and encouragement over the past three years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social Network Sites (SNS) such as Twitter and Facebook , have revolutionised communication and information exchange. Thanks to the Internet, millions of users all over the globe are able to digitally communicate using Twitter, Facebook or other social network facilities. Twitter is a microblogging service used to communicate and share ideas and opinions across different fields such as education, healthcare, business and politics [Tumasjan *et al.* 2010]. In several African countries, citizens have taken to social media platforms to actively engage in the political life and events unfolding in their respective countries. Consider the domestic uprisings that occurred in Tunisia and Egypt, which came to be known as Arab Spring, Twitter and Facebook are amongst the tools that were predominantly and effectively used by the citizens in those countries to orchestrate the downfall of their governments [Howard *et al.* 2011].

Particularly for politics, Twitter is used to provide an environment in which different stakeholders can participate and engage in a variety of political discussion [Yang and DeHart 2016; Muntean 2015]. Furthermore, Twitter is particularly attractive for text mining, since tweets occur in real-time and is representative of society on an international level, more over, these tweets are available to the pub-

lic [Russell 2013]. As such, extensive research has been done around text mining Twitter in the context of politics in countries throughout the world.

There has been a growing interest in performing sentiment analysis on tweets to gauge online political sentiment [Bakliwal *et al.* 2013]. Tumasjan *et al.* [2010] applied sentiment analysis to tweets in order to evaluate public sentiment towards political parties and politicians in the context of the 2009 German federal election. Similarly, Bakliwal *et al.* [2013] used tweets to determine public sentiment during the run-up to the Irish General Elections in 2011. On the other hand, Monti *et al.* [2013] leveraged tweets to specifically analyse public disaffection towards political parties in Italy in 2012. Furthermore, during the 2012 United States election Wang *et al.* [2012] used Twitter to track the sentiment of the general public, in real-time, toward the 2012 United States presidential candidates.

Interesting research has been done using Twitter to predict outcomes of elections [Chrzanowski and Levick 2012]. Tumasjan *et al.* [2010] investigated whether tweets could be used as a predictor of the popularity of political parties in the 2009 German federal elections. Similarly, Chung and Mustafaraj [2011] focused on using tweets to predict the outcome of the 2010 United States Senate special election in Massachusetts. In addition, Chrzanowski and Levick [2012] investigated the predictive power of Twitter in the 2012 United States Presidential Election.

Clustering is a popular data mining technique by which objects are classified into groups called clusters [Feldman and Sanger 2007]. Hence, tweet clustering is the procedure by which tweets are classified into clusters, where tweets in the same cluster are considered *"similar"*. Rosa *et al.* [2011] investigated methods to cluster tweets into six predefined categories: News, Sports, Entertainment, Science, technology, Money and "Just for fun". Whereas, Sankaranarayanany *et al.* [2009] investigated clustering tweets based on news topics in real-time. Although the work done by Rosa *et al.* [2011] and Sankaranarayanany *et al.* [2009] are not directly related to clustering tweets in a political context, they still provide valuable

insight into tweet clustering.

Association Rule Mining (ARM) is a core data mining task. It is the process in which rules that symbolize a relationship between item sets are discovered [Zhao 2012; Zaki and Hsiao 1999]. An association rule is of the form: $A \Rightarrow B$, where A and B are disjoint itemsets. Since its introduction, ARM has attracted significant interest from data mining researchers and practitioners [Zaki and Hsiao 1999]. For example, the work done by Cagliero and Fiori [2013] aimed to discover association rules from tweets collected from New York and London. On the other hand, Zingla *et al.* [2014] used association rules to extend the vocabulary of tweets in order to help users better understand tweets.

Visualization techniques such as word clouds are used to show the importance of words and help to users to quickly identify the primary content of a corpus [Zhao 2012; Wu *et al.* 2011]. Previous work done by Wang *et al.* [2012] tracked the sentiment of the public in real-time during the 2012 United States elections and used a word cloud to display useful information. Similarly, Sankaranarayanany *et al.* [2009] generated a word cloud relating to the 2009 Iranian elections. Both Wang *et al.* [2012] and Sankaranarayanany *et al.* [2009] showed that word clouds can be used effectively to visualise keywords and topics contained in a corpus of political tweets.

1.1 Problem Definition

Twitter has become a legitimate communication medium in the political arena [Tumasjan *et al.* 2010]. This was particularly evident in Barack Obama's United States presidential campaign in 2008; thereafter, numerous researchers have aimed to utilize tweets to gain insights into their respective political domains.

As mentioned earlier, not much research has been conducted in the South African political domain. Instead, opinion polls and surveys are used to gauge

the public's opinion ahead of general elections. A research institution called Ipsos measured the public's disaffection of the South African government by using polls in 2013 [Ipsos 2015]. Most of South Africa's population live in isolated rural areas. During the polling period, the opinions and views of this vast section of the country's population are often overlooked. Thanks to Twitter, the urban/rural divide prevalent in South Africa has been overcome [Worx 2012]. In this regard, Twitter is able to capture a broader spectrum of views and opinions emanating from all sources in the South African political space. Through processing and analysis, tweets can be leveraged in order to produce meaningful information, from which useful insights into the public's opinions and views on South African political events can be obtained. This information can be easily made available to the public and other interested stakeholders.

Due to the vast number of tweets generated daily it is impractical and error prone to analyse tweets manually. Text mining techniques and methodologies are particularly well suited for this task as shown by results obtained from previous research. Therefore, my research utilised text mining techniques to extract information and identify new knowledge from a corpus of tweets posted before, during and after the 2014 South African General election.

1.2 Research Significance

This research has highlighted various ways in which text mining techniques can be used to obtain insightful and meaningful information from social media network service datasets such as Twitter. The experimental results obtained from this research are promising and point to the need for further work to be done in this exciting area of research. In addition, the results are available to the public through a custom built website (https://tweetminingsa.herokuapp.com). This website can be further developed, to provide information on the state of the South African political

landscape, information that could enable the voting public to make more informed decisions when voting during elections.

1.3 Research Questions

In this research I used text mining techniques to discover patterns of useful information and knowledge from a corpus of tweets collected before, during and after the 2014 South African general election. Visualization techniques were used to render the information in a manner that is easy to understand by interested parties, particularly the general public. In addition, a Web application was developed in order to make the results of this research accessible to other researchers and to the general public.

The main research question this study sought to answer is as follows:

> ***Do text mining techniques uncover meaningful information when applied on a corpus of political tweets collected on the 2014 South African General Election?***

Q1 What sentiment is portrayed by the election tweets towards some of the parties that took part in the general election?

Q2 Does a relationship exist between the total number of positive tweets that mention a political party and the total number of votes obtained by that party?

Q3 Can clustering divide political tweets into political-party clusters, based on the corpus of tweets collected during the 2014 South African election?

Q4 How well do the association rules extracted from the political tweets generated during the 2014 South African general election characterise the relationship between the words contained in these tweets?

Q5 Can word clouds generated from political tweets, collected during the 2014 South African general election, be used to convey meaningful information regarding voter-related issues, opinions and sentiments?

1.4 Research Methods

Given the nature of the questions that were being addressed in this research, I considered a case study methodology to be a suitable approach for conducting the research. Text mining also played a significant part in this research. CRISP-DM (Cross-Industry Standard Process for Data Mining) is a popular process model that is used in data mining or text mining projects. The model consists of six phases, namely *business understanding, data understanding, data preparation, modelling, evaluation* and *deployment*. The first four phases of CRISP-DM were selected and applied in this research.

In this research different I applied text mining techniques on the dataset I collected during the 2014 South African General Election period. These techniques include sentiment analysis, correlational analysis, text clustering, association rule mining and word cloud analysis. Hutto and Gilbert [2014] have developed a parsimonious rule-based sentiment classifier called *VADER* (**V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning), suitable for use in microblogging services such as Twitter. To answer research question Q1, I applied the VADER sentiment classifier the tweet corpus I collected. Research question Q2 was answered by using the sentiment analysis results and determining the correlation between the total number of positive tweets that mention a political party and the total number of votes received by that party. The K-Means clustering algorithm was applied on the corpus in order to answer research question Q3. I used the *apriori* association rule mining algorithm to answer research question Q4. Answering research question Q5 involved generating a word cloud using the results obtained from the sentiment

analysis experiment.

Thereafter, various data metrics were used to evaluate the performance of the algorithms used during the modelling phase. The performance of the sentiment analysis classifier was measured by calculating the $F1$ score, a measure of accuracy ranging from 0 (not accurate) to 1 (perfectly accurate). It was indicated that in this research an acceptable $F1$ score would have to be significantly greater than the highest baseline score of 0.1945. The Pearson correlation coefficient (*r*) was calculated to determine the relationship between the number of positive tweets that mention a political party and the number of votes received by that party. The value for r ranges between -1 and 1 ($-1 \leq r \leq 1$). An r value of 1 indicates a strong positive relationship. On the other hand, an r value of -1 indicates a strong negative relationship. An r value of 0, for example, would indicate the absence of any correlation between the number of positive tweets that mention a political party and the number of votes received by that party. The *K-Means* algorithm was used to cluster the political tweets according to party affiliation. The tweets were labelled automatically, and each tweet label was compared to the label of the cluster containing the tweet. This way it was possible to determine the number of correctly clustered tweets. In order to ensure that only significant association rules were discovered, the support and confidence thresholds were set at 0.01 and 0.7 respectively. There is no statistical measure to test if the association rules are meaningful. Therefore, the discovered association rules were manually interpreted to determine their meaningfulness. Similarly, for word clouds, no statistical measures to test the validity or meaningfulness of a word cloud exist. Hence, the validity and meaningfulness of the word cloud had to be judged manually.

1.5 Structure of the Dissertation

Chapter 2 presents a definition of text mining, and provides a detailed description of Social Network Sites (SNS) such as Twitter. The chapter also includes a detailed review of the literature related to my research. Chapter 3 provides a detailed discussion of the methodology I followed in conducting this research. The results obtained from the application of different text mining methods on the corpus of political tweets I collected for this research are presented and analysed in Chapter 4. In Chapter 5 a detailed discussion is provided, which focuses on the findings of my research. Finally, this dissertation is concluded in Chapter 6.

Chapter 2

Background and Literature Review

2.1 Introduction

Opinion polls, surveys and questionnaires provide a useful mechanism for collecting data by sampling a cross section of the population. Opinion polls are conducted for a variety of reasons such as obtaining a view, sentiment or opinion that represents the public's attitude on a particular issue or issues. In politics opinion polls are used, among other things, to predict election outcomes. Social Network Sites (SNSs) have become a useful communication platform on which millions of users disseminate and access content. Compared to opinion polls, it is nowadays much easier to obtain SNSs data in order to gauge the public's sentiment on various issues.

In this case study research a corpus of election tweets posted by Twitter users during the 2014 South African general election were collected and analysed to obtain information regarding the users' views on different issues related to the election. In this context not much research has been done or reported in which a

detailed analysis has been performed of the data that has been generated during these elections by users of SNSs such as Twitter, Facebook, and others.

The topics addressed in this chapter are discussed in different sections as follows. In Section 2.2 a brief discussion is provided on opinion polls, including their specific application in an election context. A general discussion is held in Section 2.3, which focusses on social network sites and the different ways in which these platforms can be used. Section 2.4 discusses Twitter, the social network site that was selected for collecting the political tweets used in this study. A brief overview of data mining and text mining techniques is provided in Section 2.5. The discussion in Section 2.6 focusses on some of the work related to this research, in which tweets have been used and analysed in different ways to extract meaningful information. Topics such as sentiment analysis, election prediction, cluster analysis of tweets, association rule mining and word cloud analysis are discussed in Subsection 2.6.1 to Subsection 2.6.5.

2.2 Elections and Opinion Polls: South African Context

In 1824 an American newspaper, *The Harrisburg Pennsylvanian*, conducted what is considered to be the first known opinion poll [Smith 1990], whose results showed one of the candidates, the eventual winner, leading the other candidate in a presidential election that was held in that year. From 1916 to 1932 the *Literary Digest*, a United States magazine, was able to correctly predict the outcomes of the successive presidential elections held between these years [Smith 1990]. The predictions were made using feedback obtained from postcards returned by some of the respondents. As a result of using a very large sample of 2.4 million responses [Nations 2017], the *Literary Digest* failed dismally to correctly predict the outcome of the 1936 presidential election, an event that led to this magazine's downfall and demise. On the other hand, George Gallup adopted a scientifically-based approach

and, using a smaller, demographically representative sample, correctly predicted that Franklin D. Roosevelt would win the 1936 election [George Gallup 2000].

Within the South African context, political opinion polling is a relatively new occurrence. The whites-only general election of 1970 marked the beginning of this phenomenon in South Africa [Lever 1974]. The two polls conducted in that year focussed more on the voters' opinions on "the dominant policies and issues in South African politics", rather than predicting the outcome of that election [Lever 1974]. The first poll (the so-called "Argus poll") was conducted by the Argus group of newspaper companies, and focussed mainly on "issues and policies", and less on "party preferences and public personalities" [Lever 1974]. The second poll (the so-called "Dagbreek poll") was conducted by an Afrikaans newspaper, *Dagbreek en Sondagnuus*, and the primary aim was to obtain information from voters on issues such as "party preferences, changes in preference, and public personalities" [Lever 1974]. This poll did not focus much on election-related issues.

During the 1980s a number of private research and polling organisations in South Africa conducted surveys in which questions of interest focussed on issues such as "partisan politics, race relations and social and political change" [Mattes 2012]. The findings from these surveys were communicated to the public through different news media. Some of the private companies which took part in these surveys include Markinor and Market Research Africa [Mattes 2012]. State-funded research entities such as the Human Sciences Research Council (HSRC) were also active in the 1980s, during which period they were conducting surveys to solicit the public's opinion on issues such as apartheid [Mattes 2012].

It was during South Africa's transition to democracy in the early 1990s that opinion polling took a markedly political dimension [Mattes 2012]. Different stakeholders, including politicians, the news media, and political negotiators for the new dispensation, all engaged the services of polling organisations in order to obtain information regarding the strength of the different political parties and the

public's support for their policies [Mattes 2012]. It was vital for all these stakeholders to obtain this information, given the absence of any past information that could be used to inform decision-making in the face of competition from other role players. Since the first democratic election of 1994, South African polling organisations have developed a strong tradition of using opinion polls to obtain public opinion on political issues and to predict election outcomes [Ipsos 2016; Dufour and Calland 2016].

2.3 Social Network Sites

Social Network Sites (SNSs) such as Twitter, Facebook and others, have revolutionised communication and information exchange amongst millions of users. Thanks to the Internet these users, located all over the globe, are able to communicate and share information amongst themselves using these SNSs. In particular, Twitter enables users to communicate with one another and to exchange information, opinions and views, on a variety of topics or issues. Twitter has also found its way into the political arena, where it has become a useful platform for users to express their opinions and views on political matters [Wang *et al.* 2012; Monti *et al.* 2013].

Social networking as a concept is not new. The idea has been around for years, before the advent of the Internet [Raju and Kumar 2014]. For example, according to Raju and Kumar [2014], the various groupings or cliques found in a typical high school such as band geeks, athletes, and other cliques are an example of a basic social network. Each clique is a social group and a person can belong to one, many or none of the cliques. Internet-based social networking links users in a manner similar to the example described above. According to Boyd and Ellison [2007], social network sites are "web-based services that allow individuals to:

1. construct a public or semi-public profile within a bounded system,

2. articulate a list of other users with whom they share a connection, and

3. view and traverse their list of connections and those made by others within the system."

Kaplan and Haenlein [2010] defines SNSs as "applications that enable users to connect by creating personal information profiles, inviting friends and colleagues to have access to those profiles, and sending e-mails and instant messages between each other".

Based on the definitions provided by Boyd and Ellison [2007] and Kaplan and Haenlein [2010], an SNS must enable a user to:

1. create a personal profile

2. establish a relationship with other users

3. connect with other users via text-based messages

Each user has a unique profile, which contains user information such as photos, videos and audio files [Boyd and Ellison 2007; Kaplan and Haenlein 2010]. The visibility of a user's profile depends on the site and the user's privacy settings [Boyd and Ellison 2007]. Some SNSs allow a user's profile to be visible to anyone, even to non-users. On the other hand, some SNSs do allow users to decide whether other users may or may not view their profile [Boyd and Ellison 2007].

After joining a SNS, a user identifies other users with whom they have or wish to have a relationship [Boyd and Ellison 2007]. There are two types of relationships, namely *symmetric* and *asymmetric* relationships. Symmetric relationships require mutual acceptance; each user in the relationship needs to be aware of and confirm the relationship [Boyd and Ellison 2007]. Asymmetric relationships do not require mutual acceptance; users in the relationship need not be aware of nor confirm the relationship [Boyd and Ellison 2007]. Twitter is characterised by an asymmetric relationship model, which is discussed in Section 2.4. Each SNS provides a mechanism that facilitates communication amongst users. Text-based mes-

sages are exchanged between users, and the communication may also involve the users exchanging photos, videos and audio files [Boyd and Ellison 2007].

Many SNSs which are currently being used by millions of subscribers [Boyd and Ellison 2007]. Figure 2.1 shows the leading SNSs ranked by the number of active users as of April 2016. The most popular SNS is Facebook with approximately 1.6 billion active users [Statista 2014].

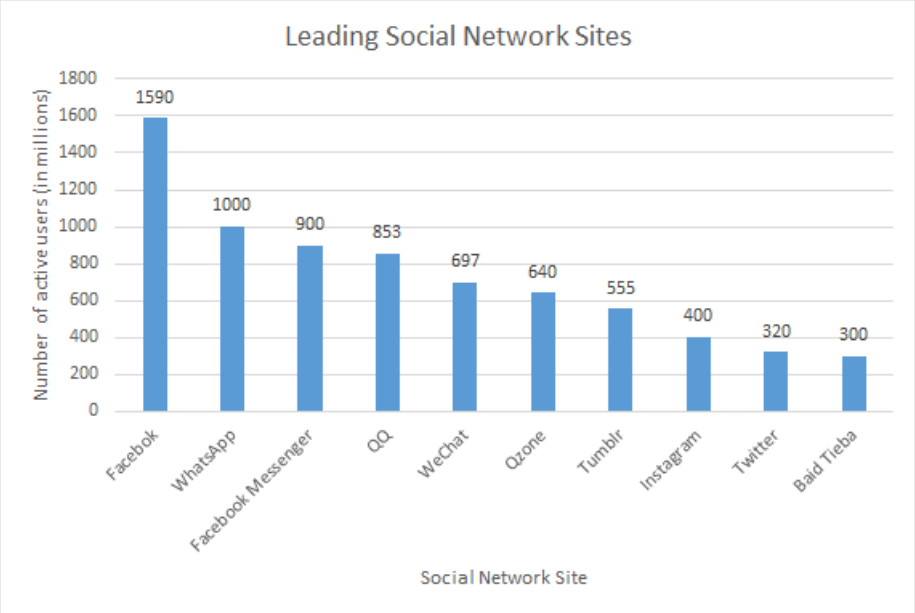

Figure 2.1: Leading social network sites ranked by the number of active users taken from Statista [2014]

Some sites are intended to be used by a diverse group of people, whilst others cater for groups of users who share common characteristics such as a common language, religion, etc. [Boyd and Ellison 2007]. To date, the following major social network categories have been identified: social connections, multimedia sharing, professional, informational, educational, hobbies and academic Raju and Kumar [2014]. Each one of these categories is described below.

Social Connections

SNSs that facilitate social connections enable users to maintain relationships with other people, such as friends and families. Facebook, Google+, Twitter and MySpace are the most popular websites for creating social connections online [Raju and Kumar 2014]. Facebook enables users to stay connected to other users, and to share and express their views and opinions on issues or topics that matter to them [Facebook 2014]. Similarly, Google+ is designed for users to create *circles* of contacts with whom they interact [Raju and Kumar 2014]. Twitter users are able to share their ideas, experiences and opinions in real time [Twitter 2014]. MySpace creates a forum for a community of artists, which allows them to showcase their work [MySpace 2014].

Multimedia Sharing

SNSs that belong to this category enable users to share video, photography and music content online [Raju and Kumar 2014]. Examples of SNSs in this category include Flickr, Picasa and YouTube. Flickr has two main goals: to enable users to organise their photos efficiently and to share them with people who matter to them [Flickr 2014]. Picasa is a Google product which enables users to organise, edit and share photos. It integrates with Google+, by allowing users to tag and share photos with their Google+ contacts [Picasa 2014]. YouTube provides a platform which enables users to discover, watch and share originally-created videos [YouTube 2014].

Professional

Professional SNSs provide a forum for professionals to connect. Some professional SNSs focus only on specific occupations or interests [Raju and Kumar 2014]. LinkedIn, Who's Who of Southern Africa and Biznik are all examples of profes-

sional SNSs. LinkedIn maintains professional users' profiles and enables users to connect with each other as professionals [LinkedIn 2014]. Likewise, Who's Who of Southern Africa assists users in creating, managing and growing their professional brand and reputation [Who 2014]. Biznik is an online community whose purpose is to provide a platform where independent business owners can gather, share resources, provide referrals and mutual support [Biznik 2014].

Informational

Informational SNSs allow users to form online communities where knowledge and resources are shared around issues of common interest [Raju and Kumar 2014]. For example, SuperGreenMe is an online community where individuals interested in living eco-friendly lifestyles can interact [SuperGreenMe 2014]. With DoItYourself users have access to an online community environment which provides home improvement and repair information [DoItYourself 2014].

Educational

Students use educational SNSs to collaborate on academic projects with other students and to interact with their teachers [Raju and Kumar 2014]. Edmodo and The Math Forum are examples of educational SNSs. Edmodo allows teachers and students to connect in online classrooms where they can share and discover resources [Edmodo 2014]. The Math Forum connects students by age group, who are specifically interested in mathematics [Raju and Kumar 2014].

Hobbies

Hobby-focused SNSs allow users to form an international online community to share ideas, opinions and resources around a shared interest [Raju and Kumar 2014]. For example, Scrapbook.com caters specifically for people interested in

scrapbooking and guides them through the memory preservation process [Scrapbook.com 2014]. Likewise, SporShouting.com provides a platform for sport fans to share their opinions on the latest sports news [SportShouting.com 2014].

Academic

Academic researchers find academic SNSs valuable for sharing their research and discovering research done by others [Raju and Kumar 2014]. Academia.edu is an example of an academic SNS that enables users to share research papers [Academia.edu 2014].

2.4 Twitter

Twitter is one of a number of Social Network Sites (SNSs) that can be accessed via the Internet. It was conceived in 2006 by Jack Dorsey, an employee at Odeo, an American podcasting company [MacArthur 2016]. Originally, Twitter was conceived as an "SMS-based communication platform" [MacArthur 2016], which is why currently, Twitter imposes a 140-character limit on communications. Apart from Jack Dorsey, other people who played an important part in the early development of Twitter are Noah Glass, Biz Stone and Evan Williams [MacArthur 2016].

Since then Twitter has evolved into a microblogging service that enables rapid and easy communication amongst users who exchange short text messages called *tweets* [Russell 2013]. Twitter has become popular amongst users because of its asymmetric *following* model. In traditional SNSs there has to be a mutual acceptance of a connection between users. Twitter, however, allows users to *follow* any other user without any mutual acceptance of the connection [Russell 2013]. In 2014 Twitter reported that approximately 500 million tweets are generated daily in real-time, are representative of society on an international level and are publicly available [Russell 2013; Haustein *et al.* 2016]. These are some of the reasons

which make Twitter attractive for text mining. However, given their characteristics and structure, applying text mining techniques on tweets can be quite challenging, an issue that is discussed in Subsection 2.4.1.

Among other uses, Twitter has become an important platform where people express their opinions and views on political issues [Wang *et al.* 2012], [Monti *et al.* 2013]. In this research I used different text mining techniques and applied them on the political tweets I collected during the period of the 2014 South African general election. These tweets are discussed in Subsection 2.4.2.

2.4.1 Tweet Anatomy

As mentioned earlier the size of a tweet is limited to a maximum of 140 characters. Given such a limit, Twitter users must compile messages that are as short as possible. This is done by using a novel syntax, very much similar to the syntax used when writing SMS (Short Message Service) messages. An example tweet is shown in Figure 2.2.

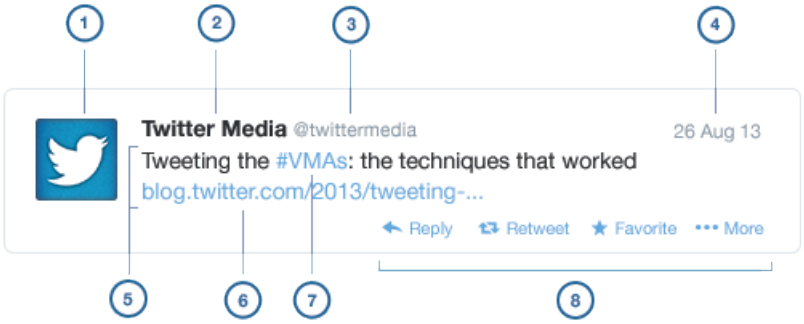

Figure 2.2: The anatomy of a tweet

The different components of this tweet are described below [Coronel-Salas and Sanmatín 2016]:

1. **Profile picture:** the personal image uploaded to a user's profile
2. **Account name:** the name of a user's Twitter account

3. **@Username:** a unique identifier for each user. The account name and @*username* do not need to be the same
4. **Timestamp/date:** indicates when the tweet was sent
5. **Tweet text:** the actual message that is limited to 140 characters
6. **Links: URLs** to other websites, articles, images or videos
7. **Hashtags:** used to reference a topic or conversation
8. **Tweet actions:** allows a user to retweet, favourite and reply to a tweet.

When a user retweets a tweet, the tweet becomes available to the user's followers. In such a case the letters "RT" are placed in front of the tweet.

2.4.2 South African election tweets

Following the 2014 South African general election of 07 May 2014, the African National Congress (ANC) emerged as the overall winner, and is therefore currently the ruling party. The Democratic Alliance (DA) and the Economic Freedom Fighters (EFF) came second and third, respectively. The DA is the official opposition party. In this research text mining techniques were applied on a corpus of political tweets collected during the election. These are tweets that:

- contain at least one of the following phrases: *african national congress, #anc, anc, democratic alliance, #da, economic freedom fighters, #eff, eff*
- were created during the period 15 April 2014 to 04 June 2014

For the purpose of this research I only collected tweets about the top three political parties mentioned above. In selecting these tweets, I chose only those tweets containing at least one of the phrases listed above. However, tweets containing the "da" phrase were ignored. This term was also found in tweets that were not necessarily related to the Democratic Alliance.

2.5 Data Mining and Text Mining

In this research data mining and text mining techniques were used to analyse and obtain information from the political tweets that were generated during the 2014 South African general election. An overview discussion of data mining is provided in Subsection 2.5.1. This is followed, in Subsection 2.5.2, by a discussion of text mining and some of the tasks that fall under it.

2.5.1 Data Mining

According to Tan *et al.* [2006], data mining "*is the process of automatically discovering useful information in large data repositories*". The need for data mining arises from the fact that vital and useful information can be extracted from the vast amount of data that is collected daily, in different fields. The following are some of the factors responsible for the growth and development of data mining as a discipline [Larose 2014]:

- Huge growth in data collection
- Storage of data in data warehouses, enabling enterprise access to a reliable data repository
- Increased data acquisition from the Web and intranet sources
- Pressure to leverage data to obtain useful information for market share and competitive advantage in a global economy
- Development of off-the-shelf proprietary data mining software
- Huge growth in computing power and storage capacity

Data mining involves different tasks such as *classification, estimation, prediction, cluster analysis* and *association analysis*. These tasks are briefly described below.

Classification

Classification uses a model to determine the class to which a particular data record or instance belongs. A classification model is first created from a set of instances that constitute a *training set*. Each instance consists of a set of input or predictor variables and a single class or target variable. In the training set the value of the target variable is known. The model produced during the training phase is then used to classify the instances in a *test set*, in which the class variable values are not known. An example of a classification task would be to categorise a particular student as being likely or not likely to plagiarise code in a programming assignment.

Estimation

Estimation, like classification, creates a model and uses to estimate the value of the target variable. Whereas in classification the target variable is categorical, in estimation this variable is continuous. An estimation model contains values for the predictor and target variables. Given a new instance, the model uses the predictor variable values to produce an estimate of the value of the new instance [Larose 2014]. As an example of estimation, Figure 2.3 illustrates the relationship between a student's graduate grade-point average (GPA) and the undergraduate GPA.

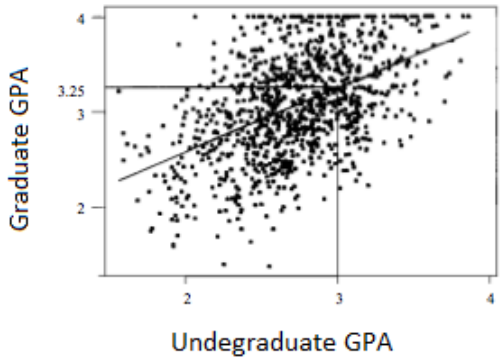

Figure 2.3: Estimating graduate GPA [Larose 2014]

Using the regression line, a student's undergraduate GPA can be used to esti-

mate the graduate GPA. For example, Figure 2.3 shows an estimated graduate GPA of 3.25, which corresponds to an undergraduate GPA of 3 [Larose 2014].

Prediction

With prediction, the aim is to forecast a future value for the target variable. For example, a prediction task may involve predicting the second semester academic performance of a student using marks obtained in the first semester. Prediction is similar to classification and estimation in that a model must be built and used to predict target variable values for unknown instances. Prediction creates a model in which the purpose is to ensure that the error between the predicted and true value of the target variable is minimised [Tan *et al.* 2006].

Cluster Analysis

Clustering, or segmentation, separates dataset instances or records into sub-groups or classes of similar instances known as *clusters*. There is a high similarity among instances in the same cluster and a low similarity among instances located in different clusters. The following are some of the differences between clustering and classification [Larose 2014; Du 2010]:

- Unlike classification, which develops a classification model, clustering focusses more on partitioning the whole data into homogeneous segments or clusters.
- Whereas classification specifies a target variable, clustering does not focus on any of the tasks related to classifying, estimating and predicting a target variable's value.
- Compared to classification, clustering does not specify any classes before a dataset is partitioned into different clusters.

Depending on how they perform the clustering task, clustering techniques are classified in different ways. There are *prototype-based, density-based, graph-based* and *model-based* clustering methods [Du 2010]:

- *Prototype-based* techniques create prototype clusters, which undergo changes during the clustering process. Examples of prototype-based clustering methods include the *k-means* and *self-organising map* (SOM) clustering techniques.
- In *density-based* clustering methods clusters are identified as areas that contain a high concentration of data points that represent the instances of a dataset. One of the most well-known methods that belong in this category is DBSCAN.
- In the *graph-based* clustering approach, a dataset is represented as a graph in which the vertices are the data points and in which the links indicate the proximity between the data points. Examples of methods that belong in this category are the minimum spanning tree (MST) and optimal partitioning methods.
- The aim behind *model-based* methods is to uncover "models that best fit the given data set" [Du 2010]. To locate clusters, the method creates a density function that indicates how the data points are distributed in space [Han *et al.* 2011]. Examples of model-based methods include COBWEB and expectation-maximisation (EM).

Association Analysis

Association analysis is also known, respectively, as *affinity analysis* and *market basket analysis*. This method focusses on the discovery of interesting relationships or significant associations among data items [Du 2010; Tan *et al.* 2006]. The term *item set* is used to refer to a set of one or more items. *Association rules* are used

to represent the relationships discovered among data items. An association rule is represented as an expression of the form:

$$X \Rightarrow Y$$

Where X and Y are mutually exclusive (disjoint) itemsets. The left-hand-side of the rule, X, is referred to as the *antecedent* of the rule and the right-hand-side, Y, is referred to as the *consequent* of the rule. Association analysis is discussed in detail in Subsection 2.6.4.

2.5.2 Text Mining

In Irfan *et al.* [2015] text mining is defined as:

> *"a knowledge discovery process used to extract interesting and non-trivial patterns from natural language".*

Other definitions of text mining can be found in [Hotho *et al.* 2005; Nasa 2012]. A vast amount of text is generated from a variety of sources such as social network sites, companies, newspaper articles, research papers, and so on [Hotho *et al.* 2005]. There is a need to make sense of and derive some meaning from this data. However, it is not easy to extract and analyse the content embedded in these unstructured texts, in order to obtain much-needed information. With the advent of text mining, an opportunity has arisen to access and analyse these vast resources of text data and, using different text mining techniques, to discover and extract meaningful information from them. Given the increase in computing power and resources, text mining techniques can be applied more efficiently.

Text Mining Process

The text mining process involves the following steps: *text preprocessing, text transformation, feature selection, text mining algorithms,* and *interpretation/evaluation*. These steps are illustrated in Figure 2.4

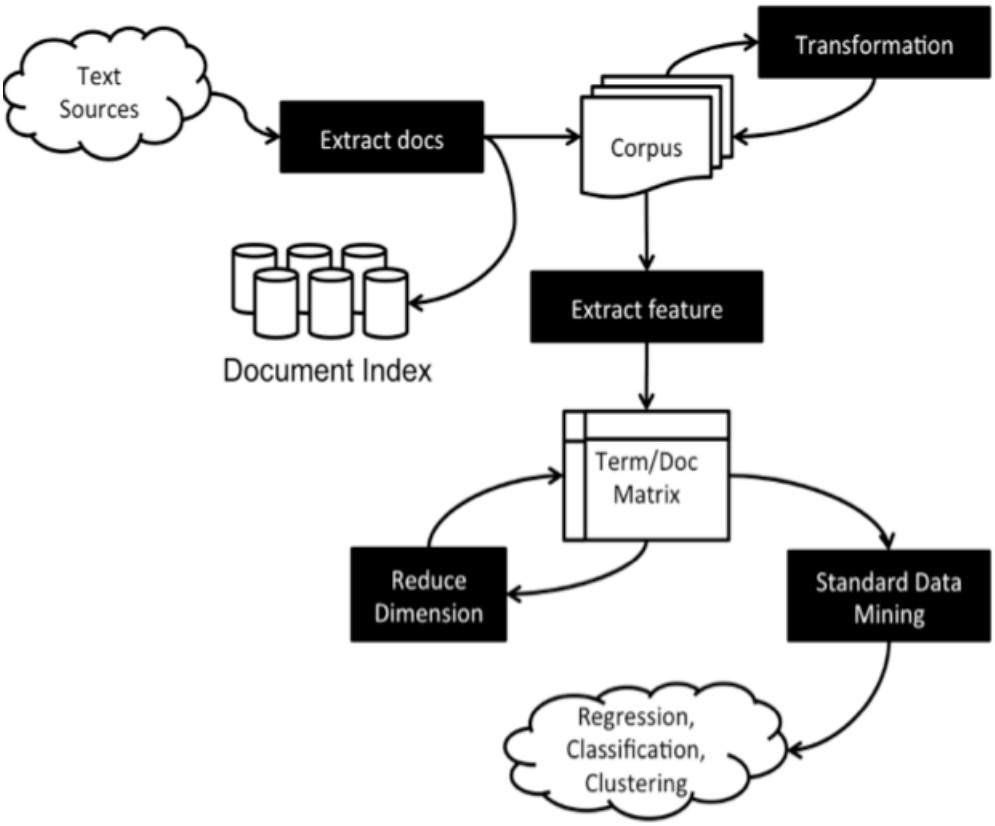

Figure 2.4: Text Mining Process [Ho 2016]

Different tasks are performed in the *text preprocessing* step. These include tokenization, stop word removal, stemming, parts of speech tagging, and word sense disambiguation.

The *text transformation* step involves tasks such as creating the term frequency document matrix (TDM) and computing frequency term counts. A document is represented using a bag of words and vector spaces [Sumathy and Chidambaram

2013].

The purpose of *feature selection* is to identify and select a subset of important variables that will be used during the modelling step. Redundant and irrelevant features are removed during this step.

In the *text mining* step selected data mining techniques are applied on the structured data prepared in previous steps. These text mining methods include document clustering, text categorisation and sentiment analysis.

In the last step, *interpretation/evaluation*, the results are analysed and checked for accuracy.

Areas of Text Mining

Text mining is an interdisciplinary field which encompasses different areas, of which some are illustrated in Figure 2.5 [Hotho *et al.* 2005; Feldman and Sanger 2007; Sumathy and Chidambaram 2013]

Figure 2.5: Text Mining Areas [Sumathy and Chidambaram 2013]

Information Extraction locates and extracts information from data stored in unstructured and/or semi structured documents. Some of the tasks involved in information extraction include term analysis, named entity recognition and fact extraction.

Information Retrieval deals with the *"representation, storage, organisation of, and access to information items, and it is concerned with providing efficient access to large amounts of unstructured contents, such as text, images, videos etc."* [Chevalier 2009]. Information retrieval systems include search engines such as *Google scholar, Yellow Book, Google* and others.

Natural Language Processing (NLP) is defined as *"a way for computers to analyse, understand, and derive meaning from human language in a smart and useful way."* [Lopez and Kalita 2017]. Some of the disciplines in which NLP activities are performed include signal processing, syntactic analysis, semantic analysis, pragmatics and others.

Data Mining is also known as *knowledge discovery in databases* (KDD). It is *"the process of discovering interesting and useful patterns and relationships in large volumes of data"* [Clifton 2009]. Data mining includes different techniques whose origins can be traced to disciplines such as numerical analysis, machine learning, and others.

Text Mining Applications

Text mining systems have been developed and implemented to address the needs of different stakeholders. There are applications in areas that include business, patent analysis, life sciences, and spam filtering [Feldman and Sanger 2007; Hotho *et al.* 2005].

- *Business Applications*
  - There are text mining applications in areas such as business intelli-

gence (market analysis) and corporate finance (bankers, analysts and consultants).

- *Patent Analysis*
  - Text mining approaches are used in patent research (patent development approaches and better use of corporate patents). Supervised and unsupervised text mining approaches are also used in analysing patent documents.
- *Life Sciences*
  - Researchers use text mining methods to study huge numbers of biomedical reports, in order "identify complex patterns of interactivities between proteins" [Feldman and Sanger 2007].These methods are also used in bio-entity recognition, where technical terms in molecular biology are identified and classified based on concepts of interest to biologists [Hotho *et al.* 2005].
- *Email Spam Filtering*
  - Spam is unsolicited electronic mail (e-mail), and is one of the major problems that affect the electronic mail system. Millions of e-mail users are inconvenienced through having to read and process messages that are of least interest to them, resulting in time and resources being wasted. One of the ways of dealing with this problem is to use machine learning-based text classification methods, capable of differentiating between spam and legitimate e-mails.

2.6 Related Work

Twitter has become an important platform where people express their opinions and views on political parties and candidates [Wang *et al.* 2012; Monti *et al.* 2013].

In countries such as Ireland, Germany and the United States research has been conducted in which text mining techniques have been used to study and analyse political tweets [Wang *et al.* 2012; Bakliwal *et al.* 2013; Tumasjan *et al.* 2010]. This section discusses related research in which political tweets have been used for different purposes, such as sentiment analysis, cluster analysis, association rule mining and word cloud analysis.

This section is structured as follows: Subsection 2.6.1 provides a discussion on sentiment analysis, a text mining technique used to classify the sentiment expressed in a sample of text. Subsection 2.6.2 discuses previous researches in which political tweets have been used to predict election outcomes. Clustering, a technique that groups together similar items, is discussed in Subsection 2.6.3. Subsection 2.6.4 discusses previous research, in which association rule mining methods have been used to generate and extract meaningful rules from Twitter corpora. Previous research focussing on the analysis of tweets using word clouds is discussed in Subsection 2.6.5.

2.6.1 Sentiment Analysis

Classification is the most common task in text mining. It involves classifying text into a prespecified set of categories [Feldman and Sanger 2007]. Sentiment analysis is a classification technique whose aim is to classify the sentiment inherent in text messages [Bifet and Frank 2010]. Tweet sentiment analysis aims to classify the sentiment of a tweet. There are two approaches to sentiment analysis: an unsupervised or a supervised approach.

An *unsupervised approach* uses pre-existing, manually created and validated sentiment lexicons [Hutto and Gilbert 2014]. These lexicons contain a list of words and their associated sentiment, either positive or negative. Examples of such lexicons include Linguistic Inquiry and Word Count (LWIC) and General Inquirer (GI). LWIC is a well-established lexicon containing 4 500 words that

have been validated by psychologists, sociologists and linguists in a process spanning over a decade [Hutto and Gilbert 2014]. GI contains more than 11 000 words that have been curated since 1996 and has been designed for content analysis [Hutto and Gilbert 2014]. Other lexicons include Affective Norms for English Words (ANEW), SentiWordNet and SenticNet. These lexicons indicate both word sentiment and sentiment intensity [Hutto and Gilbert 2014]. An unsupervised approach to sentiment analysis tends to be simpler compared a supervised approach [Cassinelli and Chen 2009]. In addition, an unsupervised approach is easily be generalized and applicable across different domains [Hutto and Gilbert 2014; Cassinelli and Chen 2009].

A *supervised approach* uses machine learning algorithms to identify essential features that contribute to the sentiment of textual data [Hutto and Gilbert 2014; Cassinelli and Chen 2009]. Often, machine learning algorithms such as Naïve Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs) are used to identify sentiment-related features in a sample of text [Hutto and Gilbert 2014]. However, supervised machine learning algorithms require labelled training sets, which can be a rather tedious and error-prone process [Hutto and Gilbert 2014]. Furthermore, supervised algorithms tend to be computationally expensive, both in terms of time and processing power [Hutto and Gilbert 2014]. Moreover, supervised algorithms are not easily applied to different domains, since they need to be retrained and validated [Cassinelli and Chen 2009].

Sentiment analysis has been applied to a variety of domains such as sociology, marketing, psychology, economics and politics [Hutto and Gilbert 2014]. In particular, there has been a growing interest in performing sentiment analysis on tweets to gauge online political sentiment [Bakliwal *et al.* 2013]. Tumasjan *et al.* [2010] applied sentiment analysis to tweets in order to gauge public sentiment towards political parties and politicians during the 2009 German federal election. Similarly, Bakliwal *et al.* [2013] used tweets to determine public sentiment during the run-up to the Irish General Elections in 2011. On the other hand, Monti *et al.* [2013]

leveraged tweets to specifically gauge public disaffection towards political parties in Italy during 2012. During the 2012 United States elections Twitter was used to track the sentiment of the general public, in real-time, toward the presidential candidates Wang *et al.* [2012].

In order to perform sentiment analysis on tweets, a corpus of tweets must be collected. Tumasjan *et al.* [2010] collected 104 003 tweets posted during 13 August 2009 and 19 September 2009, prior to the 2009 German federal elections. The tweets contained references to the following German political parties: Christian Democratic Union (CDU)/Christian Social Union (CSU), Social Democratic Party (SPD), Free Democratic Party (FDP), B90/Die Grünen and Die Linke. The tweets, which were in German, were then translated into English.

Similarly, Bakliwal *et al.* [2013] used the Twitter API to collect tweets prior to the 2011 Irish General ElectionsṪhey collected the tweets between 20 January and 25 January. The tweets mentioned the main political parties and their respective leaders: Fianna Fáil, Fine Gael, Labour, Sinn Féin and the Greens. The tweets were annotated manually, by two Irish annotators with a knowledge of the Irish political domain. Tweets labelled as sarcastic were ignored, and only tweets with either a positive, negative or neutral sentiment were considered. For their study, Bakliwal *et al.* [2013] collected a total of 2 624 labelled tweets.

Unlike the work done by Tumasjan *et al.* [2010] and Bakliwal *et al.* [2013], the study by Monti *et al.* [2013] did not specifically focus on political tweets, but rather on a mixture of political and non-political tweets. Monti *et al.* [2013] used the Twitter API to collect 120 000 tweets from 1 April 2012 to 1 June 2012. A training set consisting of 40 000 political and non-political tweets was used. The sentiment of each tweet was classified manually by 40 Italian sociology and political science students. Tweets with disagreeing sentiments were removed from the training set. The result was a final training set consisting of 28 340 tweets. Furthermore, Monti *et al.* [2013] collected a test set of 35 882 423 tweets from

167 557 Italian Twitter users, during the period 4 April 2012 to 10 October 2012. The tweets were tokenized and stemmed. Any reference to Italian political parties, politicians and political offices were removed in order to make the sentiment classifier more robust.

The work done by Tumasjan *et al.* [2010], Bakliwal *et al.* [2013] and Monti *et al.* [2013] all relied on a corpus of tweets collected before sentiment analysis was conducted. However, Wang *et al.* [2012] used tweets that were generated in real-time. A commercial Twitter data provider called Gnip Power Track was used to collect, in real time, tweets about the nine United States candidates participating in the presidential election. Wang *et al.* [2012] used 800 annotators to classify the sentiment of the tweets as either positive, negative, neutral or unsure. The result was a labelled training set consisting of 1 700 tweets.

As mentioned earlier, there are two main approaches to sentiment analysis, namely an unsupervised approach and a supervised approach. The model described in Tumasjan *et al.* [2010] adopted an unsupervised approach that used the LIWC lexicon to identify the sentiment of words within a tweet. The model classified the sentiment of a tweet into one of 12 possible sentiments: future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement and money. The authors did not provide a mathematical measure that could be used to determine the accuracy of sentiment classification. Instead, the results obtained from performing the sentiment classification were compared with anecdotal evidence from election programs and the press. Tumasjan *et al.* [2010] argued that the results of the sentiment classification closely correspond to political programs, candidate profiles and the evidence from the media coverage of the 2009 German federal elections.

The work done by Bakliwal *et al.* [2013] investigated an unsupervised and a supervised approach towards classifying the sentiment of political tweets. The unsupervised approach relied on pre-existing lexicons such as the Subjectivity Lex-

icon (SL) and the SentiWordNet (SWN) lexicon. Bakliwal *et al.* [2013] extended the SL lexicon with 341 domain-specific words, resulting in the extended SL lexicon. The extended SL and the SWM lexicons were combined to form a single lexicon. The sentiment of a word was classified by matching it to a word contained in the combined lexicon, after which the sentiment of a tweet was determined by calculating the average sentiment of the words in the tweet. This approach achieved an accuracy of 49.09%. To improve the model's accuracy, Bakliwal *et al.* [2013] extended the model by doing the following:

1. Considering only the sentiment of adjectives in a tweet
2. Stemming all inflected forms
3. Reversing the polarity of negated words
4. Matching sentiment-bearing phrases.

   Bakliwal *et al.* [2013] identified 89 sentiment-bearing phrases such as *"god save us"* and *"wolf in sheep's clothing"*. Each of these phrases were assigned a sentiment manually. If a tweet contained any of these phrases, the sentiment of the phrase was considered instead of the sentiment of each word.
5. Identifying indirect sentiment in comparisons.

   Bakliwal *et al.* [2013] determined the sentiment of a comparison by firstly identifying the sentiment of both sides of the comparison. Thereafter, the polarity of the sentiment was reversed depending on which side of the comparison a political party was.
6. Considering distance score.

   Firstly, the sentiment of a word w was determined using the process mentioned above. Thereafter, the sentiment was divided by the number of words between w and the political party mentioned in the tweet. According to Bakliwal *et al.* [2013], this indicated the political party-oriented nature of the

sentiment classification model.

The model developed using the modified unsupervised approach described in Bakliwal *et al.* [2013] achieved an accuracy of 58.96%. A supervised approach to sentiment classification was also explored in Bakliwal *et al.* [2013]. An SVM algorithm was trained on a corpus of 2 624 tweets, with the 5-fold cross validation method being used to select a random subset of data to be used during training. Firstly, a unigram approach was taken in which the top 1 000 most frequent words were used as features.

This approach achieved an accuracy of 58.92%. This result was further improved by incorporating the results of the unsupervised model. Using the combined lexicon from the unsupervised model, Bakliwal *et al.* [2013] identified 19 words that were most descriptive of sentiment. These words were added to the top 1 000 most frequent unigrams as features. This approach yielded an accuracy of 61.62%, a 2.66% improvement on the 58.96% accuracy obtained using the unsupervised model.

Whilst unsupervised and supervised methods were used by Bakliwal *et al.* [2013] in their study, the study by Monti *et al.* [2013], on the other hand, performed sentiment analysis by using only the supervised approach to specifically identify negative tweets. Monti *et al.* [2013] compared the accuracy of the following classifiers:

- ALMA (Approximate Maximal Margin Algorithm), a classifier that calculates the maximal margin hyperplane between two classes.
- OIPCAC (Online Isotropic Principal Component Analysis Classifier), a classification method that determines the linear combination of features that separates two or more classes of objects.
- Passive Aggressive, a supervised linear binary classifier.
- PEGASOS (Primal Estimated sub-GrAdient SOlver for SVM), a stochastic

sub-gradient descent algorithm that solves SVM optimization problems.

- Random Forest, a classifier that uses a set of classification trees.

Monti *et al.* [2013] obtained the best results with the Random Forest algorithm and 10-fold cross validation, which achieved an accuracy of 72.40%. However, this algorithm was found to be the most computationally expensive. On the other hand, the ALMA algorithm was found to be computationally inexpensive and obtained an accuracy of 70.30% which was similar to the accuracy obtained by the Random Forest algorithm. For these two reasons, the ALMA algorithm was selected by Monti *et al.* [2013] for use in performing sentiment classification.

Similarly, Wang *et al.* [2012] used a supervised approach to track, in real-time, the sentiment of the general public. A Naïve Bayes classifier was trained on 17 000 manually annotated tweets. The resulting model was able to classify the sentiment of a given tweet as either positive, negative, neutral or unsure. The tweets were streamed into the model in real-time and preprocessed, which involved tokenising the tweets and preserving URL's, punctuation and emoticons. Thereafter, the Naïve Bayes classifier was applied and the sentiment of individual tweets determined. The accuracy of the model was tested against a corpus of manually annotated tweets. The model achieved an accuracy of 59%.

Monti *et al.* [2013] obtained the best results using the Random Forest algorithm and 10-fold cross validation, which achieved an accuracy of 72.40%. However, this algorithm was found to be the most computationally expensive. On the other hand, the ALMA algorithm was found to be computationally inexpensive. Its accuracy was 70.30%, a value similar to the accuracy obtained by the Random Forest algorithm. For these two reasons, the ALMA algorithm was selected by Monti *et al.* [2013] and used for sentiment classification.

Similarly, Wang *et al.* [2012] used a supervised approach to track, in real-time, the sentiment of the general public. A Naïve Bayes classifier was trained on 17 000 manually annotated tweets. The resulting model was able to classify

the sentiment of a given tweet as either positive, negative, neutral or unsure. The tweets were streamed into the model in real-time and preprocessed, which involved tokenising the tweets and preserving URL's, punctuation and emoticons. Thereafter, the Naïve Bayes classifier was applied and the sentiment of individual tweets determined. The accuracy of the model was tested against a corpus of manually annotated tweets. The model achieved an accuracy of 59%.

An overview of the research discussed above is presented in Table 2.1. It can be seen that Monti *et al.* [2013] achieved the best performance with an accuracy of 72.40%. However, this model only identified negative sentiment. Furthermore, Monti *et al.* [2013] discovered a strong correlation between the negative sentiment found in tweets and the negative sentiment expressed in public opinion surveys. In this respect, the model presented by Monti *et al.* [2013] only focusses on negative sentiment. Tumasjan *et al.* [2010] provided no mathematical measure of the accuracy of their results. Instead, the authors argued that their sentiment classification results corresponded closely to political programs, candidate profiles and the evidence from the media coverage (of the 2009 German federal elections). The supervised model developed by Wang *et al.* [2012] achieved an accuracy of 59%. The supervised model presented by Bakliwal *et al.* [2013] achieved an accuracy of 61.22%, compared to the value of 58.96% achieved by the unsupervised model. Considering the simplicity of unsupervised sentiment classification models and the fact that these models can achieve accuracy levels similar to supervised models, the research I conducted adopted an unsupervised approach to sentiment classification.

Hutto and Gilbert [2014] developed a parsimonious rule-based model called *VADER* (Valence Aware Dictionary for sEntiment Reasoning) for identifying sentiment in microblog-like texts. The VADER model is able to classify the sentiment of tweets as well as the strength of the sentiment. The sentiment of each tweet can be classified as either positive, negative or neutral. The model returns

Table 2.1: Summary of previous research in political tweet sentiment analysis

	Election Year	**Country**	**Classification Approach**	**Accuracy**
Tumasjan *et al.* [2010]	2009	Germany	Unsupervised	NA
Bakliwal *et al.* [2013]	2011	Ireland	Unsupervised	58.96%
Bakliwal *et al.* [2013]	2011	Ireland	Supervised	61.22%
Monti *et al.* [2013]	2012	Italy	Supervised	72.40%
Wang *et al.* [2012]	2012	United States	Supervised	59%

a 3-tuple indicating the strength of the sentiment as negative, positive or neutral, respectively. A value of 1 indicates the strongest strength and a 0 indicates the weakest strength. The work done by Hutto and Gilbert [2014] consisted of two major parts: constructing and validating a valence-aware sentiment lexicon, and identifying heuristics used by humans to assess sentiment intensity.

The sentiment lexicon constructed by Hutto and Gilbert [2014] combined existing sentiment lexicons such as LIWC, ANEW and GI. Hutto and Gilbert [2014] extended the lexicon by incorporating sentiment-related emoticons, abbreviations and acronyms. For example, an emoticon such as "*:)*" denotes a "*smiley face*", and is associated with a positive sentiment. The "*LOL*" (Laugh Out Loud) acronym is an example of a sentiment-related acronym, which is also associated with a positive sentiment. Hutto and Gilbert [2014] also included commonly used sentiment-related slang. For example, common slang terms such as "*nah*" "*meh*" and "*giggly*" were included. This process resulted in the discovery of over 9,000 lexical candidate features. Hutto and Gilbert [2014] used 10 independent human annotators to determine the sentiment valence for each lexicon candidate feature. Each human annotator was requested to rate each word on a scale from -4 to +4. In this case, -4 meant extremely negative, 0 meant neutral and +4 meant extremely positive. The

result was a gold standard lexicon with associated sentiment valence. For example, "*okay*" was rated as 0.9 (positive) and "*sucks*" was rated at -2.2 (negative).

In order to identify the heuristics used by humans to assess sentiment intensity in text, Hutto and Gilbert [2014] used a data set of 10 000 publicly available tweets. Hutto and Gilbert [2014] used a tool called Pattern.en[1] to classify the sentiment of each tweet in the data set. The 400 most positive and negative tweets were selected. The sentiment intensity for these 400 tweets was determined by two human experts using a scale from -4 to +4. Thereafter, Hutto and Gilbert [2014] applied qualitative analysis techniques to identify characteristics that affect the text's sentiment intensity. Using this approach, Hutto and Gilbert [2014] identified five heuristics that indicate changes in sentiment intensity:

1. The exclamation mark ("!") increases the magnitude of the sentiment intensity without changing the semantic orientation. For instance, "*The service here is great!*" is more intense than "*The service here is great.*"

2. Capitalization of a sentiment-relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation. For example, "*The service here is GREAT.*" is more intense than "*The service here is great.*"

3. Degree modifiers affect sentiment intensity by either increasing or decreasing the intensity. For example, "*The food here is extremely good*" is more intense than "*The food here is good*". On the other hand "*The service here is marginally good*" reduces the intensity.

4. The conjunction "*but*" indicates a change in sentiment polarity, with the sentiment of the text following the conjunction being dominant. For example, "*The service here is great, but the food is horrible*" contains mixed sentiment,

[1] http://www.clips.ua.ac.be/pages/pattern-en#sentiment

with the phrase following the conjunction dominating the overall sentiment.

5. Examining tri-grams preceding a sentiment-laden lexical feature, Hutto and Gilbert [2014] was able to identify nearly 90% of cases where negation flips the sentiment of the text. A negated sentence would be "*The service here isn't really all that great*".

The gold standard lexicon and heuristics discussed above forms the basis of the VADER sentiment classification model. The performance of the VADER model was measured using the $F1$ score, a measure of accuracy that ranges from 0 (not accurate) to 1 (perfectly accurate). Traditionally, the $F1$ score is calculated as follows:

$$F1 = 2 \times \frac{recall \times precision}{recall + precision} \tag{2.1}$$

In this case the recall metric is the ratio of correctly classified tweets and the number of tweets with the same sentiment as identified using manual labelling. The precision metric is the ratio of correctly classified tweets and the number of tweets with the same sentiment as identified using the VADER model. The VADER model achieved an $F1$ score of 0.96. Thus, the model can accurately classify the sentiment of tweets. The VADER model is available as a Python library and can be used without needing a labelled training set. This makes it convenient to use and apply to multiple domains.

As discussed earlier in this section, sentiment analysis has been performed in countries such as Germany, Ireland, Italy and the United States to gauge the public's sentiment on a number of political issues. The unsupervised learning model developed by Hutto and Gilbert [2014] was highly accurate in classifying the sentiment expressed in tweets. I used this model my research to determine the sentiment of the tweets I generated and collected during the 2014 South African general election.

2.6.2 Predicting Election Results

Extensive research has been conducted to show how social media data can be used for predictive purposes. For example, Chung and Mustafaraj [2011] mentioned how Twitter has been used to predict box office revenue and stock market performance. Twitter has also become a vital campaigning tool for politicians [Tumasjan *et al.* 2010]. Within this context, research is being conducted in which Twitter is being used to predict election outcomes [Chrzanowski and Levick 2012]. Work done by Tumasjan *et al.* [2010] investigated whether tweets could be used to predict the popularity of political parties in the 2009 German federal elections. Similarly, the study conducted by Chung and Mustafaraj [2011] used tweets to predict the outcome of the 2010 United States Senate special election in Massachusetts. During the 2012 United States Presidential Election, Chrzanowski and Levick [2012] used Twitter to predict the outcome of that election.

The purpose of the study by Tumasjan *et al.* [2010] was to determine whether political tweets could be used to predict the popularity of the political parties that participated in the 2009 German federal elections. Prior to the 2009 German federal elections, Tumasjan *et al.* [2010] collected 104 003 tweets posted between 13 August 2009 and 19 September 2009. Six German political parties were mentioned in the tweets, namely the CDU, CSU, SPD, FDP, B90/Die Grünen and Die Linke. Prominent politicians were also mentioned. The collected tweets were posted in German and then translated into English. In order to determine whether Twitter could be used predict election results, the total number of tweets mentioning a political party was compared to that party's election results. The result of this comparison can be seen in Table 2.2.

It can be seen that the percentage of tweets that mention a political party is similar to the percentage of the votes received by that party. Tumasjan *et al.* [2010] also

Table 2.2: Comparison of total number of tweets and election results for each German political party, taken from Tumasjan *et al.* [2010]

Party	**Number of tweets**	**Share of Twitter traffic**	**Election result**	**Prediction error**
CDU	30 886	30.1%	29.0%	1.1%
CSU	5 748	5.6%	6.9%	1.3%
SPD	27 356	26.6%	24.5%	2.2%
FDP	17 737	17.3%	15.5%	1.7%
Linke	12 686	12.4%	12.7%	0.3%
Grüne	8 250	8.0%	11.4%	3.3%

calculated and obtained a value of 1.65% for the Mean Absolute Error (MAE), a measure of forecast accuracy. This value was then compared to results from various election polls. Tumasjan *et al.* [2010] showed that using the number of tweets to predict election results achieved similar accuracy to that obtained in traditional election polls.

The study by Chung and Mustafaraj [2011] followed an approach similar to that of Tumasjan *et al.* [2010] to predict the results of the 2010 United States Senate special election, held in Massachusetts. Chung and Mustafaraj [2011] used the Twitter API to collect tweets mentioning either Martha Coakly or Scott Brown, the two candidates in the election. A total of 23 467 tweets were collected from 13 January to 20 January, 2010. The tweets were preprocessed by removing hash tags, user names, URLs and emoticons.The study showed that 53.86% of the tweets mentioned Martha Coakly and 46.14% mentioned Scott Brown. However, this did not correspond to the election results, which Scott Brown won by receiving 52% of the votes, compared with Martha Coakly's 48%. This result prompted Chung and Mustafaraj [2011] to argue for tweet sentiment to be taken into account, since tweets may reflect an opposing rather than a supportive sentiment towards a candidate. Chung and Mustafaraj [2011] used an unsupervised method, in the form of the SentiWordNet classifier, to classify the sentiment of tweets. This approach achieved an accuracy of 69.03%. According to Chung and Mustafaraj [2011], the

accuracy of the sentiment classifier was not reliable enough to predict the outcome of the election.

In Chrzanowski and Levick [2012] the aim was to predict how a Twitter user would vote in the 2012 United States presidential election involving two candidates: Barack Obama and Mitt Romney. The task was to predict whether a Twitter user would vote for Barack Obama or Mitt Romney, by identifying users who have publicly endorsed either Barack Obama or Mitt Romney. The tweets were stemmed and all numbers were removed. For example, if a tweet originated from an Obama supporter, it was labelled as being a vote for him. Likewise, tweets originating from Romney's supporters were considered to be votes in his favour. A total of 7.5 million tweets were collected, with 50% coming from Obama supporters and 50% from Romney supporters. The tweets were split as follows: 80% were used for training and 20% for testing. The training set was used to train a SVM classifier. The resulting model achieved an accuracy of 94% in predicting whether a user would vote for either Barack Obama or Mitt Romney.

In this research I investigated whether a relationship existed between the number of tweets that mention a particular political party and the number of votes obtained by that party during the 2014 South African general election. In performing this task, I decided to adopt the approach described in Tumasjan *et al.* [2010] and Chung and Mustafaraj [2011], and to consider the number of political tweets posted by users and the sentiment of those tweets.

2.6.3 Tweet Clustering

Clustering assigns objects into different groups called *clusters* [Feldman and Sanger 2007]. The technique is also applied to tweets, resulting in tweets that are similar being placed in the same cluster. Several studies have been conducted in which clustering has been used to analyse Twitter datasets. In the study by Rosa *et al.* [2011], tweets were clustered into six predefined categories: News, Sports, En-

tertainment, Science, technology, Money and "Just for fun". Sankaranarayanany *et al.* [2009] investigated the clustering of tweets from news topics generated in real-time.

Rosa *et al.* [2011] used the Twitter API to collect a total of 1 107 007 tweets during the second and third week of March in 2011. This corpus contained 30 hashtags directly corresponding to six categories: News, Sports, Entertainment, Science, Technology, Money and "Just for fun". For example, hashtags such as #knicks, #heat, #nfl, #nba and #nhl were used to group the corresponding tweets under the Sport category. The tweets were preprocessed by separating each word by a white space, converting each word to lower case and removing all rare words (i.e. words with term frequency < 5).

Unsupervised and supervised clustering approaches were investigated by Rosa *et al.* [2011]. Two unsupervised algorithms, namely Latent Dirichlet Allocation (LDA) and K-Means, were applied to the subset of a corpus. The LDA algorithm achieved an $F1$ score of 0.143. Similarly, the K-Means algorithm achieved an $F1$ score of 0.143. Rosa *et al.* [2011] stated that these low $F1$ scores indicated that tweets do not naturally cluster along topics. Thereafter, Rosa *et al.* [2011] investigated a supervised approach and applied the Rocchio classifier, to classify tweets into one of the six predefined categories mentioned earlier. The corpus was automatically labelled using the hashtag in a tweet as an approximate indicator of the topic related to the tweets. Using this labelled corpus, 400 000 tweets were used for training and 50 000 for testing. This supervised approach achieved an $F1$ score of 0.685, a 54.2% improvement compared to the unsupervised approach.

In Sankaranarayanany *et al.* [2009] tweets were clustered using news topics generated in real-time. Predefined categories for the news topics were not identified. Instead, 2 000 users known for publishing news on Twitter were selected manually. All the tweets from these users were collected in real-time. In addition to the tweets obtained from these users, the Twitter API was used to stream all

public tweets. Many of the tweets that were collected did not relate to any news. These junk tweets were filtered out, and news-related tweets were retained and used. Firstly, a tweet corpus was collected and each tweet classified manually as being either news or junk. A Naïve Bayes classifier was then trained using this labelled corpus. Lastly, the model produced from the training phase was used to filter out junk tweets, with the remaining news-related tweets being used for clustering.

Sankaranarayanany *et al.* [2009] used the leader-follower clustering algorithm to cluster tweets in real time. Each tweet is represented as a feature vector using term frequency and inverse document frequency (TF-IDF) to indicate the importance of individual words in the tweet. The content and time centroids are determined during cluster formation. The content centroid is obtained by creating a weighted feature vector using TD-IDF containing all words in that cluster. The time centroid is obtained by calculating the mean publication time of the tweets in a cluster. When a tweet has to be assigned to a cluster the cosine similarity measure is used to calculate the distance between the tweet and every other cluster. The tweet is placed in the cluster closest to it, provided that the distance between the tweet and the cluster is less than a pre-defined threshold. Otherwise, the tweet is assigned to a new cluster.

Having a tweet form a single cluster by itself may lead to unrelated tweets forming singleton clusters. To avoid this situation, Sankaranarayanany *et al.* [2009] identified *seeder sources*, namely users with direct links to news sources. Only tweets from a seeder source may form a new cluster. Sankaranarayanany *et al.* [2009] have also indicated that some news events may be identified by non-seed users. In order to handle such cases, any tweet is allowed to form a new cluster Sankaranarayanany *et al.* [2009]. After K tweets have been added to the cluster and if none of the tweets can be allocated to a seeder, then the cluster has to be removed.

Another problem to avoid is the presence of duplicate clusters. If the news

event in a tweet is similar to a cluster but the distance between the cluster and the tweet is greater than the threshold a new duplicate cluster is formed. These two clusters are essentially competing with each other for tweets. In order to avoid this situation, the older cluster is labelled as the *"master"* and the duplicate cluster is labelled as the *"slave"*. So, if a tweet is closer to a slave cluster, it is added to the master cluster. Sankaranarayanany *et al.* [2009] did not provide any mathematical measure of the accuracy of the proposed clustering model. Therefore, no direct comparison may be made to the work done by Rosa *et al.* [2011].

The work done by Rosa *et al.* [2011] and Sankaranarayanany *et al.* [2009] showed that tweets can be clustered according to certain topics. In this research I used clustering techniques to analyse the corpus of tweets I collected during the 2014 South African election.

2.6.4 Association Rule Mining

Association Rule Mining (ARM) is the process of discovering rules that depict **the relationship** between the **itemsets** of a dataset [Zhao 2012]. An association rule **has** the form

$$A \Rightarrow B,$$

where A and B are disjoint **itemsets**, known respectively as the *antecedent* and *consequent* of the rule. A vast number of association rules may be discovered, which grows exponentially as the size of the dataset increases [Hipp *et al.* 2000]. Given the large number of rules that may be generated, there is a need for a measure to be used to identify only those rules that are interesting. The measures often used to identify interesting rules are *support, confidence* and *lift*. Given the disjoint **itemsets** A and B, these measures are defined as follows [Zhao 2012]:

- The *support* of the association rule $A \Rightarrow B$ is the proportion of cases that contain both A and B.

- The *confidence* of the association rule $A \Rightarrow B$ is the percentage of cases containing A that also contain B.
- The *lift* of the association rule $A \Rightarrow B$ is the ratio of confidence to the percentage of cases containing B.

Association rule mining has become one of the key data mining methods that have attracted significant interest and attention from data mining researchers and practitioners [Zaki and Hsiao 1999]. Cagliero and Fiori [2013] used the Twitter API to collect tweets in two locations, namely New York and London, from 20 March to 24 March, 2011. The aim of this study was to analyse the tweets in order to discover association rules. Three stages were involved in the model used: tweet representation, taxonomy generation and generalized pattern mining. In the tweet representation stage the tweets are modelled as records that either contain the tweet contents or contextual information such as location or time. Taxonomies were generated to help in mining association rules. They are hierarchical representations of major concepts within a domain. The taxonomies generated in Cagliero and Fiori [2013] represent the "is-a" relationship within the domain. The final stage is about discovering association rules from the tweet corpus by using the taxonomies generated earlier.

Below are examples of the association rules discovered by Cagliero and Fiori [2013] from the set of tweets collected in New York:

(i) $(Keyword_1, Obama), (Place, Washington, D.C.) \rightarrow \{(Date, 2011-03-22)\}$
$(sup = 3.6\%, conf = 100\%)$

(ii) $(Keyword_1, Congress), (Place, Washington, D.C.) \rightarrow \{(Date, 2011 - 03 - 22)\}$
$(sup = 2.2\%, conf = 97\%)$

Zingla *et al.* [2014] used association rules to provide more context on tweets. The INEX 2014 corpus, which contains 3 902 346 Wikipedia articles, was mined for association rules using the Closed Association Rule Mining (CHARM) method.

Below is an example of the rules discovered by Zingla *et al.* [2014], including the corresponding support and confidence values:

(i) $college \longrightarrow university \quad (sup = 426, conf = 0.812977)$

(ii) $motor \longrightarrow car \quad (sup = 187, conf = 0.806034)$

Thereafter, the tweets were transformed using the discovered association rules. If a word matched the left hand side of a rule, then it was replaced with the right hand side of the rule. For example, if a tweet contained the word "college" then it would be replaced with "university". Using the transformed tweet, Zingla *et al.* [2014] searched the INEX 2014 corpus for relevant Wikipedia articles. The generated summaries were then used to provide extra context around tweets. In this research I used an association rule mining algorithm to generate interesting and informative rules from the corpus of tweets I collected during the 2014 South African general election.

2.6.5 Word Clouds

Word cloud analysis uses a data visualization technique to indicate the importance of words in a corpus and to enable users to quickly identify the primary content in the corpus [Zhao 2012; Wu *et al.* 2011]. The importance of a word is often indicated by how frequently it appears in a corpus. The Term Frequency-Inverse Document Frequency (TF-IDF) is a measure that indicates the importance of a word in the document of a given corpus [Hotho *et al.* 2005]. Consider a corpus C containing N documents, then the TF-IDF for a word w is given by Equation 2.2.

$$tfidf(w) = \log\left(\frac{N}{|d \in C : w \in d|}\right) \tag{2.2}$$

That is, the TD-IDF of a word w is the logarithm of the total number of documents in the corpus divided by the number of documents that contain w.

The primary focus of the work done by Wang *et al.* [2012] was to track the sentiment of the public in real-time during the 2012 United States elections. As part of the research, an Ajax-based HTML dashboard was developed which displayed useful information such as tweet volumes and a word cloud. Similarly, Sankaranarayanany *et al.* [2009] generated a word cloud on the 2009 Iranian elections. The main purpose of the work done by Sankaranarayanany *et al.* [2009] was to develop a news processing system for identifying tweets related to news events.

Word clouds are generated from the words contained in a corpus of documents. Wang *et al.* [2012] and Sankaranarayanany *et al.* [2009] generated word clouds from a corpus of tweets. Wang *et al.* [2012] used a commercial Twitter data provider called Gnip Power Track to gather tweets, in real-time, on the nine United States republican candidates who took part in an election. Tweets that were collected within one minute were combined to form one document. Wang *et al.* [2012] collected documents over a period of 2 hours and built a corpus consisting of 120 documents that was used to generate a word cloud.

In Sankaranarayanany *et al.* [2009] 2 000 Twitter users known to publish news on Twitter were selected manually, and their tweets collected. Clustering techniques were used to group the tweets into clusters related to topics dealing with specific news events. One of the clusters identified in the study was about the 2009 Iranian elections. The tweets contained in this cluster were used to generate a word cloud.

As mentioned earlier, word clouds provide a visual representation of the importance of words within a corpus of documents. The study by Wang *et al.* [2012] used TD-IDF to identify the most important words in a corpus. These words were used to form a word cloud. The process was repeated several times, resulting in a new word cloud being generated every five minutes, with the most important words being displayed. The word cloud generated by Wang *et al.* [2012] is shown at the bottom left corner of Figure 2.6. A larger version of this image can bee seen in

Figure 1, in Appendix A.

Looking at Figure 2.6, it can be seen that words such as "Santorum", "Kansas" and "Romney" are prominently featured in the word cloud, which would indicate that these words were important at the time. A qualitative assessment of the word cloud was performed by Wang *et al.* [2012], and the TF-IDF measure was shown to be effective in identifying the most prominent words in the corpus used in the study.

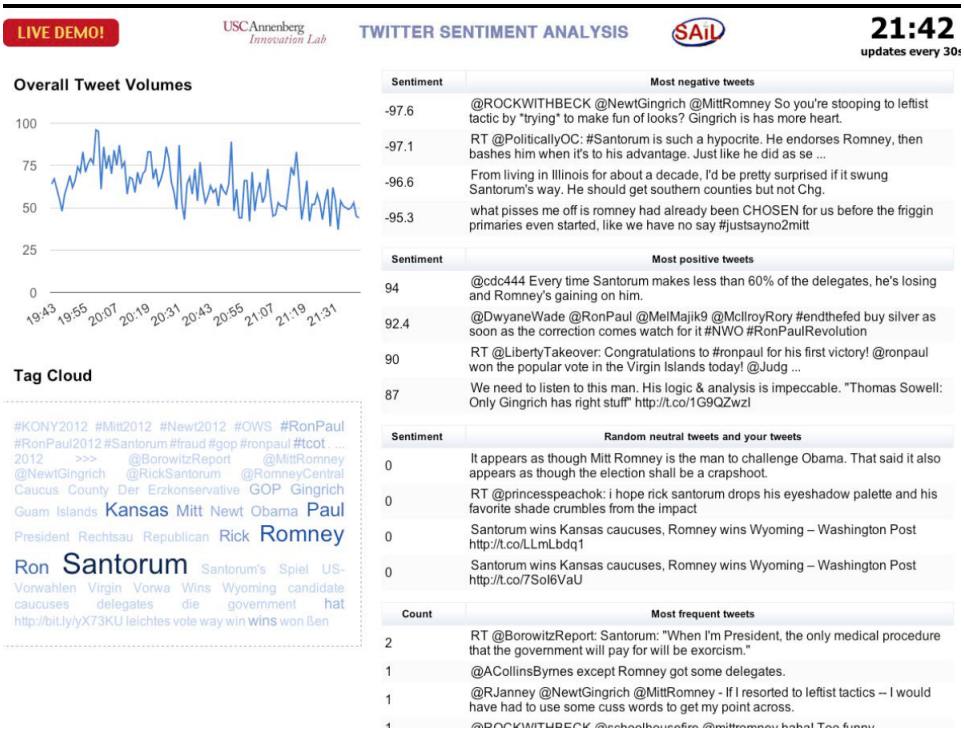

Figure 2.6: A dashboard developed by Wang *et al.* [2012] which displays a word cloud

The research by Sankaranarayanany *et al.* [2009] determined the importance of a word based on how frequently it appeared in a corpus. Using a cluster of tweets from the 2009 Iranian elections, the hashtags from these tweets were used to generate a word cloud, shown in Figure 2.7. The font size of a hashtag is directly proportional to the number of times the hashtag appears in the cluster [Sankaranarayanany *et al.* 2009]. Words such as "iranelection", "neda" and "tehran" are

displayed prominently, which indicates their importance. For example, the word "neda" is related to the shooting of Neda Agha-Soltan during a protest in the 2009 Iranian elections.

Figure 2.7: Word cloud visualizing hash tags associated with the 2009 Iranian elections taken from Sankaranarayanany *et al.* [2009]

Both Wang *et al.* [2012] and Sankaranarayanany *et al.* [2009] showed that word clouds can be used effectively to visualise keywords and topics contained in a corpus of political tweets. To determine how effective and informative word clouds would be within the context of my research, I performed word cloud analysis using a corpus of tweets posted during the 2014 South African election.

2.7 Conclusion

This chapter presented a background and discussed the related work required to understand and motivate the research that was conducted. Section 2.2 provided a discussion on the adoption and application of opinion polls to gauge the public's views during elections. Social Network Sites (SNS) such as Twitter and Facebook have revolutionised communication and information exchange. Thanks to the Internet, millions of users all over the globe are able to communicate amongst themselves using Twitter, Facebook, and many other social network facilities. There are a vast number of SNSs available today, which can be classified into seven cat-

egories: social connections, multimedia sharing, professional, informational, educational, hobbies and academic. Detailed descriptions of each of these categories were provided in Section 2.3. Since this research focused on Twitter, Section 2.4 was devoted to a detailed discussion of Twitter and the structure of tweets. The section also included a discussion on the tweets collected during the 2014 South African election. A brief background and overview discussion of the data mining and text mining techniques necessary for a better understanding of this research was provided in Section 2.5.

Section 2.6 discussed the use of text mining techniques in different countries such as the United States, Germany and Ireland. These techniques include sentiment analysis, election result prediction, tweet clustering, association rule mining and word cloud analysis. Sentiment analysis was discussed in Subsection 2.6.1, which focussed on the different methods used to determine sentiment in political tweets. Subsection 2.6.2 discussed how tweets can be used to predict election results in countries such as Germany and the United States. Although not related to any political domain, Section 2.6.3 discussed the use of clustering techniques in identifying news events from tweets. Previous work on the application of association rule mining methods in relation to tweets was presented in Subsection 2.6.4. A discussion was presented in Subsection 2.6.5 regarding the use of word clouds in visualising the information generated from tweets related to the United States and Iranian elections.

In the next chapter, Chapter 3, I present and discuss the methodology I followed in conducting this research.

Chapter 3

Methodology

3.1 Introduction

Chapter 2 provided a detailed background and literature review, focussing on the topic of text mining and some of the techniques used to extract meaningful information from unstructured text found social media platforms, particularly Twitter. This chapter presents and discusses the methodology I followed in conducting this research, in order to obtain information from a corpus of tweets that was generated during the 2014 South African general election. Twitter is used to communicate, share ideas and opinions across different fields such as education, healthcare, business, and politics [Tumasjan *et al.* 2010]. Particularly in politics, Twitter and other social media platforms are used to provide an environment in which different political stakeholders and other interested parties can participate and engage in a variety of political discussions [Yang and DeHart 2016; Muntean 2015].

In several African countries, citizens have taken to social media platforms such as Twitter and Facebook to actively engage in the political life and events unfolding in their respective countries. Consider the domestic uprisings that occurred in Tunisia and Egypt, which came to be known as the "Arab Spring", Twitter and

Facebook are amongst the tools that were predominantly and effectively used by the citizens in those countries to orchestrate the downfall of their governments [Howard *et al.* 2011]. In South Africa one of the main characteristics of the 2014 general election was the use, by different political parties, of social media platforms such as Twitter and Facebook for electoral campaigning [Reprobate 2016; Tracey 2013]. This research focusses primarily on the use of Twitter as an electoral and political campaigning tool during the 2014 South African general election.

This research was motivated by a number of factors, which are discussed in Section 3.2. A case study methodology was followed in conducting the research. In addition, given the focus of this research on the application of data mining and text mining techniques, certain phases of the *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) process model were selected and incorporated as part of the methodology I followed to conduct the research. Section 3.3 provides a detailed discussion of the methodology I used. The section also includes a discussion of the CRISP-DM process model. Section 3.4 presents and discusses the data metrics I used to evaluate the results I obtained from my research. Finally, the chapter is concluded in Section 3.5.

3.2 Motivation

Social Network Sites (SNS) such as Twitter, Facebook and others, have revolutionised communication and information exchange. Thanks to the Internet, millions of users all over the globe are able to communicate amongst themselves using Twitter, Facebook, and many other social network facilities. In particular, Twitter enables users to communicate with one another and to exchange information, opinions and views, and to engage each other on a number of issues. In addition, Twitter is used as a political discussion and communication platform [Tumasjan *et al.* 2010].

The use of Twitter as a political communication platform was particularly evident in former United States president Barack Obama's presidential campaign in 2008. Numerous researchers have become interested in studying and analysing political tweets in order to obtain better understanding and insights into the political environment, political thinking, ideas and opinions of a section of the population that communicates with other users through social media networks such as Twitter. For example, Tumasjan *et al.* [2010] have used a corpus of political tweets to predict the 2009 German election results. Larsson and Moe [2012] have been able to show the direct relationship between Twitter usage during the 2010 Swedish election and the political debates and rallies that took place during that election. Monti *et al.* [2013] have done a time-series analysis focussing on political disaffection in relation to the 2012 elections held in Italy.

This research is motivated by a number of factors. Not much evidence is available showing that active research is being conducted in South Africa which focusses on how the data generated by Twitter and other social media networks can be leveraged, specifically to obtain useful information about the political ideas, opinions and views of the voting public, including information about the campaigns and strategies used by political parties and their leaders. Currently, there is still heavy reliance in South Africa on opinion polls and surveys as a means of gauging public opinion on a number of issues, including general elections. For example, a market research institution, Ipsos, conducted an opinion poll in 2013 to determine the public's disaffection with the South African government [Ipsos 2015]. This research therefore addresses this gap by demonstrating the possibility of the kind of information and insights that can be obtained from processing, extracting and analysing Twitter data.

One of the problems facing polling organisations in South Africa has to do with the fact that significant sections of the South African population, including those living in rural areas, are not covered or catered for during the polling season. Therefore, the opinions and views of this vast section of the country's population

are often overlooked, even though a claim can be made that the surveys or polls are representative. However, thanks to Twitter and other social media platforms, this problem is being minimised, including the urban/rural divide problem that characterises countries like South Africa [Worx 2012]. Another motivation was therefore to use Twitter to access a wider audience or spectrum of the South African public, and to "harvest" their political views and opinions, specifically around issues related to the 2014 South African general election.

Given the vast number of tweets generated daily, it is impractical to analyse them manually, since the process may also be error-prone. Text mining techniques are particularly well suited for this task, as shown by the results obtained from previous and other research. The motivation for this research was therefore to see whether the selected text mining techniques could be used to leverage the tweets posted before, during and after the 2014 South African general election to produce meaningful and insightful information. It is hoped that this research will contribute to a better understanding of the role played by social network sites, especially Twitter, in the South African political landscape.

3.3 Detailed Research Methodology

3.3.1 Research Approach- A Case Study

Every five years a general election is held in South Africa. The most recent election was held in 2014, and several political parties took part in it. Before and during the election opinion polls were used to obtain the views of the voting public and this data was used, among other things, to predict the result of the election. Twitter and Facebook users were also active during this period, posting their views and opinions on the election and the political environment prevailing in the country during that period. In this study I obtained and analysed tweets that were generated before and during the 2014 general election. The questions that needed to be answered in

this study were primarily descriptive and explanatory in nature. Descriptive questions focus on "*what* happened" whilst the nature of explanatory questions is about "*how* or *why* did something happen?" [Feuer *et al.* 2002, as cited in Yin [2003]]. Given the nature of the questions that were being addressed in my research, I considered a case study methodology to be a suitable approach for conducting my research.

Case study research involves executing an empirical investigation of a contemporary phenomenon within its natural context using numerous sources of evidence [Hancock and Algozzine 2015, as cited in Yin [2003]]. There are three defining characteristics of case study research. Firstly, case study research focuses on contemporary phenomena, which represent real-life human situations. [Hancock and Algozzine 2015; Shoaib and Mujtaba 2016]. An example of a contemporary phenomenon would be the 2014 South African general election.

Secondly, the phenomenon being investigated must be placed in its natural context, bounded by space and time [Hancock and Algozzine 2015]. Removing a phenomenon from its natural setting will result in it being misunderstood [Shoaib and Mujtaba 2016]. The 2014 South African general election is an example of a contemporary phenomenon. It was held on May 7, 2014, in South Africa.

Thirdly, case study research is immensely descriptive because it is grounded in various sources of information [Hancock and Algozzine 2015]. Case studies often involve interviews, quotes and surveys from participants, an indication of the complexity inherent in the phenomenon being studied [Hancock and Algozzine 2015]. In my research, the source of information was the corpus of election tweets I collected just before and during the election.

In addition to the characteristics mentioned above, case studies can be classified according to their disciplinary orientation. In particular, a case study can be classified as: ethnographic, historical, psychological or sociological [Hancock and Algozzine 2015].

An ethnographic case study is a comprehensive study that explores the culture, behaviour and lifestyle of participants [Hancock and Algozzine 2015; Fusch and Ness 2017]. In this type of case study, the researcher may have to immerse themselves into the day-to-day lives of the participants. For example, Thome [1993] [as cited in Hancock and Algozzine 2015] mentioned a researcher that studied how children experienced gender in school. In order to answer this question, the researcher visited a multi-cultural school and observed learners going about their day-to-day activities in the school.

Historical case studies are another type of case study, which usually involve the description of a phenomenon as it evolves over time [Hancock and Algozzine 2015; Shoaib and Mujtaba 2016]. This type of case study typically focuses on individuals or organisations but may sometimes investigate events using various theories and concepts from psychology [Hancock and Algozzine 2015]. The case study Lankard and McLaughlin [2003] conducted is an example of an historical case study. They identified and described essential messages used by an environmental agency, the Wilderness Society, to promote conversation from 1964 to 2000.

Psychological literature and practices are commonly used in psychological case studies [Hancock and Algozzine 2015]. Psychological case studies focus either on individuals or groups of individuals. Utsey *et al.* [2003] used a case study approach to show how therapeutic mentoring can reduce self-destructive behaviours and promote socially adaptive behaviours amongst a group of urban African American male adolescents.

Sociological case studies focus on a variety of topics such as families, religion, politics, urbanisation, and issues relating to gender, race, status and even ageing [Hancock and Algozzine 2015]. Often, sociological case study research examines society and social relationships [Hancock and Algozzine 2015]. For example, Salamon [2003] conducted a sociological case study in the United States

of America which investigated the urbanisation of agricultural communities after World War II. Using descriptions from sociologists and anthropologists Salamon [2003] found that the urbanisation process transformed agricultural communities socially and physically, a development which threatened the uniqueness of these communities.

Case study designs may also be classified as intrinsic, instrumental, collective, exploratory, explanatory or descriptive [Hancock and Algozzine 2015]. Intrinsic case studies are conducted when the researcher is interested in finding out more details about a particular individual, group, event or organisation, without being interested in generalising any findings [Hancock and Algozzine 2015; Shoaib and Mujtaba 2016]. An intrinsic case study was conducted by Kalnins [1986] in order to gain an in-depth understanding of the contexts, processes and interactions within health care facilities that influence the residents' views on life.

Instrumental case studies are conducted when the research purpose is to better understand a general problem or theoretical question by investigating a particular phenomenon [Shoaib and Mujtaba 2016; Hancock and Algozzine 2015]. Kincannon [2002] conducted an instrumental case study to better understand the experience of university faculty when changing from face-to-face classroom lectures to Web-based lectures. The participants in the study came from a research institution in south eastern United States. In order to generalise the results beyond the participants in the group, Kincannon [2002] selected participants who represented a variety of experiences and opinions on distance education. Kincannon [2002] collected data through interviews, document reviews, lecture observations, and focus groups.

A collective case study is used when a researcher wants to obtain a general understanding and contribute to the literature base by utilising several instrumental cases [Shoaib and Mujtaba 2016; Hancock and Algozzine 2015]. Using a collective case study, Crudden [2002] examined the factors that influence the job retention of

people with vision loss. Crudden [2002] found that computer technology was a positive influence and that print access and technology were a source of stress for most participants.

Exploratory case study research aims to determine research questions for subsequent research or the validity of research methods [Hancock and Algozzine 2015]. Hancock and Algozzine [2015] provides an example of how an exploratory case study could be used to determine a company's ethical climate and how business practices occur.

Explanatory case studies are used to establish cause-and-effect relationships and to identify how events occurred [Shoaib and Mujtaba 2016; Hancock and Algozzine 2015]. For example, a teacher could conduct an explanatory case study to determine the factors in a student's home environment that affect her or his performance at school [Hancock and Algozzine 2015].

Descriptive case study research is conducted in order to obtain a complete description of a phenomenon within its natural context [Hancock and Algozzine 2015]. As an example, a descriptive case study could be used to obtain an in-depth description of a hospital's emergency room procedure when admitting incoming patients. [Hancock and Algozzine 2015].

As mentioned above, case studies are used in a variety of fields. Within the context of this work, case studies have been used in research involving Social Networking Sites. For example, Maurer and Wiegmann [2011] conducted a case study to determine the effectiveness of Facebook for marketing. A survey was conducted between November 28th, 2009 to January 7th, 2010 in Austria. A total of 2.09 million Facebook users from Austria participated. The survey contained 17 questions and was divided into 3 parts. The first part determined why the participants used Facebook. The second part determined how often the participants used Facebook. The third part aimed to determine if participants use Facebook as a source of information before making purchase decisions. Maurer and Wiegmann [2011]

found that the participants mainly used Facebook to stay in contact with friends and family. The results also showed that the majority of participants used Facebook several times a day. Most importantly, Maurer and Wiegmann [2011] found that the participants do not use Facebook as a source of information and do not purchase products because of Facebook.

Another case study conducted by Bicen and Uzunboylu [2013] aimed to investigate the effects of Facebook on education and teachers' opinions about online learning. The study involved teachers from primary and secondary schools. A group of 35 teachers conducted traditional lectures, whilst 36 teachers used Facebook to conduct online lectures. Bicen and Uzunboylu [2013] used a questionnaire consisting of 39 positive statements about Facebook, and each teacher had to indicate whether they agreed or disagreed with each statement. The participants took the survey before and after the experiment. Bicen and Uzunboylu [2013] found that Facebook could bring about positive change in teachers' opinions. In addition, the results showed that Facebook enables teachers to perform a variety of activities in an online classroom, which is not possible in the traditional classroom.

In this research the aim was to determine whether text mining techniques could uncover meaningful information when applied on a corpus of tweets posted before, during and after the 2014 South African general election. Given the exploratory, descriptive and explanatory nature of the questions that needed to be answered, a case study approach was deemed suitable for this study. Furthermore, such an approach has been shown to work successfully in studies involving Social Networking Sites such as Twitter and Facebook.

3.3.2 Research Setting and Subjects

Research Setting

This research focused on the South African general election which was held on 07 May 2014. A total of 29 political parties participated in this election, which was won by the ruling African National Congress (ANC). The ANC obtained 11 436 921 votes (62.15%) out of a total of 18 402 497 valid votes cast. The second largest party, the Democratic Alliance (DA) obtained 22.23% of the valid votes cast, followed by the recently formed Economic Freedom Fighters (EFF), which obtained 6.35% of the valid votes cast. The remaining smaller parties were left to share 9.27% of the total votes cast. Of all the national elections held so far in South Africa, the 2014 general election was characterised by the robust, active and vibrant participation of different stakeholders in the social media space [Nevill 2014]. Different political parties, leaders and members of the public used social network sites such as Facebook and Twitter to interact with and engage one another in sharing their political views, opinions and preferences, as well as participating in discussion and debate on a variety of issues related to the 2014 general election.

Research Subjects

This research did not involve direct participation by any human subjects. Instead, a corpus of 2014 South African election tweets was used. An application was developed to connect to the Twitter API, query the API for South African election tweets and then save these tweets into a database. The application ran continuously and tweets were collected during the period of 15 April 2014, 18:54 to 04 June 2014, 13:23. The final corpus contained a total of 13 946 election tweets. A more detailed discussion of how the tweets were collected is presented in Subsection 3.3.4.

3.3.3 Research Instruments

In order to answer the main research question various software components were developed, a high In order to answer the main research question various software components were developed. A high level design of these components can be seen in Figure 3.1.

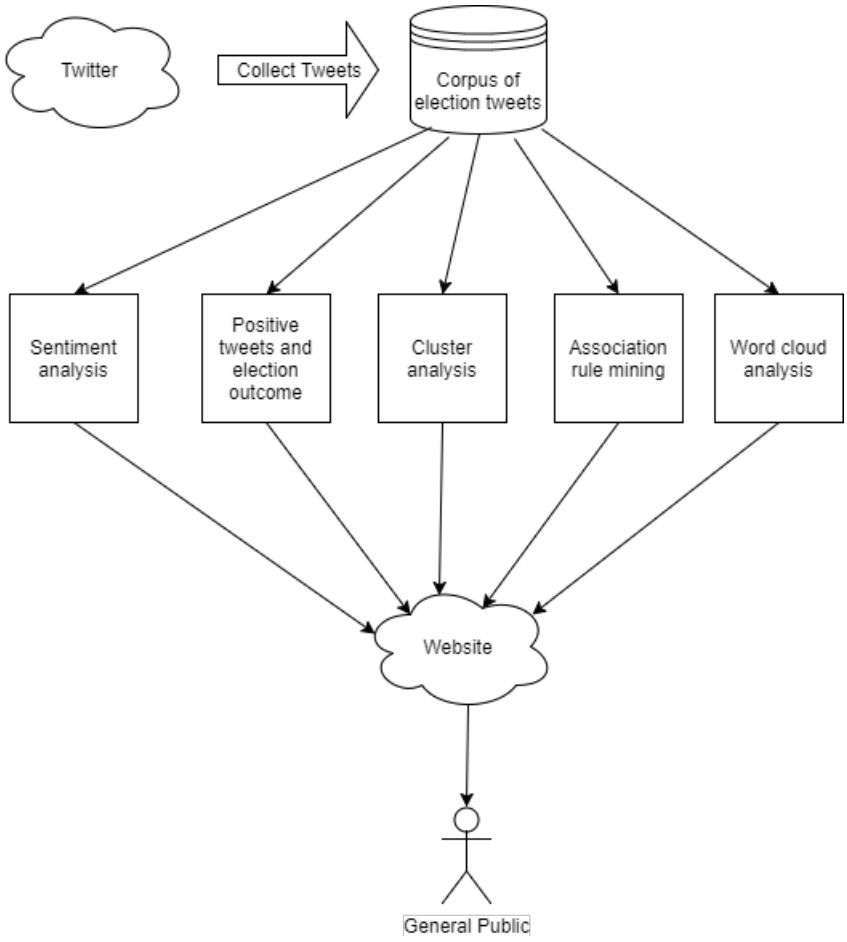

Figure 3.1: High level design of text mining application

Firstly, an application was developed to collect election tweets directly from Twitter and store them in a database. Thereafter, software was implemented to perform different tasks, such as sentiment analysis, correlational analysis, cluster analysis, association rule mining, and word cloud analysis. These components

were developed using software tools such as Python, R and WEKA. The results from this research were compiled placed on a website, which was developed using NodeJS, MongoDB, AngularJS and CSS. The website can be found at the following URL: *https://tweetminingsa.herokuapp.com*.

NodeJS is a popular open source server-side programming language that uses JavaScript to write Web servers. Due to its popularity, I used NodeJS to develop the Web server. I also used a document database, MongoDB, as a persistent data store. I chose MongoDB since it is mainly used as part of NodeJS applications.

I used HTML, Bootstrap and AngularJS to create the user interface of the application I implemented in this study. HTML is a mark-up language used to define Web pages. Bootstrap is a CSS style sheet provided by Twitter in order to style Web pages. AngualrJS provides functionality to Web pages. The functionality includes fetching and sending data between the server and Web page. These three tools used together make it possible to develop good-looking and interactive Web user interfaces. Partial screen shots of every page on the website I developed can be seen in Appendix C (see Figure 2 to Figure 7). This website can be accessed by anyone with access to the Internet.

3.3.4 CRISP-DM: Application to the Research Problem

Although the overall methodology adopted in this research was a case study, text mining was a significant part of this research and incorporated into the overall methodology. A variety of text mining tasks were discussed in Section 2.6. In data and text mining projects, a variety of machine learning algorithms are deployed to perform these tasks. These projects are conducted using different processes, such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) or SEMMA

(Sample, Explore, Modify, Model and Assess). According to a poll[1] conducted in 2014 by Dr. Gregory Piatetsky [2], CRISP-DM is the most popular process for research in data mining whilst SEMMA has experienced a decline in its adoption. I chose to apply CRISP-DM in this research for the following reasons:

- It is a robust and well-proven process
- The process is flexible
- It allows me to organise my research methodology in a logical manner
- Other researchers will be able to follow my methodology more easily since the CRISP-DM process is widely adopted
- Answering the research questions posed in Section 1.3 can be split into logical parts that correlates to the stages in the CRISP-DM process

Figure 3.2 shows the six phases of the CRISP-DM process, namely *business understanding, data understanding, data preparation, modelling, evaluation* and *deployment*. The sequence of these phases is not fixed. Often, backtracking and the repetition of previous phases are involved [Chapman *et al.* 2015]. The output of each phase determines the next phase to be executed, as illustrated by the directions of the arrows. The outer circle indicates the cyclic nature of the activities involved in data mining projects. Often, previous iterations produce new insights that benefit future iterations within the process [Chapman *et al.* 2015]. In the following sections I discuss how, within the context of this research, the first four CRISP-DM phases were implemented, including the text mining techniques that were applied on the Twitter dataset.

[1] http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[2] Dr. Gregory Piatetsky is the co-founder of ACM SIGKDD, a leading organization for knowledge discovery and data mining

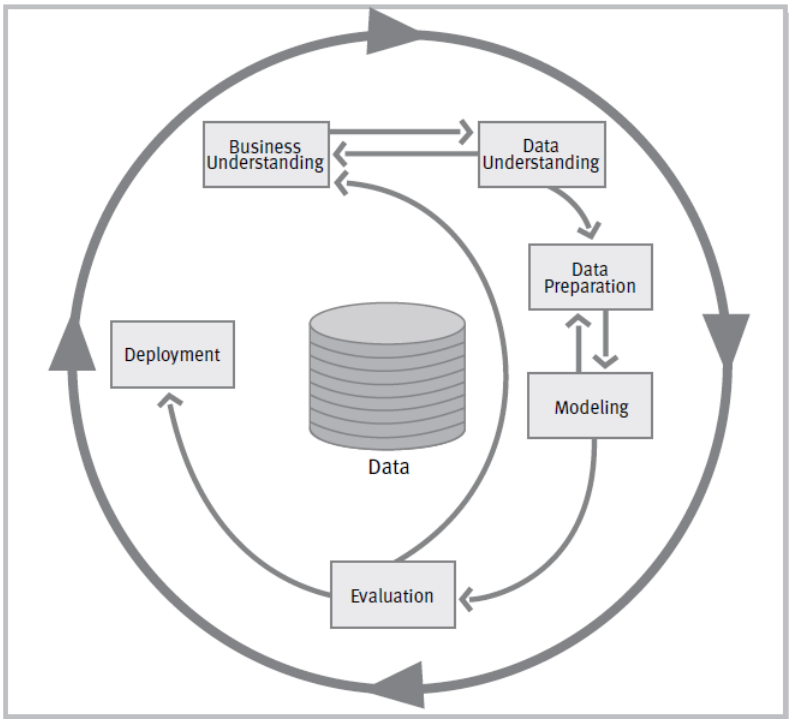

Figure 3.2: Various stages of the CRISP-DM process [Chapman *et al.* 2015]

Business Understanding

In this, the first phase of the CRISP-DM process, the primary focus is to understand clearly the project objectives and requirements, and to use this information to define the data mining problem to be addressed. This phase also looks at the development of a preliminary plan to be followed in realising the stated objectives. However, for the purpose of this project I am only concerned with understanding the project objectives **which means understanding the research objectives.**

South Africa is a multiparty democracy with two houses of parliament, namely the National Assembly and the National Council of Provinces (NCOP) [Department of Justice 1996]. There are 400 members in the National Assembly whose responsibility, amongst many others, is to pass legislation. The NCOP's mandate is to ensure that the interests of each of South Africa's nine provinces are catered for at a national level. The NCOP has 90 members or delegates. Each province is rep-

resented by 10 delegates, of which six are permanent delegates, with the remaining four being special delegates. Through its 10 representatives, who are not allowed to vote, the South African Local Government Association (SALGA) represents the interests of local government in the NCOP.

In South Africa, general elections are held every five years. The first democratic elections were held on 27 April 1994. Since then, five general elections have been held successfully in South Africa. The fifth general election was held 07 May 2014. A total of 29 political parties participated in this election, which was won by the ruling African National Congress (ANC). The ANC obtained 11 436 921 votes (62.15%) out of a total of 18 402 497 valid votes cast. The second largest party, the Democratic Alliance (DA) obtained 22.23% of the valid votes cast, followed by the recently formed Economic Freedom Fighters (EFF), which obtained 6.35% of the valid votes cast. The remaining smaller parties were left to share 9.27% of the total votes cast. The seats in the 400-member National Assembly are allocated to parties in proportion to the number of votes obtained during the elections, as follows:

- The African National Congress (ANC), the ruling party, has 249 seats.
- The Democratic Alliance (DA), the official opposition, has 89 seats.
- A recently formed party, the Economic Freedom Fighters (EFF), is the third most popular party, with 25 seats
- The remaining 37 seats are occupied by 10 parties, each having 10 or less seats.

Of all the national elections held so far in South Africa, the 2014 National Election was characterised by the robust, active and vibrant participation of different stakeholders in the social media space [Nevill 2014]. Different political parties, leaders and members of the public used social network sites such as Facebook and Twitter to interact with and engage one another in sharing their political views,

opinions and preferences, as well as participating in discussion and debate on a variety of issues related to the 2014 election. This research looks at the use of Twitter as a campaign and discussion tool by political leaders, parties and individuals during the 2014 South African national election. Thus, the primary objective of this research is to answer the following research question:

> ***Do text mining techniques uncover meaningful information when applied on a corpus of political tweets collected on the 2014 South African General Election?***

In order to answer this question five sub research questions (Q1 to Q5) were posed in Section 1.3. Furthermore, the secondary objective of this research is to develop a web-based application that will enable the general public to access the results of this research.

Data Collection and Understanding

This phase consists of the following data-related tasks: collection, description, exploration and evaluation. Generally, the data is explored using tables, graphics and other methods. In addition, the data is evaluated by determining its quality. For each task a report is produced which describes the activities performed within the task. However, for the purpose of this research only the collection and description tasks were necessary. The data for my research consisted of tweets collected during the 2014 South African national election (see Subsection 2.4.2). Twitter provides an Application Programming Interface (API) which defines a set of rules and procedures that enable users to access tweets generated by other users. Twitter also supports real time access to tweets via its Twitter Streaming API. Furthermore, this API allows one to query previous tweets, which returns tweets that satisfy a given query. In this research I used the Twitter Streaming API to collect tweets that were generated during the 2014 South African national election. Figure 3.3 shows an overview of how a corpus of South African tweets was collected.

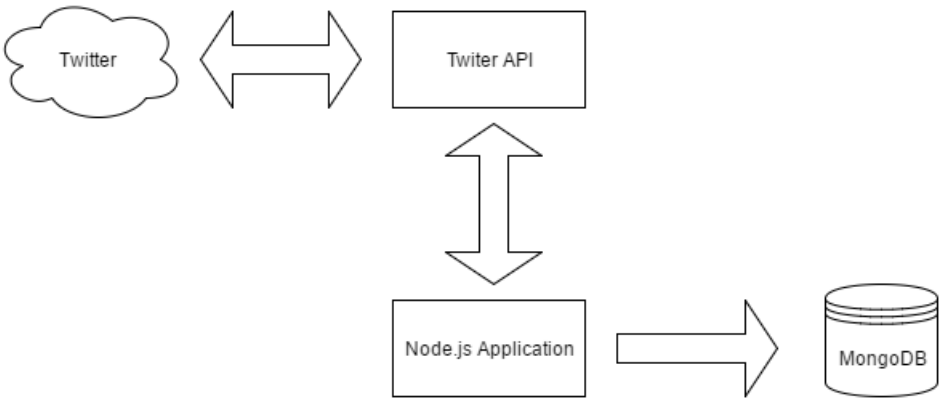

Figure 3.3: Data collection design overview

An application was developed to connect to the Twitter API, query the API for South African election tweets and then save it into a database. The application was developed using Node.js. This language was chosen due to its ease of use and existing experience using the language. In addition, a library called *Twit* was used to make connection to the Twitter API more convenient.

All tweets received from the Twitter API was stored in a MongoDB database. This database was chosen since it stores all data in JSON format which is the same format as the data received from the Twitter API, so no data conversion was needed. The Twitter API returns a vast amount of data for each tweet that satisfies the search criteria. However, in this research I selected and used the following attributes from the collected tweets:

- Tweet ID- an identifier that is unique for each tweet
- Text- the actual content of the tweet
- Time Stamp- the time when the tweet was created
- Location- the location of the user, as defined in the user's profile

A sample of the tweets collected can be seen in Table 3.1. The application ran continuously and tweets were collected during the period of 15 April 2014,

18:54 to 04 June 2014, 13:23. In this period 15 764 South African election tweets were collected. Upon closer inspection of the collected tweets, it was noticed that retweets were truncated, which means that their full contents were never recorded since the entire contents of a retweet are contained in a separate field that was never saved. However, due to the fact that some of the tweets about which I required more information had already been deleted (from Twitter's side), it was not possible to rerun my queries on the Twitter API to obtain this information. Given this situation, if a retweet was truncated, the original tweet was then located within the corpus of collected tweets and the truncated tweet replaced with the original. If no replacement was found, then the tweet was removed from the corpus. This resulted in 2 144 tweets being replaced and 1 818 tweets being removed. The final corpus of election tweets contained 13 946 tweets.

Table 3.1: Tweets collect using the Twitter streaming API

Tweet ID	**Text**	**Time Stamp**	**Location**
456143807403139070	Sometimes very nice and pretty phrases are used by the ANC to justify this diversion from the democratic road towards corruption.	2014-04-15T18:55:33	Cape Town
456144296156991000	RT @POWER987News: #NoVote Founders say there are many ANC struggle veterans who are part of the campaign and that they haven't left the par...	2014-04-15T18:57:29	Johannesburg
474169679033143000	Economic Freedom Fighters (EFF) has condemned the evictions of the people of Lwandle... http://t.co/iOAnxqCqdc	2014-06-04T12:43:55	Limpopo

Data Preparation

This phase includes all the tasks required to transform the raw data into the final dataset that will be used during the modelling phase. Usually, these tasks include data selection, cleaning, construction, integration and formatting. Some of the tasks are executed several times, with each iteration bringing the task closer to completion. For the purpose of this research, only the cleaning and formatting tasks were required. Furthermore, only one iteration of these tasks was necessary.

Due to the 140 character limit on a tweet, users must convey messages as briefly as possible. Therefore, a novel syntax very similar to the Short Message Service (SMS) syntax is used in tweets. This syntax is not suitable for use in text mining. Hence, a preprocessing step is required to render the data into a more suitable format. In this research I used the following preprocessing steps, based on the approach proposed in Agarwal *et al.* [2011]:

1. convert to lower case
2. remove URLs
3. remove re-tweet tags
4. remove punctuation
5. remove digits
6. remove stop words
7. lemmatize each tweet

Converting text to lower case is a trivial step in preprocessing, it makes string matching more convenient. All URLs have been removed as it provides little or no syntactic value, unless you inspect the content the URL is referring to. Re-tweet

tags, digits and stop words were also removed, given their negligible syntactic value. Punctuations in the form of emoticons can provide an indication of sentiment [Agarwal *et al.* 2011]. However, Agarwal *et al.* [2011] showed that emoticons can reduce the accuracy of some data mining techniques, especially Support Vector Machines (SVMs). In order to avoid negatively impacting the accuracy of any techniques used, all punctuation marks were removed. Table 3.2 shows a sample of the tweets before (first column) and after (second column) they were processed.

Table 3.2: Example of preprocessed tweets

Raw Tweet	**Preprocessed Tweet**
Sometimes very nice and pretty phrases are used by the ANC to justify this diversion from the democratic road towards corruption.	sometimes nice pretty phrases used anc justify diversion democratic road towards corruption
RT @POWER987News: #NoVote Founders say there are many ANC struggle veterans who are part of the campaign and that they haven't left the par...	novote founders say many anc struggle veterans part campaign havent left par
Economic Freedom Fighters (EFF) has condemned the evictions of the people of Lwandle... http://t.co/iOAnxqCqdc	economic freedom fighters eff condemned evictions people lwandle

Modelling

This phase involves several tasks such as selecting a modelling method, creating a model by applying the modelling technique on a training dataset, and assessing the resulting model. In this research different modelling techniques were applied on the dataset I collected during the 2014 South African general election period. The techniques included sentiment analysis, prediction of election results, tweet clustering, association rule mining and generating a word cloud to visualise predominant topics, terms and words contained in the election tweets.

Sentiment Analysis

Sentiment analysis is a classification technique that categorises text messages as conveying either a positive or a negative sentiment [Bifet and Frank 2010]. An unsupervised model for sentiment analysis would allow the model to be applied to various tweet corpora. In addition, there would be no need for a training set and no need for the tedious process of manually classifying the sentiment of each tweet in the training set. On the other hand, a supervised model would be specifically trained for a specific corpus of tweets and applying the model to a different corpora would require a labelled training set. Therefore, an unsupervised approach to sentiment analysis was taken.

Hutto and Gilbert [2014] have developed a parsimonious rule-based sentiment classifier called *VADER*(**V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning), suitable for use in microblogging services such as Twitter. In particular, Hutto and Gilbert [2014] used the VADER sentiment classifier to determine the sentiment contained in tweets. The sentiment of each tweet can be classified as either positive, negative or neutral. The VADER model returns a 3-tuple containing the probability of the sentiment being negative, positive or neutral respectively. The VADER classifier achieved an $F1$ score of 0.96, an indication of its superior accuracy in classifying tweets based on their sentiment. Given its reported effectiveness in classifying tweets, I used the VADER classifier to determine sentiment in the corpus of political tweets that were collected during the 2014 general election.

Figure 3.4: Sentiment analysis methodology

An outline of the methodology followed for classifying the sentiment of election tweets can be seen in Figure 3.4. First, the corpus of election tweets was

preprocessed using the methodology presented in Section 3.3.4. Thereafter, the VADER sentiment classifier was used to categorise each preprocessed election tweet according to its sentiment. At this stage, the sentiment of each tweet was either positive, negative or neutral. The sentiment with the highest probability value was assigned to a tweet as its overall sentiment.

Positive tweets and election outcome

Twitter has become an essential communication platform in the political arena [Tumasjan *et al.* 2010]. A number of researchers are using Twitter to obtain meaningful information and to gain some insight into how Twitter users think, discuss and share ideas and opinions on political issues. Twitter is now also being used to predict the outcome of elections. For example, thousands of political tweets posted during the 2009 elections in Germany were used to predict the outcome of that election [Tumasjan *et al.* 2010]. Each tweet can be treated as a vote. A naive interpretation would be that if a tweet mentions a particular political party, then that tweet would count as a vote in favour of that party. However, this approach does not consider the sentiment contained in the tweet. If a tweet contains a positive sentiment towards a political party then it can be considered as a vote for that party. On the other hand, a tweet with a negative sentiment would not count as a vote in favour of that party. In this research, I investigated if there was a relationship between the total number of positive tweets that mention a political party and the total number of votes received by that party during the 2014 South African general election. The methodology used to accomplish this is illustrated in 3.5.

Firstly, I preprocessed the corpus of election tweets as outlined in Section 3.3.4. Thereafter, I determined the sentiment of the preprocessed tweets, using the approach described in Section 3.3.4. Having identified the sentiment of each tweet, I proceeded to remove all the tweets with a negative or neutral sentiment, since only tweets with a positive sentiment were being considered. Finally, each tweet was

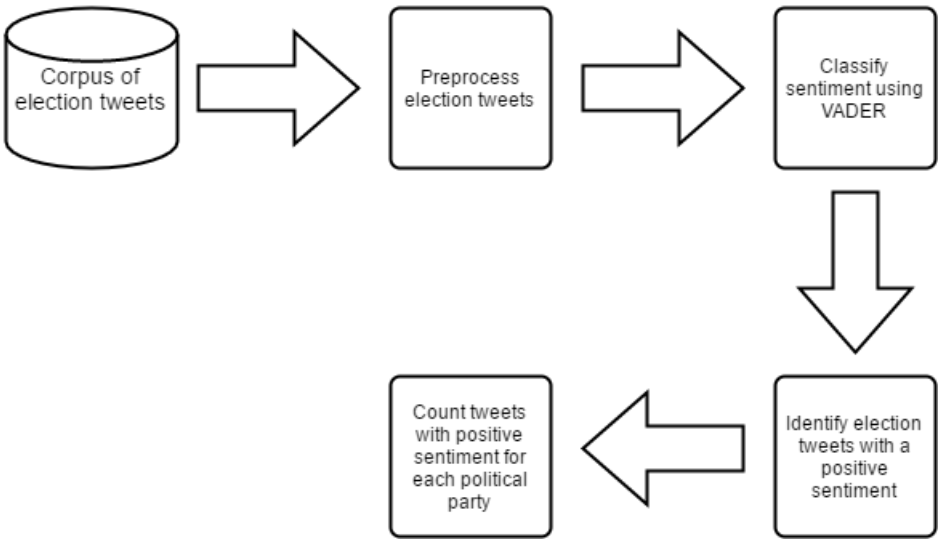

Figure 3.5: Predicting election results methodology

checked to determine whether it referenced a specific political party. That is, does the content of the tweet contain the name (full or abbreviated) of a political party? This research did not consider tweets that referenced multiple political parties.

Furthermore, I used the location data for each tweet to determine the location of the user. There are 9 provinces in South Africa: Eastern Cape, Free State, Gauteng, Kwa-Zulu Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape. If the location data of a tweet contained any of these provinces then that tweet belonged to that province. Tweets with no provinces or multiple provinces in the location data were ignored. This way I was able to determine the number of positive tweets each party received per province in South Africa.

Cluster Analysis

Clustering is the procedure by which objects are classified into groups called clusters [Feldman and Sanger 2007]. Hence, tweet clustering is the procedure by which tweets are classified into clusters, where tweets in the same cluster are considered *"similar"*. Previous research aimed to cluster tweets based on the topic of the

tweet. This research focused on clustering tweets based on the political party the tweet mentions.

Rosa *et al.* [2011] compared the performance of unsupervised and supervised clustering methods. It was shown that the supervised methods achieved a better performance than the unsupervised methods. However, supervised clustering requires a labelled training set. Rosa *et al.* [2011] used an automatic labelling approach, where the hashtag contained in the tweet was used as a topic label for the tweet. A similar approach was taken, in this research, to automatically identify the political party mentioned in a tweet. Table 3.3 show a sample of preprocessed tweets and their corresponding political label.

Table 3.3: Example of preprocessed tweets labelled and their corresponding political label

Preprocessed Tweet	**Political Label**
ronniekasrils actually encouraging ppl vote anc dathat profoundhe indirectly encouraging the	anc
earlier eff malema issue land must careful one day wake find sol kerzner bought	eff
da leadership maimaneam anthonybenadie lindimazibuko celebrating freedomday bushbuckridge mpumalanga sabcnews	da
provincial results limpopo anc da eff electionresults results change con	mto
tom udall voted increase taxes spending people congress	uk

If a tweet contains "anc" or "african national congress" then that tweet is labelled with "anc". Similarly, if a tweet contains "da" or "democratic alliance" then it is labelled with "da". Finally, if a tweet contains "eff" or "economic freedom fighters" then the it is labelled with "eff". In addition, if a tweet mentions more than one political party it is labelled with "mto" (more than one) otherwise it is labelled "uk" (unknown).

An overview of the methodology followed is presented in Figure 3.6. The cor-

pus of election tweets is preprocessed as described in Section 3.3.4. Thereafter, the political label for each tweet is automatically assigned as described above. Weka is used to apply the K-Means algorithm to the labelled set of preprocessed tweets. In addition, the number of clusters (k) was set to 5 since there are five possible political labels (anc, da, eff, mto and uk). WEKA is an acronym for Waikato Environment for Knowledge Analysis, an open source data mining toolkit that was developed at the University of Waikato, New Zealand [Garner 1995]. Weka was chosen based on its built in functionality to visualise clusters and ease of use. The K-Means clustering algorithm was chosen since it is a popular clustering algorithm and fairly simple to understand. After the clustering process, the label for each cluster needs to be identified. Each tweet has been assigned a political label, therefore, the most prevalent political label in a cluster was taken as the political label for that cluster.

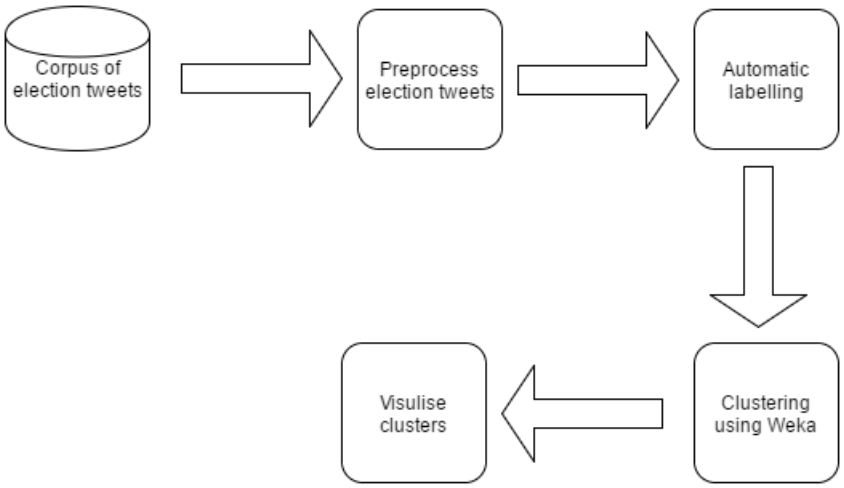

Figure 3.6: Tweet clustering methodology

Association Rule Mining

Association rule mining is the process of discovering rules that depict the relationship among the attributes of a dataset [Zhao 2012]. An association rule uses the form:

$$A \Rightarrow B$$

where A and B are disjoint itemsets, both of which are proper subsets that belong to a set of items, I. The set A is the *antecedent* and the set B the *consequent* of a rule. Two measures, namely *support* and *confidence*, are used in association with each rule. The support measure for a rule such as

$$A \Rightarrow B$$

represents the proportion of transactions in a database D that contain both A and B [Larose 2014]. The confidence of a rule indicates its accuracy, in terms of the proportion of transactions containing A as well as B. Figure 3.7 illustrates the approach I used to discover association rules using the tweets obtained during 2014 South African general elections.

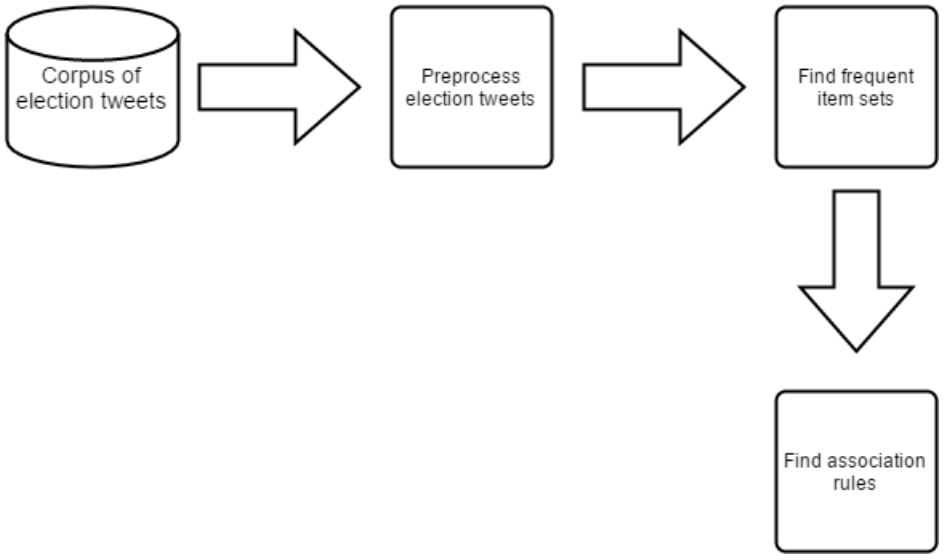

Figure 3.7: Association rule mining methodology

Firstly, the corpus of election tweets was preprocessed as described in Section 3.3.4. Using the *apriori* association rule mining algorithm, the next step was to generate frequent itemsets from the collected political tweet dataset. A frequent itemset is an itemset that occurs at or above a certain minimum support threshold. In this research frequent itemsets were obtained by generating all possible com-

binations of terms contained in the preprocessed election tweets. The search for frequent itemsets was conducted by ignoring itemsets (and their related subsets) that did not satisfy the minimum support threshold of 0.01.

Having identified all the frequent itemsets from my dataset, association rules were generated from them, by selecting only those rules that satisfied the specified support and confidence thresholds. It is possible to generate a large number of association rules, some of which may not be relevant. In this research the rules that did not meet the minimum confidence threshold were considered irrelevant and therefore ignored. The confidence threshold was set at 0.7. The support value was set at 0.01 to ensure all rules are returned regardless of how infrequent it appears in the dataset. The confidence was set at 0.07 to ensure that the rules that are identified are significant. I used an R library, *arules*, to generate association rules from the corpus of preprocessed 2014 election tweets.

Word cloud analysis of election tweets

Text-based data can be represented graphically using a *word cloud*. Frequently used words will appear more prominently in the resulting word cloud. In this research word clouds were used to visualise, analyse and interpret the content of the political tweets that were collected during the 2014 South African general election. In addition, the word cloud was generated to visualize the sentiment contained in the tweets. The methodology used is illustrated in Figure 3.8.

Firstly, the corpus of election tweets is preprocessed using the methodology presented in Subsection 3.3.4. Thereafter, the font size of each word is calculated. Previous research such as Zhao [2012] and Sankaranarayanany *et al.* [2009] visualized the importance of a word with the size of the word. That is, the more important the word is the larger it will appear in the word cloud. The importance of a word is directly proportional to the number of times the word appears in the corpus. The font size of a word, is given by formula 3.1.

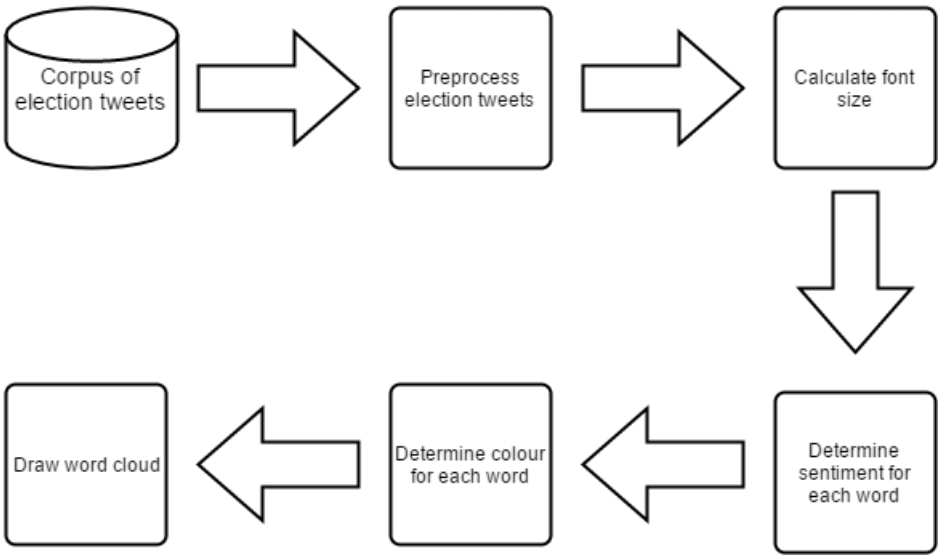

Figure 3.8: Overview of methodology used to generate a word cloud

$$font\ size = maxFontSize * \frac{frequency}{maxFrequency} \tag{3.1}$$

Here, $maxFontSize$ is the maximum font size a word can have, this was set to $853px$ which was the height of the word cloud. For each word, $frequency$ is the number of times that word appears in the corpus and $maxFrequency$ is the highest frequency in that corpus, used to normalise the $frequency$ of each word to ensure the font-size of each word is invariant of the size of the corpus.

In order to visualise the sentiment of each word, firstly the sentiment of the word needs to be determined then a colour needs to be associated to that sentiment. In order to determine the sentiment of a word, the average sentiment of each tweet the word appears in must be determined. Subsection 3.3.4 described the methodology used to determine the sentiment of a tweet. The sentiment of each tweet, t, was represented as a 3-tuple (neg, pos, neu) where neg, pos and neu is the probability of t having a negative, positive or neutral sentiment respectively.

Let T_w be the set of tweets that contain the word w where $|T_w| = k$. Then the sentiment for each tweet $t \in T_w$ is represented as a 3-tuple. The average sentiment of the tweets that contain the word w will be used as an indication of the sentiment of that word. Therefore, the sentiment of a word w is given by:

$$S_{word}(w) = (\frac{\sum_{i=1}^{k} neg_i}{k}, \frac{\sum_{i=1}^{k} pos_i}{k}, \frac{\sum_{i=1}^{k} neu_i}{k}) \tag{3.2}$$

Where neg_i, pos_i and neu_i is the probability of tweet $t_i \in T_w$ having a negative, positive or neutral sentiment respectively. It is clear to see that $S_{word}(w)$ results in a 3-tuple where each number indicates the average probability of the word w having a negative, positive or neutral sentiment. The reason for using this approach is that if a word is often found in tweets with a positive sentiment, then the word itself is likely to have a positive sentiment. This applies similarly to words found in tweets with a negative or neutral sentiment. These 3-tuples for each word can be used as RGB values to determine the colour of each word. Naturally, words that have a high probability of having a positive sentiment will contain more green. Whereas, words that are more likely to have a negative or neutral sentiment will be more red or blue respectively.

The final step involves drawing the word cloud. This was done using a Python library called *wordcloud*. This library was chosen because it calculates the font size of each word as describe by Equation 3.1. In addition, a custom colouring functions can be specified to determine the colour of each word. This allowed Equation 3.2 to be used to generate RGB values for each word. Moreover, the *wordcloud* library can overlay the generated word cloud over any image. Thus, an image containing the map of South Africa was used to overlay the word cloud over. This was done purely for aesthetics.

3.4 Data Metrics

In the modelling phase of the CRISP-DM process, techniques such as sentiment analysis, predicting election results, tweet clustering, association rule mining and word clouds were discussed. This section presents the metrics used in order to evaluate these techniques.

Selecting a Test Set

In order to test the performance/ accuracy of the various methods described in Section 3.3.4 a test set is required. The corpus used in this research contained 16 754 South African election tweets. From this corpus 1 000 distinct original tweets were randomly selected as the test set. In order to avoid the duplication in the test set which may negatively impact the evaluation results, retweets were ignored. This test was then used in the various evaluation methods described the in subsequent subsections.

Sentiment Analysis

In order to test the performance of the VADER model on the corpus of South African election tweets, a labelled test set is required. The sentiment of each election tweet in the test set described in Subsection 3.4 was manually classified. Thereafter, the manually classified sentiment and the sentiment identified by the VADER model can be compared. In order to determine the effectiveness of the VADER model, three measures were used: accuracy, precision, recall and the $F1$ score. To help define these measures a confusion matrix will be used.

Consider the confusion matrix presented in Table 3.4. The columns represent the number of tweets with the corresponding sentiment, as determined by the VADER model. For example in the "Positive" column, `TPOS` is the number of

tweets the VADER model correctly classified as positive (true positive). $\texttt{FPOS}_{neg}$ is the number of tweets the VADER model classified as positive but was actually negative. $\texttt{FPOS}_{neu}$ is the number of tweets the VADER model classified as positive but was actually neutral. Therefore, $\texttt{FPOS}_{neg} + \texttt{FPOS}_{neu}$ is the total number of tweets that were incorrectly classified as positive by the VADER model.

Table 3.4: F1 score confusion matrix

		Predicted Sentiment		
		Positive	Negative	Neutral
Actual Sentiment	Positive	TPOS	$\texttt{FNEG}_{pos}$	$\texttt{FNEU}_{pos}$
	Negative	$\texttt{FPOS}_{neg}$	TNEG	$\texttt{FNEU}_{neg}$
	Neutral	$\texttt{FPOS}_{neu}$	$\texttt{FNEG}_{neu}$	TNEU

In the "Negative" column, $\texttt{FNEG}_{pos}$ is the total number of tweets classified as negative by the VADER model was actually positive. TNEG is the total number of tweets classified as negative by the VADER model that was actually negative (true negative). $\texttt{FNEG}_{neu}$ is the total number of tweets classified as negative by the VADER model but was actually neutral. Therefore, $\texttt{FNEG}_{pos} + \texttt{FNEG}_{neu}$ is the total number of tweets incorrectly classified as negative by the VADER model.

Similarly in the "Neutral" column, $\texttt{FNEU}_{pos}$ is the total number of tweets classified as neutral by the VADER model but was actually positive. $\texttt{FNEU}_{neg}$ is the total number of tweets classified as neutral by the VADER model but was actually negative. TNEU is the total number of tweets classified as neutral by the VADER model that was actually neutral (true neutral). Therefore, $\texttt{FNEU}_{pos} + \texttt{FNEU}_{neg}$ is the total number of tweets incorrectly classified as neutral by the VADER model.

The rows represent the actual number of tweets with the corresponding sentiment, as determined by a human annotator. For example, the actual number of positive tweets would be $\texttt{TPOS} + \texttt{FNEG}_{pos} + \texttt{FNEU}_{pos}$. The actual number of negative tweets is $\texttt{FPOS}_{neg} + \texttt{TNEG} + \texttt{FNEU}_{neg}$. The actual number of neutral tweets is $\texttt{FPOS}_{neu} + \texttt{FNEG}_{neu} + \texttt{TNEU}$. With the aid of the confusion matrix in

Table 3.4, I will now present the definitions for precision, recall and the $F1$ score.

The total number of positive classifications ($\texttt{TC}_{pos}$) can be defined as follows:

$$\texttt{TC}_{pos} = \texttt{TPOS} + \texttt{FPOS}_{neg} + \texttt{FPOS}_{neu} \tag{3.3}$$

Similarly, the total number of negative classifications ($\texttt{TC}_{neg}$) and the total number of neutral classifications $\texttt{TC}_{neu}$ can be defined as:

$$\texttt{TC}_{neg} = \texttt{FNEG}_{pos} + \texttt{TNEG} + \texttt{FNEG}_{neu} \tag{3.4}$$

$$\texttt{TC}_{neu} = \texttt{FNEU}_{pos} + \texttt{FNEU}_{neg} + \texttt{TNEU} \tag{3.5}$$

Now $accuracy$ can be defined as the ratio of the total number of correct classifications to the total number of classifications. The $accuracy$ can rage between 0 and 1, the hight the accuracy the better the classifier is. The $accuracy$ can be calculated as follows:

$$accuracy = \frac{\texttt{TPOS} + \texttt{TNEG} + \texttt{TNEU}}{\texttt{TC}_{pos} + \texttt{TC}_{neg} + \texttt{TC}_{neu}} \tag{3.6}$$

Precision is defined as the ratio of the number of correctly classified tweets to the total number of tweets predicted to be in that class. Therefore, the precision for positive, negative and neutral sentiment are calculated as follows:

$$precision_{pos} = \frac{\texttt{TPOS}}{\texttt{TPOS} + \texttt{FPOS}_{neg} + \texttt{FPOS}_{neu}} \tag{3.7}$$

$$precision_{neg} = \frac{\texttt{TNEG}}{\texttt{FNEG}_{pos} + \texttt{TNEG} + \texttt{FNEG}_{neu}} \tag{3.8}$$

$$precision_{neu} = \frac{\texttt{TNEU}}{\texttt{FNEU}_{pos} + \texttt{FNEU}_{neg} + \texttt{TNEU}} \tag{3.9}$$

Now the average precision can be calculated as the average of the precision for positive, negative and neutral sentiments:

$$precision_{avg} = \frac{precision_{pos} + precision_{neg} + precision_{neu}}{3} \tag{3.10}$$

Recall is defined as the ratio of the number of correctly classified tweets to the total number of tweets that are actually in that class. Therefore, the recall for positive, negative ans neutral sentiment are calculated as follows:

$$recall_{pos} = \frac{\texttt{TPOS}}{\texttt{TPOS} + \texttt{FNEG}_{pos} + \texttt{FNEU}_{pos}} \tag{3.11}$$

$$recall_{neg} = \frac{\texttt{TNEG}}{\texttt{FPOS}_{neg} + \texttt{TNEG} + \texttt{FNEU}_{neg}} \tag{3.12}$$

$$recall_{neu} = \frac{\texttt{TNEU}}{\texttt{FPOS}_{neu} + \texttt{FNEG}_{neu} + \texttt{TNEU}} \tag{3.13}$$

Now the average recall can be calculated as the average of the recall for posi-

tive, negative and neutral sentiments:

$$recall_{avg} = \frac{recall_{pos} + recall_{neg} + recall_{neu}}{3} \tag{3.14}$$

Using the definition for precision and recall I can now define the $F1$ score which is the overall effectiveness of the classifier [Hutto and Gilbert 2014]. The $F1$ score is given by:

$$F1 = 2 \times \frac{recall_{avg} \times precision_{avg}}{recall_{avg} + precision_{avg}} \tag{3.15}$$

The $F1$ score ranges between 0 and 1, the higher the score the more effective the classifier.

Looking more closely at the 1 000 tweets in the test set, it was found that there were 389 negative, 200 positive and 411 neutral tweets. This can be seen in Table 3.5. In addition, Table 3.5 shows the $accuracy$ and $F1$ score that would be achieved if the VADER model labelled all tweets as either negative, positive or neutral respectively.

The highest accuracy of 0.411 is achieved by labelling all tweets as neutral. The second highest accuracy of 0.389 is achieved by labelling all tweets as negative. The lowest accuracy of 0.2 is achieved by labelling all tweets as positive. The difference in these accuracies is due to the fact that there are more neutral tweets than negative or positive tweets. Similarly, there are more negative tweets than positive tweets. For the purpose of my research an acceptable accuracy would be at least 0.6 or greater. However, given the uneven distribution of tweets between the positive, negative and neutral sentiments, this accuracy measure is not useful. This is why an $F1$ score is preferred in these situations.

The baseline $F1$ scores, shown in Table 3.5 are all low and shows that simply guessing the sentiment of tweet will not suffice. As described in Subsection 3.3.4,

I applied the VADER classifier to classify the sentiment contained in tweets. If the VADER model can achieve an $F1$ score significantly greater than the highest baseline of 0.1945, then it can be concluded that the VADER model classified the sentiment of 2014 South African election tweets at an acceptable level.

Table 3.5: Baseline F1 scores

	$accuracy$	$recall_{avg}$	$precision_{avg}$	$F1$
Neutral Baseline	0.411	0.3333	0.137	0.1945
Positive Baseline	0.2	0.3333	0.0667	0.111
Negative Baseline	0.389	0.3333	0.129	0.186

Positive tweets and election outcome

The methodology presented in Subsection 3.3.4 returned the count of tweets, with a positive sentiment, each political party received. The Pearson correlation coefficient (*r*), was calculated to determine the relationship between the number of positive tweets that mentions a political party and the number of votes that party receives. The value for r ranges between -1 and 1 ($-1 \leq r \leq 1$). An r value of 1 indicates a strong positive relationship, this would mean that if the number of positive tweets that mention a political party increases/decreases then so does the number of votes that party receives increases/decreases. On the other hand, an r value of -1 indicates a strong negative relationship which indicates that if the number of positive tweets that mention a political party increases/decreases then conversely the number of votes that part receives decreases/increases. An r value of 0 indicates that there is no correlation between the number of positive tweets that mentions a political party and the number of votes that party receives. The significance of the relationship between the number of positive tweets and votes can be determined by the *P-value*. A statistically significant relationship is indicated when *P-value* ≤ 0.05. A statistically insignificant relationship is indicated when

P-value ≥ 0.05.

Cluster Analysis

The corpus of South African election tweets was clustered around political parties as described in Subsection 3.3.4. Each tweet in the corpus of election tweets was automatically assigned a political label. Thereafter, the political label of each tweet was compared to the political label of the cluster. This way it was possible to determine the number of correctly clustered tweets. In addition, each cluster was visualised in a 2-dimensional space. This indicated how tightly clustered the tweets were.

Association Rule Mining

Subsection 3.3.4 presented the methodology used to discover association rules from a corpus of South African election tweets. The support and confidence thresholds were set at 0.01 and 0.7 respectively, to ensure that only significant association rules were discovered. However, there is no statistical measure to test if the association rules are meaningful. Therefore, the discovered association rules were manually interpreted to determine its meaningfulness.

Word cloud analysis of election tweets

The word cloud generated by following the methodology in Subsection 3.3.4 is a tool to visualise the importance and sentiment of words from the corpus of South African election tweets. There are no statistical measures to test the validity or meaningfulness of a word cloud. Hence the validity and meaningfulness of the word cloud was judged manually

3.5 Conclusion

Twitter has become an important platform where people express their opinions and views on political parties and candidates. As such, extensive research has been conducted that utilise tweets to gain insights into political domains in countries such as Germany, the United States and Italy. However, to the best of my knowledge, little to no research has been conducted in the context of South African politics. Thus the primary objective of this research was to investigate if text mining techniques can uncover meaningful information when applied on a corpus of political tweets collected on the 2014 South African General Election.

A case study methodology was chosen to conduct this research. The case being the 2014 South African general elections held on 07 May, 2014. Considering the various text mining techniques that was used in this research, various phases from the CRISP-DM process was adopted. The CRISP-DM process consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Only the business understanding, data understanding, data preparation and modelling phases were applied in this research.

During the business understanding phase, an in depth discussion on South African politics was provided which showed that the 3 major political parties in South Africa is the ANC, DA and EFF. Therefore, these were the only political parties that was considered in this research.

In order to collect a corpus of election tweets, an application was developed to connect to the Twitter API and query the API for tweets that mention the ANC, DA or EFF. The application ran continuously and tweets were collected during the period of 15 April 2014, 18:54 to 04 June 2014, 13:23. This resulted in corpus of election tweets which contained 13 946 tweets.

Thereafter, the corpus of election tweets was then preprocessed by converting all tweets to lower case and then removing all URLs, re-tweet tags, punctuation,

digits and stop words. In addition, each tweet was lemmatized.

In total, five text mining techniques were applied, namely: sentiment analysis, predicting election results, tweet clustering, association rule mining, and word clouds. The VADER model was used to classify the sentiment of an election tweet as either positive, negative or neutral. The number of positive tweets that mention a political party was determined and correlated to the number of votes that party receives. The K-Means clustering algorithm was used to cluster tweets around the political party the tweet mentions. Association rules were discovered by applying the apriori algorithm with support and confidence measures set at 0.01 and 0.7 respectively. In addition, word clouds were used to visualise important topics within the corpus of election tweets as well as the sentiment of the general public.

In order to test the quality of the various text mining techniques mentioned above, a test set of 1 000 tweets were randomly selected as a test set. The sentiment of each tweet was manually classified and compared to the results from the sentiment analysis and an $F1$ score was calculated using Equation 3.15. The Pearson correlation coefficient(r) was used to calculate the strength of the relationship between the number of positive tweets that mentions a political party and the number of votes that party receives. Evaluating the results of the tweet clustering was done by automatically assigning a political label to each tweet in the test set and comparing it to the political label it received after clustering. The meaningfulness of the discovered association rules and the generated word cloud had to be determined manually as there are no statistical measures that could be used.

Finally, the results obtained from the conducted research was made available to the general public through a website which can be found at this URL: *https://tweetminingsa.herokuapp.com*. The website was developed using NodeJS for the server, a MongoDB database and AngularJS and CSS was used for the front end.

In Chapter 4, the results obtained from this study is presented, to provide an-

swer to the research questions that were posed in Chapter 1.

Chapter 4

Results

4.1 Introduction

Chapter 3 provided a detailed discussion of the adopted research methodology, which is a case study. Furthermore, I stated and discussed some of the reasons why it was important to conduct this interesting research. This chapter presents the results obtained from this study, to provide answers to the research questions that were posed in Chapter 1. There are seven sections in this chapter. Each section focusses on a specific research question and presents the results obtained in relation to that question. Section 4.2 focuses on the sentiment analysis of the corpus of tweets collected during the 2014 election. Section 4.3 presents the findings on the relationship between the number of votes obtained by a political party and the number of positive tweets posted about it.

Meaningful and useful information can also be obtained by clustering tweets and analysing the resulting output. Section 4.4 presents the results obtained from the application of clustering techniques on the corpus of the 2014 general election tweets. Section 4.5 presents and discusses the association rules extracted from these election tweets, using the *apriori* association rule mining algorithm. Word

clouds are generally used to visualise the words and terms contained in a piece of text. Section 4.6 discusses how word clouds were used to analyse the 2014 election tweets and to obtain an understanding of how frequently certain words or terms were used during the election. The chapter end with a conclusion in Section 4.7.

4.2 Sentiment Analysis

Social network sites such as Twitter have become an important platform where people express their opinions and views on political parties and candidates. As such, Twitter has often been studied in the context of politics [Wang *et al.* 2012; Monti *et al.* 2013]. Extensive research has been done that utilise tweets to gauge the sentiment of the public towards political parties in countries such as the United States, Ireland, Sweden and Germany, to mention a few [Wang *et al.* 2012; Bakliwal *et al.* 2013; Larsson and Moe 2012; Tumasjan *et al.* 2010]. As far as is know, not much research has been or is currently being conducted in South Africa to mine Twitter data to identify the public's sentiment towards South African political parties or politicians. Thus, the following research question was posed:

> *What sentiment is portrayed by the election tweets towards some of the parties that took part in the general election?*

In order to answer this research question, I applied the VADER (*Valence Aware Dictionary and sEntiment Reasoner*) sentiment classifier on the preprocessed corpus of election tweets. The VADER classifier assigned a positive, negative or neutral sentiment to each tweet. Table 4.1 shows the number of positive, negative and neutral election tweets associated with each of the three major parties, namely the ANC, the DA and the EFF.

Table 4.1: Number of tweets per sentiment for each political party

	ANC	DA	EFF	Several parties	Unknown party
Positive	559	31	43	34	56
Negative	278	4	21	34	12
Neutral	8 293	499	733	2 034	1 315
Total	**9 130**	**534**	**797**	**2 102**	**1 383**

The *"Several parties"* column refers to tweets that mentioned more than one party and the *"Unknown party"* column refers to tweets whose contents did not mention any of the three major parties. The contents of Table 4.1 are presented graphically in Figure 4.1. It's clear that there is an overwhelming neutral sentiment towards the ANC, DA and EFF.

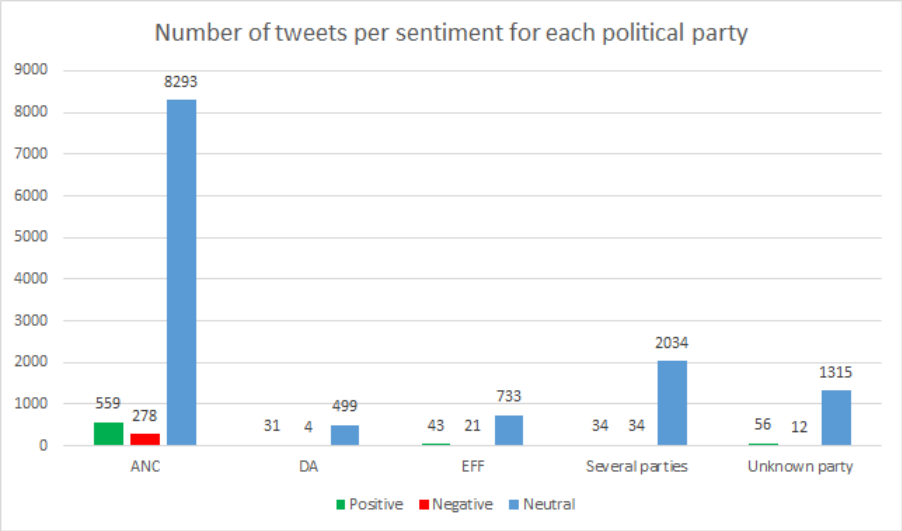

Figure 4.1: Number of tweets per sentiment for each political party

Table 4.2 shows the percentage of the total number of tweets associated with each party that are positive, negative and neutral. The percentage of positive tweets associated with the ANC was slightly higher compared to the DA and the EFF respectively. This higher positive sentiment does not by any means indicate an overwhelming endorsement of the ANC. It can also be seen that the DA was better perceived compared to the EFF, although only slightly so. In terms of negative

sentiment, the ANC was viewed most negatively across the board. The DA was the least negatively perceived party, far better than either the EFF or the ANC. Over 90% of the tweets associated with each party were classified as neutral, an indication perhaps of the noncommittal nature of the tweeting electorate on topics related to these parties during the 2014 elections.

Table 4.2: Percentage of positive, negative and neutral tweets

	ANC	**DA**	**EFF**	**Several parties**	**Unknown party**
Positive (%)	6.12	5.81	5.40	1.62	4.05
Negative (%)	3.04	0.75	2.63	1.62	0.87
Neutral (%)	90.83	93.45	91.97	96.76	95.08

As described in Subsection 3.3.4, an unsupervised sentiment analysis approach was taken to classify the corpus of tweets collected during the 2014 election period. I applied the VADER classifier on the whole corpus, which I did not divide into training and test datasets. This classifier was selected based on its ability to classify the sentiment contained in tweets. In a study conducted by Hutto and Gilbert [2014], the VADER model was shown to be superior compared to human raters. It was more accurate in classifying the sentiment of tweets than human raters, achieving an $F1$ score of 0.96 compared to the $F1$ score of 0.84 obtained by human raters. In this research the accuracy and $F1$ score was calculated to determine the VADER classifier's performance on my dataset. To determine the accuracy and $F1$ score of the VADER classifier using my dataset, a set consisting of 1 000 tweets was randomly selected and labelled manually, using the approach described in Subsection 3.4. In the test set there where 200 positive, 389 negative and 411 neutral tweets respectively. The confusion matrix presented in Table 4.3 shows the sentiment analysis results. From the set of 200 positive tweets, the VADER classifier correctly labelled 26. The VADER classifier correctly labelled 26 negative tweets out of a possible 389 negative tweets. Out of a possible 411 neutral tweets, the VADER classifier correctly labelled 398.

Table 4.3: Confusion matrix of the sentiment analysis results

		Predicted Sentiment			
		Positive	Negative	Neutral	Total
Actual Sentiment	Positive	26	3	171	200
	Negative	14	26	349	389
	Neutral	10	3	398	411
	Total	50	32	918	1000

Using the confusion matrix given in Table 4.3, the accuracy of the VADER classifier can be calculated using Equation 3.6 as follows:

$$\begin{aligned} accuracy &= \frac{\texttt{TPOS} + \texttt{TNEG} + \texttt{TNEU}}{\texttt{TC}_{pos} + \texttt{TC}_{neg} + \texttt{TC}_{neu}} \\ &= \frac{26 + 26 + 398}{200 + 389 + 411} \\ &= 0.45 \end{aligned}$$

Therefore, the VADER classifier achieved an accuracy of 0.45 on my corpus of South African election tweets. An acceptable accuracy would be at least 0.6 or greater. Hence, the VADER classifier is not accurate at classifying the sentiment contained in South African election tweets. However, this measure of accuracy is not useful since there is an uneven distribution of tweets per sentiment. A more reliable measure would be the $F1$ score.

The $F1$ score is given by Equation 3.15 presented in Subsection 3.4. The equation is given below:

$$F1 = 2 \times \frac{recall_{avg} \times precision_{avg}}{recall_{avg} + precision_{avg}}$$

The respective calculations of $recall_{avg}$ and $precision_{avg}$ were done as fol-

lows:

$$\begin{aligned} recall_{avg} &= \frac{recall_{pos} + recall_{neg} + recall_{neu}}{3} \\ &= \frac{0.13 + 0.0672 + 0.9684}{3} \\ &= 0.3885 \end{aligned}$$

$$\begin{aligned} precision_{avg} &= \frac{precision_{pos} + precision_{neg} + precision_{neu}}{3} \\ &= \frac{0.52 + 0.8125 + 0.4336}{3} \\ &= 0.5887 \end{aligned}$$

The $F1$ score was then calculated as follows:

$$\begin{aligned} F1 &= 2 \times \frac{recall_{avg} \times precision_{avg}}{recall_{avg} + precision_{avg}} \\ &= 2 \times \frac{0.3885 \times 0.5887}{0.3885 + 0.5887} \\ &= 0.4828 \end{aligned}$$

From these results it can be seen that the overall performance of the VADER model, using my dataset, was rather low. Table 4.4 shows the baseline accuracy and $F1$ scores when all tweets were labelled neutral, positive or negative respectively. It is clear that the VADER model achieved a greater accuracy and $F1$ score than the baselines. However, the accuracy achieved by the VADER model was only marginally larger. Since there was an uneven distribution of tweets per sentiment in the test set, the accuracy measure was unreliable. Instead, an $F1$ score was used. The highest baseline $F1$ score is 0.1945, which was obtained by labelling all tweets neutral. However, the VADER model achieved an $F1$ score of 0.4828,

which is significantly larger than the highest baseline. Thus, the VADER model adequately classified the sentiment contained in South African election tweets.

Table 4.4: F1 scores achieved for classifying sentiment of election tweets

	$accuracy$	$recall_{avg}$	$precision_{avg}$	$F1$
VADER Classifier	0.45	0.3885	0.5887	0.4828
Neutral Baseline	0.411	0.3333	0.137	0.1945
Positive Baseline	0.2	0.3333	0.0667	0.111
Negative Baseline	0.389	0.3333	0.129	0.186

The sentiment of the general public towards the ANC, DA and EFF was overwhelmingly neutral. This could perhaps indicate the noncommittal nature of the tweeting electorate on topics related to these parties during the run-up to the 2014 elections. There was a slightly higher positive sentiment towards the ANC compared to the DA and the EFF respectively. It was shown that the DA was better perceived compared to the EFF, although only slightly so. In terms of negative sentiment, the ANC was viewed most negatively across the board. The DA was the least negatively perceived party, far better than either the EFF or the ANC.

4.3 Positive tweets and election outcome

One of the interesting questions explored in this research was to investigate whether the number of positive tweets associated with a political party was in any way related to the election outcome, as reflected by the total number of votes obtained by that party. The following question was posed to study this relationship in some detail:

Does a relationship exist between the total number of positive tweets

that mention a political party and the total number of votes obtained by that party?

In order to answer this question, firstly, I selected and used only those tweets that contained a positive sentiment. Secondly, the political party mentioned in the tweet was identified. For example, a tweet containing the name "anc" or "african national congress" was simply labelled as "anc". Similarly, if a tweet contained the name "da" or "democratic alliance" it was labelled as "da". Finally, if a tweet contained "eff" or "economic freedom fighters" then it was labelled as "eff". Also, if a tweet mentioned more than one political party or if the political party mentioned in the tweet could not be determined, then the tweet was ignored.

Thirdly, the province associated with a tweet was determined. There are 9 provinces in South Africa: Eastern Cape, Free State, Gauteng, Kwa-Zulu Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape. If the location data of tweet contained any reference to any of these provinces, then it was associated to that province. For example, if the location data of a tweet contained "Gauteng" then the tweet was associated with the Gauteng province. If a tweet mentioned multiple provinces or if it had fictitious location data then it was ignored. For example, some tweets had "*In Wonderland*" or "*Mars*" as a location. Some tweets tweets had no location data at all, so those tweets had to be ignored also.

Figure 4.2 shows the number of positive tweets related to each political party per province in South Africa. In total 336 tweets with positive sentiment and valid location data were identified. Between the ANC, DA and EFF, the ANC received the most positive tweets (303) which correlates to the fact that the ANC received the highest number of votes (11 436 921). The EFF received more positive tweets (22) than the DA (11), however, the DA received more votes (4 091 584) than the EFF (1 169 259). The Majority of the ANC's positive tweets originated from Gauteng, with the least number of positive tweets from Limpopo. Once again, the majority of the positive tweets originated from Gauteng for both the EFF and DA.

In addition, for both the EFF and DA, there were no positive tweets identified in the Free State, Limpopo, Mpumalanga, North West and Northern Cape.

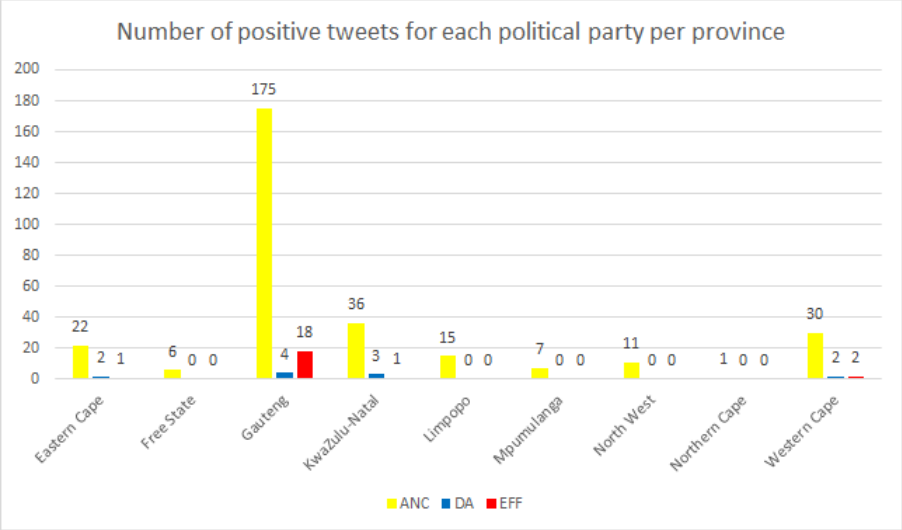

Figure 4.2: Number of positive tweets for each political party per province

In order to determine if there is a relationship between the total number of positive tweets that mention a political party and the total number of votes received by that party, I calculated the Pearson correlation coefficient(r). However, due to the lack of data for the EFF and DA, I could not conduct any correlation analysis. Fortunately, there was enough data to calculate the Pearson correlation coefficient(r) for the ANC.

Table 4.5: Number of votes and associated positive tweets for the ANC per province

Province	**Number of votes**	**Number of positive tweets**
Eastern Cape	1 528 345	22
Free State	708 720	6
Gauteng	2 348 564	175
KwaZulu-Natal	2 475 041	36
Limpopo	1 149 348	15
Mpumalanga	1 045 409	7
North West	733 490	11
Northern Cape	272 053	1
Western Cape	697 664	30

Table 4.5 shows the total number of votes received by the ANC and the number of positive tweets that mention the ANC. The highest number of votes the ANC received was from the Kwa-Zulu Natal province but the highest number of positive tweets the ANC received was from the Gauteng province. The lowest number of votes for the ANC was from the Free State province but the lowest number of positive tweets the ANC received was from the Northern Cape province. Using the data presented in Table 4.5, I conducted a paired t-test between the number of positive tweets that mention the ANC and the number of votes the ANC received. The results of the t-test can be seen in Table 4.6. It can be seen that the Pearson correlation coefficient (r) = 0.67, which means that there is a fairly strong positive correlation between the number of positive tweets that mention the ANC and the number of votes the ANC receives. A positive relationship means that if the number of positive tweets that mentions the ANC increases/decreases then the number of votes the ANC receives also increases/decreases. In Figure 4.3, the trend line (in red) has a positive gradient which shows this positive relationship. Furthermore, the P(T$\leq$ t) two tail = $0.0014 \leq 0.05$. This means that the relationship between the number of positive tweets that mention the ANC and the number of votes the ANC receives is statistically significant.

Therefore, it can be concluded that there was a fairly strong positive correlation between the number of positive tweets that mentioned the ANC and the number of votes the ANC received. However, not conclusion can be made for the EFF and DA due to the lack of data.

Table 4.6: Results of a paired t-test conducted on the number of positive tweets and number of votes for the ANC

	Number of positive tweets	Number of votes
Mean	**33.6667**	**1217626.0000**
Variance	2942.0000	581102735313.5000
Observations	9	9
Pearson Correlation (r)	**0.6738**	
Hypothesized Mean Difference	0	
Degrees of freedom	8	
t Statistic	**-4.7918**	
P(T<=t) one-tail	0.0007	
t Critical one-tail	1.8595	
P(T<=t) two-tail	**0.0014**	
t Critical two-tail	**2.3.060**	

4.4 Cluster Analysis

Clustering plays an important role in identifying similar items and grouping them together into a cluster. Items belonging to the same cluster are more similar to each other, but different from items in other clusters. The corpus of tweets I collected for this research contained the names of different political parties such as the ANC, the DA and the EFF. In this part of the research I investigated whether the election tweets could be clustered based on the names of the political parties mentioned in those tweets. The following research question was posed:

> *Can clustering divide political tweets into political-party clusters, based on the corpus of tweets collected during the 2014 South African election?*

Some research has been conducted on the clustering of tweets [Rosa *et al.* 2011; Sankaranarayanany *et al.* 2009]. However, to the best of my knowledge, not much work has been done using clustering techniques to analyse and extract meaningful information from South African political tweets. In order to answer the research question posed above, the popular K-Means clustering algorithm was

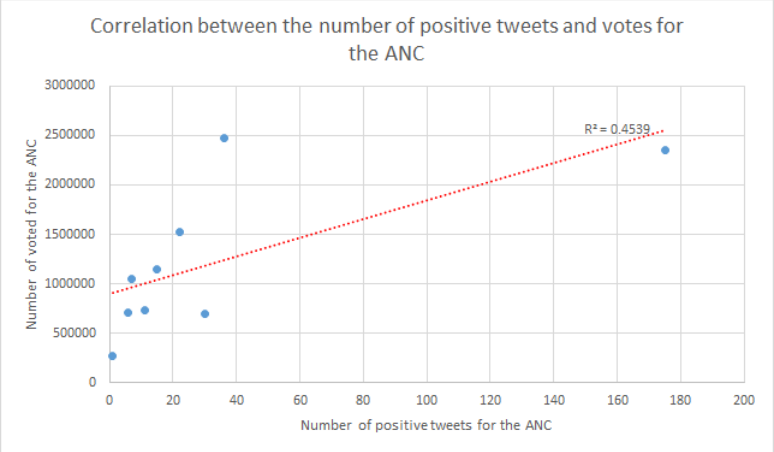

Figure 4.3: Correlation between the number of positive tweets and votes for the ANC

applied on the collected corpus of election tweets. Firstly, the tweets were pre-processed and each tweet was automatically assigned a political party label. For example, a tweet containing either the word "anc" or the word "african national congress" was labelled as an "anc" tweet. Similarly, a tweet containing "da" or "democratic alliance" was labelled a "da" tweet. Also, a tweet containing "eff" or "economic freedom fighters" was labelled a "eff" tweet. A tweet mentioning more than one political party was labelled "mto" (more than one), otherwise it was labelled "uk" (unknown).

I then applied the WEKA toolkit's K-Means algorithm on the tweet dataset. Since five political party labels were being used, the K-Means algorithm was executed by setting $k = 5$, to specify the number of clusters to be generated. The corpus was split into a training (60%) and test (40%) set, respectively. The maximum iteration and seed setting in WEKA was set an 1000 and 100 respectively. The value of these settings were obtained empirically. Figure 4.4 shows the clusters produced after applying the K-Means algorithm on the test set.

Cluster 0 predominantly contains the tweets labelled as "mto". Cluster 1 and Cluster 2 was an ANC cluster. Most of the tweets in these clusters referred to

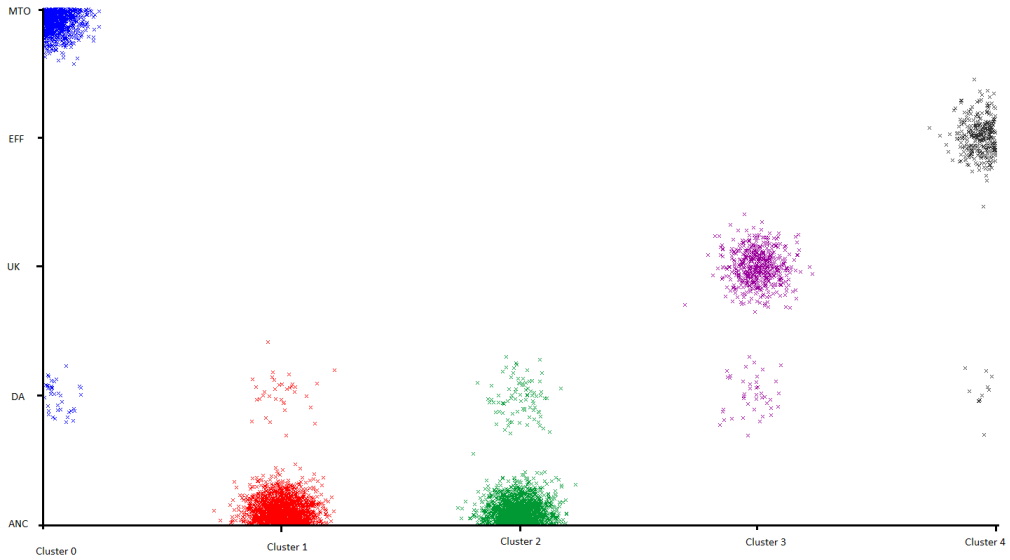

Figure 4.4: Results of K-Means clustering on the corpus of election tweets

the word "anc". The Majority of tweets in Cluster 3 were labelled "uk". Cluster 4 mostly consisted of tweets labelled "eff". There is no cluster that contains the tweets labelled "da", instead portions of these tweets are contained in every other cluster.

Further investigation of each cluster revealed useful information such as the total number of tweets per cluster, the sentiment of the different tweets contained in the cluster, the average number of characters per tweet and the political party associated with each cluster. All this information is presented in Table 4.7.

Table 4.7: Interesting details about each cluster

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Total number of tweets	849	1915	1891	584	340
Positive tweets	2.12%	8%	4.44%	4.62%	5.88%
Negative tweets	1.53%	2.66%	3.44%	0.86%	2.1%
Neutral tweets	96.35%	89.35%	92.1%2	94.52%	92.1%
Number of retweets	123	303	296	50	43
Average number of characters	71.53	67.46	70.61	76	69.63
Political party	MTO	ANC	ANC	UK	EFF

It is important to note that the sentiment displayed in Table 4.7 was obtained by excluding the neutral sentiment from mixed sentiment tweets. The average length of tweets in each cluster was relatively similar, except in the case of Cluster 0 and Cluster 1, where the average tweet length was exactly 122 characters in both clusters.

Cluster 0 was the third largest cluster, containing 849 tweets of which 14.49% were retweets. This cluster contained all tweets that were labelled "mto" and some tweets labelled "da". In fact, there 810 tweets labelled "mto" and only 39 tweets labelled "da". Looking more closely at the tweets labelled "mto", it was observed that the term "myanc" appears frequently. Similarly, this term also appears in the tweets labelled "da". This would explain why these tweets were clustered together. Given that there are more "mto" tweets, this cluster was labelled as the "several" political party cluster. This cluster consisted of 2.12% positive, 1.53% negative and 96.35% neutral tweets. This was a predominantly neutral cluster. Thus, this cluster was classified as the *neutral several political party cluster*.

Cluster 1 was the largest cluster, containing 1915 tweets, of which 15.82% were retweets. This cluster contained 1880 tweets labelled "anc" and 35 tweets labelled "da". Upon further inspection of the tweets contained in this cluster, it was observed that the majority of the tweets labelled "da" actually contained the term "ancs" which is very close to the term "anc". This is a possible explanation why these tweets were placed in Cluster 1. This cluster consisted of 8% positive, 2.66% negative and 89.35% neutral tweets respectively. Thus, this cluster was predominantly neutral. Therefore, this cluster was classified as the *neutral ANC cluster*.

Cluster 2 was the second largest cluster containing 1891 tweets with 15.65% retweets. This cluster contains tweets that mentioned the ANC and DA. There were 87 DA tweets and 1804 ANC tweets. The term "gautenganc" and "ancyouth" ap-

peared frequently in the tweets from this cluster, even the DA tweets. Furthermore, this cluster contained 4.44% positive, 2.66% negative and 89.35% neutral tweets. Similar to Cluster 1, this cluster was classified as the *neutral ANC cluster*. There is no apparent reason why, the ANC tweets were clustered into 2 separate clusters. However, Cluster 2 does contain significantly less positive tweets compared to Cluster 1. In addition, this Cluster 2 has a higher percentage of neutral tweets than Cluster 1.

Cluster 3 was the second smallest cluster containing 584 tweets with 8.56% retweets. This cluster contains every tweet that was labelled "uk" and some tweets that mentioned the DA. In this cluster there were 32 DA tweets and 552 tweets labelled as "uk". Upon further inspection of the tweets in this cluster, it was observed that these tweets often contained words such as "daanc", "noweff". These are misspelt references to political parties. In the case of "daanc" this is a reference to the "da" and the "anc" which is concatenated. This could explain why these tweets were clustered together. Furthermore, this cluster contained 4.62% positive, 0.86% negative and 94.52% neutral tweets. Thus, this cluster was classified as the *neutral unknown political party cluster*.

Cluster 4 was the smallest cluster. It contained 340 tweets, of which 12.65% were retweets. This cluster contained every tweet relating to the EFF and some DA tweets. There were 329 EFF tweets and only 11 DA tweets. Moreover, this cluster contained 5.88% positive, 2.1% negative and 92.1% neutral tweets. Therefore, this cluster was classified as the *neutral EFF cluster*.

It's no surprise that all clusters are overwhelmingly neutral. This is due to the high volume of neutral tweets in the data set. It was surprising to find that there were two separate ANC clusters (Cluster 1 and Cluster 2) of similar size. There is no apparent reason why these clusters are separate and further investigation is needed. However, it can be seen from Cluster 1 that the ANC had a higher concentration of positive tweets than the EFF in Cluster 4. The highest concentration of

negative tweets was found in Cluster 2 which is also an ANC cluster. The tweets relating to the DA was scattered across all the other clusters. This was due to the common terms found in the DA tweets and the tweets in the other clusters. This highlights possible uncertainty amongst the DA supporters, since tweets relating to the DA also mentioned, in some form, another political party.

Since there was no DA cluster, all the tweets relating to the DA was incorrectly classified. Thus there were only 204 misclassifications from 5579 tweets, which means that the K-Means algorithm achieved an accuracy of 96.34%. Therefore, it can be concluded that clustering successfully divided a corpus of 2014 South African election tweets into political-party clusters.

4.5 Association Rule Mining

Association rule mining algorithms are applied on different types of datasets in order to identify and extract interesting and meaningful relationships or associations amongst the dataset items. I used the apriori association rule mining algorithm to study and analyse the relationship amongst the words contained in the political tweets that were collected during the 2014 South African general election. This part of the research was conducted based on the following research question:

> *How well do the association rules extracted from the political tweets generated during the 2014 South African general election characterise the relationship between the words contained in these tweets?*

This question was answered by applying the *apriori* algorithm on my dataset of political tweets and discovering associations between the terms contained in the tweet dataset. Association rule mining often generates a large number of rules, not all of which may be interesting or meaningful. As discussed in Subsection 3.8.4 two measures, namely *support* and *confidence*, can be used to discover meaningful rules. In this research the support and confidence values were set at 0.01 and

0.7, respectively. The support value was set at 0.01 to ensure that all rules were returned, regardless of how infrequent they may appear in the dataset. The confidence was set at 0.07 to ensure that the generated rules were significant.

Firstly, I preprocessed the tweets as discussed in subsection 3.3.4, after which I applied the *apriori* algorithm contained in the *arules* library of the statistical package R, in order to discover association rules linking the terms used in these tweets. The following tweet generated a large number of rules:

> *But you cant use the same BIS bought with NSFAS to tweet "ANC HAS DONE NOTHING". Uxokelani?.*

I decided to remove this tweet from the corpus of preprocessed tweets. Thereafter, I reapplied the *apriori* algorithm on the tweet dataset. A total of 137 rules were generated, with $0.01 \leq support \leq 0.7974$ and $0.7023 \leq confidence \leq 1$. A complete list of the association rules generated from this experiment can be seen in Appendix B. In addition to the support and confidence metrics, I used the *lift* metric to determine the strength of the association between the terms contained in the tweets. The greater the lift, the stronger the association. For a particular rule, if $lift \gg 1$, then this indicates that in their occurrence the antecedent and consequent of the rule depend on each other. For the association rules generated in this experiment the lift values ranged from 0.8808 to 78.3964. Table 4.8 shows 20 of the strongest association rules generated using the *apriori* algorithm. The rules are sorted in descending order according to the lift metric. They are also numbered for convenience.

The strongest rule generated was $anc, ruling, south \Rightarrow africas$ with a lift of 78.3964, which indicates a strong association between *anc, rulling, south* and *africa*. It has a support of 0.0104 which means the words "anc", "ruling" and "south" do not appear often in tweets together. More specifically, it appears in 1.04% of tweets in the corpus. However, it has a confidence of 0.9793 which

Table 4.8: Association rules with the largest lift

Rule	Support	Confidence	Lift
1. anc,ruling,south ⇒ africas	0.0104	0.9793	78.3964
2. ruling,south ⇒ africas	0.0106	0.9732	77.9036
3. africas,anc,south ⇒ ruling	0.0104	0.8554	41.9709
4. africas,south ⇒ ruling	0.0106	0.8529	41.8491
5. africas,anc ⇒ ruling	0.0104	0.8503	41.7195
6. africas ⇒ ruling	0.0106	0.8480	41.6044
7. anc,western ⇒ cape	0.0162	0.9911	29.6866
8. western ⇒ cape	0.0169	0.9830	29.4442
9. bjp ⇒ congress	0.0103	0.9592	25.2505
10. bjp,people ⇒ congress	0.0102	0.9589	25.2432
11. africas,ruling ⇒ south	0.0106	1.0000	21.4898
12. africas,anc,ruling ⇒ south	0.0104	1.0000	21.4898
13. africas ⇒ south	0.0124	0.9942	21.3641
14. africas,anc ⇒ south	0.0121	0.9940	21.3611
15. africa,anc ⇒ south	0.0175	0.8989	19.3167
16. africa ⇒ south	0.0214	0.8960	19.2554
17. counted ⇒ votes	0.0112	0.7512	14.4228
18. anc,counted ⇒ votes	0.0112	0.7512	14.4228
19. bjp ⇒ people	0.0107	0.9932	11.3678
20. bjp,congress ⇒ people	0.0102	0.9929	11.3645

means that whenever the words "anc", "ruling" and "south" appear in a tweet together, there is a 97.93% chance that the word "africas" will also appear in the same tweet. From the association rules in Table 4.8, it is clear that there are strong associations between "anc", "ruling", "south" and "africa". In addition, association rules 9, 10, 19 and 20 from Table 4.8 contained words such as "bjp" and "congress". Upon further investigation, I realised that the association rules generated from these three words were derived from tweets about the Bharatiya Janata Party (BJP), the political party currently governing in India, and possibly the opposition Indian National Congress (INC, also known as "Congress"). These tweets do not relate in any way to the South African General elections. However, they would have passed the Twitter API filtering since it contains the word "national congress", which partially matches the ANC (African National Congress).

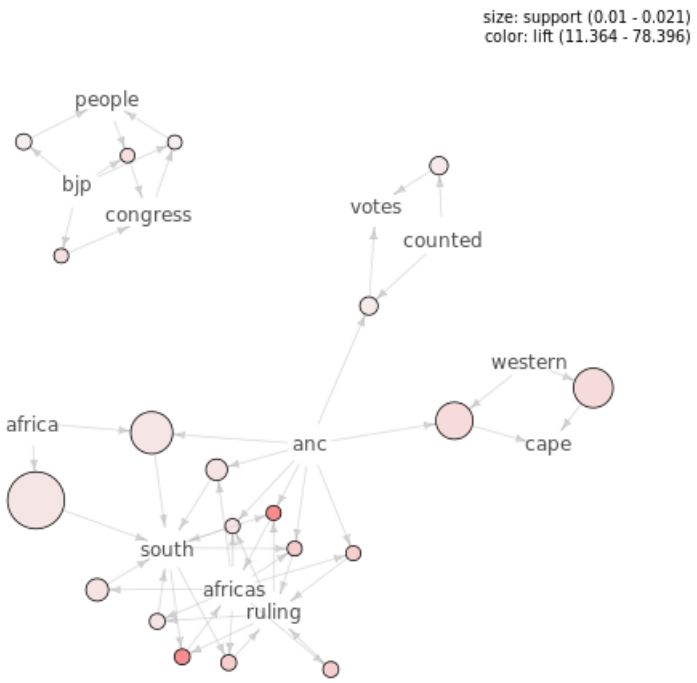

Figure 4.5: Association rules visualised as a graph

Association rules can be visualised using a directed graph, as shown in Figure 4.5 for the association rules in Table 4.8. In this graph words that have edges **into** a node form the *left hand side* (LHS) of a rule, and words that have edges that lead **out of** a node form the *right hand side* (RHS) of the rule. For example, the words "africa" and "anc" have edges pointing to the same node, whose edge points to the word "south", resulting in the rule

$$africa, anc \Rightarrow south$$

Worth noting is the fact that Figure 4.5 actually contains two graphs. In the top left corner of the figure there is a smaller graph consisting of the words "people", "bjp" and "congress". This graph is completely disconnected from the bigger graph in the figure. As mentioned previously, the BJP and INC are political parties from India and in no way relate to the South African election. This is confirmed by the

fact that this graph is disjoint from the graph consisting of words used in the South African election tweets. The larger graph also shows the interconnectedness of the words "anc", "south", "africa" and "ruling", an indication of the strong associations between these words.

Apart from the association rules presented in Table 4.8 I also found some of the other rules to be interesting and informative, even though their lift values were somewhat low. For example, two of these rules are shown in Table 4.9.

Table 4.9: Interesting association rules

Rule	**Support**	**Confidence**	**Lift**
nkandla ⇒ anc	0.0172	0.8489	1.0647
corruption ⇒ anc	0.0206	0.8319	1.0433

The first rule

$$nkandla \Rightarrow anc$$

has a low support value of 0.0172. However, if the word "nkandla" appears in a tweet there is a 84.89% chance that the word "anc" will also appear in that tweet. This rule has a lift value of 1.0647, which near to 1. This means the occurrence of "nkandla" in a tweet has no effect on the occurrence of "anc" in a tweet. In the South African political context, "nkandla" refers to the private homestead of Mr. Jacob Zuma (the president of South Africa and of the ruling party, the ANC). The second rule in Table 4.9 was also found to be very interesting.

$$corruption => anc$$

This rule rule has a lift of 1.0433 which is also near to 1. This indicates that the occurrence of "corruption" in a tweet has no effect on the occurrence of "anc" in a tweet. However, this rule does have a high confidence of 0.8319, which means that if "corruption" appears in a tweet there is a 83.19$\%$ chance of the word "anc" appearing in the tweet. These two rules highlight the discussions on twitter

regarding corruption and the ANC.

Table 4.10: Other interesting association rules

Rule	**Support**	**Confidence**	**Lift**
majority ⇒ anc	0.0104	0.9226	1.1571
lead ⇒ anc	0.0130	0.9223	1.1567
mbalulafikile ⇒ anc	0.0142	0.9559	1.1988
love ⇒ anc	0.0135	0.9024	1.1318
cosatu ⇒ anc	0.0132	0.9095	1.1407
leads ⇒ anc	0.0148	0.9018	1.1310
leading ⇒ anc	0.0185	0.9806	1.2298
lead ⇒ anc	0.0130	0.9223	1.1567
win ⇒ anc	0.0149	0.8908	1.1172
freedom ⇒ anc	0.0134	0.7023	0.8808
ruling ⇒ anc	0.0191	0.9355	1.1732
government ⇒ anc	0.0182	0.8356	1.0479
hellenzille ⇒ anc	0.0184	0.8103	1.0162
da ⇒ anc	0.1267	0.7448	0.9341
million,votes ⇒ anc	0.0121	0.9763	1.2245

Other rules were also found to be interesting, such as the ones shown in Table 4.10. In the vast majority of all the rules that were generated the consequent consisted primarily of the term *anc*. Some of the rules in Table 4.10 showed the ruling party, the ANC, leading during the 2014 election, being in the majority and eventually winning these elections. The party was also correctly identified as the ruling or governing party. The relationship between the ANC and the Congress of South African Trade Unions (COSATU), a member of the tripartite alliance that includes the ANC and the South African Communist Party (SACP), was also correctly identified. The adversarial relationship between the main opposition party the Democratic Alliance (DA) and the ANC was also depicted, as was the equally adversarial relationship between the then leader of the opposition Ms. Hellen Zille, and the ANC. The Minister of Sports, Mr. Fikile Mbalula, was also correctly identified as having a relationship with the ANC.

Amongst the numerous interesting association rules that were discovered, there were rules that maybe considered *redundant*. For example, consider the rules

$A \Rightarrow B$ and $B \Rightarrow C$ then the rule $A \Rightarrow C$ would be considered redundant as the first two rules imply the third rule. In addition, if $A \Rightarrow B$ and $C \Rightarrow B$ then the rule $A, C \Rightarrow B$ is redundant as it could formed by joining the first two rules. Some of the ruled that were identified as redundant are shown in Table 4.11.

Table 4.11: Redundant rules

Rule	**Support**	**Confidence**	**Lift**
1. africas ⇒ south	0.0124	0.9942	21.3641
2. south ⇒ anc	0.0416	0.8933	1.1203
3. africas ⇒ anc	0.0122	0.9766	1.2248
4. africas,south ⇒ anc	0.0121	0.9765	1.2246

Consider rule 1 and 2, shown below:

$$africas \Rightarrow south$$
$$south \Rightarrow anc$$

Then rule 3:

$$africas \Rightarrow anc$$

is redundant, since by transitivity, rule 1 and 2 imply rule 3. Furthermore, consider rules 2 and 3:

$$south \Rightarrow anc$$
$$africas \Rightarrow anc$$

Then rule 4:

$$africas, south \Rightarrow anc$$

is redundant. It is clear that rule 3 and 4 in Table 4.11 can be determined using rule 1 and 2. However, discovering redundant rules did not impact the significance or meaningfulness of the other rules that were discovered. It may be beneficial to remove redundant rules, in order to avoid a large number of association rules being discovered. Although, using appropriate support and confidence measures can reduce the number of rules that are discovered.

Concerning the question that had to be answered in this section, it is clear that association rule mining has produced insightful and meaningful information from the corpus of tweets that were collected during the 2014 general election period.

4.6 Word cloud analysis of election tweets

A *word cloud* is a visualization technique that indicates the frequency of the words used in a sample of text, thereby enabling users to identify the main content of the text Zhao [2012]; Wu *et al.* [2011]. The more frequently a particular term or word is used in the text, the more prominently it appears in the corresponding word cloud. In this research the following question was posed which investigated how word clouds could be used to convey the information contained in the election tweets collected during the 2014 South African general election:

> *Can word clouds generated from political tweets, collected during the 2014 South African general election, be used to convey meaningful information regarding voter-related issues, opinions and sentiments?*

Sankaranarayanany *et al.* [2009] used a word cloud to visualise hash tags that were used in the 2009 Iranian elections. In this research I used word clouds to learn how frequently certain words were used in the selected corpus of election tweets. Another idea was to visualise the sentiment associated with each word.

In order to answer the stated research question, I used the Python *wordcloud* library to generate a word cloud from the collected corpus of election tweets. Firstly, the results obtained from the sentiment analysis experiments performed in Section 4.2 were used to generate different word clouds. The font size of each word is directly proportional to its frequency in the corpus of election tweets. The font size of a word is calculated as shown in Equation 3.1, Subsection 3.3.4. The colour of each word corresponds directly to the word's sentiment. Red, blue and green colours were used to indicate negative, neutral and positive sentiments, respec-

tively. For each word its sentiment is the average sentiment of every tweet the word appears in, as shown Equation 3.2, Subsection 3.3.4. The generated word clouds were displayed using the shape of the South African map. To avoid producing illegible and overcrowded word clouds, a decision was taken to select only the top 1 000 most frequent words to appear in these word clouds.

The word cloud generated from the whole corpus of the 2014 South African general election tweets is shown in Figure 4.6. This image conveys different kinds of information. Clearly, the most dominant terms, *anc* and *da*, refer to the two major political parties in South Africa. The term *vote* was also equally dominant and, given the election context in which these tweets were collected, it does make sense for this term to be prominent. The smaller terms, namely *election* and *voting*, should also be understood within this context. The term *eff* represents the third largest political party in South Africa. This was also seemingly reflected by the smaller size of this term compared to the terms *anc* and *da*. A much smaller political party, *cope*, was also featured amongst these parties. The image also contains information about political party leaders, namely *zuma*, *helenzille* and *malema*, a reflection again of their dominant party status. Highly topical issues were also covered, as exemplified by the use of words such as *nsfas*, *corruption* and *nkandla*.

The image also contained the word *uxokelani*, used in South Africa amongst Xhosa language speakers. Depending on the context in which it is used, it either means *why is he/she lying?* (third person) or *why are you lying?* (first person). There is no doubting the fact that, during the election period, such a word may have been targeted at one or more politicians or parties who may have been considered to be untrustworthy or perhaps economical with the truth. The tone in the majority of the words in Figure 4.6 was neutral, as shown by their colour, which is a darker shade of blue. This observation is not surprising, since most of the tweets were shown in Section 4.2 to be neutral in sentiment. Words such as *party*, *win*, *freedom*, *thank* and *well* all appeared to be a shade of turquoise, indicating a positive sentiment amongst tweets containing these words.

Figure 4.6: Word cloud visualising public sentiment

Given the large number of neutral tweets that dominated in my corpus, it was decided to ignore the neutral sentiment allocated to tweets, and to focus instead on positive and negative tweets, which were then used to generate positive and negative sentiment word clouds, respectively. Figure 4.7 is a word cloud dominated by positive tweets.

Figure 4.7 shows the most dominant terms to be *anc*, *win* and *thank*. As expected all the words appeared to be a shade of green since only tweets with a positive sentiment were used to generate this word cloud. However, several words such as *anc*, *zuma* and *cape* were much darker than others. This indicated that the tweets these words appear in have a very strong positive sentiment. Although words such

Figure 4.7: Word cloud visualising positive public sentiment

as *killing* and *corruption* are generally associated with a negative sentiment, they appeared prominently in this word cloud. How they came to be included in this image can only be understood by looking at the positive tweets in which they appear, to better understand the context in which they were used. It is also possible that the tweets containing these terms were misclassified as having a positive, rather than a negative sentiment, which may be why these terms appear in this word cloud.

As far as political parties are concerned, the terms *anc, eff* and *da* reflect the strong positive sentiment associated with these three major parties, the largely dominant ANC being followed by the EFF and the DA. Only one of the smaller political parties is represented, namely the IFP (*ifp*), which emerged after the elec-

tions as the largest party amongst the smaller parties. Leaders of the three major political parties are viewed positively, with more positive sentiment being directed toward *zuma*, followed by *malema* and *hellenzille*, respectively. The term *vavi* is also included. It reflects the positive tweets mentioning Zwelinzima Vavi, the former Secretary General of the Congress of South African Trade Unions (COSATU), who was highly critical before and during the elections on the state of affairs in the country. Positive words such as *freedom, loyalty, love* and *economy* were used to refer to some of the themes that dominate in this word cloud. There was an interest in the economy and on economic issues (*economic*). There were also values that people hold highly and are positive about, such as freedom, loyalty and love.

I was also interested to see which terms would dominate or appear amongst the negative tweets that were posted during the election period. Figure 4.8 shows a word cloud generated from these negative tweets. The dominant words were *anc*, *eff* and *fear*, followed by *zuma*, *da*, *people* and *vote*. A red colour was assigned for negative tweets. The most dominant words such as *anc*, *eff*, *zuma* and *da* were shown in a dark red colour, an indication that the tweets containing these words had a very strong negative sentiment. It is interesting to note that words such as *thank* and *better*, which are generally associated with a positive sentiment, also appeared in this word cloud. Again, a better understanding of why they appeared in this word cloud can be obtained by looking at the tweets from which they originate, to understand the context in which they were used. It is worth noting, however, that not many of such words appeared amongst the majority of words whose sentiment in this word cloud was actually negative.

Looking at the political parties, there was a strong negative sentiment against the ANC and the EFF, with the DA also being negatively perceived, but less so. This is not surprising, since the ANC in particular has had to deal with numerous charges, which included corruption, flouting the constitution, the Nkandla debacle, and arrogance. At least visually, it would seem that the sentiment towards the ANC was as positive as it was negative (see Figure 4.7 and Figure 4.8). For the EFF and

Figure 4.8: Word cloud visualising negative public sentiment

the DA, the negative sentiment towards these two parties seemed to be stronger than the positive sentiment associated with them. Looking at the words that referred to party political leaders, it is clear that *zuma* was by far more negatively perceived compared to either *malema* or *helenzille*. Given the current political climate in South Africa, this does not come as a surprise. The positive sentiment for the word *zuma* (Figure 4.7) was surpassed by the negative sentiment of the tweets associated with it. For *malema* and *helenzille*, there was not much difference between them in terms of positive or negative sentiment, even though *helenzille* appeared to be the least negatively perceived word, compared to *zuma* and *malema*. Figure 4.8 also contained a number of words associated with negativity, such as *corrupt, killing,*

racist, stupidity, hate, worst, unemployment and even profanity. This reflects some of the issues that were the source of the political bickering, divisions and conflicts that were witnessed during the election period.

The effectiveness of words clouds in illustrating the importance of tweet terms and their related sentiment was judged manually. Figure 4.6 is the image that was obtained when all the tweets, regardless of sentiment type, were combined. I found the word cloud to be effective in highlighting the dominant words. For example, the dominance off the three major parties, namely the ANC, DA and EFF was confirmed. However, the generated word cloud was not effective in highlighting the different types of sentiment, since almost all the words appeared as blue, indicating a dominant neutral sentiment. In other words, the majority of tweets in the election corpus were neutral in sentiment. I also generated a word cloud consisting of terms contained only in positive tweets, which is shown in Figure 4.7. This word cloud was able to identify the importance of the ANC, freedom and the economy. Moreover, the various shades of green indicated how strongly associated a word is with a positive sentiment. Similarly, I also generated a word cloud consisting of terms contained in negative tweets (see Figure 4.8). The various shades of red indicated how strongly associated a word is with a negative sentiment. In addition, it highlighted the importance of the ANC, EFF, fear and racism. Overall, a word cloud is an effective tool for visualising public sentiment towards political parties.

4.7 Conclusion

The primary objective of this research was to investigate how text mining techniques can be used to discover information and new knowledge from a corpus of tweets posted before, during and after the South African General Election of May 2014. In order to do this various research questions were posed. This chapter presented the results obtained in order to answer these research questions.

The sentiment of the general public towards political parties was explored in Section 4.2. It was discovered that there was an overwhelming neutral perception of the ANC, DA and EFF during the 2014 South African elections, which may indicate the noncommittal nature of the tweeting electorate on topics related to these parties. There was a slightly higher positive sentiment towards the ANC compared to the DA and the EFF respectively. It was shown that the DA was better perceived compared to the EFF, although only slightly so. In terms of negative sentiment, the ANC was viewed most negatively across the board. The DA was the least negatively perceived party, far better than either the EFF or the ANC.

Secondly, Section 4.3 investigated the relationship between the total number of positive tweets that mention a political party and the total number of votes received by that party during the 2014 South African general election. The ANC had the most positive tweets with a total of 303. The Majority of the ANC's positive tweets originated from Gauteng, with the least number of positive tweets from Limpopo. The EFF and DA had 22 and 11 positive tweets respectively. Once again, the majority of the positive tweets originated from Gauteng for both the EFF and DA. In addition, for both the EFF and DA, there were no positive tweets identified in the Free State, Limpopo, Mpumalanga, North West and Northern Cape. Due to the lack of data for the EFF and DA, no correlation analysis could be conducted for these political parties. However, the number of votes and positive tweets for the ANC was correlated on a provincial level. The results showed that there is a fairly strong positive correlation between the number of positive tweets that mention the ANC and the number of votes the ANC receives. A positive relationship means that if the number of positive tweets that mentions the ANC increases/decreases then the number of votes the ANC receives also increases/decreases. No conclusions could me made regarding the EFF and DA.

In Section 4.4 the K-Means clustering algorithm was used to examine if clustering could successfully divide political tweets into predominantly political-party clusters, based on the corpus of tweets collected during the 2014 South African

election. The corpus of election tweets was automatically labelled with five possible labels: "anc", "da", "eff", "mto"(more than one) and "uk"(unknown). It's no surprise that all clusters are overwhelmingly neutral. This is due to the high volume of neutral tweets in the data set. It was surprising to find that there were two separate ANC clusters (Cluster 1 and Cluster 2) of similar size. There is no apparent reason why this clusters are separate and further investigation is needed. There was a distinct EFF cluster which contained all the EFF tweets. The tweets relating to the DA was scattered across all the other clusters. This highlighted possible uncertainty amongst the DA supporters, since tweets relating to the DA also mentioned, in some form, another political party. Thus, there were only 204 misclassifications from 5579 tweets, which means that the K-Means algorithm achieved an accuracy of 96.34%. Therefore, it can be concluded that clustering successfully divided a corpus of 2014 South African election tweets into political-party clusters.

Association rule mining was applied, in Section 4.5, in order to determine how well the association rules extracted from the political tweets generated during the 2014 South African general election characterise the relationship between the words contained in these tweets. In total 137 association rules were discovered. These rules correctly identified the ANC as the ruling party in South Africa. Furthermore, the rules highlighted how often the public tweets about the ANC and corruption, even though this was a weak association. The discovered association rules truly reflect the perception and the reality currently prevailing in South Africa's political landscape. It was clear that association rule mining has produced insightful and meaningful information from the corpus of tweets that were collected during the 2014 general election period.

Section 4.6 investigated how useful word clouds are as indicators of voter opinion and sentiment. The effectiveness of words clouds in illustrating the importance of tweet terms and their related sentiment was judged manually. Figure 4.6 is the image obtained when all the tweets, regardless of sentiment type, were combined. I found the word cloud to be effective in highlighting the dominant

words. For example, the dominance off the three major parties, namely the ANC, DA and EFF was confirmed. However, the generated word cloud was not effective in highlighting the different types of sentiment, since almost all the words appeared as blue, indicating a dominant neutral sentiment. In other words, the majority of tweets in the election corpus were neutral in sentiment. I also generated a word cloud consisting of terms contained only in positive tweets, which is shown in Figure 4.7. This word cloud was able to identify the importance of the ANC, freedom and the economy. Moreover, the various shades of green indicated how strongly associated a word is with a positive sentiment. Similarly, I also generated a word cloud consisting of terms contained in negative tweets (see Figure 4.8). The various shades of red indicated how strongly associated a word is with a negative sentiment. In addition, it highlighted the importance of the ANC, EFF, fear and racism. Overall, it was concluded that a word cloud is an effective tool for visualising public sentiment towards political parties.

In Chapter 5, the results presented here will be discussed in the context of related research. Furthermore, Chapter 5 will discuss in detail how all the research questions tie together to answer the main research question.

Chapter 5

Discussion

Twitter provides a platform that enables users to express their views and opinions on a variety of issues, including political topics [Wang *et al.* 2012; Monti *et al.* 2013]. Extensive research has been conducted in countries such as Ireland, Germany and the United States, in which text mining techniques have been used to obtain information from politically oriented tweets [Wang *et al.* 2012; Bakliwal *et al.* 2013; Tumasjan *et al.* 2010]. However, not much is known about similar work being conducted within the South African context. The purpose of this research was to contribute in addressing this gap, by using different text mining techniques to analyse the political tweets that were collected during the 2014 South African General Election. In order to provide a clear focus on the research that was being conducted, the following research question was posed:

> *Do text mining techniques uncover meaningful information when applied on a corpus of political tweets collected on the 2014 South African General Election?*

In order to answer this question five sub-questions were posed, each focussing on a different aspect of the main research question. To answer these sub-questions, a total of 15 764 tweets were collected, of which 13 946 tweets were selected and

analysed. This research focused on the three major political parties in South Africa namely, the African National Congress (ANC), the Democratic Alliance (DA) and the Economic Freedom Fighters (EFF). The results obtained in this study point to the potential and usefulness of Twitter data, in the South African political context, to provide insightful information to different stakeholders such as political role players, political parties, research and marketing organisations, and academics. This chapter provides a discussion of the research findings in relation to the five sub-questions that were posed. The discussion is also placed within the context of related work, in which Twitter datasets have been used to investigate and derive a better understanding of the issues involved in political and election environments.

Sentiment analysis was used to gauge the sentiment of the general public towards the three major political parties given a corpus of 2014 South African election tweets. The results obtained were presented in Section 4.2. Using the VADER sentiment classifier developed by Hutto and Gilbert [2014], the tweets were classified into positive, negative and neutral categories.

As shown in Table 4.2, over 90% of the tweets assigned to the ANC, DA and EFF were neutral tweets. This is an indication perhaps that the people who posted these tweets were uncertain about whether or not to pledge their support to any of these political parties. In the case of positive tweets, the ANC was only slightly ahead of the DA and the EFF. This slight lead of the ANC compared to the DA and EFF is surprising considering that the ANC has a larger support base than the DA and EFF.

Sentiment analysis was used to gauge the sentiment of the general public towards political parties given a corpus of 2014 South African election tweets. The results obtained were presented in Section 4.2. The VADER sentiment classifier developed by Hutto and Gilbert [2014] was used to classify the sentiment of each tweet in the corpus of election tweets that was collected. It was discovered that the general public showed an overwhelmingly neutral sentiment towards the ANC,

DA and EFF. This may indicate the noncommittal nature of the tweeting electorate on topics related to these parties. This neutrality highlights the uncertainty of the public towards each of the political parties.

There was a clear indication from the negative tweets that the ANC was perceived more negatively than either the DA or the EFF. As shown in Table 4.2 the DA was the least negatively perceived party. It is perhaps not surprising that from these tweets, sentiment towards the ANC was most negative. A number of reasons could be responsible for this poor perception of the party. During the election period, the party had to deal with service delivery protests, numerous allegations of corruption against some party members and ministers, and the leader of the party himself Hamil [2014]. There could be a number of reasons for the negative sentiment against the EFF. The new party has proposed radical economic changes such as nationalising the private sector and expropriating land without compensation. These and other factors may have contributed towards the negative sentiment uncovered in the tweets associated with the EFF. Compared to the ANC and the EFF, the negative sentiment towards the DA was the lowest. However, the DA had to grapple with issues related to poor service delivery in the Western Cape province, a situation that may have contributed to the negative sentiment towards the party.

Previous work done by Tumasjan *et al.* [2010] used a corpus of 104 003 tweets collected during the 2009 German federal election. One of the aims of the study was to determine the sentiment contained in the tweets, about political parties and politicians. Tumasjan *et al.* [2010] showed that the sentiment contained in the tweets was closely related to the political stances actually taken by politicians and their respective parties. Similarly, my research has shown how the sentiment contained in the 2014 general election tweets mirrored the social and political events actually taking place in South Africa during that election.

Twitter was also used during the 2012 United States presidential election to track the sentiment of the general public, in real-time, towards the two presidential

candidates [Wang *et al.* 2012]. A relationship was found between the sentiment of the tweets and the electoral events taking place. This finding by Wang *et al.* [2012] is similar to the finding in my study, which was also able to establish a connection between tweet sentiment and the political events taking place in society at the time. Unlike Wang *et al.* [2012], whose study determined tweet sentiment in real time, in my research the tweets were classified offline. Monti *et al.* [2013] have used Twitter data to look into the phenomenon of *political disaffection*, within the context of the 2012 Italian election. The authors define political disaffection as *"negative sentiment towards the political system in general, rather than towards a particular politician, policy or issue"*. Machine learning methods were applied on a 35-million tweet corpus to generate time-series information on political disaffection. A strong correlation was discovered between political disaffection and the factors related to political disaffection in opinion polls. This finding is somewhat similar to the result obtained in my research, in which a relationship was observed between the negative tweets and the negative political events that were taking place in South African society. However, the difference between my research and the work of Tumasjan *et al.* [2010] is that the sentiment analysis results I obtained was not directly correlated to opinion polls.

The VADER sentiment classifier developed by Hutto and Gilbert [2014] was used to classify the sentiment of each tweet in the corpus of election tweets I collected. According to Hutto and Gilbert [2014] the VADER model achieved an $F1$ score of 0.096 on their test set. Using my dataset, the model produced an $F1$ score of 0.4828, significantly larger than the highest baseline of 0.1945. However, I am mindful of the fact that my results cannot be compared directly to those of Hutto and Gilbert [2014], since the latter used a corpus of randomly selected tweets, whereas the tweets used in my research were specifically related to the 2014 South African general election. The VADER model uses the lexicon gathered by Hutto and Gilbert [2014], which may perhaps not be suitable for the South African English dialect. The study by Bakliwal *et al.* [2013] used Twitter data to determine

public sentiment during the run-up to the 2011 Irish General Elections. Supervised and unsupervised sentiment classification approaches were compared. The supervised approach was slightly more accurate (61.2%) compared to the unsupervised approach (58.96%). In my research the tweets were classified using the VADER model developed by Hutto and Gilbert [2014], which is implemented through an unsupervised algorithm. Perhaps a supervised approach to sentiment classification of South African election tweets can be explored as part of future work.

Another interesting question explored in this research was to determine whether a relationship existed between the total number of positive tweets that mention a particular political party and the total number of votes received by that party. In order to determine this relationship, a correlation analysis was conducted. However, due to the lack of sufficient data, the analysis did not include the EFF and the DA. Fortunately, there was enough data to perform the analysis for the ANC. The correlation coefficient value ($r = 0.67$) obtained from this limited analysis suggested a fairly strong positive correlation between the number of positive tweets that mention the ANC and the number of votes received by the party. However, given that there was insufficient data for the other parties, this result cannot be used to state conclusively that a relationship exists between the number of tweets that mention a party and the number of votes the party receives.

Similar work done by Tumasjan *et al.* [2010] investigated whether tweets could be used to predict party popularity during the 2009 German federal elections. The study showed that the volume of tweets that mention a political party corresponds to the number of votes obtained by that party. Different results were obtained in my research. Of the three major parties whose positive tweets were used, the number of positive tweets for a party did not correspond to the number of votes obtained by that party. In the case of the ANC, which received the majority of positive tweets, the number of votes it received were also in the majority. The DA received the least number of positive tweets but the second largest number of votes. On the other hand, the EFF received the second largest number of positive

tweets but the least number of votes.

The study conducted by Chung and Mustafaraj [2011] produced results similar to mine. Chung and Mustafaraj [2011] used tweets to predict the outcome of the 2010 United States Senate special election held in Massachusetts. Only two candidates participated in that election, namely Martha Coakly and Scott Brown. A total of 53.86% of the tweets mentioned Martha Coakly and 46.14% mentioned Scott Brown. However, Scott Brown won the election, receiving 52% of the vote, compared to Martha Coakly's 48%. Chung and Mustafaraj [2011] have argued that tweet sentiment must be taken into account, since tweets may reflect an opposing rather than supportive sentiment towards a candidate. However, Chung and Mustafaraj [2011] did not provide the results from their research. They stated that the accuracy of the sentiment classifier used in their research was not good enough. Overall, their study indicated that the number of positive tweets that mention a political party cannot be used as an indicator of election outcome. In my research there was not enough data to make any conclusions for the EFF or DA. Thus, future work can improve on this by collecting tweets with valid location data. Furthermore, I only considered the most prominent parties in South Africa (the ANC, DA and EFF), perhaps, future work can extend on this and consider more South African political parties.

The third sub-question in this research was intended to determine if clustering could successfully divide political tweets into clusters associated with the parties they mentioned. Each tweet was automatically labelled as belonging to one of five categories, namely "anc", "da", "eff", "mto" and "uk" (for ANC, DA, EFF, several parties and an undetermined party, respectively). The *K-Means* clustering algorithm was used, and since there were five tweet categories, k (number of clusters) was set to 5.

Surprisingly, two of the five clusters were distinctly ANC-related clusters. Due to the limited time available, it was not possible to determine the reason for having

these two, distinct ANC-related clusters. This is an area that could be explored as part of future work. However, what was not surprising, given the large support base of the ANC, was to notice that the population of tweets in the ANC-related clusters was the largest, compared to both the DA and EFF. Even though the DA has a larger support base than the EFF, there was no distinct DA cluster. Instead, DA-related tweets were present in all the other clusters. This could be seen as a sign of uncertainty amongst the DA supporters, since the content in DA-related tweets also mentioned other political parties.

Considering the clusters, the first ANC-related cluster contained the largest number of positive tweets compared to the EFF-related cluster, which may have to do with the larger support base of the ANC. The second ANC-related cluster contained the highest concentration of negative tweets. This may be related to the fact that the ANC had to deal with numerous allegations of corruption, fraud and other malpractices. The *K-means* algorithm used to cluster the tweets was 96.34% accurate, an indication that cluster analysis was largely successful in assigning the tweets into respective party-related clusters.

The purpose behind the fourth sub-question was to discover interesting relationships among the keywords of the tweet contents. To determine these relationships I used the *apriori* association rule mining algorithm. The *support* and *confidence* values were set at 0.01 and 0.7, respectively. The algorithm, generated a total of 137 rules, the majority of which contained the "anc" keyword as part of the consequent. These rules highlighted the fact that the ANC is the governing party of South Africa. Other relationships that were correctly identified by the rules were those between the ANC and the other two members of the tripartite alliance, namely the Congress of South African Trade Unions (COSATU) and the South African Communist Party (SACP). The adversarial relationship between the main opposition party DA and the ruling party (ANC) was also depicted, as was the equally adversarial relationship between the then leader of the DA, Ms. Hellen Zille, and the ANC. The Minister of Sports, Mr. Fikile Mbalula, was also correctly

identified as being linked to the ANC.

Some of the rules were discovered to be redundant. However, they had no impact on the significance or meaningfulness of the rules that were discovered to be interesting. Perhaps future work may focus on ways of removing redundant rules, in order to avoid vast numbers of association rules being discovered. Another way in which the tweets were used was to create an association rule graph (ARG), to depict the strong relationships that were discovered between the tweet content keywords. The ARG shown in Figure 4.5 accurately depicts the rules that were generated during the association rule mining phase. It is clear from the results obtained in this part of the research that association rule mining techniques, applied on textual datasets such as Twitter data, can produce interesting, informative and useful rules.

The last sub-question in this research aimed to determine the usefulness of word clouds as visual and informative representation of the prominent keywords contained in the political tweets that was collected during the 2014 election. Three word clouds were created. In the first word cloud (see Figure 4.6), the most dominant terms are *anc* and *da*, which refer to the two major political parties in South Africa. The term *vote* was also equally dominant and, given the election context in which these tweets were being collected, it does make sense for this term to be as prominent as it is. The *eff* term is also dominant, but less prominent compared to either the *anc* or *da* terms. This reflects the current political reality in South Africa, where the EFF is the third largest party after the ANC and the DA. A much smaller term, *cope*, also features in the word cloud. It represents one of the smaller opposition parties that participated in the election. Looking at this word cloud, one is able to obtain clear and correct information about which parties dominate the South African political scene and what their relative strengths are. There are also terms that represent the names of political party leaders, such as *zuma*, *helenzille* and *malema*. These are the leaders of the three major parties. It is interesting to notice that these terms seem to reflect the relative sizes of the parties associated

with them. Terms that represent major issues that were topical during the election period also appear in this word cloud. For example, there are terms such as *nsfas*, *corruption* and *nkandla*. Again, the word cloud is providing useful information about the prominent issues and events that were tweeted about, and which might have been taking place during this period. The tone conveyed in the majority of the terms is neutral, as shown by the colour, which is a darker shade of blue. This is not surprising, since the sentiment in the majority of the tweets was found to be neutral. Given the large number of neutral tweets that dominated in the corpus, it was decided to generate positive and negative sentiment word clouds.

A positive word cloud was created (see Figure 4.7) from the tweets that were classified with a positive sentiment. In total, 633 tweets were used to create the word. These were ANC, DA and EFF-related positive tweets. There are dominant terms in this word cloud, such as *anc*, *win* and *thank*. A green colour was used to show positive tweets. However, several words such as *anc*, *zuma* and *cape* are much darker than others, an indication of a very strong positive sentiment in the tweets containing these words. Although words such as *killing* and *corruption* are generally associated with a negative sentiment, they appear prominently in this word cloud. How they came to be included in this image can only be understood by looking at the positive tweets in which they appear, to better understand the context in which they were used. It is also possible that the tweets containing these terms were misclassified as having a positive, rather than a negative sentiment, resulting in their inclusion in this word cloud. The strong positive sentiment reflected by the *anc* term can be related to how strong and dominant the party was during the election period. By comparison, the *da* and *eff* terms are much smaller. Other positive words that were used in this word cloud included *win, freedom, loyalty, love* and *economy*.

The positive word cloud (see Figure 4.7) showed the most dominant were words *anc*, *win* and *thank*. As expected all the words appear to be a shade of green since only tweets with a positive sentiment were used to generate this word cloud.

However, several words such as *anc*, *zuma* and *cape* are much darker than others. This indicates that the tweets these words appear in have a very strong positive sentiment. Although words such as *killing* and *corruption* are generally associated with a negative sentiment, they appear prominently in this word cloud. How they came to be included in this image can only be understood by looking at the positive tweets in which they appear, to better understand the context in which they were used. It is also possible that the tweets containing these terms were misclassified as having a positive, rather than a negative sentiment, which may be why these terms appear in this word cloud.

A negative word cloud was also created, to investigate what information could be obtained from it. The word cloud was created from a total of 303 negative election tweets, drawn from ANC, DA and EFF-related tweets. The tweets in this word cloud were coloured in red (see Figure 4.8). It can be seen in the figure that the most dominant party-related terms were *anc* and *eff*, and to a lesser extent, the term *da*. What this means is that the ANC and the EFF were the most negatively perceived parties during the election period compared to the DA. The possible reasons for this negative sentiment towards all three parties were provided in my discussion earlier, on sentiment analysis. At individual level the term *zuma* was by far negatively perceived compared to either *malema* or *hellenzille*, both of which are present in the word cloud. Again, possible reasons were provided for this negative perception towards the politicians represented by these terms, in particular Zuma and Malema. It is interesting to note that words such as *thank* and *better*, which are generally associated with a positive sentiment, also appear in this word cloud. Again, a better understanding of why they appear in this word cloud could be obtained by looking at the tweets from which they originate, to understand the context in which they were used. Figure 4.8 also contains a number of words that are associated with a negative sentiment, such as *corrupt, killing, racist, stupidity, devil, hate, worst* and *unemployment*. All these words point to a toxic environment prevailing in South Africa during the election period and which, sadly, continues

to prevail.

In previous word by Wang *et al.* [2012], word clouds were generated in order to track public sentiment in real-time, during the 2012 United States elections. In Sankaranarayanany *et al.* [2009] clustering was used to identify tweets related to news events. Sankaranarayanany *et al.* [2009] identified a cluster related to the 2009 Iranian elections and used its tweets to generate a word cloud . Both Wang *et al.* [2012] and Sankaranarayanany *et al.* [2009] were able to demonstrate the effectiveness of word clouds in visualising the keywords and topics contained in a corpus of tweets. In this respect, their work is similar to the research I conducted on word clouds. However, unlike my research, the keywords contained in the word clouds generated by Wang *et al.* [2012] and Sankaranarayanany *et al.* [2009] were not classified according to sentiment. In an article by Byler [2014], word clouds were used to visualise the 100 most common words used by George W. Bush and Barak Obama in speeches they gave relating to the 9/11 attacks in 2001. The word clouds revealed that both George W. Bush and Barak Obama used similar words in their speeches, such as *"will", "day", "America", "nation"* and *"people"*. However, the meaning of these words varied between George W. Bush and Barak Obama. Thus, Byler [2014] pointed out that word clouds are not effective an conveying context.

In the results of this research, it was found that word clouds effectively visualized the prominent keywords contained in the political tweets that was collected during the 2014 election. The results have also demonstrated the usefulness of Twitter data as a resource that can be used in text mining and to extract information from it.

Chapter 6

Conclusion and Future Work

Millions of people across the world use Social Network Sites such as Twitter to share their views and opinions on a variety of topics, such as politics. The primary objective of this research was to determine how text mining techniques can be utilized to discover meaningful information from a corpus of tweets posted before, during the South African General Election of May 2014. This research focused on the 3 most prominent political parties in South Africa; the African National Congress (ANC), the Democratic Alliance (DA) and the Economic Freedom Fighters (EFF).

Sentiment analysis was used to gauge the sentiment of the general public towards political parties given a corpus of 2014 South African election tweets. The results revealed an overwhelming neutral sentiment of the general public towards the ANC, DA and EFF. This result was rather unexpected, given the expectation that users who were tweeting during the election period would have decided how positively or negatively they felt about these parties. The neutral sentiment towards the DA and the EFF was slightly higher (93.45% and 91.97%, respectively), compared to the ANC (90.83%). The results showed that the ANC, DA and EFF were almost equally perceived, in terms of positive sentiment. Although there was a

slightly larger positive sentiment towards the ANC compared to the DA and EFF. In terms of negative sentiment, the ANC was perceived more negatively compared to the other two major parties, with the DA being the least negatively perceived party. What made this result significant was the fact that the sentiment identified in the results corresponded to events that occurred during the run up to the election. The results have shows the usefulness of sentiment analysis for extracting meaningful information from political tweets.

This research also investigated whether there exists a relationship between the total number of positive tweets that mention a party and the total number of votes received by that party. In the case of the ANC, the results showed that there is a fairly strong positive relationship between the total number of positive tweets that mention the ANC and the total number of votes received by the ANC. Unfortunately, due to the lack of data for the DA and EFF no conclusions can be made for these parties. Work done by Gaber [2016] found that there is a relationship between the Twitter activity of politicians and their election results in the 2015 General Election held in the United Kingdom. Similarly, the study done by DiGrazia *et al.* [2013] found a relationship between the tweet mentions of candidates and their performance in the 2012 Unites Sates mid-term elections. Due to the lack of data for the DA and EFF, no general conclusion can be made as to whether or not a relationship exists between the total number of positive tweets that mention a party and the total number of votes received by that party.

The K-Means clustering algorithm was used to successfully cluster tweets into political party clusters with an accuracy of 96.34%. Upon analysis, the resulting clusters were found to contain useful and relevant information. For example, the largest clusters were the ANC clusters which correlates to the fact that the ANC has a larger support base than both the DA and EFF. The results also showed that there was no distinct DA cluster which may be due possible uncertainty amongst the DA supporters.

The application of association rule mining to the collected corpus of tweets resulted in the discovery of highly informative rules, which clearly depicted the reality prevailing in South Africa's political landscape at the time of the elections. These rules highlighted the adversarial relationship between the main opposition party the DA and the ANC. In addition, the adversarial relationship between the then leader of the DA Ms. Hellen Zille, and the ANC was also identified from the association rules. The results of this research has shown that association rule mining to be useful in extracting meaningful information from tweets.

Another interesting question posed in this research was to explore how effective word clouds could be at conveying meaningful information regarding voter-related issues, opinions and sentiments. The generated words clouds visualised the positive and negative sentiments towards the ANC, DA and EFF respectively. It was pointed out by Byler [2014] that there is a lack of context within a word cloud. However, in my research word clouds were used effectively to visualise information from a corpus of tweets.

This chapter would not be complete without highlighting some of the limitations of this study. In conducting this research, a case study methodology was followed. On the whole the results obtained were very informative and meaningful. However, given the nature of a case study methodology, these results are not generalisable. They are specific to the 2014 South African general election context, which meant that in this regard the research was limited. Using the VADER sentiment classifier was another limitation, in the sense that the lexicon used in the classifier does not cater for some of the uniquely South African English words that form part of the South African English lexicon. An unforeseen limitation was the lack of data collected for the DA and EFF. Due to this the sub-question on correlation analysis was not fully explored. Another limitation was that this research only applied the *K-Means* and the *apriori* algorithms for clustering and association rule mining respectively. Perhaps other algorithms could have been used and the results compared.

During the course of this research several observations were made, pointing to the need for further work to be done in several aspects of this research. The following are some of the interesting ideas for future work that may be explored:

- Using a supervised machine learning approach to sentiment analysis, and building a lexicon that caters for some of the words used in the South African English dialect.
- Obtaining and using sufficient Twitter data, in order to fully address the sub-question on correlation analysis.
- Improving the quality of generated rules by identifying and removing redundant association rules.
- Including more political parties as part of the study.
- Investigate other algorithms for clustering and association rule mining

This research has shown that text mining techniques can uncover meaningful information from a corpus of political tweets. In addition, the results from the five sub-research questions have shown Twitter to be a rich resource for text mining in the context of South African politics.

Appendices

A View of dashboard from Twitter Sentiment Analysis Tool

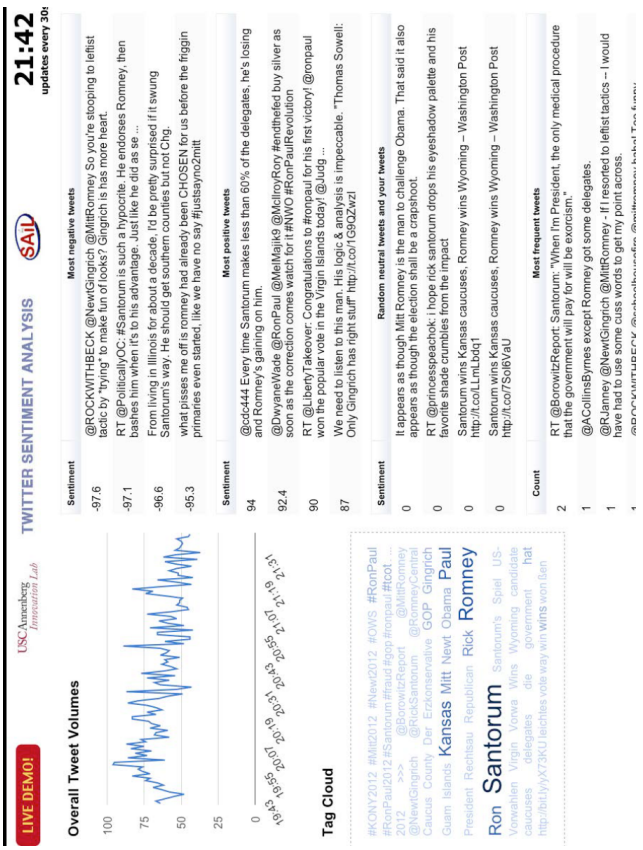

Figure 1: A dashboard developed by Wang *et al.* [2012] which displays a word cloud

B Complete List of Association Rules

Table 1: Complete list of association rules

No.	Rule	Support	Confidence	Lift
1	anc,ruling,south ⇒ africas	0.010373292	0.979310345	78.39637024
2	ruling,south ⇒ africas	0.010592446	0.973154362	77.90356764
3	africas,anc,south ⇒ ruling	0.010373292	0.855421687	41.97085115
4	africas,south ⇒ ruling	0.010592446	0.852941176	41.84914611
5	africas,anc ⇒ ruling	0.010373292	0.850299401	41.71952868
6	africas ⇒ ruling	0.010592446	0.847953216	41.60441426
7	anc,western ⇒ cape	0.016217401	0.991071429	29.68660128
8	western ⇒ cape	0.016874863	0.982978723	29.444192
9	bjp ⇒ congress	0.010300241	0.959183673	25.2505102
10	bjp,people ⇒ congress	0.01022719	0.95890411	25.24315068
11	africas,ruling ⇒ south	0.010592446	1	21.48979592
12	africas,anc,ruling ⇒ south	0.010373292	1	21.48979592
13	africas ⇒ south	0.01241873	0.994152047	21.3641246
14	africas,anc ⇒ south	0.012126525	0.994011976	21.36111451
15	africa,anc ⇒ south	0.017532325	0.898876404	19.31667049
16	africa ⇒ south	0.021404047	0.896024465	19.25538289
17	counted ⇒ votes	0.011249909	0.751219512	14.42278247
18	anc,counted ⇒ votes	0.011249909	0.751219512	14.42278247
19	bjp ⇒ people	0.010665498	0.993197279	11.36779059
20	bjp,congress ⇒ people	0.01022719	0.992907801	11.36447734
21	congress ⇒ people	0.028636131	0.753846154	8.62826087
22	anc,eff,results ⇒ da	0.01183432	0.975903614	5.738464166
23	eff,results ⇒ da	0.012272628	0.96	5.644948454
24	anc,eff,elections ⇒ da	0.011030755	0.915151515	5.381232427
25	anc,electionresults ⇒ da	0.010884652	0.892215569	5.246365516

Table 1: continued

Table 1: continued

Table 1: continued

Table 1: continued

C Screen shots of website

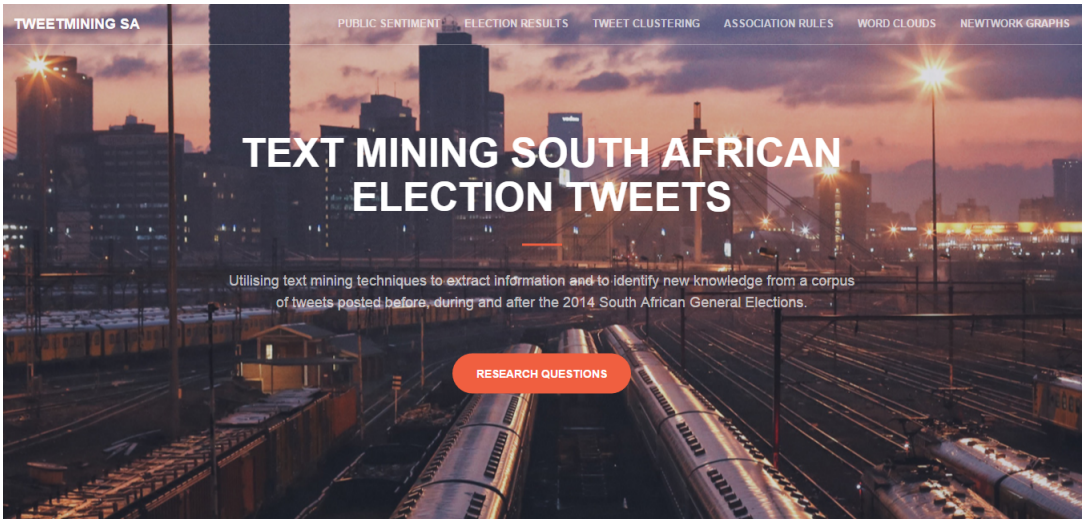

Figure 2: Partial screen shot of the home for the website

Figure 3: Partial screen shot of the page displaying the sentiment analysis results

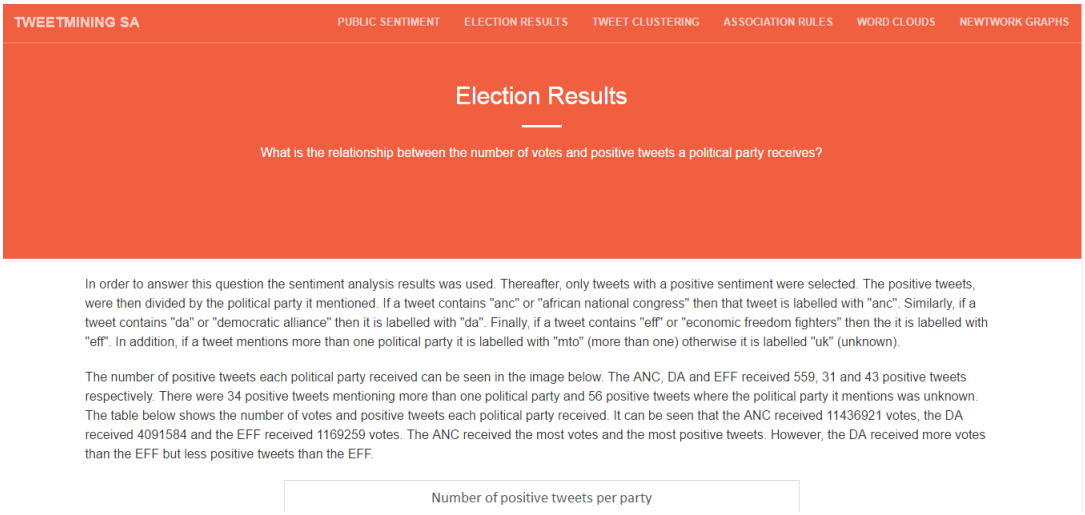

Figure 4: Partial screen shot of the page displaying the results from investigating the relationship between the number of positive political tweets and votes a political party receives

Figure 5: Partial screen shot of the page displaying the tweet clustering results

No.	Rule	Support	Confidence	Lift
26	electionresults ⇒ da	0.010884652	0.86627907	5.093854891
27	anc,provincial ⇒ da	0.010446344	0.748691099	4.402419442
28	provincial ⇒ da	0.010446344	0.737113402	4.334340791
29	counted ⇒ anc	0.014975528	1	1.254145671
30	da,electionresults ⇒ anc	0.010884652	1	1.254145671
31	da,provincial ⇒ anc	0.010446344	1	1.254145671
32	counted,votes ⇒ anc	0.011249909	1	1.254145671
33	da,leading ⇒ anc	0.010811601	0.986666667	1.237423729
34	nationally ⇒ anc	0.015121631	0.985714286	1.236229304
35	provincial ⇒ anc	0.013952809	0.984536082	1.234751666
36	leading ⇒ anc	0.018481993	0.980620155	1.229840522
37	million ⇒ anc	0.017386223	0.979423868	1.228340205
38	africas,ruling ⇒ anc	0.010373292	0.979310345	1.22819783
39	africas,ruling,south ⇒ anc	0.010373292	0.979310345	1.22819783
40	africas ⇒ anc	0.012199576	0.976608187	1.22480893
41	africas,south ⇒ anc	0.012126525	0.976470588	1.224636361
42	million,votes ⇒ anc	0.012053474	0.976331361	1.22446175
43	ruling,south ⇒ anc	0.010592446	0.973154362	1.220477331
44	electionresults ⇒ anc	0.012199576	0.970930233	1.217687948
45	another ⇒ anc	0.012199576	0.970930233	1.217687948
46	da,results ⇒ anc	0.017678428	0.968	1.21401301
47	opposition ⇒ anc	0.014391117	0.965686275	1.211111261
48	da,eff,results ⇒ anc	0.01183432	0.964285714	1.209354754
49	da,eff,elections ⇒ anc	0.011030755	0.961783439	1.206216537
50	cape,western ⇒ anc	0.016217401	0.961038961	1.205282853
51	cape ⇒ anc	0.031996494	0.958424508	1.202003947
52	mbalulafikile ⇒ anc	0.014245014	0.955882353	1.198815715
53	western ⇒ anc	0.016363504	0.953191489	1.19544098
54	thank ⇒ anc	0.012199576	0.948863636	1.190013222
55	eff,results ⇒ anc	0.012126525	0.948571429	1.189646751

No.	Rule	Support	Confidence	Lift
56	said ⇒ anc	0.010008036	0.944827586	1.184951427
57	done ⇒ anc	0.018262839	0.939849624	1.178708337
58	ruling ⇒ anc	0.019066404	0.935483871	1.173233047
59	cape,da ⇒ anc	0.011396011	0.934131737	1.171537274
60	sa ⇒ anc	0.030900723	0.933774834	1.171089666
61	ifp ⇒ anc	0.010154138	0.932885906	1.169974821
62	nothing ⇒ anc	0.010154138	0.932885906	1.169974821
63	results ⇒ anc	0.032434802	0.928870293	1.164938657
64	votes ⇒ anc	0.048140843	0.924263675	1.159161286
65	voted ⇒ anc	0.015048579	0.923766816	1.158538154
66	majority ⇒ anc	0.010446344	0.922580645	1.157050522
67	lead ⇒ anc	0.013003141	0.922279793	1.15667321
68	wc ⇒ anc	0.010373292	0.916129032	1.14895926
69	zuma ⇒ anc	0.03937468	0.915110357	1.147681692
70	come ⇒ anc	0.011688217	0.914285714	1.146647471
71	voters ⇒ anc	0.010811601	0.913580247	1.145762712
72	doesnt ⇒ anc	0.011761268	0.90960452	1.140776571
73	cosatu ⇒ anc	0.013222295	0.909547739	1.140705359
74	thats ⇒ anc	0.013003141	0.908163265	1.138969028
75	da,eff ⇒ anc	0.041785375	0.905063291	1.135081209
76	love ⇒ anc	0.013514501	0.902439024	1.131789996
77	leads ⇒ anc	0.014756374	0.901785714	1.13097065
78	campaign ⇒ anc	0.01994302	0.898026316	1.126255816
79	south ⇒ anc	0.041566221	0.893249608	1.120265128
80	win ⇒ anc	0.014902476	0.890829694	1.117230205
81	election ⇒ anc	0.029293593	0.887168142	1.112638084
82	well ⇒ anc	0.015998247	0.886639676	1.111975312
83	b ⇒ anc	0.011907371	0.881081081	1.105004024
84	da,votes ⇒ anc	0.01672876	0.877394636	1.100380685
85	vote ⇒ anc	0.087223318	0.876651982	1.099449289

No.	Rule	Support	Confidence	Lift
86	better ⇒ anc	0.013441449	0.876190476	1.098870493
87	members ⇒ anc	0.012857038	0.875621891	1.098157404
88	gauteng ⇒ anc	0.013368398	0.875598086	1.098127549
89	da,elections ⇒ anc	0.021623201	0.868035191	1.088642577
90	news ⇒ anc	0.01183432	0.861702128	1.080699993
91	country ⇒ anc	0.012710936	0.861386139	1.080303697
92	far ⇒ anc	0.013076193	0.860576923	1.079288823
93	dont ⇒ anc	0.030243261	0.858921162	1.077212257
94	know ⇒ anc	0.016436555	0.858778626	1.077033496
95	myanc ⇒ anc	0.017970633	0.857142857	1.074982004
96	back ⇒ anc	0.021330996	0.856304985	1.07393119
97	voting ⇒ anc	0.025641026	0.856097561	1.07367105
98	think ⇒ anc	0.012345679	0.853535354	1.070457669
99	lol ⇒ anc	0.014391117	0.852813853	1.069552802
100	parties ⇒ anc	0.011907371	0.848958333	1.064717419
101	nkandla ⇒ anc	0.01724012	0.848920863	1.064670426
102	even ⇒ anc	0.016947914	0.846715328	1.061904364
103	via ⇒ anc	0.010373292	0.845238095	1.060051698
104	say ⇒ anc	0.029147491	0.843551797	1.057936835
105	good ⇒ anc	0.020965739	0.841642229	1.055541958
106	president ⇒ anc	0.012345679	0.84079602	1.054480689
107	want ⇒ anc	0.015925195	0.838461538	1.051552909
108	time ⇒ anc	0.013222295	0.837962963	1.050927623
109	government ⇒ anc	0.018189787	0.83557047	1.047927088
110	may ⇒ anc	0.015121631	0.834677419	1.046807072
111	never ⇒ anc	0.017020966	0.832142857	1.043628362
112	corruption ⇒ anc	0.020600482	0.831858407	1.04327162
113	must ⇒ anc	0.020089123	0.825825826	1.035705885
114	party ⇒ anc	0.035356856	0.824531516	1.034082632
115	one ⇒ anc	0.016947914	0.819787986	1.028133554

No.	Rule	Support	Confidence	Lift
116	africa,south ⇒ anc	0.017532325	0.819112628	1.027286557
117	africa ⇒ anc	0.019504712	0.816513761	1.024027199
118	da,vote ⇒ anc	0.02103879	0.813559322	1.020321902
119	im ⇒ anc	0.015852144	0.812734082	1.019286931
120	elections ⇒ anc	0.050551538	0.812206573	1.018625357
121	economy ⇒ anc	0.010081087	0.811764706	1.018071192
122	like ⇒ anc	0.027905618	0.81104034	1.017162731
123	helenzille ⇒ anc	0.018408941	0.810289389	1.01622093
124	go ⇒ anc	0.01183432	0.805970149	1.010803974
125	us ⇒ anc	0.022718971	0.805699482	1.010464517
126	get ⇒ anc	0.018555044	0.80126183	1.004899055
127	"" ⇒ anc	0.797355541	0.797355541	1
128	still ⇒ anc	0.019796917	0.794721408	0.996696413
129	eff,vote ⇒ anc	0.012857038	0.792792793	0.994277649
130	see ⇒ anc	0.010373292	0.78021978	0.97850926
131	would ⇒ anc	0.013368398	0.778723404	0.976632586
132	national ⇒ anc	0.019285558	0.754285714	0.945984163
133	cant ⇒ anc	0.016436555	0.745033113	0.934380053
134	da ⇒ anc	0.12667105	0.744845361	0.934144585
135	u ⇒ anc	0.018993352	0.732394366	0.918529224
136	years ⇒ anc	0.018920301	0.719444444	0.902288136
137	freedom ⇒ anc	0.013441449	0.702290076	0.880774059

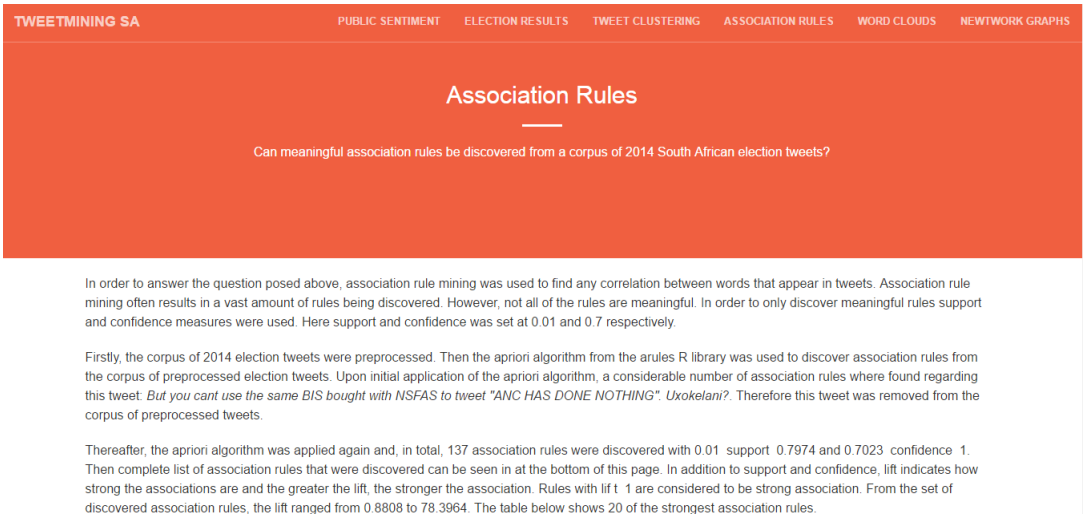

Figure 6: Partial screen shot of the page displaying the association rules discovered

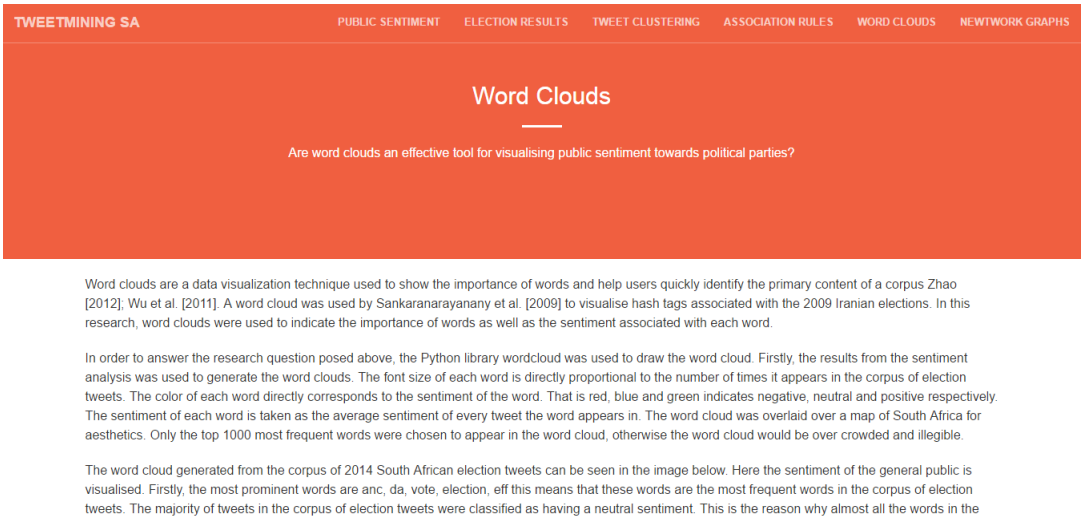

Figure 7: Partial screen shot of the page displaying the generated word clouds

References

[Academia.edu 2014] Academia.edu. *About Academia.edu*. `http://www.academia.edu/about`, 2014. Accessed: 2014/08/24.

[Ackerman 2016] Dan Ackerman. *Twitter beats national polls for election predictions, prof claims*. `https://www.cnet.com/news/professor-says-twitter-is-better-than-polls-for-election-predictions/`, 2016. Accessed: 2017/08/13.

[Agarwal *et al.* 2011] Apoorv Agarwal, Boyi Xie, Ilia Vovsh, Owen Rambow, and Rebecca Passonneau. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[Bakliwal *et al.* 2013] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. In *Proceedings of the Workshop on Language in Social Media*, pages 49–58, 2013.

[Beauchamp 2017] Nicholas Beauchamp. Predicting and Interpolating State-Level Polls Using Twitter Textual Data. *American Journal of Political Science*, 61(2):490–503, 2017.

[Bicen and Uzunboylu 2013] Hüseyin Bicen and Hüseyin Uzunboylu. The Use of Social Networking Sites in Education: A Case Study of Facebook. *J. UCS*, 19(5):658–671, 2013.

[Bifet and Frank 2010] Albert Bifet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. In *International Conference on Discovery Science*, pages 1–15. Springer, 2010.

[Biznik 2014] Biznik. *Why Biznik*. `http://biznik.com/about`, 2014. Accessed: 2014/08/24.

[Boundless 2016] Boundless. *Limited Government*. `https://www.boundless.com/political-science/textbooks/boundless-political-science-textbook/the-constitution-and-the-founding-of-america-2/the-constitution-26/limited-government-160-1849/`, 2016. Accessed: 2017/08/13.

[Boyd and Ellison 2007] Danah M. Boyd and Nicole B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[Byler 2014] David Byler. *9/11 Anniversary Speeches: What Text Analysis Tells Us*. `https://www.realclearpolitics.com/articles/2014/09/10/911_anniversary_speeches_what_text_analysis_tells_us_123913.html`, 2014. Accessed: 2017/08/13.

[Cagliero and Fiori 2013] Luca Cagliero and Alessandro Fiori. Discovering generalized association rules from Twitter. *Intelligent Data Analysis*, 17(4):627–648, 2013.

[Cassinelli and Chen 2009] Andrés Cassinelli and Chih-Wei Chen. *CS224N Final Project Boost up! Sentiment Categorization with Machine Learning Techniques*, 2009.

[Chapman *et al.* 2015] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM*

1.0. Step-by-step data mining guide. http://www.the-modeling-agency.com/crisp-dm.pdf, 2015. Accessed: 2015/01/11.

[Chevalier 2009] Max Chevalier. *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling: Techniques for Improved User Modeling*. IGI Global, 2009.

[Chrzanowski and Levick 2012] Mike Chrzanowski and Daniel Levick. *Using Twitter to Predict Voting Behavior*. https://core.ac.uk/display/22365395, 2012. OAI Identifier: oai:CiteSeerX.psu.10.1.1.278.3856.

[Chung and Mustafaraj 2011] Jessica Elan Chung and Eni Mustafaraj. Can Collective Sentiment Expressed on Twitter Predict Political Elections? In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, volume 11, pages 1770–1771, 2011.

[Clifton 2009] Christopher Clifton. *Data Mining*. https://www.britannica.com/technology/data-mining, 2009. Accessed: 2017/08/13.

[Coronel-Salas and Sanmatín 2016] Gabriela Coronel-Salas and Catalina Mier Sanmatín. Impact of Ibero-American Science and Technology in Twitter. *Revista Latina de Comunicación Social*, (71):668, 2016.

[Crudden 2002] Adele Crudden. Employment after Vision Loss: Results of a Collective Case Study. *Journal of Visual Impairment & Blindness (JVIB)*, 96(09), 2002.

[Department of Justice 1996] Department of Justice. *The Constitution Of The Republic Of South Africa, 1996*. http://www.justice.gov.za/legislation/constitution/SAConstitution-web-eng.pdf, 1996. Accessed: 2018/02/10, ISBN: 978-0-621-39063-6.

[DiGrazia *et al.* 2013] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and

Fabio Rojas. More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLOS ONE*, 8(11), 2013.

[DoItYourself 2014] DoItYourself. *About DoItYourself*. `http://www.doityourself.com/aboutus`, 2014. Accessed: 2014/08/24.

[Du 2010] Hongbo Du. *Data Mining Techniques and Applications: An Introduction*. Cengage Learning, 2010.

[Dufour and Calland 2016] Nathan Dufour and Richard Calland. *South Africa Local Government Election 2016: A Three-Layer Extrapolation for 2019 - Mining the depth & scope of ANC decline*. `https://www.eisa.org.za/eu/eu2016extrapolation.htm`, 2016. Accessed: 2017/08/13.

[Edmodo 2014] Edmodo. *Edmodo*. `https://www.edmodo.com`, 2014. Accessed: 2014/08/24.

[Facebook 2014] Facebook. *Facebook Info*. `https://www.facebook.com/facebook/info`, 2014. Accessed: 2014/08/24.

[Feldman and Sanger 2007] Ronen Feldman and James Sanger. *THE TEXT MINING HANDBOOK: Advanced Aapproaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[Feuer *et al.* 2002] Michael J. Feuer, Lisa Towne, and Richard J. Shavelson. Scientific Culture and Educational Research. *Educational researcher*, 31(8):4–14, 2002.

[Flickr 2014] Flickr. *About Flickr*. `https://www.flickr.com/about`, 2014. Accessed: 2014/08/24.

[Fusch and Ness 2017] Gene E. Fusch and Lawrence R. Ness. How to Conduct a Mini-Ethnographic Case Study: A Guide for Novice Researchers. *The Qualitative Report*, 22(3):923–941, 2017.

[Gaber 2016] Ivor Gaber. Twitter: A Useful Tool for Studying Elections? *Convergence*, 2016.

[Garner 1995] Stephen R Garner. WEKA: The Waikato Environment for Knowledge Analysis. In *Proceedings of the New Zealand Computer Science Research Students Conference*, pages 57–64. Citeseer, 1995.

[Gupta and Lehal 2009] Vishal Gupta and Gurpreet S. Lehal. A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies In Web Intelligence*, 1(1):60–76, 2009.

[Hamil 2014] James Hamil. Zuma's Scandals Threaten ANC, South Africa With 'Lost Decade'. *World Politics Review*, pages 2–5, 2014.

[Han *et al.* 2011] Jiawei Han, Jian Pei, and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[Hancock and Algozzine 2015] Dawson R. Hancock and Bob Algozzine. *Doing Case Study Research: A Practical Guide for Beginning Researchers*. Teachers College Press, 2015.

[Haustein *et al.* 2016] Stefanie Haustein, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1):232–238, 2016.

[Hipp *et al.* 2000] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for Association Rule Mining A General Survey and Comparison. *ACM SIGKDD Explorations Newsletter*, 2(1):58–64, 2000.

[Ho 2016] Ricky Ho. *Common Text Mining Workflow*. `https://dzone.com/articles/common-text-mining-workflow`, 2016. Accessed: 2017/08/13.

[Hotho *et al.* 2005] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief Survey of Text Mining. In *Ldv Forum*, volume 20, pages 19–62, 2005.

[Howard *et al.* 2011] Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil M. Hussain, Will Mari, and Marwa Maziad. *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?* Social Sciences Research Network (SSRN), 2011. Project on Information Technology and Political Islam.

[Hutto and Gilbert 2014] Clayton J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[Ipsos 2015] Ipsos. *Ipsos Poll South Africans dissatisfied with government*. http://www.ipsos.co.za/SitePages/IpsosPollSouthAfricansdissatisfiedwithgovernmentperformanceinkeypolicyareas.aspx, 2015. Accessed: 2015/08/15.

[Ipsos 2016] Ipsos. *Ipsos: Statement by Mari Harris on the Election Outlook: Ipsos "poll of polls"*. http://www.polity.org.za/article/ipsos-statement-by-mari-harris-on-the-election-outlook-ipsos-poll-of-polls-2016-08-02, 2016. Accessed: 2017/08/13.

[Irfan *et al.* 2015] Rizwana Irfan, Christine K. King, Daniel Grages, Sam Ewen, Samee U. Khan, Sajjad A. Madani, Joanna Kolodziej, Lizhe Wang, Dan Chen, Ammar Rayes, Nikolaos Tziritas, Cheng-Zhong Xu, Ablert T. Zomaya, Ahmed S. Alzahrani, and Hongxiang Li. A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(2):157–170, 2015.

[Kalnins 1986] ZG Kalnins. An exploratory study of the meaning of life as de-

scribed by residents of a long-term care facility. 1986. Project proposal, Peabody College of Vanderbilt University.

[Kaplan and Haenlein 2010] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1):59–68, 2010.

[Kincannon 2002] Joyce M. Kincannon. From the Classroom to the Web: A Study of Faculty Change. In *American Educational Research Association Annual Meeting*, 2002.

[Lankard and McLaughlin 2003] Annemarie Lankard and William J. McLaughlin. Marketing an Environmental Issue: A Case Study of The Wilderness Society's Core Messages to Promote National Forest Conservation from 1964 to 2000. *Society & Natural Resources*, 16(5):415–434, 2003.

[Larose 2014] Daniel T. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, 2014.

[Larsson and Moe 2012] Anders Olof Larsson and Hallvard Moe. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, 14(5):729–747, 2012.

[Lever 1974] Henry Lever. Opinion Polling in South Africa: Initial Findings. *The Public Opinion Quarterly*, 38(3):400–408, 1974.

[LinkedIn 2014] LinkedIn. *About LinkedIn*. `http://press.linkedin.com/products`, 2014. Accessed: 2014/08/24.

[Liu *et al.* 1998] Bing Liu, Wyne Hsu, and Yiming Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998.

[Lopez and Kalita 2017] Marc Moreno Lopez and Jugal Kalita. Deep Learning applied to NLP. *arXiv:1703.03091*, 2017.

[MacArthur 2016] Amanda MacArthur. *The Real History of Twitter, In Brief*. https://www.lifewire.com/history-of-twitter-3288854, 2016. Accessed: 2017/08/13.

[Mattes 2012] Robert Mattes. Opinion Polls and the Media in South Africa. In *Opinion Polls and the Media*, pages 175–197. Springer, 2012.

[Maurer and Wiegmann 2011] Christian Maurer and Rona Wiegmann. Effectiveness of Advertising on Social Network Sites: A Case Study on Facebook. In *ENTER*, pages 485–498, 2011.

[Monti *et al.* 2013] Corrado Monti, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson, and Elanor Colleoni. Modelling Political Dissaffection from Twitter Data. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 3. ACM, 2013.

[Muntean 2015] Alina Muntean. The Impact of Social Media Use of Political Participation. August 2015. MA in Corporate Communication, Aarhus University.

[MySpace 2014] MySpace. *About MySpace*. https://myspace.com/pressroom/aboutmyspace, 2014. Accessed: 2014/08/24.

[Nasa 2012] Divya Nasa. Text Mining Techniques- A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(4):50–54, 2012.

[Nations 2017] Polling The Nations. *A Brief History of Polling*. http://poll.orspub.com/static.php?type=about&page=briefhistory2, 2017. Accessed: 2017/08/13.

[Nevill 2014] Glenda Nevill. *The vote on South Africa's social media elections*. http://themediaonline.co.za/2014/05/the-vote-

on-south-africas-social-media-elections/, 2014. Accessed: 2016/09/16.

[Piatetsk 2014] Gregory Piatetsk. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, 2014. Accessed: 2016/08/24.

[Picasa 2014] Picasa. *Picasa*. http://picasa.google.com, 2014. Accessed: 2014/08/24.

[Raju and Kumar 2014] K. Venkateswara Raju and Dr. D. Prasanna Kumar. Social Media Marketing and its Role in Enhancing a Brand's Equity. *SAARANSH RKG Journal of Management*, 5:64–76, 2014.

[Reprobate 2016] Reprobate. *South Africa Election 2014: the social media wars*. http://reprobate.co.za/south-africa-election-2014-the-social-media-wars/, 2016. Accessed: 2016/08/22.

[Rosa *et al.* 2011] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gersham, and Robert Frederking. Topical Clustering of Tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.

[Russell 2013] Matthew A. Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, Inc., 2013.

[Salamon 2003] Sonya Salamon. From Hometown to Nontown: Rural Community Effects of Suburbanization. *Rural Sociology*, 68(1):1–24, 2003.

[Sankaranarayanany *et al.* 2009] Jagan Sankaranarayanany, Hanan Samety, Benjamin E. Teitlery, Michael D. Liebermany, and Jon Sperlingz. Twitterstand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International*

Conference on Advances in Geographic Information Systems, pages 42–51. ACM, 2009.

[Scrapbook.com 2014] Scrapbook.com. *About Scrapbook.com*. `http://www.scrapbook.com/about`, 2014. Accessed: 2014/08/24.

[Shoaib and Mujtaba 2016] Shandana Shoaib and Bahaudin G. Mujtaba. Use It or Lose It: Prudently Using Case Study as a Research and Educational Strategy. *American Journal of Education and Learning*, 1(2):83–93, 2016.

[SportShouting.com 2014] SportShouting.com. *About SportShouting.com*. `http://www.sportshouting.com`, 2014. Accessed: 2014/08/24.

[Statista 2014] Statista. *Global social networks ranked by number of users 2014*. `http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, 2014. Accessed: 2014/09/06.

[Sumathy and Chidambaram 2013] K.L. Sumathy and M. Chidambaram. Text Mining: Concepts, Applications, Tools and Issues - An Overview. *International Journal of Computer Applications*, 80(4), 2013.

[SuperGreenMe 2014] SuperGreenMe. *About SuperGreenMe*. `http://www.supergreenme.com/go-green-environment-eco:About-Us`, 2014. Accessed: 2014/08/24.

[Tan *et al.* 2006] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education India, 2006.

[Thome 1993] Barrie Thome. Political Activist As Participant Observer: Conflicts Of Commitment In A Study Of The Draft Resistance Movement Of The 1960's. *Contemporary Field Research: A Collection of Readings*, pages 216–234, 1993.

[Tracey 2013] Lauren Tracey. *Will social media influence election campaign-*

ing in South Africa? https://www.issafrica.org/iss-today/will-social-media-influence-election-campaigning-in-south-africa, 2013. Accessed: 2016/08/22.

[Tumasjan *et al.* 2010] Andranik Tumasjan, Timm O. Sprenger, and Isabell M. Welpe Philipp G. Sander. Predicting Elections with Twitter: What 140 Characters reveal about Political Sentiment. *Proceedings of the Fourth International AAI Conference on Weblogs and Social Media*, 10(1):178–185, 2010.

[Twitter 2014] Twitter. *About Twitter*. https://about.twitter.com, 2014. Accessed: 2014/08/24.

[George Gallup 2000] George Gallup. *The Gallup Poll*. Scholarly Resources Inc, 2000.

[Smith 1990] Tom W. Smith. The first straw? A study of the origins of election polls. *Public Opinion Quarterly*, 54(1):21–36, 1990.

[Utsey *et al.* 2003] Shawn O. Utsey, Alexis Howard, and Otis Williams III. Therapeutic Group Mentoring with African American Male Adolescents. *Journal of Mental Health Counseling*, 25(2):126–139, 2003.

[Wang *et al.* 2012] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[Who 2014] Who's Who. *About Who's Who of Southern Africa*. http://whoswho.co.za/about, 2014. Accessed: 2014/08/24.

[Worx 2012] World Wide Worx. *Social Media breaks barriers in SA*. http://www.worldwideworx.com/socialmedia2012/, 2012. Accessed: 2015/08/15.

[Wu *et al.* 2011] Yingcai Wu, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma. Semantic-Preserving Word Clouds by Seam Carving. In *Computer Graphics Forum*, volume 30, pages 741–750. Wiley Online Library, 2011.

[Yang and DeHart 2016] Hongwei "Chris" Yang and Jean L. DeHart. Social Media Use and Online Political Participation Among College Students During the US Election 2012. *Social Media + Society*, 2(1):1–18, 2016.

[Yin 2003] Robert K. Yin. *Case Study Research: Design and Methods*. Sage publications, 2003.

[YouTube 2014] YouTube. *About YouTube*. `https://www.youtube.com/yt/about`, 2014. Accessed: 2014/08/24.

[Zaki and Hsiao 1999] Mohammed J. Zaki and Ching-Jui Hsiao. *CHARM: An Efficient Algorithm for Closed Association Rule Mining*. Technical report, Technical Report 99, 1999.

[Zhao 2012] Yanchang Zhao. *R and Data Mining: Examples and Case Studies*. Academic Press, 2012.

[Zingla *et al.* 2014] Meriem Amina Zingla, Mohamed Ettaleb, Chiraz Latiri, and Yahya Slimani. INEX2014: Tweet Contextualization Using Association Rules between Terms. In *CLEF (Working Notes)*, pages 574–584, 2014.