Sentiment classification and an approach to sentiment visualisation

Zoë-Mae Adams

Report presented in partial fulfilment

of the requirements for the degree of

MCom Statistics

at the University of Stellenbosch

Supervisor: Dr Johané Nienkemper-Swanepoel

Degree of confidentiality: A

December 2022

PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.
2. I agree that plagiarism is a punishable offence because it constitutes theft.
3. I also understand that direct translations are plagiarism.
4. Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.
5. I declare that the work contained in this assignment, except otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

Z Adams	08 June 2022
Initials and surname	**Date**

ACKNOWLEDGEMENTS

I would like to express deepest appreciation to my supervisor, Dr Johané Nienkemper-Swanepoel. I am grateful for her patience, understanding and valuable feedback. This endeavour would not have been possible without the generous support from the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS).

I would like to extend my sincere thanks to my parents, Ferdi and Lucretia, for their support at home and understanding my decision to pursue an academic career. I am also thankful for all four of my siblings for checking in and knowing when to ask about my progression on this study. Lastly, I'd like to mention my best friends, my found family for offering stellar emotional support and encouraging me to not give up.

SUMMARY

The social media platform, Twitter, presents a great amount of text data regarding social interactions from the Tweets posted by users. The user-generated text data contains opinions and sentiments that are considered to be biased towards the users' individual and community experiences. In this study, text data related to the COVID-19 pandemic is procured from Twitter. The Tweets are utilised in two respective case studies. The first case study uses Tweets posted from three South African cities and the second case study uses Tweets posted from three countries. The selected cities are Cape Town, Durban and Johannesburg. The selected countries are South Africa, Australia and the United Kingdom. The subjective nature of the text leads to the use of sentiment classification to gain insight from the observed text data as well as expose the meaning and context. Sentiment classification entails matching the pre-processed text (i.e. text elements) to terms and phrases in a sentiment lexicon to determine their sentiment polarities. This study considers two sentiment lexicons: Bing and AFINN. Sentiment visualisation is concerned with summarising the content and underlying meaning within the text as well as displaying the distinct sentiments. This study explores and enhances two existing text visualisation tools: word clouds and multiple correspondence analysis (MCA) biplots. These visualisations are used to analyse the content and gauge the underlying sentiment within the text. Word clouds provide an overview of the occurrences of words in a given context. The word clouds are systematically enhanced by colour coding words according to their associated sentiment categories to reveal not only the most relevant topics in the text, but also the overall sentiment. In order to evaluate the dominant sentiment of the text, the word clouds are further enhanced to only display the words that are matched to the Bing sentiment lexicon. Considering that fear and uncertainty were identified as relevant topics related to the pandemic, the overall sentiment within the Tweets is reflected as negative. The sentiment classification results along with additional relevant categorical variables are compiled into a categorical dataset suitable for MCA and biplot visualisation. In its simplest form, a biplot is regarded as a generalised scatterplot which allows the visualisation of observations on more than two variables simultaneously. In this study, the MCA biplot will enable the investigation of the relationships among the Tweets and the levels of the categorical variables. The proximity of points in the biplot display suggests similar response profiles and associations between the category levels under investigation. The categorical variables considered in the case studies, include the location the Tweet was sent from, the overall sentiment categories per Tweet and the number of words in each Tweet classifiable by the sentiment lexicon. The standard MCA biplot is enhanced through word embedding which additionally displays the classifiable words along with the levels of the categorical variables. The number of words considered for classification is found to influence the overall

sentiment classification of the Tweet. The embedded word MCA biplot confirmed the consistency of the sentiment classification through the close proximity of category levels representing similar sentiment scores. Words with similar sentiment are also located in close proximity which eases the interpretation of the underlying meaning of the Tweets. Overall, the biplots reveal that the number of words influence the strength of the sentiment classification, seeing that a larger number of classifiable words in the Tweets is more likely to lead to a neutral sentiment due to the averaging of sentiment scores to determine the overall sentiment of a particular Tweet. The methodology enables the visualisation of a quantified measure of sentiment along with the associated words. These promising results therefore add to the developing field of sentiment visualisation through the enhancement of existing text visualisation tools to visualise sentiments within the text.

Keywords:

biplots; multiple correspondence analysis; Twitter; sentiment classification; sentiment visualisation; word clouds

OPSOMMING

Die sosiale media platform, Twitter, bied ‘n groot hoeveelheid teks data aan met betrekking tot die sosiale interaksies vanaf die *Tweets* wat geplaas word deur gebruikers. Die teks data gegenereer deur die gebruikers bevat menings en sentimente wat beskou word as bevooroordeeld teenoor die gebruikers se individuele en gemeenskapservarings. Die teks data vir hierdie studie, wat verband hou met die COVID-19 pandemie, word verkry vanaf Twitter. Die *Tweets* word gebruik in twee onderskeie gevallestudies. Die eerste gevallestudie gebruik *Tweets* geplaas vanaf drie Suid-Afrikaanse stede en die tweede gevallestudie gebruik *Tweets* geplaas vanaf drie lande. Kaapstad, Durban en Johannesburg is gekies as die stede vir die eerste gevallestudie. Suid-Afrika, Australië en die Verenigde Koninkryk is gekies as die lande vir die tweede gevallestudie. Die subjektiewe aard van die teks data lei tot die gebruik van sentiment klassifikasie om insig te verkry uit die waargenome teks data asook om die betekenis en konteks in die teks te onthul. Sentiment klassifikasie behels die vergelyking van die vooraf verwerkte teks (d.w.s. teks elemente) met terme en frases in ‘n sentiment leksikon om die sentiment kategorieë te bepaal. Hierdie studie oorweeg twee sentiment leksikons: Bing en AFINN. Sentiment visualisering behels die opsomming van die inhoud en die onderliggende betekenis in die teks asook die voorstelling van die afsonderlike sentimente. Hierdie studie verken en verbeter twee bestaande metodes vir teks visualisering: woord wolke en meervoudige ooreenkomsanalise (Eng. MCA) bi-stippings. Hierdie visualiseringsmetodes word gebruik om die inhoud te ontleed en die onderliggende sentiment in die teks te meet. Woord wolke bied ‘n oorsig van die aantal kere wat woorde voorkom in ‘n gegewe konteks. Die woord wolke word sistematies verbeter deur die woorde se kleure te kodeer volgens hul ooreenstemmende sentiment kategorieë om die mees relevante onderwerpe en die omvattende sentiment te openbaar. Om die oorheersende sentiment van die teks te evalueer, word die woord wolke verder verbeter deur slegs die woorde voor te stel wat ooreenstemmende inskrywings het in die Bing sentiment leksikon. Aangesien vrees en onsekerheid geïdentifiseer word as relevante onderwerpe wat verband hou met die pandemie, word die algehele sentiment in die *Tweets* gereflekteer as negatief. ‘n Kategoriese datastel geskik vir MCA en bi-stipping visualisering word saamgestel vanuit die resultate van die sentiment klassifikasie en addisionele relevante kategoriese veranderlikes. ‘n Bi-stipping in sy eenvoudigste vorm word beskou as ‘n veralgemeende verspreidingsdiagram wat die gelyktydige visualisering van waarnemings vir meer as twee veranderlikes moontlik maak. In hierdie studie, word die MCA bi-stipping gebruik om die verwantskappe tussen die *Tweets* en die vlakke van die kategoriese veranderlikes te ondersoek. Die nabyheid van die punte in die bi-stipping is ‘n aanduiding van soortgelyke responsprofiele en verhoudings tussen die kategorie vlakke wat ondersoek word. Die kategoriese veranderlikes wat oorweeg word in die

gevallestudies sluit die ligging waarvandaan die *Tweet* gestuur is in, die algehele sentiment kategorieë per *Tweet* en die aantal woorde wat deur die sentiment leksikon geklassifiseer kan word. Die standaard MCA bi-stipping word verbeter deur woordinbedding wat die klassifiseerbare woorde saam met die vlakke van die kategoriese veranderlikes vertoon. Die ingebedde woord MCA bi-stipping bevestig die konsekwentheid van die sentiment klassifikasie deur die nabyheid van kategorie vlakke wat gelyksoortige sentiment tellings verteenwoordig. Woorde met gelyksoortige sentimente is ook in nabyheid geleë wat die interpretasie van die onderliggende betekenis van die *Tweets* vergemaklik. In die algemeen, onthul die bi-stippings dat die aantal woorde die sterkte van die sentiment klassifikasie beïnvloed, aangesien ‘n groter aantal klassifiseerbare woorde in die *Tweets* meer geneig is om ‘n neutrale sentiment te hê omdat die gemiddelde sentiment telling gebruik word om die algehele sentiment van ‘n spesifieke *Tweet* te bepaal. Die metodologie maak dit moontlik om die gekwantifiseerde sentiment saam met die geassosieerde woorde te visualiseer. Hierdie belowende resultate dien as ‘n bydrae tot die ontwikkelende navorsingsveld van sentiment visualisering deur die verbetering van bestaande teksvisualiseringsmetodes om die sentiment in teks te visualiseer.

Sleutelwoorde:

bi-stipping; meervoudige ooreenkomsanalise; Twitter; sentiment klassifikasie; sentiment visualisering; woord wolk

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
CA	Correspondence Analysis
COVID-19	Coronavirus Disease 2019
HTML	HyperText Markup Language
MCA	Multiple Correspondence Analysis
MDS	Multi-dimensional Scaling
NRC	National Research Council Canada
POS	Parts of speech
SVD	Singular Value Decomposition
TDM	Term Document Matrix
URL	Uniform Resource Locators

CHAPTER 1
INTRODUCTION

1.1 RATIONALE

The Coronavirus Disease 2019 (COVID-19) pandemic has significantly altered how people communicate due to the regulations such as lockdowns and social distancing. Advancing digital technology and social media increasingly aid in human interaction and virtual communication through text, speech and images shared online. This results in the increased availability of digital text data. The meaning of the text and its underlying context provides useful information to interested parties, since this useful information may contain subjective and factual information regarding, for example politics, social events, marketing campaigns and product preferences. Online platforms allow for information to be shared more frequently, which results in increased exposure to popular topics (Cambria, Schuller, Xia, *et al.*, 2013; Kaur & Gupta, 2013). An advantage of these online platforms is the accessibility. Since any user can share their thoughts, however, it introduces the disadvantage of diminished credibility of the information being shared. There is also the risk of the information being false, especially if the user is not an expert on the matter (Hui & Gregory, 2010). Considering this, it is safe to assume that the posts shared consist of opinions and subjective information. The micro-blogging service, Twitter, allows masses of users to communicate through short messages directly from any web-based service. Throughout this dissertation, the term “Tweet” will be used to refer to these short messages posted on Twitter. This aids users to generate information in real time to a large audience (Mendoza, Poblete & Castillo, 2010).

A large collection of Tweets provides a mass of text data that contains valuable information. One of the methods of gaining insight from the text is through classification. Text classification refers to assigning the text to a set of categories such as sentiment categories (Jurafsky, 2012). Sentiment analysis is a field of study which focuses on the detection, analysis and evaluation of sentiments. This is done by extracting opinions, attitudes, sentiments and emotions from the observed text (Yadollahi, Shahraki & Zaiane, 2017). Sentiment analysis can thus be used to determine the overall sentiment of text data to provide an overview on the content and context within the text.

Granted that the sentiment classification provides adequate information for the text data and the underlying sentiments, text data visualisation is used to analyse high-dimensional text data by transforming the data into a lower-dimensional space while retaining the interactions of the content within the text (Kim & Lee, 2014). However, simply using text data visualisation methods for sentiment analysis will not account for the meaning in the text. Sentiment visualisation forms apart of text data visualisation and focuses on visualising the sentiments

in the text. Limited sources of literature on sentiment visualisation exists given that it is still a developing research field (Kim & Lee, 2014; Kucher, Paradis & Kerren, 2018).

This study therefore combines gauging sentiment from available text data by performing sentiment classification and the promotion of visual exploration of sentiments by contributing to the research on sentiment visualisation.

1.2 PROBLEM STATEMENT

The abundance of user-generated content available on online platforms, such as Twitter, results in a plethora of available text data containing sentiment and opinions. The content ranges over various topics and are posted in real time. Due to the nature, frequency and reduced credibility of the available content, it becomes challenging to filter through and extract sentiments and opinions in terms of computational complexity and limited knowledge bases.

Although the increased exposure of popular topics and accessibility to online social interaction is advantageous on an individual and societal level, the disadvantage lies within separating fact from fiction. Additionally, these social media platforms (including those of news outlets) are progressively designed to cater to the users' interests, which results in their opinions and sentiments being biased towards their individual and community experience.

Therefore, Twitter not only provides researchers with large amounts of text data, but the opportunity to observe social patterns, societal trends, mental health indicators and other subjective information. The subjective nature of the text leads into the use of sentiment classification to gain insight from the observed text data as well as expose the meaning and context. The sentiments conveyed by the Tweets can be summarised by means of sentiment visualisation. Different from tabular or textual summaries of data, the visualisation of the data allows for the display of previously unseen trends, patterns and variation in the data (Friendly, 2008). Sentiment visualisation entails obtaining a lower-dimensional display of the text data while preserving the associations among the words in the collections of Tweets and allows for the graphical display of the distinct sentiments.

1.3 DATA

This study focuses on Tweets, which are regarded as text data and mainly consists of free-flowing text in the form of paragraphs, sentences and words where the number of characters per Tweet is limited. This study uses the term corpus to refer to a collection of Tweets.

Twitter is a social media platform that allows users to individually post information (as well as images, videos and links) about a variety of topics, limited to 280 characters (Kwartler, 2017). The data for this study was extracted using Twitter's application programming interface (API). The API provides a user with access to a specified number of random Tweets. The word,

Tweet, acts as a verb and a noun, representing the action of posting on the Twitter platform and the resulting post.

As a pilot study, one thousand Tweets were obtained with the keyword “covid” from Twitter for three different locations and were posted between 09/09/2021 and 16/09/2021. The locations were specified using coordinates and included all areas within a 50-mile radius. The objective was to obtain Tweets about the COVID-19 pandemic from users residing or situated in three different cities in South Africa. The selected cities were Johannesburg, Cape Town and Durban. To increase the sample size, an additional three thousand Tweets were obtained with the keyword “covid” for three different countries over an extended period and were posted between 03/10/2021 and 01/11/2021. The selected countries were South Africa, Australia and the United Kingdom.

These Tweets are used as the pieces of text considered for the sentiment classification and visualisations. The first case study will utilise the Tweets from South African cities to demonstrate the research methodology in CHAPTER 3 and the second case study will present the application of the methodology on the Tweets obtained from the different countries in CHAPTER 4.

The underlying sentiment of the Tweets will be based on the sentiment classification of two sentiment lexicons, namely the Bing lexicon and the AFINN lexicon. The Bing lexicon consists of 6786 existing subjective or opinion related terms with corresponding classifications into negative or positive sentiment categories. There are 4781 negative terms and 2005 positive terms (Hu & Liu, 2004). The latest version of the AFINN lexicon consists of 2477 terms with associated scores ranging between -5 and 5, where -5 indicates a term with a very negative sentiment and 5 indicates a term with a very positive sentiment (Nielsen, 2011).

1.4 AIM AND STUDY OBJECTIVE

This study aims to gain insight from text data, specifically Tweets, through summarising the content and visualising the sentiment classification.

This will be achieved by means of two study objectives. First, a simple text visualisation tool, word clouds, will be explored and enhanced by incorporating the sentiment classification to gain insight from the summarised text.

World clouds are frequently used to provide a quick summary of text data. These visualisations are useful to provide an overview of the occurrences of words in a given context, however, the underlying meaning is forfeited. In order to gain insight from text this popular tool will be enhanced to incorporate the sentiment of the words along with the frequency at which they occur.

The second objective is to adapt an advanced multivariate visualisation tool, multiple correspondence analysis (MCA) biplots. The multivariate display enables the simultaneous representation of the Twitter users and multiple factors relating to their Tweets. The distances between coordinates in the multivariate display expresses similarity between response patterns. This will allow the insight into the similarity of the sentiment of Twitter users, based on their Tweets, considering additional factors included in the input data matrix.

1.5 CHAPTER OUTLINE

The first chapter is an introduction to the study, which explains the rationale, problem statement, the type of data used in the study and the objectives to achieve the aim of the study. The second chapter provides a literature review of the background applicable to the study before focusing on sentiment classification and the visualisation of the sentiment within the text. The complete methodology is presented in the third chapter, which includes the explanation of the data procurement from Twitter, pre-processing and the required visualisation techniques. Although the standard word clouds and MCA biplots are suitable for visualisation of the text data, enhancements are considered to highlight additional information on the sentiments and address the objectives of this study. The results and graphs obtained after the application of all techniques explained in the research methodology are presented and discussed in the fourth chapter. The fifth chapter contains the conclusion and further recommendations given the results of the study.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

This chapter contains a reflection of the relevant literature applicable to the study, where Section 2.4 and Section 2.5 are of special interest. These sections explain the use of sentiment classification to gain insight from text data and the methods in which the results can be graphically displayed while preserving the meaning contained within the text.

2.2 THE ADVANTAGE OF SOCIAL MEDIA PLATFORMS

Naturally occurring unstructured data generated by people are available on social media platforms. This data is thus accessible to a potentially vast, geographically diverse participant pool in real-time at a low cost for researchers. The social nature of the users' interactions results in the observed text containing natural language (Ignatow & Mihalcea, 2018).

The social networking platform, Twitter, was launched in July 2006 (History.com, 2019) and has become an important source of information that is readily available to a mass of users. The popularity of Twitter, and other similar networks such as Facebook and LinkedIn, resulted in researchers launching studies ranging from understanding the messages to determining the geographical reach of the network. Java *et al.* (2007) researched the intentions of users and the community structure the social network provides. Their text classification study reveals that people use Twitter for sharing their daily routines and activities, commenting on existing posts, sharing links to other websites and reporting current events. Another study using data obtained on Twitter by Kwak *et al.* (2010) explored the rate at which Tweets become popular based on the number of retweets (reposted Tweets), the influence of the user who initially posted the Tweet and whether it is a news headline.

Twitter enables researchers to observe user-generated texts containing information on each user's opinion about their social reality. This is considered to be an accurate reflection of opinions, since users can voluntarily share their interests and attitudes on the platform (Ignatow & Mihalcea, 2018; Jansen, Zhang, Sobel, *et al.*, 2009).

Since there is a character limit imposed on Tweets, this prompts users to become innovative in articulating the meaning behind their thoughts, which results in text loaded with meaning being condensed into fewer sentences (Hong & Davison, 2010). Considering that the users have the freedom to choose the way in which they articulate the meaning in their Tweets, this could lead to misinterpretation of the underlying meaning in the Tweet by other users or the researcher. The timely and accessible nature of Twitter results in the network being used effectively as a source of broadcasting information on events such as natural disasters and

pandemics, but still presents the issue of credibility raised in Section 1.1 (Lazard, Scheinfeld, Bernhardt, *et al.*, 2015; Mendoza *et al.*, 2010). Twitter is thus a powerful source of subjective text data, for which sentiment classification is an appropriate technique for analysis.

2.3 THE ANALYSIS OF TEXT DATA

According to Ignatow & Mihalcea (2018), formal statistical methods and human interpretation are combined in text analysis. Text analysis involves the extraction of useful insight from unstructured text data (Feldman & Sanger, 2007; Kwartler, 2017; Radovanovic & Ivanovic, 2008). There are six main approaches currently being adopted in the field of text analysis. One is conversation analysis, which focuses on using language to describe daily situations and settings. The second and third approaches are analysis of discourse positions and critical discourse analysis. These methods concentrate on linking the observed text to their social spaces or other spaces. Foucauldian analysis, the fourth approach, identifies the meaning of the text concerning similar discourses. The fifth approach is analysing text as social information, which treats the text as a reflection of the author's practical knowledge since user-generated texts contain current information about social reality. Lastly, the sixth approach is content analysis. Content analysis gains information that goes beyond the literal presentation of the text. This can also be used to identify attitudes, views and interests of not only individuals but groups of authors (Ignatow & Mihalcea, 2018). The approaches that are applicable to this study are the fifth and sixth approaches. The observed Tweets in this study are analysed as social information as well as for the identification of attitudes, views and interests related to the COVID-19 pandemic.

The origin of text analysis can be traced back to a team of friars (male members of the Dominican Roman Catholic order) led by Hugh of Saint-Cher that created the first cross-listing of terms and concepts in the Bible in the early 1200's (Ignatow & Mihalcea, 2018). Krippendorff (2004) states that the first systematic studies of text were conducted in the 1600's by academics in theology. These theological studies analysed the content of newspapers in fear of the circulation of nonreligious printed matter. In the 1700's, the content of a collection of popular hymns was investigated due to controversy in Sweden (Dovring, 1954). The expansion of the field consists mainly of research in social sciences and humanities. Although the examination of text originated in the social sciences, the development of text analysis techniques advanced within the field of computer science as the complexity of text mining tasks increased. Text mining typically involves acquiring the text, tagging the parts of speech (POS) in the text, determining context and extracting attitudinal information. (Ignatow & Mihalcea, 2018).

According to Kwartler (2017), the process of text analysis starts with collecting the data that is appropriate to the study. For example, if the study aims to analyse text from scientific reviews, the text would be observed from academic journals, unlike observing text data from social media platforms for a study on opinions about a popular topic. There are two ways in which text can be analysed. The first is the bag-of-words method, where every word is treated as a separate element in the piece of text without taking word order and POS into account. The analysis is then performed on these elements. Word clouds (cf. Section 2.5.1) are an example of a technique that treats the words in the text as elements ignoring word order and POS. The second method is referred to as syntactic parsing. This method tags each word element as POS to identify the words themselves in a linguistic context. Tagging the words as their POS creates the sentence components which represent the data utilised for analysis. Sentiment classification uses syntactic parsing by identifying adjectives that encompass attitude or emotion. After the pieces of text are deconstructed into elements for the analysis, pre-processing takes place. Pre-processing includes changing all the text to lowercase and removing punctuation. It also allows text data to be converted from a sequence of characters into patterns of encoded bits that can be analysed by computers and removes tags and unwanted elements found in online text. Examples are HTML tags that refer to keywords containing information on how the content on a web page is displayed on a web browser and Uniform Resource Locators (URL's) that allows users to be directed to a specific resource such as a web page (Mhatre, Phondekar, Kadam, *et al.*, 2017). Following the pre-processing step, the text is analysed according to the particular research objective. To address the first research objective (cf. Section 1.4), both the bag of words method for the word cloud visualisation and the syntactic parsing method for the sentiment classification is utilised. This is discussed further in Sections 2.4 and 2.5.1.

In 2015, some studies were conducted using text mining on observed Twitter data to reveal insights on public concerns about diseases. Lazard *et al*. (2015) aimed to understand the public concern after the first diagnosis of Ebola in the United States by grouping Tweets into related topics to identify whether these themes indicate overall concern or fear. Shepherd *et al*. (2015) assessed the manner in which Twitter can be used to communicate mental health concerns among users as well as provide feedback to mental health service providers by using content analysis to categorise the Tweets into themes such as the impact of diagnosis on personal identity and crisis planning. Similarly, this dissertation aims to use a form of text mining to explore the public concern about the COVID-19 pandemic using a sample of Tweets (cf. Section 1.3).

2.4 SENTIMENT CLASSIFICATION

Given a set of pre-defined categories, standard classification focuses on determining which category a quantitative observation belongs to. Examples of quantitative measures that can be classified into categories are test scores and observed temperatures. Test scores can be categorised into grade symbols A, B, C, D or F and temperatures can be classified into categories such as cold, cool, mild, warm, or hot. This study focuses on analysing text data, for which standard classification is not appropriate. Text classification refers to assigning the text to a set of categories, for example assigning a set of documents to a range of topics, different languages or different sentiment categories (Jurafsky, 2012). According to Pang, Lee and Vaithyanathan (2002), the majority of research on text classification focuses on classifying text documents according to their topics. The growing field of sentiment analysis (Mäntylä, Graziotin & Kuutila, 2018) is a specific type of text classification where the categories are specific to emotions or attitude. Public opinion or community sentiment has been relevant since ancient times when leaders were interested in measuring their popularity, for example voting in Athenian democracy, which dates back to the fifth century before the Common Era (Thorley, 2004). Traditionally, public opinion was measured using surveys and polls and related research started around the late 1930's and was political in nature. Early applications of modern sentiment analysis focused on product and movie reviews in the mid-2000's. Sentiment analysis has evolved to contribute to the prediction of financial markets, crisis control, sports, education, tourism and healthcare (Akcora, Bayir, Demirbas, *et al.*, 2010; Kim & Lee, 2014; Mäntylä *et al.*, 2018; Shayaa, Jaafar, Bahri, *et al.*, 2018). Current and future research in sentiment analysis encompasses improving algorithms using advancing computational power, building more resources, accounting for deeper emotional nuances in the text and visualising the results of sentiment analysis (Cambria *et al.*, 2013; Medhat, Hassan & Korashy, 2014). This study contributes to the current state of sentiment analysis by providing an additional method to visualise the results of sentiment classification.

Sentiment analysis aims to identify and extract emotional intent from text. Therefore, the meaning and context of the text becomes more essential to the analysis (Kwartler, 2017). Sentiment analysis is a field of study which focuses on the detection, analysis and evaluation of sentiments. This is done by extracting opinions, attitudes, sentiments and emotions from observed communications, such as writings, speech and facial expressions (Yadollahi *et al.*, 2017). This study focuses on analysing the sentiment of text from writings, in the form of Tweets.

Since sentiment analysis entails classifying text into categories, it is thus considered as a sentiment classification problem (Medhat *et al.*, 2014). The analysis starts with pre-processing, which focuses on the extraction and selection of some text elements and removal

of the text elements that have no impact on the general orientation of the classification (Ganesh, 2019; Mhatre *et al.*, 2017). The individual words in the text are extracted along with the frequency at which those individual words occur. The first type of element extraction involves tagging the words in the text as POS and extracting adjectives, since POS information primarily aids in determining which meaning of a word is activated when used in a particular context. Adjectives often indicate sentiment since it describes how users feel about the noun they are describing and can thus guide the process to classify the sentiment (Cambria *et al.*, 2013). Commonly used opinion words are the second type of text feature to be extracted. Examples of commonly used opinion words are “like”, “dislike”, “terrible” and “great”. Negative words, such as “bad” and “no” are another type of text element to be extracted since negative words can alter the orientation of the opinion stated in a sentence (Medhat *et al.*, 2014). The next step focuses on the application of a method which can perform classification while also taking into account how the words in the text are associated with one another (Ganesh, 2019). This study uses MCA to analyse the results of the sentiment classification as well as other categorical variables. MCA deals with examining the associations among the sample points and variables, similar to how sentiment classification is concerned with the associations among the text elements (Section 2.5.2). The remaining words thus contain subjective and relative information and are compared to a sentiment lexicon. The sentiment lexicon contains terms and phrases along with their known sentiment scores and polarities. The text elements are matched to terms in the sentiment lexicon to identify their contextual polarity and subjectivity (Cambria *et al.*, 2013). Sentiment lexicons differ with respect to the terms included therein. The effectivity of the sentiment lexicon is dependent on the language being used in the text it is being matched to (Hu & Liu, 2004; Nielsen, 2011). There are various sentiment lexicons available and further discussion will be directed to the sentiment lexicons available in the `tidytext` package in `R`. The lexicons available in the `tidytext` package are the Bing, AFINN, Loughran-McDonald and the NRC Word-Emotion Association lexicons (Silge & Robinson, 2016). The Bing sentiment lexicon has been compiled by Bing Liu since 2004 in a study that used sentiment classification to summarise customer reviews (Hu & Liu, 2004). The terms in the Bing sentiment lexicon are either categorised as positive or negative. The AFINN lexicon was initially created and manually scored by Nielsen (2011) in 2009 for a sentiment analysis for Tweets related to the UN Climate Conference. The terms in the AFINN lexicon have corresponding sentiment scores ranging between -5 and 5, where -5 indicates a very negative sentiment and 5 indicates a very positive sentiment. Loughran & McDonald (2011) created a word list to account for the differences in the meanings of words in a financial context. The Loughran-McDonald sentiment lexicon classifies terms into one of six categories, namely: positive, negative, constraining, litigious, superfluous and uncertainty. The National Research Council in Canada funded the creation of a sentiment lexicon that identifies the

emotions associated with a list of common words (Mohammad & Turney, 2010). This lexicon will henceforth be referred to as the NRC sentiment lexicon. The terms in the NRC sentiment lexicon thus contain the list of words and the corresponding emotion or sentiment that they evoke. The categories for the NRC lexicon are eight basic emotions and two sentiments, which are anger, fear, anticipation, trust, surprise, sadness, joy, disgust, negative and positive. After consideration of the various sentiment lexicons, this study will utilise both the Bing and the AFINN sentiment lexicons (cf. Section 1.3). The Tweets concerning the COVID-19 pandemic can be observed as user reviews about the topic of the pandemic, therefore, the Bing sentiment lexicon is suitable since it was initially created to categorise customer reviews. Similar to the study by Nielsen (2011), this study performs sentiment classification on Tweets related to an event (the COVID-19 pandemic), thus the AFINN sentiment lexicon is appropriate for application in this study.

2.5 VISUALISATION

Simply using text data visualisation methods for sentiment analysis will not account for the meaning in the text. Sentiment visualisation aims to differentiate among the sentiments within the text and to graphically display the conceptual similarity of the sentiments in a lower-dimensional space. Words with similar sentiments will be displayed in close proximity and words with different sentiments will be displayed further apart (Cambria, Song, Wang, *et al.*, 2014; Kim & Lee, 2014; Kucher *et al.*, 2018). Research in sentiment visualisation is developing and thus limited sources of literature exist. Kim and Lee (2014) proposed a semi-supervised nonlinear dimensionality reduction method to increase the accuracy of the sentiment classification as well as visualise the sentiments of Amazon reviews. Cambria *et al*. (2014) and Osgood *et al*. (1975) applied multi-dimensional scaling (MDS) to observe the semantic similarity among different terms and display those words as points in a lower-dimensional space where the distances between the points represent the similarities. This study proposes two approaches to sentiment visualisation by enhancing existing visualisation techniques. These are word clouds (cf. Sections 2.5.1) and MCA biplots (cf. Sections 2.5.2). These sentiment visualisation tools will allow us to summarise the sentiments conveyed by the Tweets and evaluate the results of the sentiment classification.

2.5.1 Simple displays

The use of word clouds originated in a study by Stanley Milgram nearly 50 years ago. Subjects had to submit hand-drawn maps of Paris containing all the city's elements that they could think of immediately. The objective was to obtain a map of Paris containing the names of each element shown in a size proportional to the frequency at which the subjects included it in their

hand-drawn maps (Lee, Riche, Karlson, *et al.*, 2010; Milgram & Jodelet, 1976). The original map is shown in Figure 2.1:

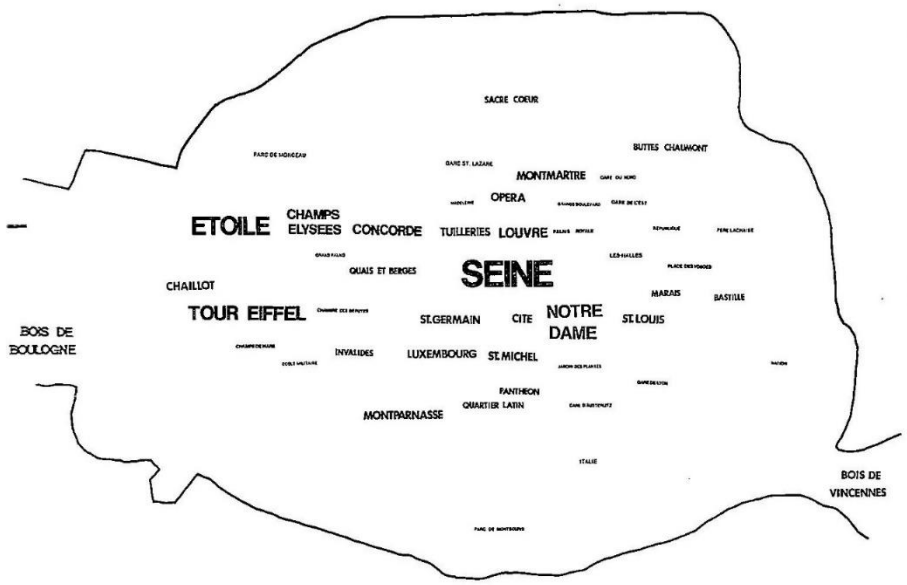


Figure 2.1: The most frequently cited elements of Paris according to the hand-drawn maps submitted by the subjects[1]

Word clouds provide users with a summary of the content within a large piece of text data (Seifert, Kump, Kienreich, *et al.*, 2008). This visualisation of text data consists of a cloud of words, in which the words are depicted in different font sizes and colours, which represent different attributes of the words in the data. (Halvey & Keane, 2007; Kwartler, 2017). Traditionally, the term frequency is represented by the size and location of the word in the cloud. The larger and bolder the word appears, the more important or relevant the word is in the text source (or corpus). Font sizes are selected given the constraints of the size of the display (i.e. available plotting area) and legibility of the text (Seifert *et al.*, 2008).

Word clouds are also referred to as tag clouds, wordles or weighted lists in the available literature. Although word clouds are a popular visualisation tool which simplifies the comprehension of text to general audiences, the tool has become misused and used too frequently. The removal of some words during pre-processing could lead to biased conclusions as a consequence of the varied interpretation between viewers (Kwartler, 2017). The length of the word becomes conflated with its size, which results in longer words appearing smaller in size relative to shorter words. Additionally, the typical word cloud layout does not capture the meanings or contexts of the words and only considers the frequency of the words in the text. Since words with similar meaning are not deliberately positioned in close

[1] Source: (Milgram & Jodelet, 1976)

proximity in the display, some words with similar meaning might be easily overlooked (Hearst & Rosner, 2008).

The methodology to construct a world cloud of the observed Tweets to provide a comprehensive summary, will be discussed in Sections 3.3 and 3.4.2.

2.5.2 Multivariate displays

Firstly, consider scatterplots, which are well-known visualisations for continuous bivariate variables. These visualisations are typically used to visually inspect the relationships among sample points (typically the rows of the matrix) for two continuous variables (typically the columns of the matrix) in a single display (Friendly & Denis, 2005). If the number of variables exceed two, the same analysis is possible by producing an array of scatterplots. This array results in a difficult inspection as the individual scatterplots become cumbersome to read as the number of variables increase. A remedy to this problem is to construct a biplot. Gabriel (1971) published a paper in which the use of the biplot as a visual evaluation of the characteristics of a matrix was presented. Each row and column are represented by respective vectors and graphed onto a representative plane. A biplot thus allows for the simultaneous visual inspection of samples as well as variables. The sample points and variables are seen as the two modes highlighted in the prefix, “bi”, in biplot. The biplot can be regarded as a generalised scatterplot in which more than two variables can be displayed as non-orthogonal axes. The biplot reveals the similarity among the points based on the distance between the points. The interrelationships among the variables are added onto the same plotting area by either additional points or a set of axes as the second mode of the biplot. The angles between the axes are determined by the correlation between the variables (Johnson & Wichern, 2014; Jolliffe, 2002). Biplots can be constructed for various data types and can include various features depending on the required outcome (Gower, Gardner-Lubbe & Le Roux, 2011).

In the context of the analysis, the metric space which represents meanings of linguistic units (words) is known as the semantic space (Leopold, 2007). The semantic space, like any k-dimensional mathematical space, consists of objects treated as coordinates with the relationships between these points, defining the nature of the semantic space (Tsirelson, 2018). The origin of the semantic space refers to a point that lacks any meaning or semantic quality. Meanings of words can be represented by a meaning vector from the origin, where the length represents the degree of meaningfulness and the direction the semantic quality. Additional to the length and direction of the meaning vectors, reference coordinates for the position in the k-dimensional space are required (Osgood *et al.*, 1975). This idea aligns with the classic biplot approach by Gabriel (1971).

MDS is a method of dimension reduction that attempts to retain the global structure of the data by ensuring that most of the variation in the data is represented in a lower-dimensional space. This is achieved by obtaining an optimal visual representation in a lower-dimensional space with approximate distances between points to represent the observed dissimilarities (i.e. distances or proximities) (Cox & Cox, 2001; Heiser & Meulman, 1983). MDS maps differ with regard to input data, distance measure and the resulting output (Van der Klis & Tellings, 2021). Biplots are related to MDS, but differ in the sense that the aim is to approximate the data matrix and not the dissimilarities in a lower-dimensional visualisation (Gower *et al.*, 2011).

Initially, biplots were developed for continuous data with further generalisations by Gower (1992) enabling the use of biplots for categorical variables and mixed data, which contains both categorical and continuous variables and this study focuses on categorical data. Categorical biplot displays consist of two types of coordinates which represent the responses and the category levels per variable, respectively.

The construction of biplots, the plotting positions of the rows and columns, relies on singular value decomposition (SVD) of the data. The difference lies within the nature of the data matrix the SVD is calculated from.

The sentiment classification results can be summarised in the format of a multivariate categorical data matrix (cf. Section 3.5.1). Categorical data consisting of two variables can be visualised by means of a correspondence analysis (CA) map. For CA the SVD is found for a matrix of residuals and points are plotted to optimally approximate chi-squared distances (Jolliffe, 2002). According to Greenacre (2017), the theory of CA was originally published amid the 1960's by Jean-Paul Benzécri and his co-workers. CA allows a low-dimensional graphical representation of a frequency table by redefining the dimensions of the space such that the maximum variance is captured by the reduced number of dimensions. Coordinate points in a CA map that appear in close proximity indicate similar profiles, or context (Blasius, Greenacre, Gower, *et al.*, 2006; Johnson & Wichern, 2014).

CA can be extended to MCA for the application on multivariate categorical data. MCA allows one to study the magnitude and direction of the interrelationships between responses and variables, represented by the rows and columns of a data matrix, respectively (Greenacre, 2017). In this project, MCA will be performed by applying CA on an indicator matrix. The indicator matrix consists of zeros and ones, to specify the responses per category level. The technique relies on SVD of the weighted indicator matrix to visualise the high dimensional data set in two dimensions (Gower *et al.*, 2011).

The coordinates in a biplot can be expressed as either principal or standard coordinates (cf. Section 3.5.2). Symmetric biplots result from expressing the rows and columns of the data

matrix as the same type of coordinate (i.e. standard or principal) in the display. It is typical to visualise CA maps as symmetric biplots, since there is no true difference in the mode of the rows and columns of a two-way contingency table. Whereas for a multivariate categorical data set, the rows and columns represent different modes and should be expressed in an asymmetrical biplot with different coordinates used for the samples and variables (Blasius *et al.*, 2006; Gower *et al.*, 2011; Greenacre, 2017)

The inherent high dimensionality of text data introduces difficulty when performing sentiment analysis (Gentzkow, Kelly & Taddy, 2019). The initial classification problem considers many words of which some express sentiment and some that do not. Words with no sentiment become irrelevant to the sentiment classification, which introduces redundancy and unwanted repetitive information. This can be addressed by extracting the irrelevant text elements, which simplifies the data and as a result improves the performance of the classification (Mhatre *et al.*, 2017; O'Keefe & Koprinska, 2009; Sharma & Dey, 2012). This study uses the information gained from the text data as input to obtain a lower-dimensional solution with sample coordinates representing extracted words from the data and variable coordinates representing the sentiment categories. The MCA biplot allows the visual exploration of the interrelationships among categorical variables by representing the sentiment category levels (and additional variables cf. Table 3.5) as a set of points along with a set of points representing the cases within the dataset. The analysis consists of identifying groups of words along with an associated sentiment category level to determine which words are associated with each sentiment category level (Greenacre, 2010, 2017). The position of the words relative to other words could reflect similarity of context and meaning (Lowe, 2001).

2.6 CONCLUSION

This chapter reviewed and reflected upon the relevant literature to highlight the objectives of the study. Considering the need for more research on sentiment classification on Twitter data and sentiment visualisation, the following CHAPTER 3 will discuss the research methodology of sentiment classification and the enhancement of existing visualisation techniques as an approach to sentiment visualisation.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 INTRODUCTION

This section will include relevant information on the process of obtaining the data, data preparation, sentiment classification, dimension reduction and visualisation. In order to reproduce the methodology described in this section, refer to the `R` Code and functions in the APPENDIX B.

3.2 DATA PROCUREMENT AND PRE-PROCESSING

The Tweets are obtained using the methods described in Chapter 1.3. The `R` package, `twitteR` provides the user with an interface to the Twitter web API which allows the extraction of Tweets that are not older than a week at the time of extraction (Gentry, 2015; R Core Team, 2021). The `R` package, `tm`, offers functionality for managing text documents, condenses the process of document manipulation and eases the usage of heterogenous text formats in `R` (Feinerer & Hornik, 2020). The typical steps in pre-processing will be shown by means of an example.

Consider the i^{th} Tweet, denoted by $\mathbf{t}_i$.

@xxx: Actually you are wrong. The majority of people dying of Covid are not vaccinated. You need to get your facts right https://t.co/xXxyYzZ

After removing punctuation marks, special characters and links, the remaining words are recoded into lower case and presented below.

actually you are wrong the majority of people dying of covid are not vaccinated you need to get your facts right

The process of tokenisation refers to breaking down a string of words into separate elements or tokens. These tokens then become a collection of text elements to be analysed. The `R` package, `tidytext`, allows for the conversion of text in each Tweet into tokens as shown below (Silge & Robinson, 2016):

“actually”	“you”	“are”	“wrong”	“the”	“majority”	“of”
“people”	“dying”	“of”	“covid”	“are”	“not”	“vaccinated”
“you”	“need”	“to”	“get”	“your”	“facts”	“right”

For consideration in the classification, the stop words are removed from this collection. Stop words refer to common words in the English language that aid in forming the sentences but do not contribute to the underlying meaning of the sentence. Examples of stop words are “the” and “and” (Leskovec, Rajaraman & Ullman, 2019). The stop words in $\mathbf{t}_i$ are identified as the words covered with a red cross below:

“actually”	“you”	“are”	“wrong”	“the”	“majority”	“of”
“people”	“dying”	“to”	“covid”	“are”	“not”	“vaccinated”
“you”	“need”	“to”	“get”	“your”	“facts”	“right”

Note that the word “not” is considered a stop word. The word, however, also operates as a negation cue, which can shift the sentimental orientation of the terms observed in the text (Reitan, Faret, Gambäck, *et al.*, 2015). This term and other negational cues such as “no” are thus significant for sentiment classification, which will be discussed in Section 3.4. Although this step aids in formatting the data, human intervention is still required to ensure a more accurate classification. The block below contains the remaining words after the stop words are removed.

“wrong”	“majority”	“people”	“dying”
“covid”	“vaccinated”	“facts”	“right”

The i^{th} Tweet thus becomes a character vector of word tokens w_{ij}, for $j = 1, \dots, m$, where m represents the number of remaining words after pre-processing as indicated in $\mathbf{t}_i$.

$$\mathbf{t}_i = \begin{bmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{ij} \\ \vdots \\ w_{im} \end{bmatrix} = \begin{bmatrix} \text{wrong} \\ \text{majority} \\ \text{people} \\ \text{dying} \\ \text{covid} \\ \text{vaccinated} \\ \text{facts} \\ \text{right} \end{bmatrix}$$

By means of pre-processing the initial piece of text can be converted to a vector which is suitable for further analysis.

3.3 WORD CLOUDS

The first exploratory analysis considers the frequency of words by means of word clouds (cf. Section 2.5.1). A separate notation is considered for this section and should not be confused with the notation in Section 3.2. Consider a collection of Tweets extracted from one city or

country, T^{LOC}, where the set of Tweets is denoted by T and the location by the superscript LOC . The words, w_l^{LOC}, in T^{LOC} are each a function of the string of characters, w_l^s, and the relevance of that word, denoted, w_l^r.

$$T^{LOC} = \{w_l^{LOC} = f(w_l^s, w_l^r)\}$$

$$\text{for } l = 1, \dots, M$$

The value of M is the size of the corpus created from T^{LOC} (cf. Sections 1.3 and 2.5). The `R` function, `wordcloud`, requires the collection of the pre-processed terms in the corpus and the frequencies at which they occur therein (Fellows, 2018).

The layout of the word cloud is created subject to the constraints related to font size and the length of the strings. These constraints are accounted for by introducing the following parameters:

- v_{min}: the minimum font size allowed
- v_{max}: the maximum font size allowed
- r_{min}^0: the initial value for the font size assigned to the least relevant word
- $r_{max}^0 = v_{max}$: the initial value for the font size assigned to the most relevant word, which is simply equal to the maximum font size allowed

It should be noted that the relevance of the words is represented by the frequencies at which they occur in the text T^{LOC}. The font size of the words cannot be too small, since it will affect the legibility of the word in the visualisation. A font size too large will result in the words taking up too much space in the layout, which reduces the number of words to be displayed in the word cloud.

The relevance of the word is determined by first constructing a term document matrix (TDM) from a corpus of T^{LOC}. The rows of the TDM represent each word in the corpus and the columns each Tweet in T^{LOC}. The TDM contains the frequencies at which w_l^{LOC} occurs in $\mathbf{t}_i$.

$$\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & \dots & t_n \end{array} \\ \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \end{bmatrix} \end{array}$$

The frequency at which each word, w_l, appears in T^{LOC} is calculated as the row totals of the TDM and subsequently be referred to as the relevance values.

Table 3.1: Example of words and corresponding relevance values obtained from row sums of the term document matrix

Word	**Relevance value**
vaccine	24
people	15
news	15
⋮	
microscopic	1
sanlam	1
variants	1

Table 3.1 is ordered according to the relevance values in descending order, where the word with the largest value is the most relevant and the word with smallest value the least relevant.

The optimal layout of the word cloud is achieved heuristically by taking as input $w_l^{LOC} = f(w_l^s, w_l^r)$, v_{min}, v_{max} and $r_{min}^0$and changing the parameters until the following criteria are met:

- The most relevant word is the largest in size and in the centre of the layout with the lesser relevant words surrounding it
- The words do not overlap in the display
- The font sizes correspond to the relevance of the word

The most relevant word is placed in the centre of the cloud and the remaining words placed around it in decreasing relevance. An example of a resulting word cloud is presented in Figure 3.1 below. Take note that the identified cuss words and strong language are censored in all graphs constructed for this study.

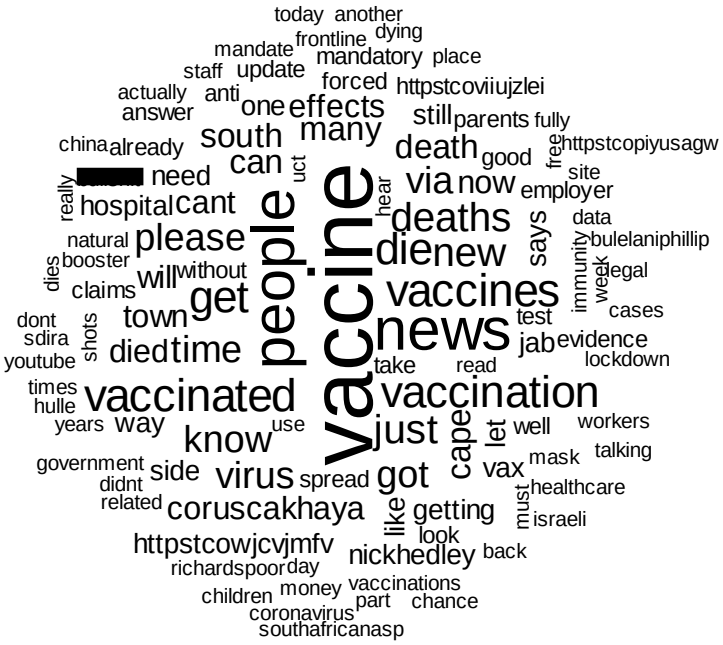


Figure 3.1: An example word cloud for COVID-19 related Tweets

A visual interpretation of a word cloud entails identifying the largest words in the cloud as the most relevant word in the collection of Tweets. The smaller words in the cloud are thus identified as the words occurring less frequently in the text. In the context of the study, the smaller words therefore occur less in the Tweets extracted using the keyword (i.e. covid, cf. Section 1.3). In the same way, the larger words then occur more frequently in the extracted Tweets. The word cloud thus acts as an initial summary of the text as a whole. As can be observed in Figure 3.1, the term “httpstcowjcvjmfv” has not been removed during the pre-processing step of the analysis. This could be due to a difficulty in which the nature of the character string hindered the process in which the string was to be identified containing “https”. Similarly, some usernames remained in the corpus as users can decide on any combination of terms to identify themselves on social media platforms. Often, usernames include sentiment words, which will incorrectly be considered for sentiment classification. This is a limitation in the sense that pre-processing still requires human intervention.

Although this word cloud provides information on the relevant words in the set of Tweets, it does not provide insight into the emotive quality of the words. An extra dimension of colour is incorporated to expose the sentiment of the words to enhance the standard word cloud (cf Section 3.4.3).

3.4 SENTIMENT CLASSIFICATION

Section 3.2 introduced the pre-processing and formatting of the Tweets into vectors that can be used for analysis. The Tweet $\mathbf{t}_i$ contains the words (tokens) w_{ij}, for $i = 1, \dots, n$ and $j = 1, \dots, m$. The word cloud alone is an initial summary, but sentiment classification is performed

to gain insight regarding the underlying meaning and emotive quality of the words. As mentioned in Section 2.4, words are compared to terms in a sentiment lexicon to determine a sentiment score and consequently classify the word as positive, negative or neutral. The tokens created during the pre-processing step in Section 3.2 are matched to the words in the sentiment lexicon using the functions in the `R` package, `tidytext` (Silge & Robinson, 2016). This study performs sentiment classification based on two different sentiment lexicons: the AFINN lexicon and the Bing lexicon.

3.4.1 AFINN lexicon

Let L^A refer to the AFINN sentiment lexicon which consists of 2477 existing subjective and opinion related terms with corresponding sentiment scores ranging between -5 and 5 (cf. Section 2.4). Scores lower than 0 refer to negative sentiments and scores higher than 0 refer to positive sentiments. Scores equal to 0 are classified as neutral sentiments. Let the classes therefore be denoted by c_{-5}^A, c_{-4}^A, c_{-3}^A, c_{-2}^A, c_{-1}^A, c_0^A, c_{+1}^A, c_{+2}^A, c_{+3}^A, c_{+4}^A and c_{+5}^A. Consider an example where the pre-processed Tweet y, $\mathbf{t}_y$, contains the following words: “adequate”, “fair”, “approved”, “avoid” and “doubted” (Table 3.2). These words are then matched to terms in L^A and classified accordingly:

Table 3.2: Matching words in $\mathbf{t}_y$ to terms in L^A

Word	Sentiment score
adequate	1
fair	2
approved	2
avoid	-1
doubted	-1

To obtain an overall score for $\mathbf{t}_y$, a weighted mean is calculated and rounded to the nearest integer relating to an AFINN score. Let the frequency of words assigned to each class be denoted by f_{-5}^A, f_{-4}^A, f_{-3}^A, f_{-2}^A, f_{-1}^A, f_0^A, f_{+1}^A, f_{+2}^A, f_{+3}^A, f_{+4}^A and f_{+5}^A. The overall sentiment score for $\mathbf{t}_i$ is therefore determined by Equation 3.1:

Equation 3.1: Calculating the overall sentiment score using the AFINN sentiment lexicon

$$s_i^A = \sum_{\mathrm{j}=1}^{\mathrm{m}} c_j^A f_j^A \Big/ \sum_{j=1}^{m} f_j^A$$

Considering the example in Table 3.2, the weighted mean score is calculated as follows:

$$s_2^A = \frac{(-5\times 0)+(-4\times 0)+(-3\times 0)+(-2\times 0)+(-1\times 2)+(0\times 0)+(1\times 1)+(2\times 2)+(3\times 0)+(4\times 0)+(5\times 0)}{0+0+0+0+2+0+1+2+0+0+0}$$

$$= \frac{0+0+0+0-2+0+1+4+0+0+0}{5} = \frac{3}{5} = 0.6 \approx 1$$

Therefore, the overall sentiment score is one ($s_y^A = 1$), which indicates a positive sentiment score for Tweet y. Since categorical multivariate visualisations will be applied (cf. Section 2.5.2), the AFINN scores s^A will be categorised into five category levels centred around zero: $[-5,-3]$, $(-3,-1]$, 0, $[1,3)$ and $[3,5]$. This will be elaborated on in Section 3.5.1.

3.4.2 Bing lexicon

Let L^B refer to the Bing lexicon which consists of 6786 existing subjective and opinion related terms with corresponding classifications into negative or positive sentiment categories (cf. Section 2.4).

In this study, Tweets are classified into the categories c_-^B (negative), $c_\bullet^B$ (neutral) and c_+^B (positive). Consider the same set of words from the example in Section 3.4.2 in Table 3.2Table 3.2 .

Table 3.3: Matching words in $\mathbf{t}_y$ to terms in L^B

Word	Classification
adequate	positive
fair	positive
approved	positive
avoid	Not in lexicon
doubted	negative

The overall score, s_i^B, for $\mathbf{t}_y$ is obtained from the majority classification. Let the frequency of words assigned to each class be denoted by f_+^B and f_-^B.The overall sentiment score for $\mathbf{t}_i$ is therefore determined by Equation 3.2:

Equation 3.2: Calculating the overall sentiment score using the Bing sentiment lexicon

$$s_i^B = \begin{cases} c_+^B & \text{if } f_+^B > f_-^B \\ c_-^B & \text{if } f_+^B < f_-^B \\ c_\bullet^B & \text{if } f_+^B = f_-^B \end{cases}$$

Considering the example in Table 3.3, the frequencies of the word classification are $f_-^B = 1$ and $f_+^B = 3$. Since $f_+^B > f_-^B$, $s_y^B = c_+^B$. Tweet y is thus classified as positive.

3.4.3 Enhanced word clouds

The sentiment classification (cf. Section 3.4.2) can be incorporated in the standard word cloud by the addition of colour. This will enhance the visual interpretation of the underlying sentiment of the classified words.

Table 3.1 contains the words in the corpus and their corresponding relevance values relative to the other words in T^{LOC}. Each w_l^{LOC} is searched for in L^B to determine its sentiment category. Words that are not elements of L^B will be displayed in grey and cannot be classified. The classification rule and corresponding colours are summarised in Table 3.4.

Table 3.4: Colours used in enhanced word cloud based on sentiment category

Sentiment category of w_l^{LOC} according to L^B	Colour of the word in the word cloud	
w_l^{LOC} not found in L^B	grey	
positive	green	
negative	red	

The enhanced word cloud presented in Figure 3.2 does not only provide a summary of the relevant words used in the text, but also expose their emotive nature of the relevant words by utilising the classification according to the Bing lexicon (cf. Section 3.4.2). Apart from the terms “didnt”, “dont” and ”cant”, the unclassifiable words do not have distinctive polarity which confirms that the Bing lexicon effectively classifies the sentiments in the presented example. The term “dont” is an example of the pre-processing step modifying contraction terms due to the removal of the punctuation mark, the apostrophe, which results in the term not being considered for matching to the lexicon. Other notable unclassified words are “dying” and “dies”, which should be classified along with “death”, “die” and “died”, but are not contained in L^B. Contractions and sentiment words that are not contained in the lexicons present the limitations of the sentiment lexicons that require further development.

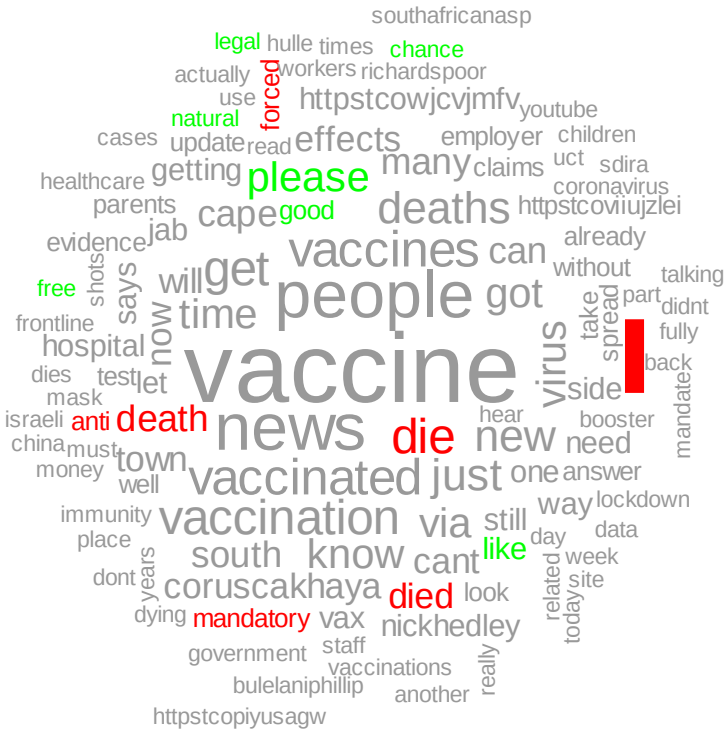


Figure 3.2: Enhanced word cloud with relevant words in corpus

To visualise the relevance of the classified words, a word cloud can also be created to exclude the unmatched words as presented in Figure 3.3. This type of word cloud highlights which of the emotive text in the corpus is more relevant to the text. The context of the raw data is however forfeited by excluding the unclassified words.

Figure 3.3: Enhanced word cloud with emotive relevant words in corpus

The word clouds now provide an improved summary of the relevant words in the text, while also considering the meaning and impact of these words. The word cloud containing only the emotive words indicate whether sentiments in the text tend to be positive or negative by identifying whether the majority of the words in the cloud are green or red.

The enhancements provide more insight to the text but does not take context or individual Tweets into account. The focus will now be on the multivariate exploration of the Tweets by explaining the methodology of how MCA is used to analyse and visualise the text in the Tweets.

3.5 MULTIPLE CORRESPONDENCE ANALYSIS

3.5.1 Data preparation

In this section the procedure of MCA will be discussed by first explaining the construction of the categorical multivariate data set in Section 3.5.1. The methodology of the MCA approach followed in this study will be presented in Section 3.5.2 for a general case, which will be followed by the notation applicable to this study. The construction of the MCA biplot along with discussions to aid interpretation will be presented.

Since MCA is applied on a multivariate categorical data set (cf. Section 2.5.2), the following categorical variables are created to represent the Tweets along with the sentiment classifications as discussed in the previous sections of this chapter. The results of the sentiment classification for the two lexicons are regarded as two categorical variables with the

Bing classification resulting in three category levels (cf. Section 3.4.2) and the AFINN classification in five category levels (cf. Section 3.4.1). The specified location at which the Tweet was created is a category level of a variable representing the location. The categorical data set was built systematically on sentiment classification using L^A due to its larger classification range (cf. Figure 4.6 and Figure 4.7). Each $\mathbf{t}_i$ consists of words that can be classified to either sentiment lexicon or words that cannot be classified. Therefore, the words classifiable according to the AFINN sentiment lexicon and the number of these words per Tweet are presented. The number of words is treated as a nominal categorical variable in this analysis in which the category levels represent the number of words. The sentiment classification using L^B was used additionally to evaluate the similarity of the sentiment classification categories. Table 3.5 is an excerpt of the data set that illustrates the construction of the multivariate categorical data set from the Tweet information (cf. Section 1.3) and sentiment classification (cf. Section 3.4).

According to Section 1.3, one thousand Tweets were procured from the cities and three thousand Tweets were procured from the countries. The data sets in the case studies, however, only contain the Tweets that were successfully classified according to both L^A and L^B. The sample sizes are thus noticeably smaller than the number of Tweets initially procured.

Table 3.5: Categorical data set obtained from Tweets' information and results from sentiment classification

Case	Original Tweet	Words considered for classification	Number of words considered for classification	Location	s_i^A	s_i^B
1	@xxx: Natural immunity after covid is far better than both vaccined	better_natural	2	CT	[1,3)	positive
2	@xxx: nice to know. I've just recovered from covid and clear. early detection and treatment saved me. No vaccine. natural Immunity.	clear_natural_nice_no_saved	5	DBN	[1,3)	positive
…						
n	@xxx: Anti-vaxx leader's COVID death could have been avoided, says brother - The Jerusalem Post	anti_avoided_death	3	JHB	(-3,-1]	negative

Table 3.5 contains six columns excluding the indices of the Tweets. The last four columns are considered as categorical variables. Emotive words are extracted from each Tweet during the pre-processing stage (cf. Section 3.2) and stored within the vector $\mathbf{t}_i$. These words are presented in the second column (Words considered for classification) of the data set separated by an underscore to facilitate the visualisation discussed in the Section 3.6. The third column contains the number of the words presented in the second column. For example, the words "better" and "natural" were extracted from $\mathbf{t}_1$ in Table 3.5. Since two emotive words are extracted and considered for sentiment classification, the number of words is equal to two.

The second categorical variable refers to the specific location from which the user sent the Tweet. A possible separation between users from different locations is envisioned due to the differences in healthcare and infrastructure in different cities in the same country.

The third and fourth categorical variables represent the overall sentiment scores using L^A (s_i^A) and L^B (s_i^B). Using the data excerpt from Table 3.5 as an example, the resulting scores from the classifications L^A and L^B agree as Tweets with a positive Bing score is expected to be associated with a positive AFINN score. For example, the n^{th} Tweet ($\mathbf{t}_n$) in Table 3.5 has an AFINN score of $(-3, -1]$ and a negative Bing score. The first Tweet ($\mathbf{t}_1$) in Table 3.5 contains two words considered for sentiment classification. According to Equation 3.2 and Equation 3.1, two positive words will result in an overall positive score, two negative words will result in an overall negative score and a combination of words with opposite sentiments will result in a neutral score or a score equal to 0. Although, L^A contained both the words "better" and "natural" and their corresponding sentiment scores 2 and 1. The resulting sentiment score for $\mathbf{t}_1$ using "better" and "natural" is calculated using Equation 3.1 is equal to 2. L^B, however, does not contain the term "natural" and therefore only calculates the sentiment score based on the term "better", that is classified as a positive term. When examining the Tweet that the words were extracted from, it would seem that the user has a positive sentiment toward natural immunity. This corresponds to the sentiment classification results from both lexicons. This affirms the effectivity of the sentiment classification since objectively the combination of these words would lead to an overall positive sentiment. Although both sentiment lexicons lead to similar results, L^A has a wider range of sentiment categories and can therefore indicate both the sentiment as well as some level of intensity compared to the two available categories presented in L^B.

Compared to $\mathbf{t}_1$, the second Tweet in Table 3.5, $\mathbf{t}_2$, contains five words considered for classification. Acknowledging the effect of the number of words on the classification, one can identify that a larger number of words will influence the overall sentiment score since it considers the overall sentiment of the majority of the words. The words "clear", "natural",

"saved" and "nice" have individual sentiment scores of 1 and larger, but the individual score for the word "no", that is equal to -1, brings down the overall score to a value that is negative. The n^{th} Tweet, $\mathbf{t}_n$, contains three negatively associated words and could have had a much lower score that falls within the bracket $[-5,-3]$, but the scores for "anti" and "avoided", that have higher individual scores, result in a higher sentiment score that falls within the bracket $(-3,-1]$. At this stage, one can foresee that an increase in the number of words can cause the discrimination among the sentiment categories to become less clear.

The results in Table 3.5 corroborates the effectiveness of the sentiment classification, but still reveals some limitations in capturing the context of the original Tweet. The "no" in the second Tweet indicates negation since the user is referring to "no vaccine", which does not indicate a negative sentiment. This is a limitation of sentiment analysis since the process of sentiment classification commences with tokenisation (cf. Section 3.2) and separates the "no" terms and automatically considers them as negative terms (Pang & Lee, 2008).

Despite the results presented in Table 3.5 leading to an insightful interpretation, visualisation as a tool would provide an overall summary of the results and allow one to explore additional trends within the data. The MCA biplot will thus be explained in the following sections.

3.5.2 MCA background

Let $\mathbf{X}$ be a multivariate categorical data matrix with n rows and q columns, where n represents the Tweets and q the categorical variables (cf. Section 3.5.1).

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

Let each categorical variable be denoted by Q_k for $k = 1, \ldots, q$. This categorical dataset is then converted into an indicator matrix, $\mathbf{G}$, which consists of the combined indicator matrices for each Q_k, denoted by $\mathbf{G}_k$. The category levels, J_k, determines the number of columns of each $\mathbf{G}_k$. The value one is indicated in the column representing the response for the particular variable (Q_k) and zeros are indicated in the remaining columns. The indicator matrix $\mathbf{G}$ will therefore have $J = \sum_{k=1}^{q} J_k$ columns.

Consider the following toy data matrix, $\mathbf{X}_{(3\times 3)}$:

$$\mathbf{X} = \begin{bmatrix} B & B & B \\ A & A & B \\ C & C & A \end{bmatrix}$$

The toy matrix, X, consists of three rows ($n = 3$) and three categorical variables (Q_k, where $k = 1, 2, 3$). Suppose that the first two variables have three category levels ($J_1 = J_2 = 3$) and

the third variable consists of two category levels ($J_3 = 2$). This will result in an indicator matrix, $\mathbf{G}$, with three rows and eight columns as shown in the matrices below:

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \mathbf{G}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \mathbf{G}_3 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\mathbf{G} = [\mathbf{G}_1 \quad \mathbf{G}_2 \quad \mathbf{G}_3]$$

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Two of the various approaches to perform MCA include performing CA on either the indicator matrix $\mathbf{G}$ or the Burt matrix, which is calculated as $\mathbf{G}'\mathbf{G}$.

This study applies MCA by performing CA on the indicator matrix. The following discussion explains the steps of performing CA on the indicator matrix. Ultimately, SVD is performed on a weighted indicator matrix, which will become apparent throughout this discussion.

First, consider the correspondence matrix $\mathbf{P}$:

$$\mathbf{P} = \frac{1}{nq}\mathbf{G}_{n \times J}$$

Each row of the $\mathbf{P}$ matrix will contain q nonzero entries for which the row total for each sample is calculated as

$$q \times \frac{1}{nq} = \frac{1}{n}$$

It then follows that the row totals of the $\mathbf{P}$ matrix is contained in the following vector, $\mathbf{r}_{n \times 1}$:

$$\mathbf{r} = \mathbf{P}\mathbf{1}_J = {}^1\!/_n\, \mathbf{1} = \begin{bmatrix} {}^1\!/_n \\ {}^1\!/_n \\ \vdots \\ {}^1\!/_n \end{bmatrix}$$

The following matrix $\mathbf{D}_r^{-1/2}$ is an inverted square root diagonal matrix of the vector $\mathbf{r}$:

$$1 \div \sqrt{{}^1\!/_n} = 1 \div {}^1\!/_{\sqrt{n}} = \sqrt{n}$$

$$\mathbf{D}_r^{-1/2} = \begin{bmatrix} \sqrt{n} & 0 & \cdots & 0 \\ 0 & \sqrt{n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{n} \end{bmatrix}$$

The column totals of $\mathbf{P}$ are represented in the vector $\mathbf{c}$. Since the column totals of $\mathbf{G}$ convey the frequencies for each level of the q categorical variables, the elements of $\mathbf{c}$ are calculated as these frequencies divided by nq.

$$\mathbf{c} = \mathbf{P}'\mathbf{1}_n = \begin{bmatrix} \sum_{i=1}^{n} p_{i1} \\ \sum_{i=1}^{n} p_{i2} \\ \vdots \\ \sum_{i=1}^{n} p_{iJ} \end{bmatrix} : J \times 1$$

$$\mathbf{D}_c^{-1/2} = \begin{bmatrix} 1/\sqrt{\sum_{i=1}^{n} p_{i1}} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{\sum_{i=1}^{n} p_{i2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{\sum_{i=1}^{n} p_{iK}} \end{bmatrix}$$

Similar to $\mathbf{D}_r^{-1/2}$, $\mathbf{D}_c^{-1/2}$ is an inverse square root diagonal matrix of the vector $\mathbf{c}$, where the diagonal contains the values $1/\sqrt{\sum_{i=1}^{n} p_{ij}}$, for $j = 1, \ldots, K$.

The matrices above are used to calculate the matrix of standardised residuals $\mathbf{S}$, of which the SVD is calculated:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \sqrt{n}(\mathbf{P} - \tfrac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{P})\mathbf{D}_c^{-1/2}$$

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

The coordinates for the sample points and the levels of the categorical variables in the first two dimensions are obtained by the first two columns of the matrices $\mathbf{F}$ and $\mathbf{\Gamma}$ respectively. The matrix $\mathbf{F}$ contains the principal coordinates for the sample points and the $\mathbf{\Gamma}$ matrix the standard coordinates for the category levels. The principal coordinates are obtained by multiplying the standard coordinates with the singular values ($\mathbf{D}$).

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}$$

$$\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$$

The `R` package, `ca`, was used to perform MCA and obtain the coordinates for the MCA biplots (Nenadic & Greenacre, 2007).

3.5.3 MCA of the COVID-19 related Tweets

As mentioned in Section 1.3, two case studies are conducted, where the first case study uses the data from South African cities to illustrate the research methodology and the second case study applies those techniques to the Tweets from different countries.

Case study 1:

$$n = 334$$

$$q = 4$$

Case study 2:

$$n = 1664$$

$$q = 4$$

$Q_1 =$ Number of words considered for classification (κ). The words considered for classification are the remaining subjective text elements after pre-processing which were matched to terms in both sentiment lexicons. The maximum number of words considered for classification are observed as five. $Q_1 = \{1,2,3,4,5\}$.

$Q_2 =$ Location. This variable simply indicates the city or country from which the Tweet was sent. Case study 1: $Q_2 = \{CT, DBN, JHB\}$ or Case study 2: $Q_2 = \{RSA, AUS, UK\}$.

$Q_3 =$ Categories for the overall AFINN sentiment score. $Q_3 = \{[-5,-3], (-3,-1], 0, [1,3), [3,5]\}$

$Q_4 =$ Categories for the overall Bing sentiment score. $Q_4 = \{positive, negative, neutral\}$

$$\mathbf{X} = \begin{bmatrix} 2 & CT & [1,3) & positive \\ 5 & DBN & [1,3) & positive \\ \vdots & \vdots & \vdots & \vdots \\ 3 & JHB & (-3,-1] & negative \end{bmatrix}$$

$$\mathbf{x}_1' = [2 \quad 5 \quad \cdots \quad 3]$$

$$\therefore \mathbf{G}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{G} = [\mathbf{G}_1 \quad \mathbf{G}_2 \quad \mathbf{G}_3 \quad \mathbf{G}_4]$$

Indicator matrix $\mathbf{G}$: $n \times 16$

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$nq = 334 \times 16 = 5344$$

The indicator matrix $\mathbf{G}$ is used to obtain the weighted indicator matrix and coordinates for the MCA biplot by applying the methodology detailed in Section 3.5.2.

3.5.4 MCA biplot

The MCA biplot allows us to investigate relationships among the sample points, represented by the grey circles (•) as well as the relationships among the category levels (blue symbols). Sample points, representing Tweets, appearing in close proximity indicate Tweets that are similar in context and meaning. Category level points occurring in close proximity indicate that those levels are associated.

If a grouping of sample points is in close proximity to a group of category levels, it suggests that those Tweets are associated with those levels of the categorical variable.

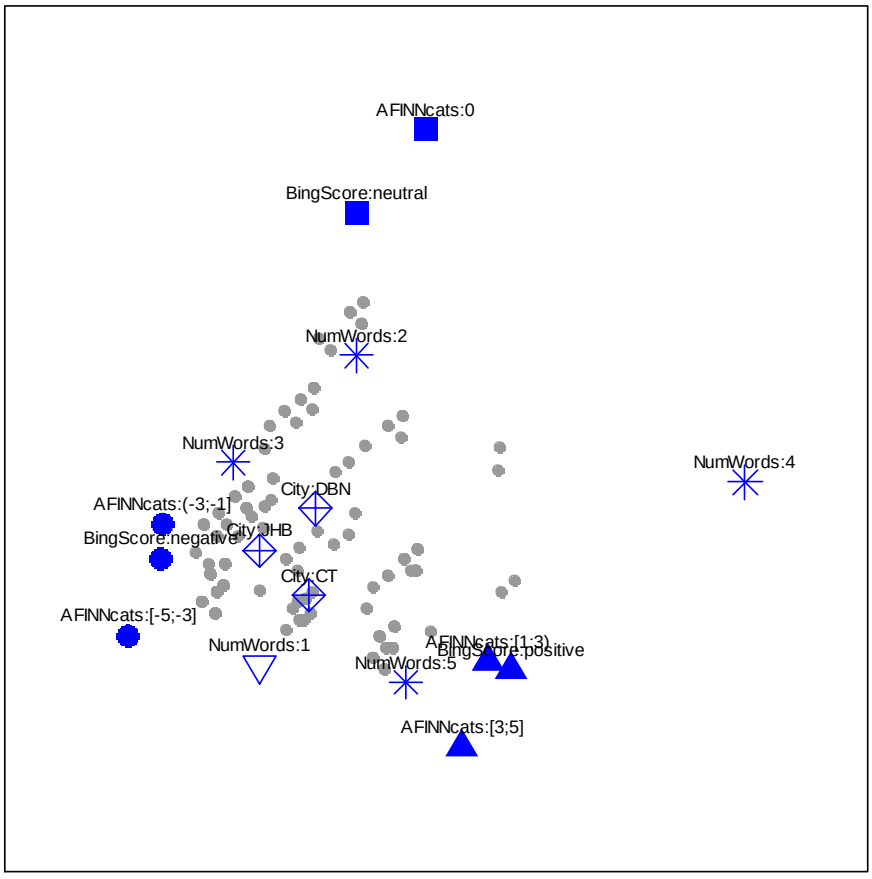


Figure 3.4: MCA Biplot Example

The category levels for the Location variable are represented by the blue diamond symbols (◈). According to the Figure 3.4, the category levels of the Location variable are grouped adjacent to one another. This implies that there is minimal distinction among the Tweets with regards to the location the Tweet was sent from.

The blue upwards pointing symbols (▲) represent the category levels of the s^A and s^B variables associated with positive sentiments, that is Bing scores that are positive and AFINN scores that are either between $+1$ and $+3$ or between $+3$ and $+5$. The blue upwards pointing symbols are grouped together in Figure 3.4, which indicates that Tweets classified as positive using L^B are also classified as positive using L^A. Similarly, the category levels associated with negative and neutral sentiments are represented by the blue circles (●) and squares (■) respectively. Since these category levels that are associated with the sentiments are grouped together as well, this indicates that when comparing the use of the two lexicons, the results of the sentiment classification are in agreement.

When the number of words considered for classification are examined in collaboration with the other category levels in Figure 3.4, we can determine whether longer or Tweets with more classifiable words are associated with more distinct sentiment classifications. The category level where only one word is considered for classification is represented by the blue downward pointing symbol (▽) and the remaining category levels, where two, three, four or five words are considered for classification are represented by the blue asterisk symbol (✳). When only one word is considered for sentiment classification in the Tweet, the word is either classified as positive or negative, which means that the overall sentiment category is solely dependent on the one classifiable word. Thus, the category level *NumWords:1* appears between the groups formed by the positive and negative sentiment categories.

As the number of words considered for classification increases, the possibility of the sentiment classification of the words balancing each other out, increases. If two words of opposing sentiments are considered for classification, one word could be positive and the other negative. The overall classification would be considered neutral, therefore, *NumWords:2* appears in close proximity to the neutral sentiment categories. The category levels for three to five words considered for classification also tend towards the neutral sentiment category group.

3.6 EMBEDDED WORD MCA BIPLOT

Based on the additional insight gained from the MCA biplot, we can identify which category levels are similar, as well as groups of Tweets associated with those groups of category levels. However, it is still unknown which Tweets or which words in the Tweets are associated with the category levels. It is therefore of interest to reveal the classifiable words contained in the

Tweets represented by grey circles in the MCA biplot in Figure 3.4. These additional insights can be gained from an enhanced version of the MCA biplot (cf. Section 3.5.4).

The embedded word MCA biplot enhances the MCA biplot by displaying the words considered for classification as it appears in the third column of Table 3.5 in grey text (abc). The category levels represented by the blue text (**abc**) and sample points represented by red circles (•), are in the same locations as they were in the initial MCA biplot in Figure 3.4. Firstly, Figure 3.5 presents the MCA biplot without the embedded words, where the red circles represent the Tweets and the blue text represent the category levels. The embedded word MCA biplot is displayed in Figure 3.6. This is the exact same MCA solution as observed in Figure 3.4 and Figure 3.5, the embedded word MCA biplot is therefore an enhancement of the existing MCA biplot by displaying additional information regarding the classifiable words in each Tweet.

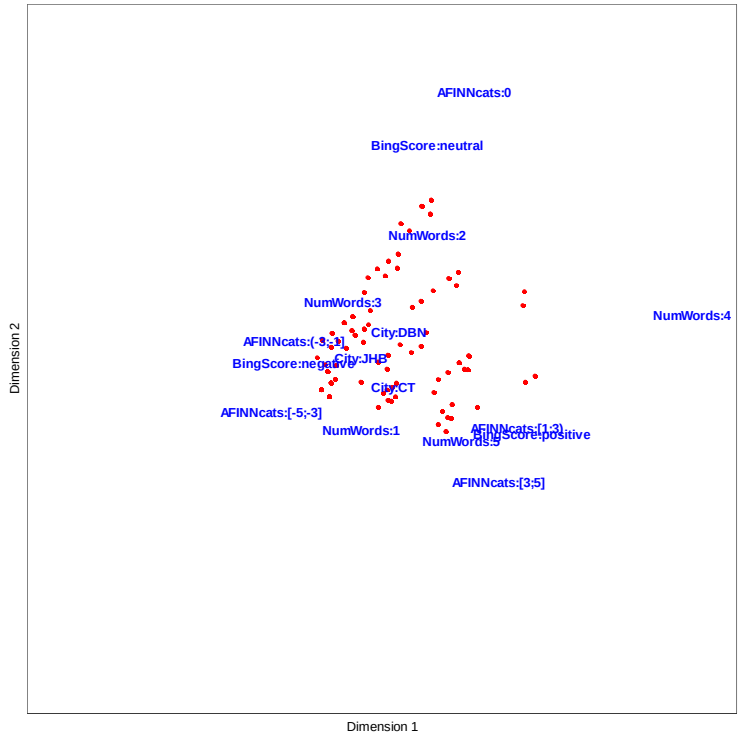


Figure 3.5: MCA biplot without embedded words

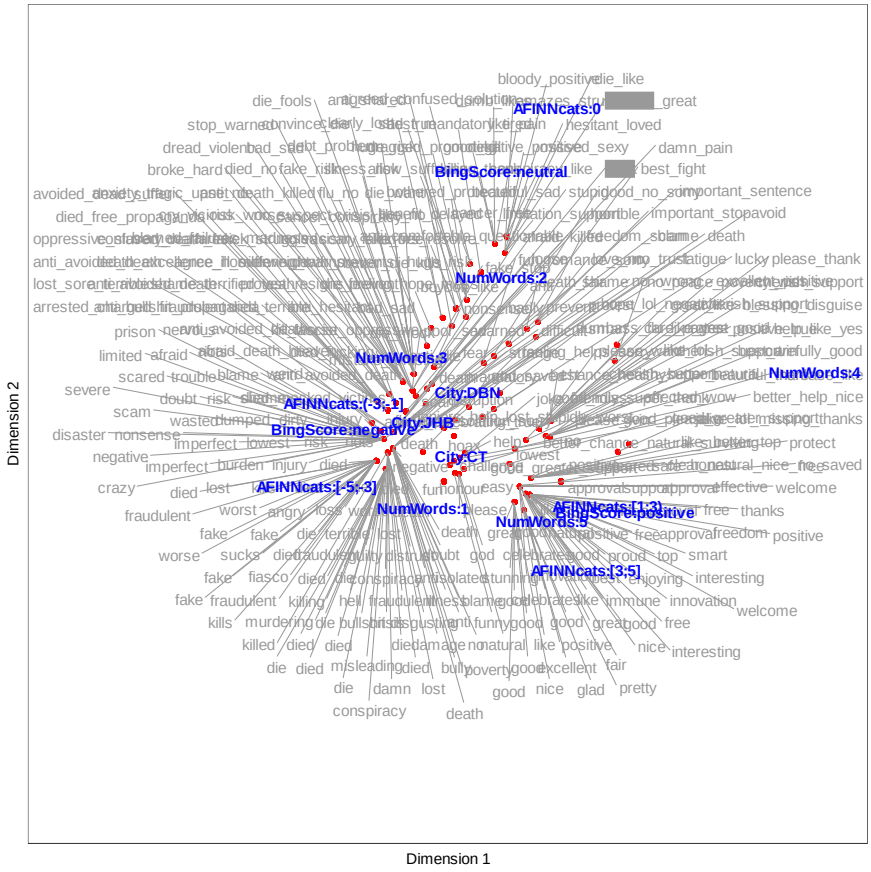


Figure 3.6: Embedded word MCA biplot

The embedded word MCA biplot now provides an opportunity to verify which of the words considered for classification are associated with the category levels. This however results in an overplotted display which is not legible and requires some additional intervention. The display in Figure 3.6 is constructed using the `geom_text_repel` function in the `ggrepel` package in `R` (R Core Team, 2021; Slowikowski, 2021). This function adds text directly to the plot and displays the label of the sample point further away, especially where the sample point coordinates overlap. This facilitates the legibility within the display. Some labels are discarded from the display where there are too many overlaps.

The categorical variable Q_1, which represents the number of words considered for sentiment classification (cf. Section 3.5.3) can be displayed as different colours instead of one of the categorical variables in the MCA solution. The main focus of interpretations in Figure 3.4 and Figure 3.6 were the associations among the category levels. Associations among the category levels of Q_1, Q_3 and Q_4 were revealed in Section 3.5.4. It was emphasised that the extracted number of classifiable words from the processed Tweet (*NumWords*) impacts the success of the classification of the Tweets. To summarise, the length of $\mathbf{t}_i$ and the sentiment category levels show strong associations among certain levels due to their close proximity. Based on these findings, to simplify the MCA solution, the number of words will now not be utilised as

an active variable in the MCA, but rather be used to visualise subsets of the MCA solution as will be presented in Figure 3.8.

The Tweets previously displayed as grey circles in Figure 3.4 are now represented by the classifiable words in each Tweet (cf. the second column of Table 3.5) and the levels of Q_1 are represented by the five different colours specified in Table 3.6.

Table 3.6: Colours used in embedded word MCA biplot for different levels of Q_1

Number of words considered for classification (Q_1)	Colour
1	
2	
3	
4	
5	

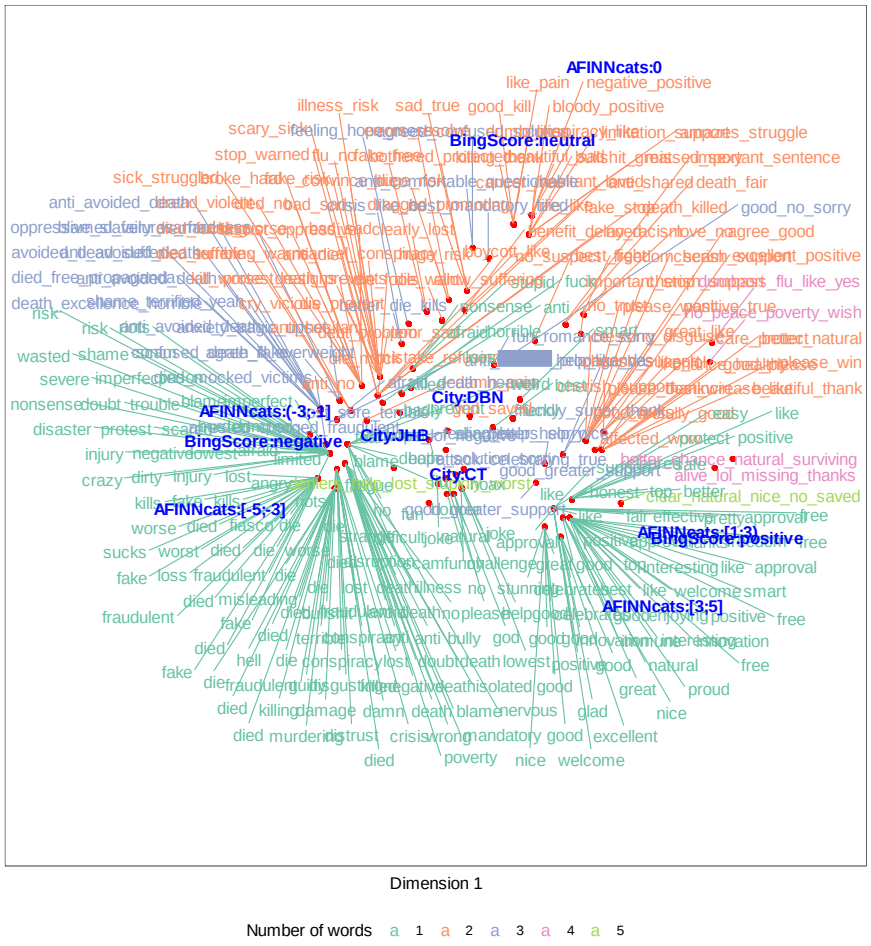


Figure 3.7: Embedded word MCA biplot where colour is used as an additional dimension

The size of the data and length of the words still influence the legibility of the plot, thus individual plots can be presented for each level of Q_1. This allows one to read the sets of words associated with each number of words as well as the other category levels. The embedded word MCA biplot (Figure 3.7) and the individual embedded word MCA biplots are displayed in Figure 3.8. Displaying the plots individually allows one to identify groups of words associated with the specific category levels. As the frequencies of the fourth and fifth category levels are lower, the words in these categories are combined in one display. The purpose of this section is to introduce the capabilities of the embedded word MCA biplots. Therefore, the plots in Figure 3.8 are presented for illustration purposes and a detailed interpretation follows in Section 4.5.

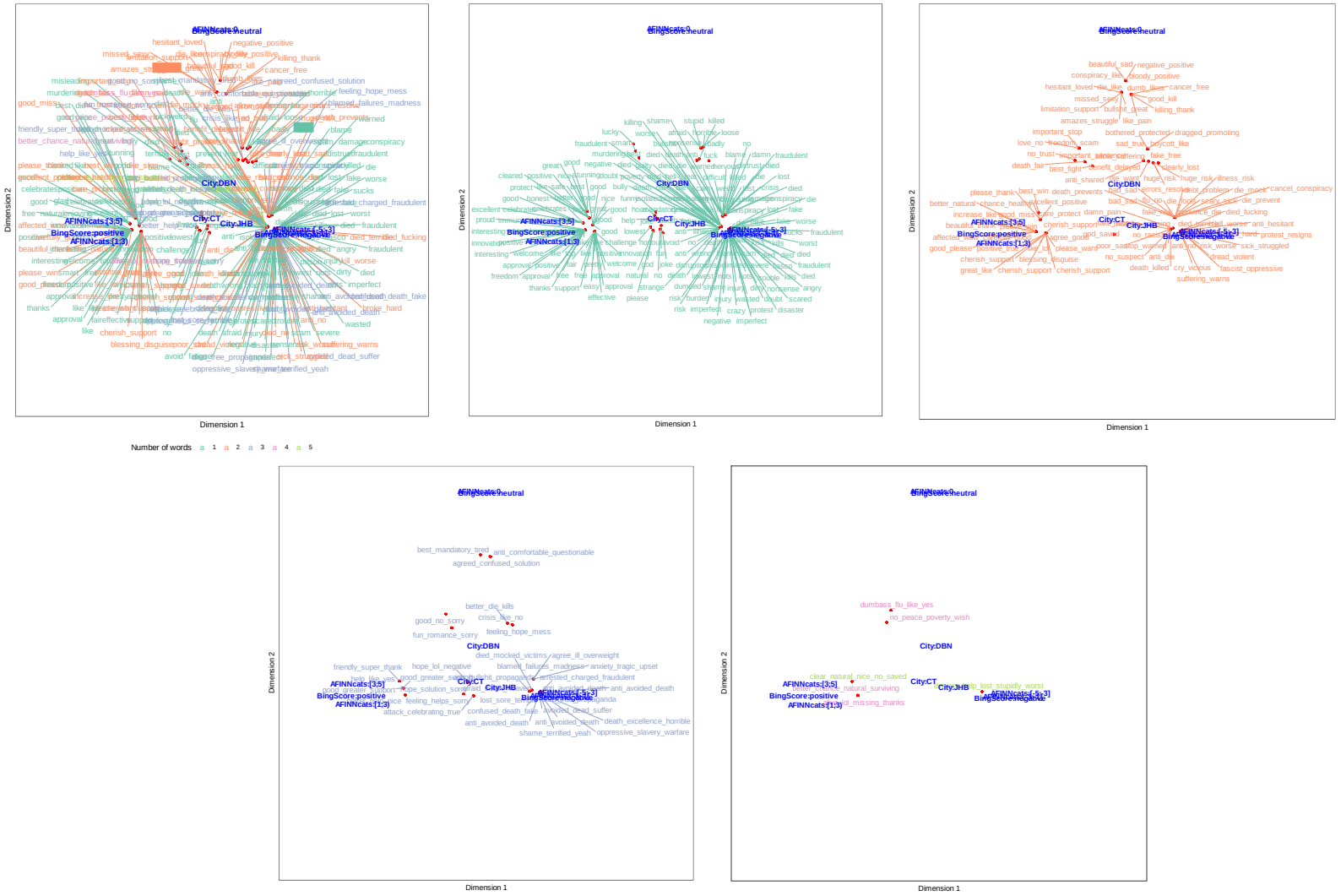


Figure 3.8: Complete and individual embedded word MCA biplots

3.7 CONCLUSION

This chapter explained the research methodology used to perform sentiment classification and graphically display the obtained results.

The pre-processing step introduced some challenges. During the stop word removal step, there was some difficulty in identifying the words in context that would influence the overall sentiment of the observed text. Human intervention is thus still required during pre-processing.

The Bing (L^B) and AFINN (L^A) sentiment lexicons were used to classify the Tweets into sentiment categories. Both lexicons are appropriate for classifying the Tweets since L^B contains a large number of terms with their scores and L^A was constructed for the sentiment classification of blog posts. The limitation presented during the sentiment classification is that some of the terms in L^B are not contained in L^A and that some of the context was lost while determining a sentiment score for each individual classifiable word.

The main advantage of the word clouds is that it provides a comprehensive summary of the words contained in the text. The information provided by the word clouds is insightful, but the emotive quality of the words is not perceived. The word clouds are enhanced by displaying the words in different colours according to their corresponding sentiment categories. This allows for the understanding of the text itself but does not consider the context or individual Tweets.

The MCA biplot identified groups of the respective sentiment categories represented by their category level points. The category levels associated the distinct sentiment categories are in close proximity, which highlights the success of the sentiment classification. Additionally, the MCA biplot displayed the particular category level points associated with each number of classifiable words. Despite the fact that associations among category level points were ascertained by utilising the MCA biplot, the Tweets represented by the sample points are still not identified.

The MCA biplot is enhanced by displaying the classifiable words per Tweet on top of their corresponding sample points. This allows one to read the sets of words associated with each number of words as well as the other category levels. The subsequent additional insight and interpretation from both the enhanced word clouds and embedded word MCA biplot confirms that these enhancements explored in this study are significant contributions to text classification as well as the visualisation of text data.

CHAPTER 4
FINDINGS

4.1 INTRODUCTION

This chapter presents the results of the sentiment classification in the form of word clouds and MCA biplots. The techniques discussed in the research methodology (cf. CHAPTER 3) were applied to Tweets filtered by the keyword “covid”. As mentioned in Section 1.3, the two sets of Tweets are utilised in respective case studies. The first case study uses the Tweets from different South African cities. This data set is referred to as “South African cities”. The results are based on the comparison of three cities in South Africa. The chosen cities are Cape Town, Durban and Johannesburg. The second case study uses the Tweets from three countries resulting in a larger sample referred to as the “countries” data set. The chosen countries are South Africa, Australia and the United Kingdom. These analyses aim to visualise the results of the sentiment classification and to determine if there is a difference in sentiments toward the COVID-19 pandemic for a fixed time period based on the location of the user.

4.2 WORD CLOUDS

Word clouds provide an overall summary of the content within the set of obtained Tweets. The colours used to display the words in the layout indicate its sentiment category according to the Bing sentiment lexicon as positive or negative (cf. Section 3.4.2 and Section 3.4.2). Enlarged images of the word clouds in Figure 4.1 and Figure 4.2 are available in APPENDIX A.

4.2.1 Word clouds with all words in the corpus

Cape Town **Durban** **Johannesburg**

Figure 4.1: Word clouds with all words for three cities in South Africa

The word clouds for the three different cities indicate that the most relevant term occurring for the keyword “covid” relate to the vaccine. These Tweets were sampled in early September 2021, amidst the vaccine rollout in South Africa. It is thus reasonable that this is the most frequently occurring topic of interest related to conversations of the COVID-19 pandemic.

Table 4.1: Relative frequencies of words in each sentiment category for three cities

City	Cape Town	Durban	Johannesburg
Relative frequency of unmatched words	90.5891%	91.0753%	89.1911%
Relative frequency of positive words	4.5142%	3.7308%	4.2233%
Relative frequency of negative words	4.8967%	5.1939%	6.5855%

The words displayed in red (negative sentiments) and green (positive sentiments) do not occur as frequently in the overall text and are displayed in smaller font sizes. The values in the first row of Table 4.1 indicate that between approximately 89% and 92% of words of the respective cities are not matched to terms in the Bing sentiment lexicon. These words, however, account for the most frequently occurring words in the word clouds, which is reflected by the larger font sizes. The relative frequencies of negative words range between approximately 4% and 7% for all three word clouds. In comparison, the relative frequencies of the positive words are lower than those of the negative words and range between approximately 3% and 5%. Despite the fact that the number of words matched to the Bing sentiment lexicon are much lower than those that are not matched, the number of negative words still exceed the amount of the positive words which suggests that the Twitter users have negative sentiment towards the topic.

South Africa **Australia** **United Kingdom**

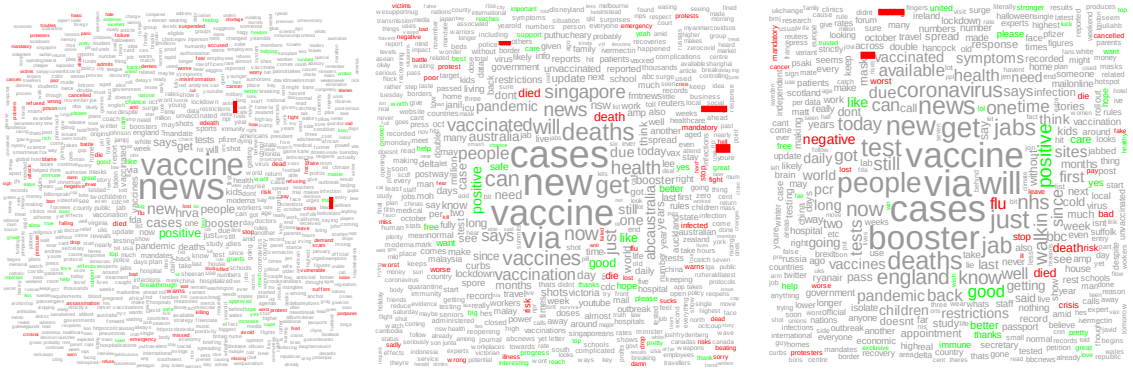


Figure 4.2: Word clouds with all words for South Africa, Australia and the United Kingdom

The “countries” data set consists of a larger number of Tweets (cf. Section 1.3), which results in denser word clouds. The smaller font size introduces some difficulty with regard to the legibility of the text in the word clouds. Considering all the word clouds in Figure 4.1 and Figure 4.2, it is revealed that Twitter users in South Africa, Australia and the United Kingdom deem the COVID vaccine an important discussion point. However, it is worth noting that the Tweets for the second set of results were created during October 2021, thus there might be slight differences in secondary topics when examining the word clouds in Figure 4.2. Also, countries have different rollout programmes and the phase of vaccine administration differ between countries. Additional topics such as the additional vaccine dosage (booster), mandates and travelling appear in the second set.

Table 4.2: Relative frequencies of words in each sentiment category for three countries

Country	**South Africa**	**Australia**	**United Kingdom**
Relative frequency of unmatched words	92.2764%	91.8567%	91.0729%
Relative frequency of positive words	2.9027%	3.3255%	3.4149%
Relative frequency of negative words	4.8209%	4.8177%	5.5122%

The words displayed in red and green appear more frequently in Figure 4.2 in comparison to Figure 4.1, which could be due to the larger sample of Tweets.

The values in Table 4.2 indicate that each country’s word cloud contains between approximately 91% and 93% of words not matched to terms in the Bing sentiment lexicon. The relative frequencies of negative words range between approximately 4% and 6% for all three word clouds. In comparison, the relative frequencies of the positive words are lower than those of the negative words and range between approximately 2% and 4%. Similar to the observation made from Figure 4.2, the number of negative words still exceed the number of the positive words. Figure 4.3 compares the relative frequencies of positive words and the relative frequencies of negative words for the word clouds in Figure 4.1 and Figure 4.2.

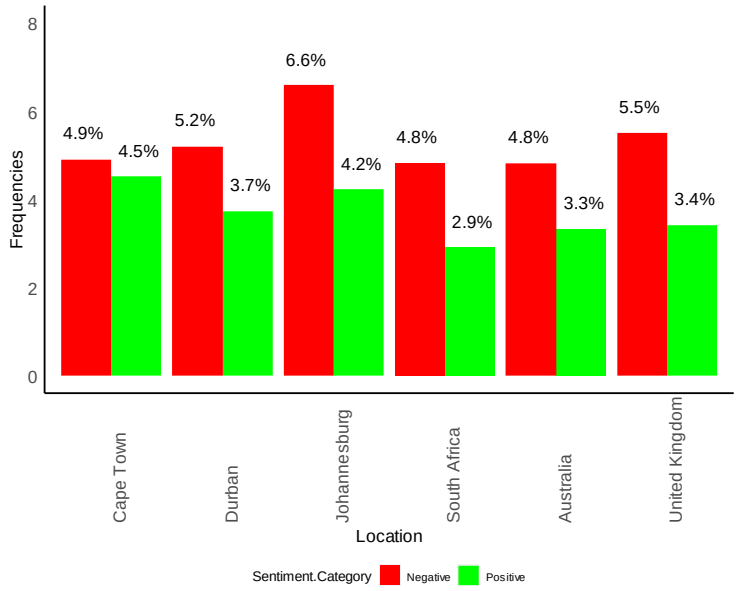


Figure 4.3: Comparison of relative frequencies of negative and positive words in Figure 4.1 and Figure 4.2

The frequencies presented in Figure 4.3 confirm the observations that there are more words matched to negative terms in the Bing sentiment lexicon than the positive terms in the lexicon. This suggests that the overall sentiment of the Tweets for all locations is likely negative. The Tweets created in Johannesburg have the highest relative frequency of negative words and the Tweets created in Australia have the lowest relative frequency. The Tweets created in the city of Cape Town and the Tweets created in the country of South Africa have the highest and lowest relative frequencies of positive words, respectively, compared to the other locations.

Since the display is cluttered by the unmatched words, it is still unclear which sentiment dominates discussions related to COVID-19. In order to enhance the visual interpretation, word clouds are presented in Section 4.2.2, which only focus on words in the text that are matched to the sentiment lexicon.

4.2.2 Word clouds with words matched in the sentiment lexicon

Figure 4.4: Word clouds with words matched to sentiment lexicon for three cities in South Africa

The word clouds suggest that the sentiments of South African citizens are overwhelmingly negative since the words displayed in larger fonts are also displayed in red. This is supported by the values in Table 4.1 as well as the interpretation of Figure 4.3, which show that the number of words in the corpus matched to negative terms exceed the number of words in the corpus matched to positive terms. The most frequently occurring word, however, is “no”, which highlights one of the known challenges of sentiment analysis of negation. Since the text is formatted into tokens, it is unknown whether these occurrences of “no”, co-occur in contexts in which they indicate negation. Another challenge is the underlying sentiment of informal language, such as the term “lol”. This is an abbreviation for “laugh out loud”, which could be a facetious comment and not necessarily a positive sentiment as it has been classified in the word clouds for Durban and Johannesburg (cf. Figure 4.4).

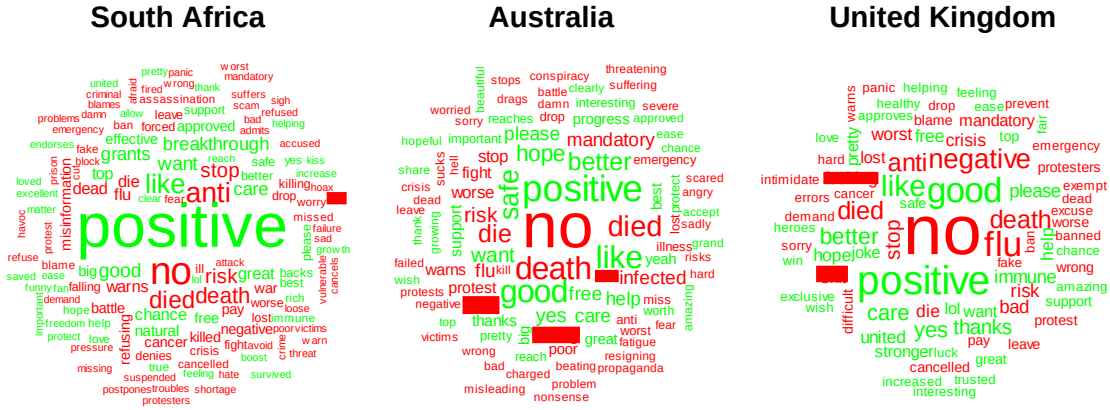


Figure 4.5: Word clouds with words matched to sentiment lexicon for South Africa, Australia and the United Kingdom

According to the word clouds in Figure 4.5, the words with negative sentiments outnumber the words with positive sentiments. This suggests that most of the Tweets used in the sample

contain negative sentiments when Tweeting about the COVID-19 pandemic. This is supported by the values in Table 4.2 as well as the interpretation of Figure 4.3 which indicate that the relative frequencies of negative words exceed that of positive words. In comparison to the word clouds for the South African cities, the word clouds for the countries show more variety of classified terms.

Although these word clouds provide us with information contained in the Tweets, it still reveals a drawback of the word cloud visualisation technique, which is that the context is unclear. In the context of COVID-19, the word “positive” could indicate that the user is referring to a positive test result for the virus. The sentiment lexicon would code the word “positive” as a green sentiment category, as illustrated in all panels of Figure 4.5.

The occurrence of words such as “death” and “dying” in all the word clouds confirms the fatal nature of the coronavirus and according to the sample, it appears that the overall attitude towards the pandemic is fear. Although less frequent than the negative sentiments, terms such as “like”, “good”, “better” and “support” appear with high frequencies as well. This is a promising result for our case study since the enhanced word clouds as a sentiment visualisation tool can reveal the overall sentiment toward a topic of interest from observed text data. The positioning of the words in word clouds is based on the frequency at which they appear within the Tweets and does not display the similarity among the words.

4.3 SENTIMENT CLASSIFICATION

As mentioned in Section 3.4, this study performs sentiment classification based on two different sentiment lexicons, namely the Bing (L^B) and AFINN (L^A) lexicons. The Bing lexicon consists of 6786 terms and the AFINN lexicon consists of 2477 terms. Table 3.3 and Table 3.2 revealed that although both L^A and L^B are comprised of large collections of words with corresponding sentiment scores or sentiment categories, some words that are included in L^A are not included in L^B. The bar graphs presented in Figure 4.6 and Figure 4.7 illustrate the number of words that was classified using the two lexicons. The first bar represents the percentage of words that was matched to the AFINN lexicon and the second bar represents the percentage of words that was matched to the Bing lexicon. These percentages can then be compared to the third bar that expresses the percentage of words that was matched to both lexicons.

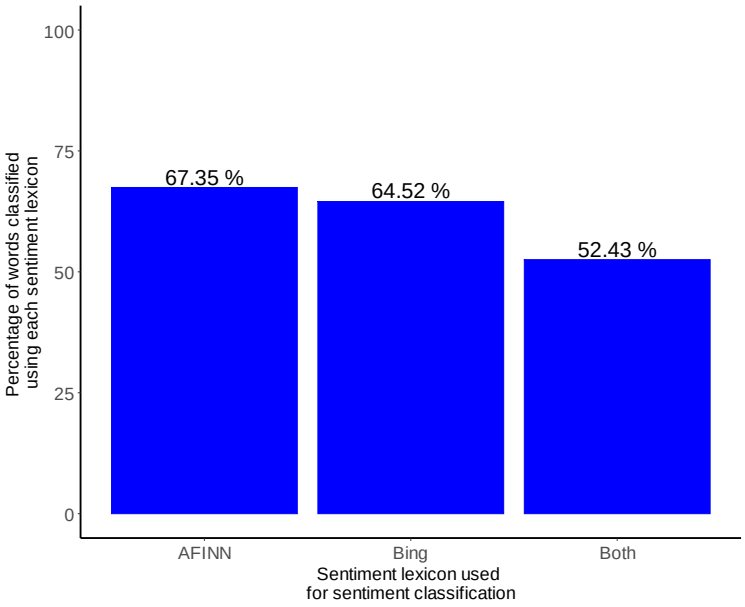


Figure 4.6: Comparison of sentiment classification based on number of words classified by each sentiment lexicon (South African cities)

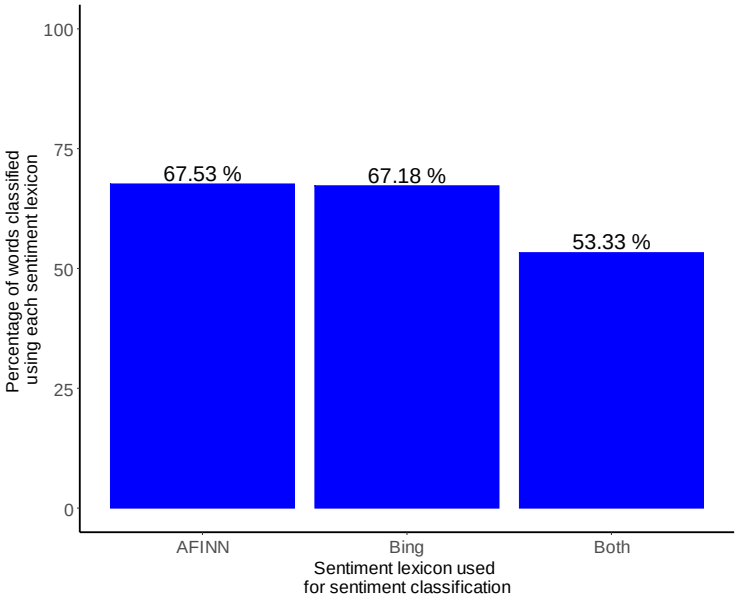


Figure 4.7: Comparison of sentiment classification based on number of words classified by each sentiment lexicon (countries)

The percentage at which the number of words classified by the AFINN lexicon exceeds the number of words classified by the Bing lexicon is at most approximately 3%, which suggests that the words in the text are equally likely to be matched by either lexicon. The number of words in the texts that are matched to both lexicons range between roughly 52% and 54%.

Sentiment classification is performed on each Tweet to obtain the sentiment scores, s_i^B and s_i^A, for $i = 1, \ldots, n$, which are considered as two categorical variables in the formatted categorical dataset (cf. Section 3.5.1). The constructed data set (cf. Table 3.5) only includes

Tweets where the words are classified using both lexicons so that each Tweet has an entry for a Bing sentiment score as well as an AFINN sentiment score. The similarity between the classifications of the two sentiment lexicons will be compared by means of visualisation in Section 4.4.

4.4 MCA BIPLOT

The MCA biplot aims to investigate the relationships among the levels of the categorical variables as well as identify groups of Tweets associated with those category levels. The technique also allows us to evaluate the accuracy of the classification by inspecting whether the sentiment category levels for the classification using the Bing lexicon (cf. Section 3.4.2) is displayed in close proximity to those of the classification using the AFINN lexicon (cf. Section 3.4.1). Figure 4.8 presents the MCA biplot of the “South African cities” data set, followed by the “countries” MCA biplot in Figure 4.9.

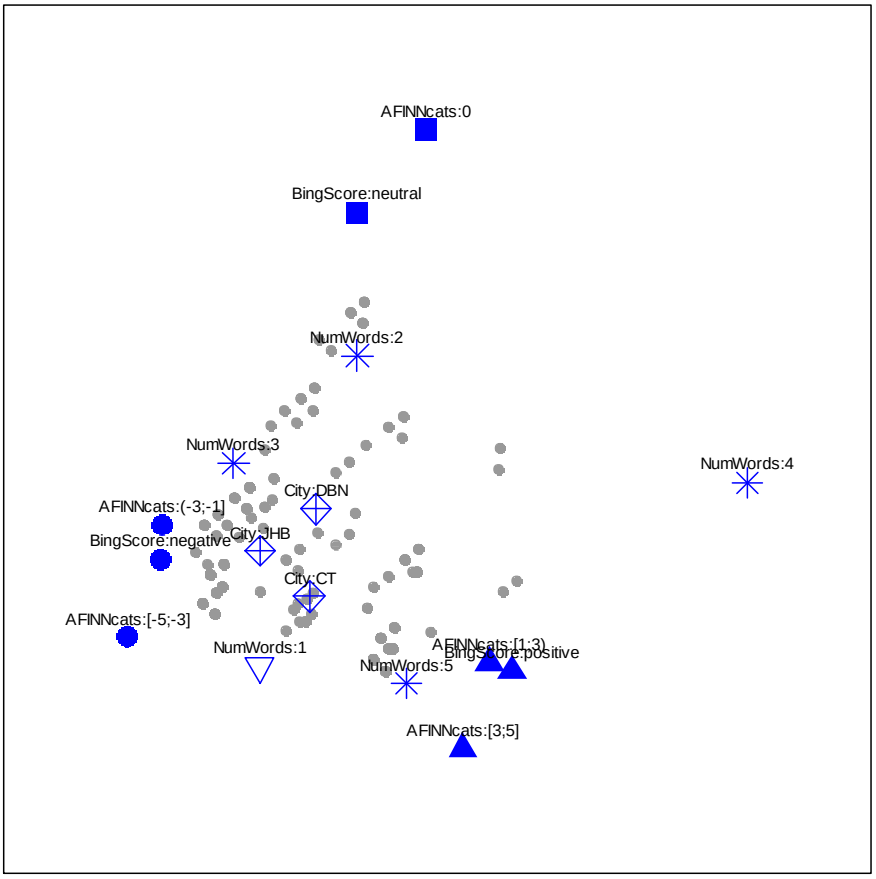


Figure 4.8: MCA biplot for results of sentiment analysis performed on Tweets created in three South African cities

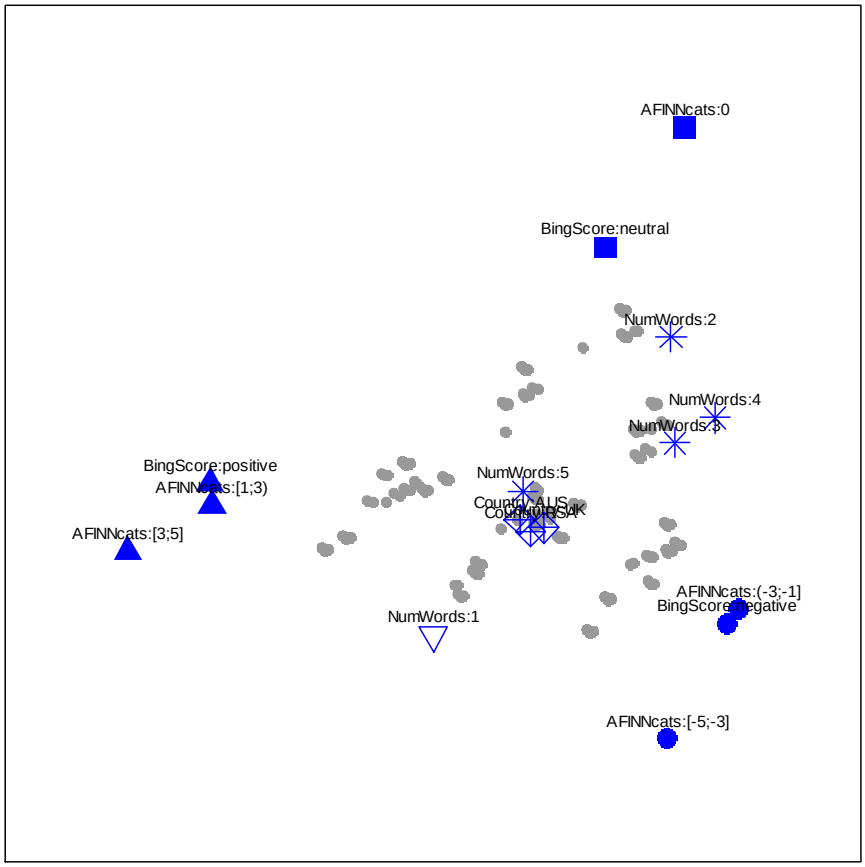


Figure 4.9: MCA biplot for results of sentiment analysis performed on Tweets created in South Africa, Australia and the United Kingdom

The interpretation of the MCA biplots will be focused on the category level points, which will be followed by the interpretation of the samples, representing the processed Tweets, interpreted in the embedded word MCA biplots (cf. Section 4.5). The description of the symbols and the initial interpretation is discussed in Section 3.5.4. The location levels, displayed by the ⌖ symbols, appear to form their own group. This indicates that the response profiles of these category levels are similar. The processed Tweets (cf. Section 3.2) for the different cities (Figure 4.8) or countries (Figure 4.9) are therefore barely distinguishable, especially in Figure 4.9. This also supports the interpretation of the similar content within the word clouds in Figure 4.1 and Figure 4.2. Although, if we compare the proximity of the Johannesburg level in Figure 4.8 to the other groups of category levels, we can note that it is positioned towards the blue circle symbols (●), which represents the negative sentiment category group. This is a possible indication that Tweets created in Johannesburg tend to have negative overall sentiments. Similarly, the Cape Town level is positioned towards the blue upwards pointing symbols (▲), representing the positive sentiment category group, which suggests that Tweets created in Cape Town tend to have positive overall sentiments. This relates to the interpretation of Figure 4.3 in which Johannesburg Tweets contained the highest

relative frequency of negative words and the Cape Town Tweets contained the highest relative frequency of positive words.

In both Figure 4.8 and Figure 4.9, the *NumWords:1* (▽) category level lies between the groups of positive and negative sentiment categories. This confirms that the overall sentiments of the processed Tweets with only one word considered for classification are determined by the sentiment category of that one word.

According to both MCA biplots, Tweets containing two words considered for classification are associated with neutral sentiments since the asterisk symbol (✳) for the category level *NumWords:2* is positioned near the group of neutral sentiment categories. The overall sentiments of these Tweets are determined by whether both words are either positive or negative or whether the words are of opposing sentiment categories. If the words in the Tweet are of opposing categories, the sentiment classification of the words balance each other out and results in a neutral sentiment category for the Tweet.

In Figure 4.8, the *NumWords:3* category level is located near the negative sentiment categories and the *NumWords:5* category level is located near the positive sentiment categories. According to Equation 3.2 and Equation 3.1, the overall sentiment scores are calculated by classifying the Tweet into the sentiment category in which majority of the words are classified. The resulting overall sentiment scores of Tweets that contain two or more words considered for classification is therefore dependent on a majority score. As mentioned in Section 3.5.4, if the Tweet contains three classifiable words of which two are of the same sentiment, the Tweet will be associated with that sentiment category (positive or negative). The asterisk symbol (✳) for *NumWords:3* seems to be associated with negative sentiments in Figure 4.8 and equally as interesting, the *NumWords:5* and *NumWords:4* category levels are located either near the positive sentiment categories or on the far right thereof. This implies that Tweets created in the three cities of South Africa that contain three or more classifiable words more likely contain at least two negative words. In contrast, the Tweets that contain four or five classifiable words contain a majority of positive words. Figure 4.9 reveals that the larger sample size available in the “countries” data could have led to the category levels of *NumWords* (excluding *NumWords:1*) being in closer proximity than the levels in Figure 4.8. The *NumWords:2*, *NumWords:3* and *NumWords:4* category levels are positioned closer to the neutral sentiment category group since the increase in the number of classifiable words results in a higher probability of the sentiment classification of the words balancing each other out. The *NumWords:5* category level is located somewhat further away than the other *NumWords* category levels, but closer to the *NumWords:1* category level, which is positioned between the positive and negative sentiment category groups. This could be caused by the small number

of Tweets that contain five classifiable words. This applies to the “South African cities” data set as well (cf. Figure 4.8).

Sentiment category groups are formed by the category levels of positive, negative and neutral sentiment categories. The positive sentiment category group is represented by the group of upwards pointing symbols (▲). The group of blue circles (●) represents the negative sentiment category group and the blue squares (■) represent the category levels associated with neutral sentiments. This confirms that the sentiment classification performed using the Bing lexicon and the AFINN lexicon produced similar results. Correspondingly, the interpretation of Figure 4.6 and Figure 4.7 suggest that the chance of words in the Tweet being matched to either lexicon is nearly equal. This is a promising result for the study since it validates the second objective to obtain a multivariate sentiment visualisation tool that displays the sentiment categories near the words of the Tweets associated with them. Additionally, the category levels associated with the same sentiment are also grouped together, which confirms consistency of classification across two lexicons.

The MCA biplot is thus an effective visualisation tool to illustrate similarity among category level points in a categorical data set. The study aims to visualise the results of the sentiment classification through the MCA biplot by determining whether the words in each Tweet are positioned near the correct sentiment category. The above MCA biplots succeed in displaying the results of the sentiment classification through similarity of category levels.

The MCA biplots presented in Figure 4.8 and Figure 4.9 will now be enhanced by embedding the processed words for each Tweet in the display. This will enable simultaneous interpretation of matched words and their related sentiments. It will also expose the similarity of the polarity of classified words.

4.5 EMBEDDED WORD MCA BIPLOT

The embedded word MCA biplot is an enhanced version of the MCA biplot in Section 4.4, which displays the combination of words considered for classification for each sample point (grey solid circles in Figure 4.8 and Figure 4.9). The number of words variable is now suppressed and used as levels of colour in the plot so that the variable still contributes to the MCA solution. The interpretation of the relationships among the remaining variables should be unchanged.

As expected, the groups of variables identified in the previous interpretations are still valid, since it is still based on the same MCA solution, with labels containing processed words included for the sample coordinates and indicating the number of words that is classified in a

specific colour. This is promising for the analysis, since it verifies that the sentiment classification results in adequate overall sentiment scores for each Tweet.

The words in Figure 4.10 and Figure 4.11 can be inspected and associated with the sentiment category levels in closest proximity to them. The current format of the embedded word MCA biplot complicates the identification of any words and interpretation of the displays, especially Figure 4.11, due to overplotting.

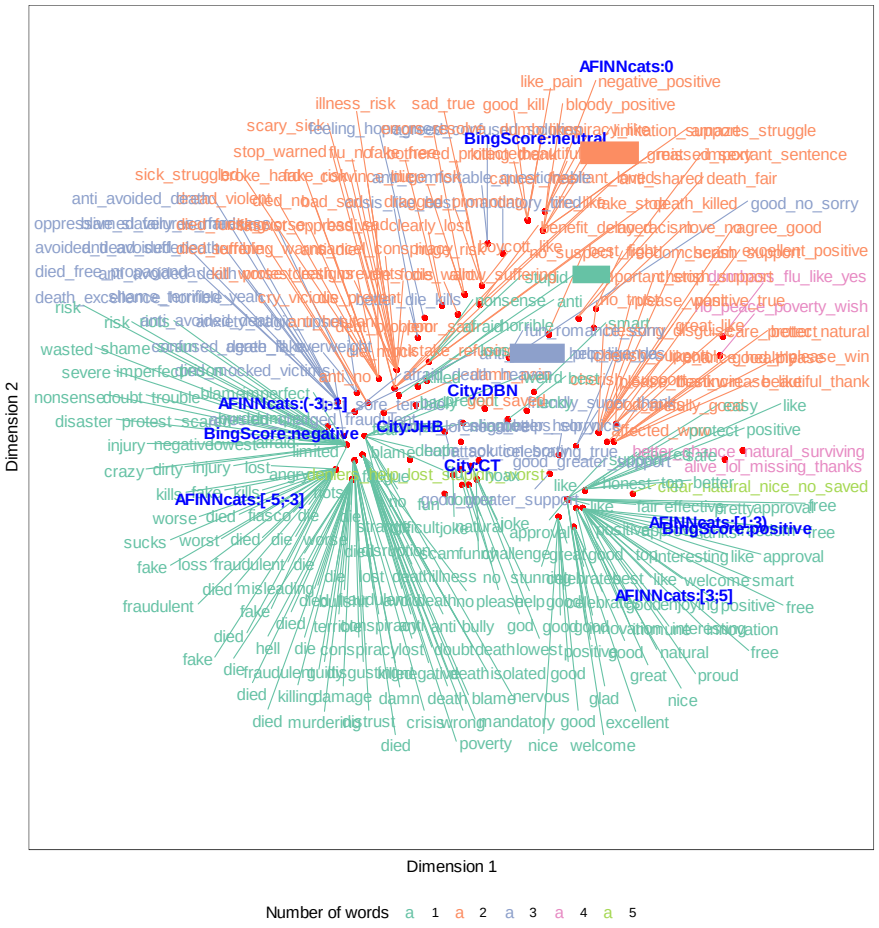


Figure 4.10:Embedded word MCA biplot for results of sentiment analysis performed on Tweets created in three South African cities

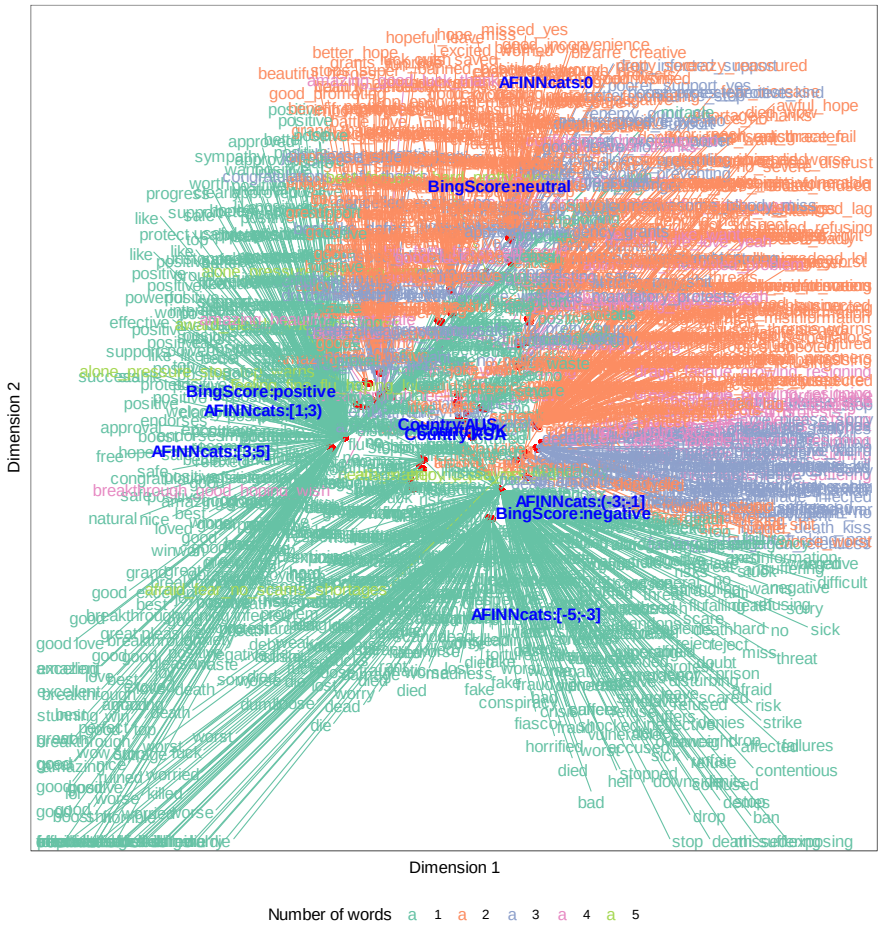


Figure 4.11: Embedded word MCA biplot for results of sentiment analysis performed on Tweets created in three different countries

The effectiveness of the embedded word MCA biplot is hindered by the cluttering of words, which undermines the legibility. We thus examine individual embedded word MCA biplots for each level of *NumWords*. The discussion for Figure 4.12 to Figure 4.19 will precede the visualisations that are displayed on the following pages.

The individual graphs for one word considered are displayed in Figure 4.12 and Figure 4.13 for the “South African cities” and “countries” case studies, respectively. As expected, large groups are formed either associated with negative or positive sentiment categories and situated far away from the neutral sentiment categories. This is due to the distinct classification that can be made for one word at a time, since the word will be classified into a specific category. Examples of words in the Tweets associated with negative sentiments include “die”, “dumped”, “murder” and “negative”. Consider the word “innovation” located in the lower right quadrant in Figure 4.12. It represents the only classifiable word in that particular Tweet and has a positive sentiment score according to both the sentiment lexicons. As expected, the point that represents the Tweet is located near the positive sentiment category group in the right of the display. According to the AFINN sentiment lexicon, the word “big” has a sentiment

score of 1 and is located in the lower left quadrant of Figure 4.13 below the *AFINNcats:[3;5]* category level. It appears to be less associated with either sentiment category group and the Tweet is possibly considered as neutral overall since the sample point is located towards the middle of the display.

According to Figure 4.14 and Figure 4.15, more Tweets tend towards the neutral sentiment categories, when two words are considered for classification. The combinations of words associated with neutral sentiment categories contain more classifiable words. Considering the combination of words “beautiful_sad” for the Tweet located towards the top of the display, underneath the neutral sentiment category group in Figure 4.14, “beautiful” is classified as positive and “sad” is classified as negative. This would thus result in a neutral sentiment score since the words balance each other out. Tweets containing two classifiable words are thus more likely to be associated with neutral sentiments although Tweets associated with strong positive or negative sentiments still occur in both figures. A Tweet containing the two classifiable words “care” and “rich” is located towards the far left of the display in Figure 4.15. Both of these words have positive sentiment scores, which leads to the overall sentiment of the Tweet being positive and thus occurring near the positive sentiment category group.

Figure 4.16 and Figure 4.17 display the individual graphs for three words considered. Tweets containing three words considered for classification are associated with the sentiment category where the majority of the three words are classified into. The overall sentiment scores are calculated to consider the sentiments of the majority of the words as stated in Equation 3.2 and Equation 3.1 (cf. Section 3.4). For example, the Tweet located towards the middle of the display in Figure 4.16 contains the classifiable words “better”, “die” and “kills”. The words “die” and “kills” are negative and the word “better” is positive. The overall sentiment of the Tweet will be negative since the majority of the words are negative. The positioning of the Tweet tends toward the middle of the display, because the one positive word prevents the Tweet from being overwhelmingly negative. Another example is the Tweet that contains the words “better”, “engage” and “supports”, which is located at the top section of the lower left quadrant in Figure 4.17. All three words in this Tweet are positive, which results in an overwhelmingly positive sentiment score. This explains why the Tweet is located near the positive sentiment category group.

According to Figure 4.18 and Figure 4.19, Tweets containing four or more words considered for classification are not common in the observed text. This could be due to the character limit imposed on Twitter (cf. Section 1.3), as well as the large proportion of unclassifiable words. Similar to the case of three classifiable words, these Tweets are classified to the sentiment category wherein the majority of the words are. However, when four words are considered for classification, the Tweet can also be treated in the same way as when two words are

considered since the even number of classifiable words can also result in the sentiment classification of the words balancing each other out. For example, the Tweet containing the words “hahaha”, “shame”, “struggle” and “yeah” is located towards the left of the top right quadrant in Figure 4.19. The words “hahaha” and “yeah” are positive and the words “shame” and “struggle” are negative. This results in the sentiment classifications of the words balancing each other out and leads to the position of the Tweet tending towards the neutral sentiment category group.

The results and interpretations presented with the embedded word MCA biplots are promising given that the enhanced versions of the MCA biplots provided additional insight into the sentiment of the Tweets. In addition to examining the relationships among the category levels from the variables in the two data sets, the embedded word MCA biplot displays the classifiable words corresponding to each Tweet to determine the groups of Tweets associated with each category level. This allows for an additional evaluation of the consistency of the sentiment classification results since one can compare each word to its associated sentiment. Compared to the enhanced word clouds presented in Section 4.2, the embedded word MCA biplots not only summarises the content in each set of Tweets, but also indicates similarity within the visualisation.

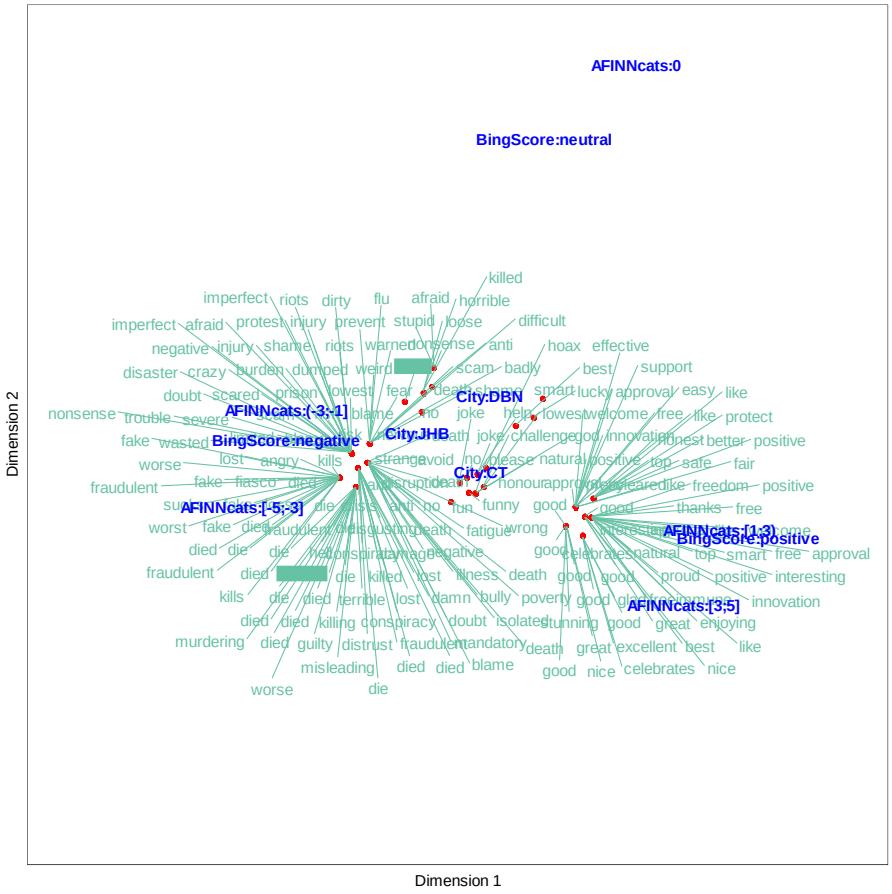


Figure 4.12: Individual embedded word MCA biplot where one word in the text is considered for classification (MCA on South African cities)

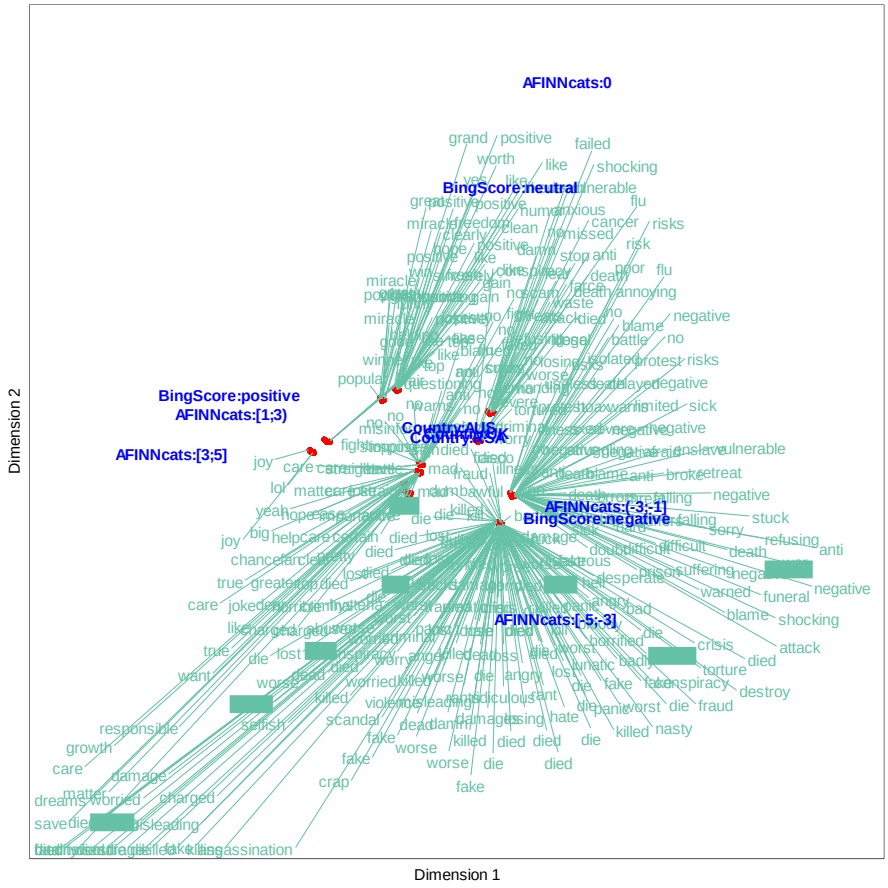


Figure 4.13: Individual embedded word MCA biplot where one word in the text is considered for classification (MCA on different countries)

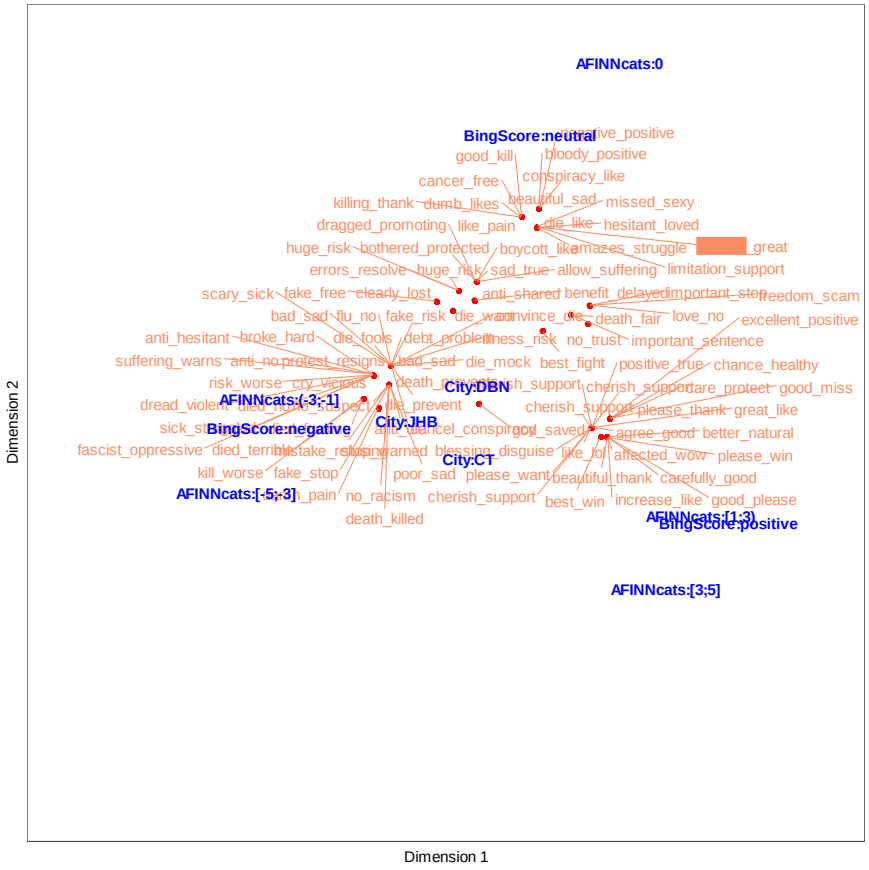


Figure 4.14: Individual embedded word MCA biplot where two words in the text are considered for classification

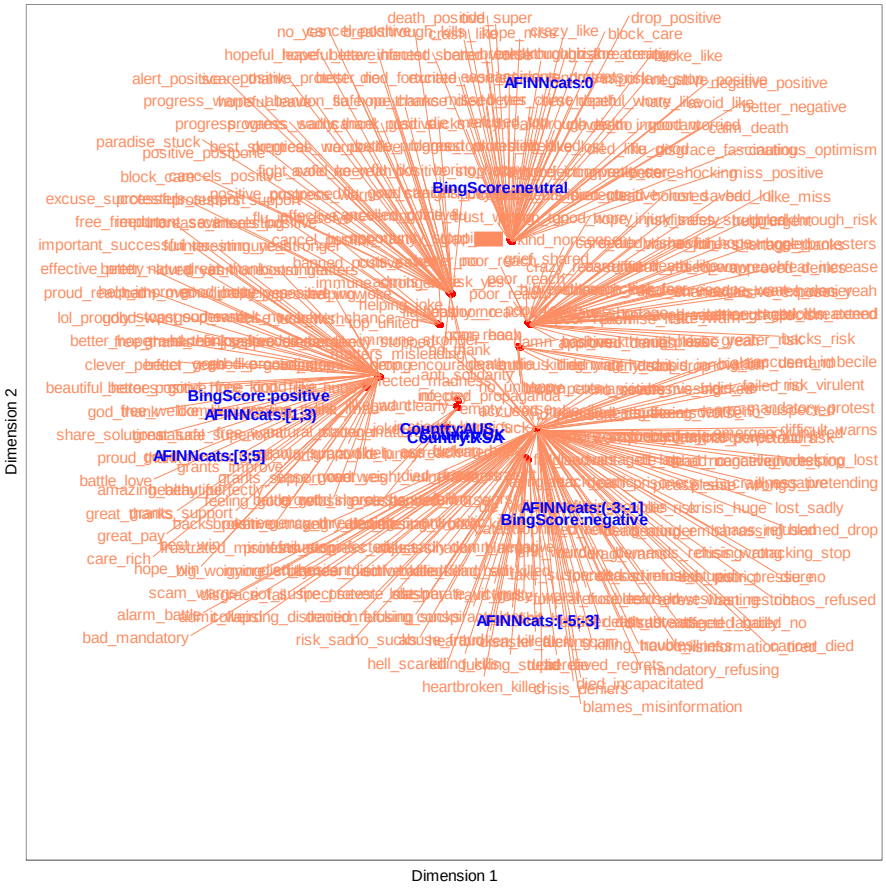


Figure 4.15: Individual embedded word MCA biplot where two words in the text are considered for classification (MCA on three different countries)

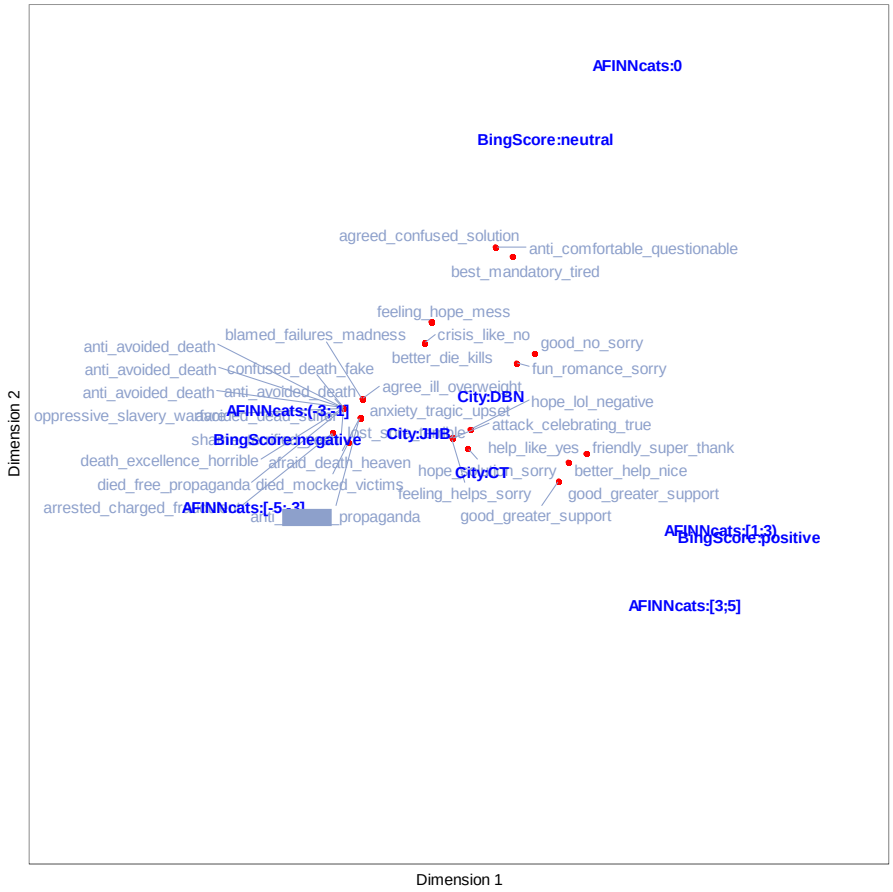


Figure 4.16: Individual embedded word MCA biplot where three words in the text are considered for classification (MCA on South African cities)

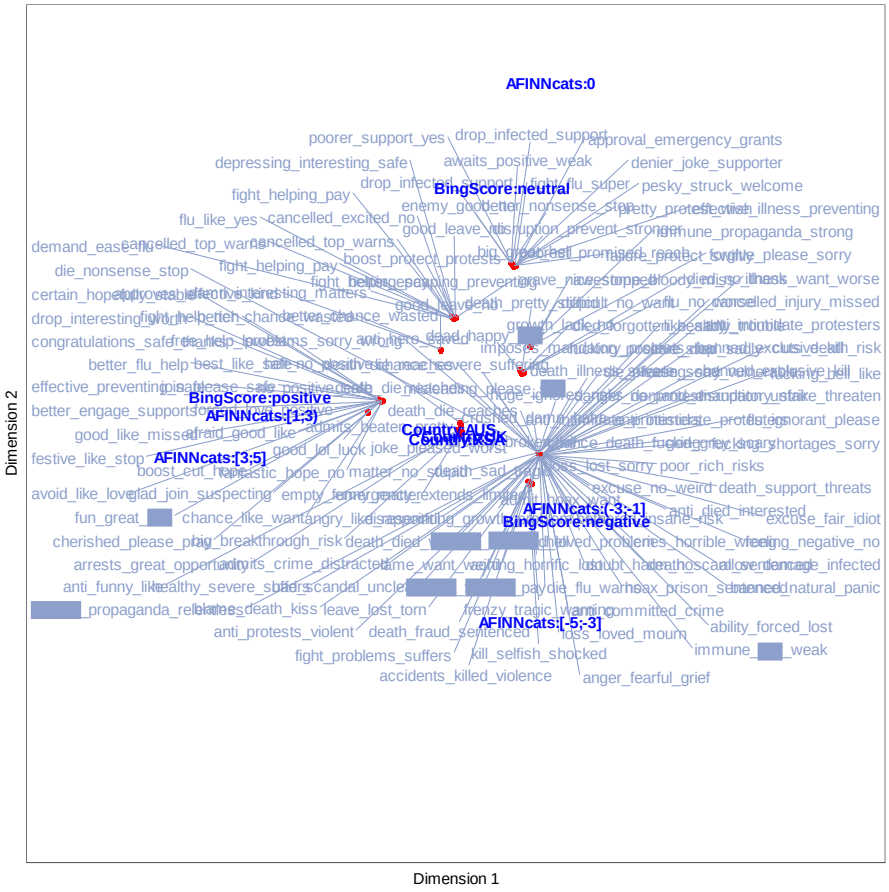


Figure 4.17: Individual embedded word MCA biplot where three words in the text are considered for classification (MCA on three different countries)

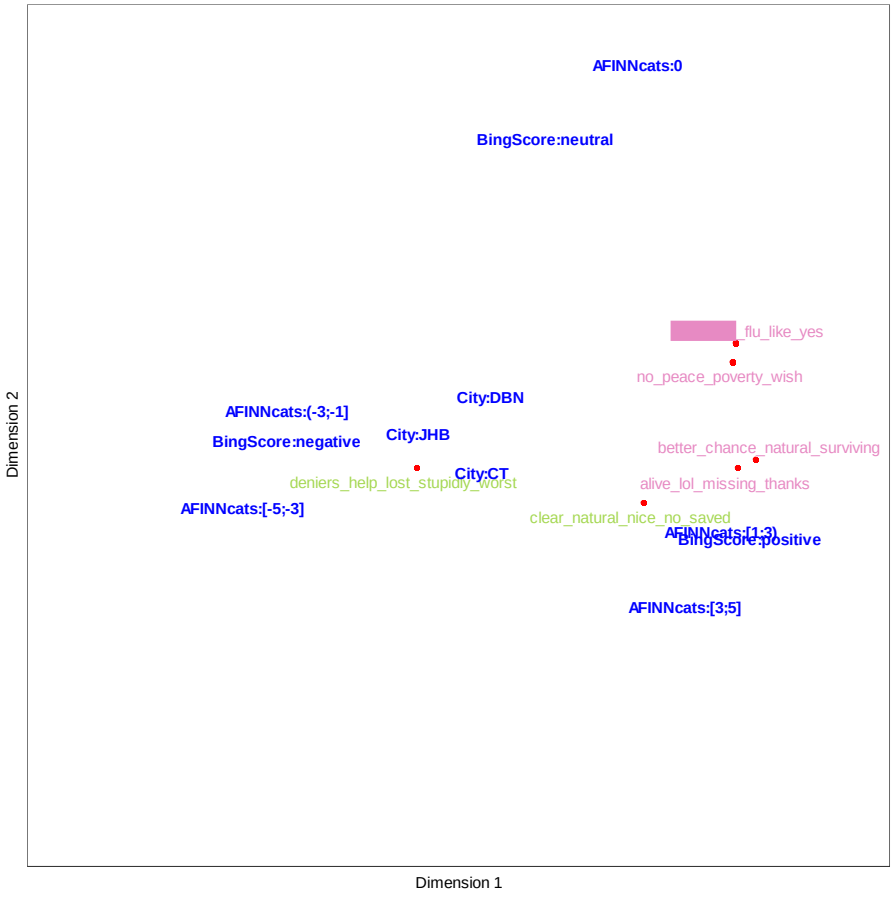


Figure 4.18: Individual embedded word MCA biplot where four to five words in the text are considered for classification (MCA on South African cities)

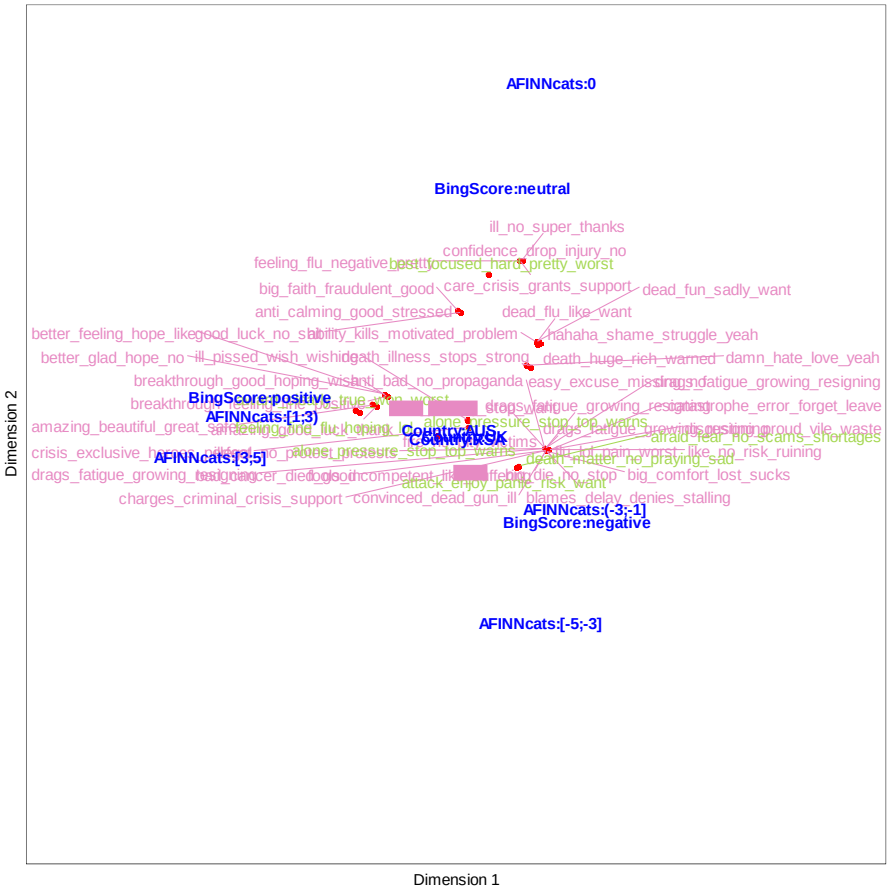


Figure 4.19: Individual embedded word MCA biplot where four to five words in the text are considered for classification (MCA on three different countries)

4.6 CONCLUSION

To conclude this section, the results obtained from the sentiment classification as well as the sentiment visualisation confirm that the study aim and objectives were achieved. Despite the size difference between the two sentiment lexicons, the sentiment classification produced similar results.

The first set of enhanced word clouds successfully summarised the content within the Tweets as well as provided information about the overall sentiment within each collection of Tweets. The majority of the words in the word clouds were not classified according to the Bing sentiment lexicon, but still provided a summary of the more relevant topics regarding the COVID-19 pandemic such as “vaccination”, “death” and “fear”. Given that the first set of word clouds mainly consisted of unmatched words, the second set of word clouds focused on the words that are matched to the Bing sentiment lexicon to determine which of the sentiment words are more relevant within the Tweets and the overall sentiment of the words. These word clouds indicate that the overall sentiment within the Tweets is negative considering the uncertainty and fear related to the pandemic.

Word clouds, however, do not take context or negation into account, which results in a biased summary of the text. For example, the words “positive” and “negative” in this context could most likely refer to the results from testing for the virus and not positive or negative outlooks. The theme of the COVID-19 pandemic therefore introduced a new vocabulary, which is challenging for the overall interpretation of the word clouds. Although the enhanced word clouds contribute to the gauging of sentiment within the text, the similarity among the words are not displayed.

The MCA biplot enables the exploration of the relationships among the Tweets as well as the relationships among category levels of the categorical variables through the positioning of the points based on their similarity. A limitation of the presented standard MCA biplot, is that the words within each Tweet are not displayed. As an enhancement, the embedded word MCA biplot displays the same MCA solution as the MCA biplot, embeds the classifiable words in each Tweet onto the sample points within the biplot.

Groups of category level points are identified in the MCA biplots. This indicates that each group of category levels results in similar interpretations. The displays suggest that there are little significant differences regarding the location the Tweet was sent from. The category levels of the *City* or *Country* variables form their own group of points located in the centre surrounded by the other category level groups. The number of words considered for classification is found to influence the overall sentiment classification of the Tweet. The sentiment classification of Tweets with one classifiable word is solely dependent on the sentiment category of that one word and the category level point is thus located in the centre between the positive and sentiment category groups. Tweets with two or an even number of classifiable words present a likelihood that the sentiment classifications of the words balance each other out since a set of two words considered could have opposing sentiments or a set of four words could have two positive words and two negative words. This would result in these Tweets being associated with neutral sentiments. The Tweets with three or more words considered for classification are classified into a sentiment category in which majority of the words are classified into. The category levels for the *NumWords* variable where there are two or more words considered for classification therefore tend towards the neutral sentiment category. Groups of sentiment categories are formed by identifying category level points of similar sentiment in close proximity. This verifies that the sentiment classification results from both sentiment lexicons are similar. The embedded word MCA biplot confirms the success of the sentiment classification by identifying embedded words of similar sentiment in close proximity to the respective sentiment categories.

The study objectives and research aim were achieved, through the word clouds providing a summary of the content of the data as well as exposing the overall sentiment contained within

the text by including the sentiment classification. The embedded word MCA biplots enabled the visualisations of the sentiment classification results by simultaneously displaying the classifiable words in Tweets along with categorical levels associated with the recorded Tweets. This research enhanced existing approaches by successfully combining sentiment classification and visualisation.

To conclude this dissertation, a summary of the main findings, together with some final thoughts on the research is presented in the following CHAPTER 5. The final chapter also contains suggested research topics to be considered for future research.

CHAPTER 5
CONCLUDING REMARKS

5.1 INTRODUCTION

This final chapter serves as the conclusion to the study. It focuses on summarising the main findings revealed during the analysis and suggesting further research advancing from the study.

5.2 MAIN FINDINGS

This study aimed to gain insight from text through the quantification of the meaning and underlying context from available text data. This was achieved by enhancing existing visualisation techniques to graphically display the sentiment within the text.

The quantification of the underlying meaning and context in the text was determined through sentiment classification on each of the pre-processed Tweets in the procured data. The classification of the two sentiment lexicons, Bing and AFINN, were found to be in agreement. The MCA biplots revealed that the sentiment classification was found to be consistent by reason of the associated sentiment scores displayed in close proximity. Additionally, the consensus of the sentiment classification using the two different sentiment lexicons is confirmed by the percentages of words classified by the respective lexicons being almost equal (cf. Figure 4.6 and Figure 4.7).

It was shown systematically that the standard word clouds can be enhanced by adding colour coding to match the classification of negative and positive words according to their sentiment classification. Displaying the words in the cloud according to their sentiment classification provided a comprehensive summary of the relevant themes in the text as well as an assessment of the overall sentiment classification using the Bing lexicon. In order to expose the relative frequencies and sentiments of the extracted words, only the classifiable words were visualised in separate word clouds. This confirmed that the first study objective to enhance simple visualisations with the addition of sentiment classification was achieved.

Visualisation of the distinct sentiments within the text was also achieved by the second study objective, the standard MCA biplot. The biplots constructed from the two case studies exposed that that the Twitter users from different locations are not notably different, which was reflected by the close proximity for these category levels in the displays. As previously mentioned, the similarity in the sentiment classification from the lexicons were visually confirmed.

The relevance of the length of the processed Tweets (*NumWords*) to the sentiment classification is highlighted by the strong associations among some of the levels of the

NumWords variable and the sentiment category levels. The sentiment classification for Tweets with one classifiable word was found to be solely dependent on the sentiment category of that word. Tweets with an uneven number of words considered for classification were classified into the sentiment category in which a majority of the words within the Tweet were classified into. As soon as the number of classifiable words exceed one, the overall sentiment score is obtained as an “average” of the words under consideration, which results in a loss of strength of polarity.

The standard MCA biplot is thus an appropriate visualisation tool to evaluate the success of the sentiment classification as well as to examine the associations between the sentiment category levels, other categorical variables and groups of Tweets. Although the explicit interpretation of these Tweets was addressed in the enhancement of the MCA biplot, the embedded word MCA biplot continues examining the associations among the categorical variables as well as displaying the classifiable words per Tweet. This display provided promising results for the case studies since it not only grouped similar Tweets together, but grouped them in relation to the sentiment categories and lengths of Tweets they were associated to. The embedded word MCA biplot achieved both the aims of gauging sentiment within text as well as successfully visualising the distinct sentiments.

5.3 FURTHER RESEARCH

A natural progression of these case studies is to analyse the sentiment within Tweets regarding the COVID-19 pandemic over time. A repetition of the study over time could lead to further understanding of the sentiment towards this theme and could reveal more insight regarding the long-term effect of the pandemic.

During the pre-processing step, some limitations were revealed. Firstly, as stop words were removed, this included the deletion of some negational cues that influence the sentimental orientation of the observed text. An example would be to consider the pair of words “not” and “wrong”. Both of these words would be classified as negative and the observed text would be considered as negative, but if the words appeared together in the observed text as “not wrong”, the resulting classification would be considered as positive. This is a known limitation of sentiment classification as contractions and the informal language contained in Tweets are not recognised by the sentiment lexicon. As the sentiment classification procedure involves matching tokens to a sentiment lexicon, we suggest considering using a sentiment lexicon that contains pairs of words as well as individual words for a more accurate sentiment classification. Further to this, certain terms such as “positive” and “negative” were classified correctly according to the sentiment lexicons. However, in the context of the COVID-19 pandemic, these terms could indicate the results of a COVID-19 test, which would have been

associated with an opposite sentiment. The sentiment lexicons have been created for general sentiment classifications, whereas the theme of the case studies in this dissertation revolves around a specific pandemic. This demonstrates the difficulty of sentiment classification to adapt to diverse topics.

This study considered elements of both supervised and unsupervised classification, the individual words are matched to an existing sentiment lexicon, but the pre-processed Tweets were not labelled before the classification. Liu et al. (2013) proposed classifying Tweets with an adaptive multi-class support vector machine model in which the sentiment classification becomes a semi-supervised classification problem and could gauge the sentiment more accurately.

The MCA biplots revealed an association among the number of words considered for classification and the sentiment categories. Future studies might further explore the impact of the length of the pre-processed Tweets on the sentiment classification, since further interpretation is beyond the scope of the study. This dissertation focused on the visualisation of sentiment classification results and the number of classifiable words acted as an additional variable to perform MCA. The classifiable words used for the MCA biplots were matched from the AFINN sentiment lexicon. Some words included in the AFINN lexicon are not included in the Bing lexicon. The use of a different sentiment lexicon to extract the classifiable words could result in a different set of classifiable words. It would be possible to add another embedded layer to display the extracted words classified by the Bing sentiment lexicon. In this study the classification by the two lexicons were illustrated. Further research could be useful to validate the association among the category levels utilising a set of Tweets with a larger variety of lengths.

Despite the insight provided regarding the associations of words with sentiment categories, the legibility of the embedded word MCA biplot is hindered by the cluttering of words. This study made use of individual embedded word MCA biplots by suppressing different levels of the *NumWords* variable and displaying one particular level per display. Further research could explore interactive graphical displays which would display the initial MCA biplot, but allows the user to view the Tweet and extracted words related to a particular. The interactive display could suppress the levels of categorical variables in the data instead of obtaining separate displays for each level. These plotting capabilities could be explored by using for example the `R` package, `plotly` (Sievert, 2020).

5.4 IMPACT OF THE STUDY

The methodology presented in this study adds to the developing field of sentiment visualisation. In particular, the multivariate exploratory analysis of Tweets including the enhancement of embedded information enables the simultaneous interpretation of information related to procured Tweets along with their associated sentiment scores and additional available information. This study will aid as a primer for further research on sentiment classification and visualisation.

REFERENCES

Akcora, C.G., Bayir, M.A., Demirbas, M. & Ferhatosmanoglu, H. 2010. Identifying breakpoints in public opinion. *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. 62–66.

Blasius, J., Greenacre, M., Gower, M. & De Leeuw, J. 2006. Introduction. in *Multiple correspondence analysis and related methods* M. Greenacre & J. Blasius (eds.). Chapman & Hall/CRC M. Greenacre & J. Blasius (eds.). 1–76.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*. 28(2):15–21.

Cambria, E., Song, Y., Wang, H. & Howard, N. 2014. Semantic multidimensional scaling for open-domain sentiment analysis. *IEEE Intelligent Systems*. 29(2):44–51.

Cox, T.F. & Cox, M.A.A. 2001. *Multidimensional Scaling*. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Dovring, K. 1954. Quantitative semantics in 18th century Sweden. *Public Opinion Quarterly*. 18(4):389–394.

Feinerer, I. & Hornik, K. 2020. tm: Text Mining Package. *https://CRAN.R-project.org*. R package(version 0.7-8). [Online], Available: https://cran.r-project.org/package=tm.

Feldman, R. & Sanger, J. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Fellows, I. 2018. wordcloud: Word Clouds. *https://CRAN.R-project.org*. R package(version 2.6). [Online], Available: https://cran.r-project.org/package=wordcloud.

Friendly, M. 2008. The golden age of statistical graphics. *Statistical Science*. 23(4):502–535.

Friendly, M. & Denis, D. 2005. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*. 41(2):103–130.

Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*. 58(3):453–467.

Ganesh, P. 2019. *Sentiment Analysis : Simplified.* [Online], Available: https://towardsdatascience.com/sentiment-analysis-simplified-ac30720a5827 [2021, June 21].

Gentry, J. 2015. twitteR: R Based Twitter Client. *https://CRAN.R-project.org/*. R package(version 1.1.9). [Online], Available: https://cran.r-project.org/package=twitteR.

Gentzkow, M., Kelly, B. & Taddy, M. 2019. Text as data. *Journal of Economic Literature*.

57(3):535–574.

Gower, J.C. 1992. Generalized biplots. *Biometrika*. 79(3):475–493.

Gower, J., Gardner-Lubbe, S. & Le Roux, N. 2011. *Understanding biplots*. John Wiley & Sons, Ltd.

Greenacre, M. 2010. Multiple correspondence analysis biplots II. in *Biplots in Practice* Fundación BBVA. 99–108.

Greenacre, M. 2017. *Correspondence analysis in practice*. 3rd ed. Boca Raton: CRC Press.

Halvey, M.J. & Keane, M.T. 2007. An assessment of tag presentation techniques. *16th International World Wide Web Conference, WWW2007*. 1313–1314.

Hearst, M.A. & Rosner, D. 2008. Tag clouds: Data analysis tool or social signaller? *Proceedings of the Annual Hawaii International Conference on System Sciences*. 160.

Heiser, W.J. & Meulman, J. 1983. Empirical measures of distance are mostly called. *Journal of Econometrics*. 22:139–167.

History.com. 2019. *Twitter Launches*. [Online], Available: https://www.history.com/this-day-in-history/twitter-launches.

Hong, L. & Davison, B.D. 2010. Empirical Study of Topic Modeling in Twitter. *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. 80–88.

Hu, M. & Liu, B. 2004. Mining and summarizing customer reviews. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 168–177.

Hui, P. & Gregory, M. 2010. Quantifying sentiment and influence in blogspaces. *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. 53–61.

Ignatow, G. & Mihalcea, R. 2018. *An Introduction to Text Mining: research design, data collection, and analysis*. Thousand Oaks: SAGE Publications.

Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*. 60(11):2169–2188.

Java, A., Song, X., Finin, T. & Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* San Jose: Association for Computing Machinery, New York, NY, United States. 56–65. [Online], Available: http://www.pownce.com.

Johnson, R. & Wichern, D. 2014. *Applied multivariate statistical analysis*. 6th ed. Harlow: Pearson Education Limited.

Jolliffe, I.T. 2002. *Principal Component Analysis, Second Edition*. 2nd ed. Springer.

Jurafsky, D. 2012. Text Classification and Naïve Bayes. *Lecture slides CS224n: Natural Language Processing with Deep Learning*. [Online], Available: http://spark-public.s3.amazonaws.com/nlp/slides/naivebayes.pdf.

Kaur, A. & Gupta, V. 2013. A survey on sentiment analysis and opinion mining techniques. *Journal of Emerging Technologies in Web Intelligence*. 5(4):367–371.

Kim, K. & Lee, J. 2014. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. *Pattern Recognition*. 47(2):758–768.

Krippendorff, K. 2004. *Content analysis: an introduction to its methodology*. 2nd ed. SAGE Publications.

Kucher, K., Paradis, C. & Kerren, A. 2018. The State of the Art in Sentiment Visualization. *Computer Graphics Forum*. 37(1):71–96.

Kwak, H., Lee, C., Park, H. & Moon, S. 2010. What is Twitter, a social network or a news media? in *Proceedings of the 19th international conference on world wide web* (WWW '10). Raleigh: ACM. 591–600.

Kwartler, T. 2017. *Text Mining in pratice*. 1st ed. Hoboken: John Wiley & Sons.

Lazard, A.J., Scheinfeld, E., Bernhardt, J.M., Wilcox, G.B. & Suran, M. 2015. Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *American Journal of Infection Control*. 43(10):1109–1111.

Lee, B., Riche, N.H., Karlson, A.K. & Carpendale, S. 2010. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*. 16(6):1182–1189.

Leopold, E. 2007. Models of semantic spaces. in *Aspects of automatic text analysis* Berlin: Springer Science+Business Media. 117–138.

Leskovec, J., Rajaraman, A. & Ullman, J. 2019. Data mining. in *Mining of Massive Datasets* 3rd ed. Cambridge University Press. 1–17.

Liu, S., Li, F., Li, F., Cheng, X. & Shen, H. 2013. Adaptive Co-Training SVM for Sentiment Classification on Tweets. in *International conference on Information & Knowledge Management*. 2079–2088.

Loughran, T., Mcdonald, B., Battalio, R., Easton, P., Fuehrmeyer, J., Gao, P., Harvey, C., Hirschey, N., et al. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*. 66(1):35–65.

Lowe, W. 2001. Towards a theory of semantic space. *Proceedings of the annual meeting of the cognitive science society*. 23(23):276–281.

Mäntylä, M. V., Graziotin, D. & Kuutila, M. 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*. 27:16–32.

Medhat, W., Hassan, A. & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. 5(4):1093–1113.

Mendoza, M., Poblete, B. & Castillo, C. 2010. Twitter under crisis: Can we trust what we RT? *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. 71–79.

Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A. & Ghag, K. 2017. Dimensionality reduction for sentiment analysis using pre-processing techniques. in *International Conference on Computing Methodologies and Communication (ICCMC)* Institute of Electrical and Electronics Engineers Inc. 16–21.

Milgram, S. & Jodelet, D. 1976. Psychological Maps of Paris. in *Environmental Psychology: People and their physical settings* 2nd ed. H.M. Proshansky, W.H. Ittelson, & L.G. Rivlin (eds.). New York: Holt, Rinehart and Winston H.M. Proshansky, W.H. Ittelson, & L.G. Rivlin (eds.). 104–124.

Mohammad, S.M. & Turney, P.D. 2010. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. in *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* Los Angeles: Association for Computational Linguistics. 26–34.

Nenadic, O. & Greenacre, M. 2007. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*. 20(3):1–13. [Online], Available: http://www.jstatsoft.org.

Nielsen, F.Å. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CEUR Workshop Proceedings*. 718:93–98.

O'Keefe, T. & Koprinska, I. 2009. Feature selection and weighting methods in sentiment analysis. *ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium*. 67–74.

Osgood, C.E., May, W.H. & Miron, M.S. 1975. *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

Pang, B. & Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2(1–2):1–135.

Pang, B., Lee, L. & Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. in *Conference on empirical methods in natural language processing*.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. [Online], Available: https://www.r-project.org/.

Radovanovic, M. & Ivanovic, M. 2008. Text Mining: Approaches and Applications. *Novi Sad Journal of Mathematics*. 38(3):227–234.

Reitan, J., Faret, J., Gambäck, B. & Bungum, L. 2015. Negation scope detection for twitter sentiment analysis. *6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings*. (Wassa):99–108.

Seifert, C., Kump, B., Kienreich, W., Granitzer, G. & Granitzer, M. 2008. On the beauty and usability of tag clouds. *Proceedings of the International Conference on Information Visualisation*. 17–25.

Sharma, A. & Dey, S. 2012. A Comparative Study of Feature Selection and Machine Learning Techniques for Sentiment Analysis. in *ACM Research in Applied Computation Symposium*. 1–7.

Shayaa, S., Jaafar, N.I., Bahri, S., Sulaiman, A., Seuk Wai, P., Wai Chung, Y., Piprani, A.Z. & Al-Garadi, M.A. 2018. Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*. 6:37807–37827.

Shepherd, A., Sanders, C., Doyle, M. & Shaw, J. 2015. Using social media for support and feedback by mental health service users: Thematic analysis of a twitter conversation. *BMC Psychiatry*. 15(1).

Sievert, C. 2020. plotly: Interactive Web-Based Data Visualization with R, plotly, and shiny. *https://CRAN.R-project.org*. R package(version 4.9.3). [Online], Available: https://plotly-r.com.

Silge, J. & Robinson, D. 2016. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*. 1(3).

Slowikowski, K. 2021. ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”. *https://CRAN.R-project.org*. R package(version 0.9.1). [Online], Available: https://cran.r-project.org/package=ggrepel.

Thorley, J. 2004. *Athenian democracy, second edition*. 2nd ed. E.J. Evans & P.D. King (eds.). New York: Routledge: Taylor & Francis Group.

Tsirelson, B. 2018. Spaces in mathematics. *WikiJournal of Science*. 1(1):2.

Van der Klis, M. & Tellings, J. 2021. Generating semantic maps through multidimensional scaling: linguistic applications and theory. *Corpus Linguistics & Linguistic Theory*.

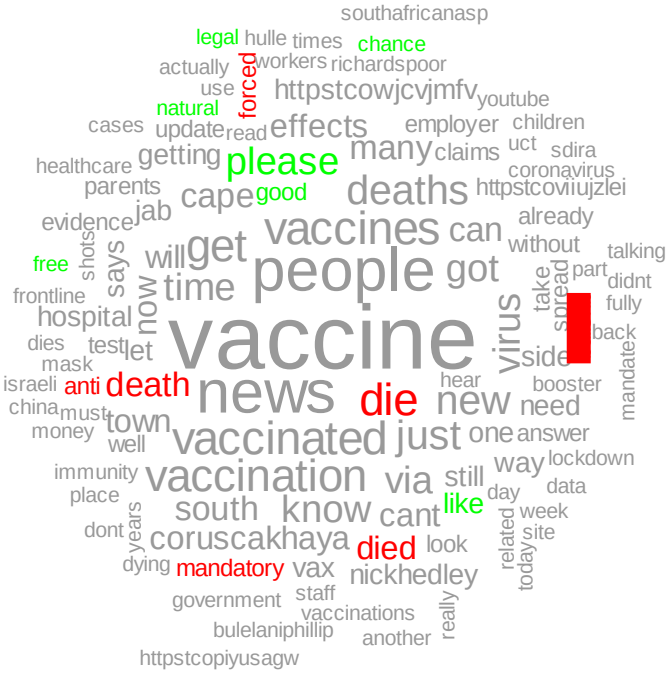
Yadollahi, A., Shahraki, A.G. & Zaiane, O.R. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*. 50(2).

APPENDIX A DETAILED IMAGES

This appendix contains more detailed images of the word clouds in CHAPTER 3 and CHAPTER 4.

A.1 Enhanced word clouds containing all words in corpus

A.1.1 Cape Town

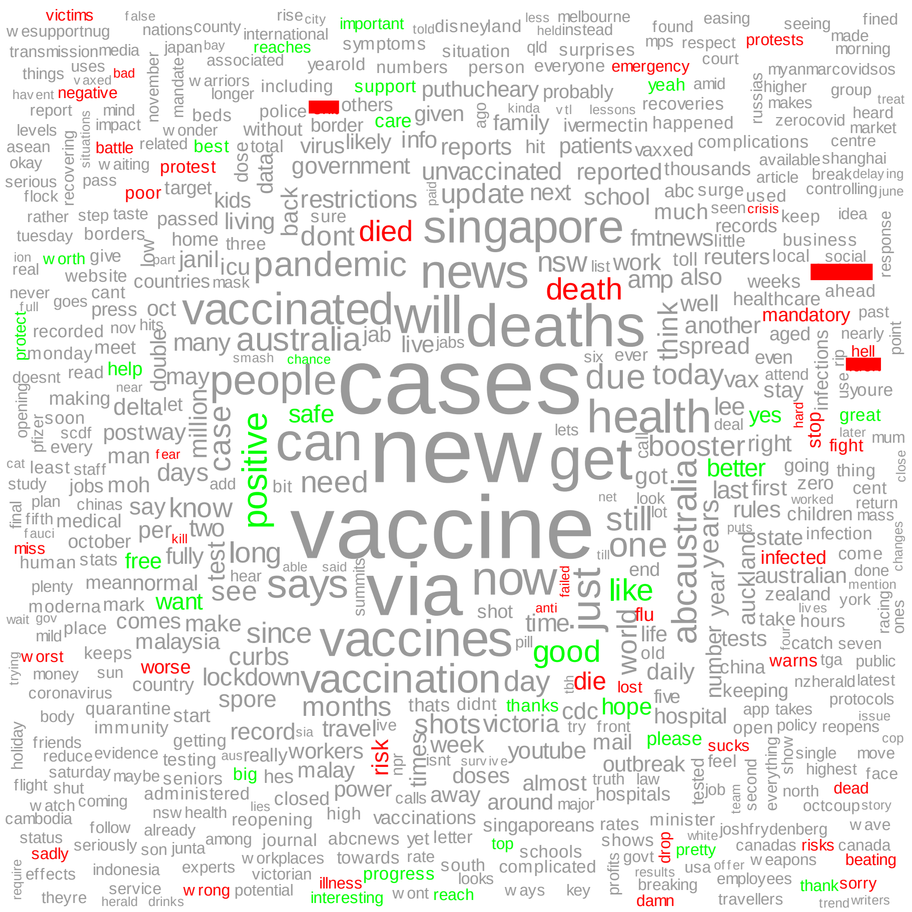


A.1.2 Durban

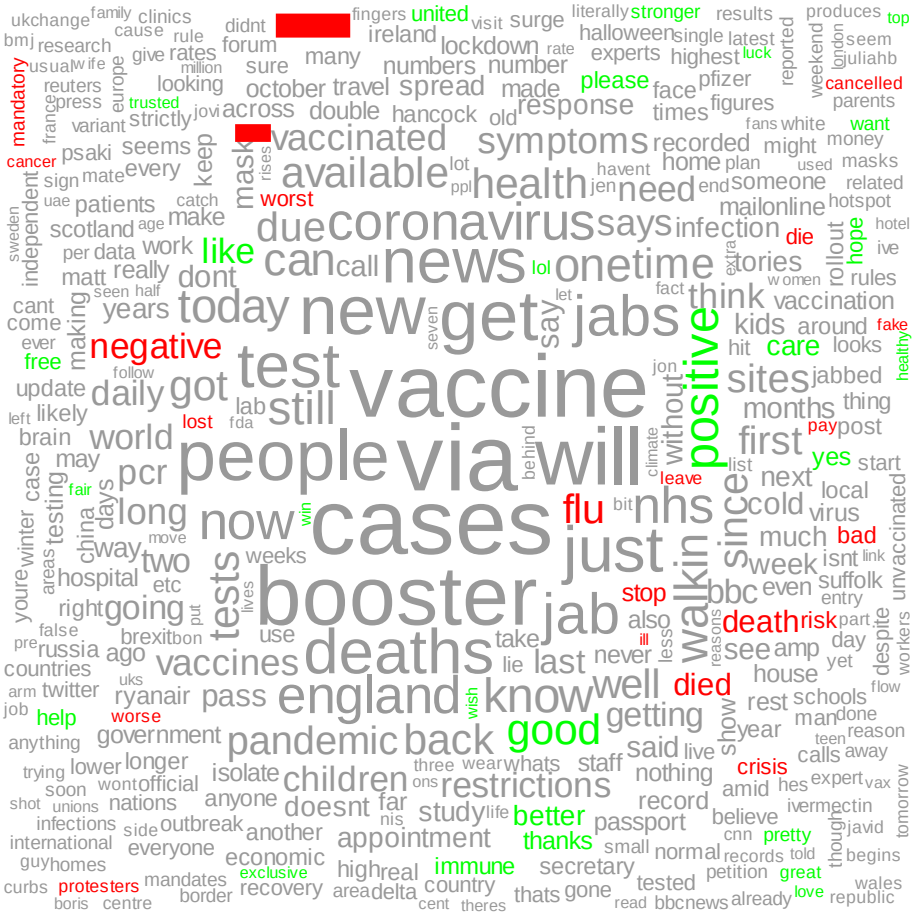
A.1.3 Johannesburg

A.1.4 South Africa

A.1.5 Australia



A.1.6 United Kingdom



A.2 Enhanced word clouds containing words matched to L^B

A.2.1 Cape Town

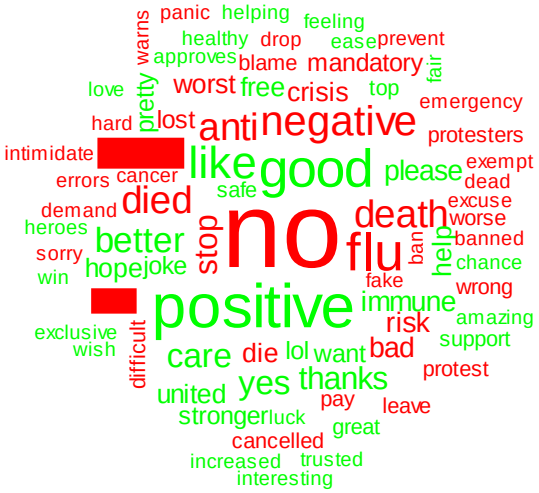
A.2.2 Durban

A.2.3 Johannesburg

A.2.4 South Africa

A.2.5 Australia

A.2.6 United Kingdom



APPENDIX B R CODE

The appendix for this study contains the R code used for the analysis as described in CHAPTER 3. The code is separated by headings which specifies the aim and subject of the functions used in each section. Thus, each section consists of the heading, required libraries and subsequent functions used to obtain the results. The first two sections involve the data procurement described in Section 1.3 as well as formatting the obtained data as a precursor to the analyses in the following sections. The provided code will produce the results for the "South African cities" case study and can be applied to the "countries" data to produce the results for the second case study. Various functions from the R packages, tidyverse and tidytext (Silge & Robinson, 2016) in combination with modified functions are used to perform the sentiment classification for the two different sentiment lexicons (cf. Sections 3.4.2 and 3.4.1). The word clouds section introduces a function which creates produces three individual word clouds given a set of Tweets and a keyword. The resulting word clouds detailed in Sections 3.3 and 3.4.2 provide enhanced visual aid to summarising the data while taking the sentiment polarity of the content into account. The second last section involves performing MCA and producing an MCA biplot discussed in Section 3.5. The last section of code produces the embedded word MCA biplots described in Section 3.6. As mentioned before, this R code and accompanying functions can be used to reproduce the methodology detailed in CHAPTER 3.

```
#####################################################################
#ACQUIRING DATA FROM TWITTER#
#####################################################################
library(twitteR)
#Connecting to Twitter
setup_twitter_oauth(consumer_key =
"gVyiusNI0DEo0efb3JYRgxv52",consumer_secret =
"GD6cDy3uZI6eOzIEF3n7qlszV4Q00VLJpGewmz3wvlXodGF5v9",access_token =
"941736808072843264-1vLZYXNZNVTmy4QtpHJ7iRzbJTUwBLK",access_secret =
"5ux9wxm0I4YQU7yT438X4uKc7FD843NbjUrxQlQNsMefE")
#Obtaining Tweets
covid.ct <- searchTwitter('covid', n = 1000,geocode = '-
33.9082139,18.4039413,50mi')
#Compiling Tweets over time (Applied to countries)
rsa.all<-c(covid.rsa,covid.rsa.2,covid.rsa.3,covid.rsa.4)
```

```
#################################################################
#FORMATTING DATA FROM TWITTER#
#################################################################
#Extracting text from lists of Tweets
CT.dat<-data.frame("Tweet" = sapply(X = covid.ct,FUN = function(x)
x$getText()),"City" = rep("CT",1000))
#Combining Tweets from different locations
##Table format for Table xx
SACity<-rbind(CT.dat,DBN.dat,JHB.dat)
##List format for sentiment classification
SACity.list<-c(covid.ct,covid.durb,covid.jhb)
#################################################################
#SENTIMENT CLASSIFICATION#
#################################################################
library(tidyverse)
library(tidytext)
#Sentiment lexicons
##AFINN lexicon
L_A<-get_sentiments(lexicon = c("afinn"))
##Bing lexicon
L_B<-get_sentiments(lexicon = c("bing"))

#FUNCTIONS#
##Sentiment classification using AFINN sentiment lexicon
afinn.sc<-function(tweetx){

#################################################################
#################################################################
  #Information
  #This function performs sentiment classification using the AFINN
sentiment lexicon for a Tweet obtained using Twitter API
#################################################################
  #Arguments
  #"tweetx" a Tweet obtained using Twitter API
#################################################################
  #Value
  #"Text" the text contained in the Tweet
  #"Sentiment_words" words considered for sentiment classification
  #"AFINN_Score" overall sentiment score for the Tweet
```

```
  #"NumWords" number of words considered for sentiment
classification
  #"SentTab" results from the sentiment classification before
determining overall score
#####################################################################
#####################################################################
  #Obtain text from Tweets
  textdat<-tweetx$getText()
  #Pre-processing
  textdat<-iconv(textdat,'UTF-8','ASCII')
  #Tokenisation
  tokens<-data_frame(text = textdat) %>% unnest_tokens(word,text)
  #Sentiment classification
  senti.tab<-tokens %>% inner_join(L_A) %>% count(value)
  #return(senti.tab)
  #Overall sentiment score
  afinn.score<-
round((sum(senti.tab$value*senti.tab$n))/sum(senti.tab$n),0)
  #return(textdat)
  #Sentiment words in text
  word.vec<-tokens %>% inner_join(L_A) %>% count(word)
  word.vec<-word.vec[,1]
  nw<-nrow(word.vec)
  #return(nw)
  #word.vec<-paste(word.vec$word,sep = "",collapse = "")
  word.vec<-paste(word.vec$word,sep = "_",collapse = "_")
  #Output
  list("Text" = tweetx$getText(),"Sentiment_words" =
word.vec,"AFINN_Score" = afinn.score,"NumWords" = nw,"SentTab" =
senti.tab)
}


##Sentiment classification using Bing sentiment lexicon
bing.sc<-function(tweetx){
#####################################################################
#####################################################################
  #Information
  #This function performs sentiment classification using the Bing
sentiment lexicon for a Tweet obtained using Twitter API
#####################################################################
```

```
  #Arguments
  #"tweetx" a Tweet obtained using Twitter API
#####################################################################
  #Value
  #"score" overall sentiment score for the Tweet
  #"senti.tab" results from the sentiment classification before
determining overall score
#####################################################################
#####################################################################
  #Obtain text from Tweets
  textdat<-tweetx$getText()
  #Pre-processing
  textdat<-iconv(textdat,'UTF-8','ASCII')
  #Tokenisation
  tokens<-data_frame(text = textdat) %>% unnest_tokens(word,text)
  #Sentiment classification
  senti.tab<-tokens %>% inner_join(L_B) %>% count(sentiment)
  ns<-nrow(senti.tab)
  sB<-ifelse(test = ns==1,yes = senti.tab$sentiment,no = ifelse(test
=
ns>1&&senti.tab$n[senti.tab$sentiment=="positive"]==senti.tab$n[sent
i.tab$sentiment=="negative"],yes = "neutral",no = ifelse(test =
ns>1&&senti.tab$n[senti.tab$sentiment=="positive"]>senti.tab$n[senti
.tab$sentiment=="negative"],yes = "positive",no = ifelse(test =
ns>1&&senti.tab$n[senti.tab$sentiment=="positive"]<senti.tab$n[senti
.tab$sentiment=="negative"],yes = "negative",no = NA))))
  list("score" = sB, "senti.tab" = senti.tab)
}


#Performing sentiment classification on Tweets
##Creating variables per Tweet
sentab.sac<-list() #List of sentiment classification results per
Tweet
words.sacity<-NULL #Words considered for classification
afscore.sacity<-NULL #Aggregated sentiment score using AFINN lexicon
bingscore.sacity<-NULL #Aggregated sentiment score using Bing
lexicon
nw.sacity<-NULL #Number of words considered for classification
##Filling in variable values per Tweet
for(i in 1:nrow(SACity)){
```

```
  words.sacity[i]<-afinn.sc(tweetx =
SACity.list[[i]])$Sentiment_words
  afscore.sacity[i]<-afinn.sc(tweetx = SACity.list[[i]])$AFINN_Score
  bingscore.sacity[i]<-bing.sc(tweetx = SACity.list[[i]])
  nw.sacity[i]<-afinn.sc(tweetx = SACity.list[[i]])$NumWords
  sentab.sac[[i]]<-afinn.sc(tweetx = SACity.list[[i]])$SentTab
}
##Adding variables to data matrix
SACity$Words<-words.sacity
SACity$AFScore<-afscore.sacity
SACity$BingScore<-bingscore.sacity
SACity$NumWords<-nw.sacity

##Categorising overall AFINN scores into one of five levels
afinn.cut<-function(out){
#########################################################################
#########################################################################
  #Information
  #This function converts the overall AFINN scores per Tweet into
one of five category levels
#########################################################################
#########################################################################
  #Arguments
  #"out" the aggregated sentiment score using the AFINN lexicon
#########################################################################
#########################################################################
  #Value
  #"new.col" categorised AFINN sentiment score per Tweet
#########################################################################
#########################################################################
  #Numeric AFINN sentiment score per Tweet
  out <- as.numeric(out)
  #Creating an empty character object
  new.col <- character(length(out))
  #Categorising the numeric score
  new.col[out<= -3] <- "[-5;-3]"
  new.col[out > -3 & out <= -1] <- "(-3;-1]"
  new.col[out == 0] <- "0"
  new.col[out >= 1 & out < 4] <- "[1;3)"
  new.col[out >= 3 & out <= 5] <- "[3;5]"
```

```
  #Output
  new.col
}


SACity$AFINNcats<-afinn.cut(out = SACity$AFScore)
#Determining the percentage of Tweets that were classified using
both sentiment lexicons
afinn.match.sa<-
round((sum(!is.na(SACity$AFScore))/sum(SACity$NumWords))*100,2)
bing.match.sa<-
round((sum(!is.na(SACity$BingScore))/sum(SACity$NumWords))*100,2)
SACity$Match<-rep(0,nrow(SACity))
for(i in 1:nrow(SACity)){
  SACity$Match[i]<-ifelse(test =
(!is.na(SACity$AFScore[i])&&!is.na(SACity$BingScore[i]))==TRUE,yes =
1,no = 0)
}
both.match.sa<-round((sum(SACity$Match)/sum(SACity$NumWords))*100,2)
##Summarise matches using a bar graph
library(ggplot2)
plot.sc.city<-data.frame("Name" =
c("Bing","AFINN","Both"),"Percentages" =
c(bing.match.sa,afinn.match.sa,both.match.sa))
ggplot(data = plot.sc.city,mapping = aes(x = plot.sc.city[,1], y =
plot.sc.city[,2]))+geom_bar(stat = "identity", fill = "blue",color =
"blue")+geom_text(mapping = aes(label = paste(Percentages,"%"),vjust
= -0.25),size = 6)+scale_y_continuous(limits = c(0,100))+labs(x =
"Sentiment lexicon used for sentiment classification",y =
"Percentage of words classified using each sentiment
lexicon")+theme_bw()+theme(axis.text = element_text(size = 14))


#####################################################################
#WORD CLOUDS#
#####################################################################
library(tm)
library(wordcloud)
library(expss)
#FUNCTION FOR BOTH WORD CLOUDS
wordcloud_sc<-function(tweetlistx,keyword){
  library(tidyverse)
  library(tidytext)
  library(tm)
```

```
  library(wordcloud)
  library(expss)
  #Obtaining text from list of Tweets
  textx<-sapply(X = tweetlistx,FUN = function(x) x$getText())
  textx<-iconv(textx,'UTF-8','ASCII')
  #Creating a corpus based on all words in the text
  corp<-Corpus(x = VectorSource(x = textx))
  #Creating term document matrix
  tdm.x<-TermDocumentMatrix(x = corp, control =
list(removePunctuation = TRUE,stopwords =
c(keyword,"@","http",stopwords("english"),"https"),removeNumbers =
TRUE,tolower = TRUE))
  tdm.x<-as.matrix(tdm.x)
  #Colors
  lexicon.afinn<-get_sentiments(lexicon = 'afinn')
  lexicon.afinn$col<-ifelse(test = lexicon.afinn$value>0,yes =
"green",no = ifelse(test = lexicon.afinn$value<0,yes = "red",no =
"black"))
  #Wordcloud with all words
  word.freq <- sort(rowSums(tdm.x),decreasing = TRUE)
  dm.x <- data.frame(word = names(word.freq),freq = word.freq)
  tokens.dm<-data_frame(text = dm.x$word) %>%
unnest_tokens(word,text)
  full.set<-tokens.dm %>% full_join(lexicon.afinn)
  #full.set$freq<-vlookup(lookup_value = full.set$word,dict =
dm.x,result_column = 2,lookup_column = 1)
  full.set$col<-rep(0,nrow(full.set))
  #fsc.1<-NULL
  na.ind<-which(is.na(full.set$value))
  pos.ind<-which(full.set$value>0)
  neg.ind<-which(full.set$value<0)
  neu.ind<-which(full.set$value==0)
  full.set$col[na.ind]<-"gray60"
  full.set$col[pos.ind]<-"green"
  full.set$col[neg.ind]<-"red"
  full.set$col[neu.ind]<-"black"
  #return(full.set)
  rescol<-which(colnames(full.set)=="col")
  lookcol<-which(colnames(full.set)=="word")
```

```
  dm.x$col<-vlookup(lookup_value = dm.x$word,dict =
full.set,result_column = rescol,lookup_column = lookcol)

  windows()

  wordcloud(words = dm.x$word,freq = dm.x$freq,random.order = FALSE,
colors = "black")

  windows()

  wordcloud(words = dm.x$word,freq = dm.x$freq,random.order =
FALSE,ordered.colors = TRUE,colors = dm.x$col)

  #Wordcloud with sentiment words only

  tokens<-data_frame(text = textx) %>% unnest_tokens(word,text)

  words.count<-tokens %>% inner_join(lexicon.afinn) %>% count(word)

  words.count<-words.count[order(words.count$n,decreasing = TRUE),]

  words.col<-words.count %>% inner_join(lexicon.afinn)

  windows()

  wordcloud(words = words.count$word,freq =
words.count$n,random.order = F,random.color = F,ordered.colors =
TRUE,colors = words.col$col)

  #Frequencies of words and colours

  freq.col.<-table(dm.x$col)

  total<-sum(freq.col.)

  freq.col.perc<-round((freq.col./total)*100,4)

  #Output

  list("colourtab" = freq.col.perc, "TDM" = dm.x)

}


WC.CT<-wordcloud_sc(tweetlistx = covid.ct,keyword = "covid")


##Extracting relative frequencies

ct.rf<-WC.CT$colourtab


##Plot of positive vs negative relative frequencies

#rf.tab<-"Sentiment "=c(rep("Positive",6),rep("Negative",6))

posi.rf<-
c(ct.rf["green"],dbn.rf["green"],jhb.rf["green"],rsa.rf["green"],aus
.rf["green"],uk.rf["green"])

names(posi.rf)<-c("Cape Town","Durban","Johannesburg","South
Africa","Australia","United Kingdom")

negi.rf<-
c(ct.rf["red"],dbn.rf["red"],jhb.rf["red"],rsa.rf["red"],aus.rf["red
"],uk.rf["red"])
```

```
names(negi.rf)<-c("Cape Town","Durban","Johannesburg","South
Africa","Australia","United Kingdom")

#rf.tab<-as.data.frame(rbind(posi.rf,negi.rf))

rf.tab<-data.frame("Sentiment
Category"=c(rep("Positive",6),rep("Negative",6)),"Location" =
c("Cape Town","Durban","Johannesburg","South
Africa","Australia","United Kingdom"),"Frequencies" =
c(posi.rf,negi.rf))

ggplot(data = as.data.frame(rf.tab),mapping = aes(x = Location,y =
Frequencies,fill = Sentiment.Category))+geom_bar(stat =
"identity",position = position_dodge())+geom_text(mapping =
aes(label = paste(round(Frequencies,1),"%",sep = "")),position =
position_dodge(1),vjust = -1.5,size = 5)+scale_y_continuous(limits =
c(0,8))+scale_x_discrete(limits = c("Cape
Town","Durban","Johannesburg","South Africa","Australia","United
Kingdom"))+scale_fill_manual(values =
c("red","green"))+theme(legend.position = "bottom",axis.line =
element_line(size = 0.75,color = "black"), axis.text.y =
element_text(size = 14),axis.text.x = element_text(size = 14,angle =
90,hjust = -0.00005),panel.background = element_blank(),panel.grid =
element_blank(),axis.title = element_text(size = 14),axis.ticks =
element_blank())


#####################################################################
#MCA BIPLOT#
#####################################################################
library(ca)
#MCA on data using variables City, BingScore, NumWords and AFINNcats
SACity$NumWords<-as.factor(SACity$NumWords)
mca.dat.sacity<-SACity[,c(2,5,6,7)]
mca.sacity<-mjca(obj = mca.dat.sacity)


biplFig <- function (CLPs, Zs, Lvls=NULL, Z.col="grey37",
CLP.col="forestgreen", Z.pch=1, CLP.pch=17,title="",
Z.cex=1,CLP.cex=1.7,exp.factor = 0.2)
{

#####################################################################
#####################################################################
  #Information
  #This function constructs a biplot after MCA

#####################################################################
  #Arguments
  #"CLPs" category level points (standard coordinates for variables)
```

```
  #"Zs" sample principal coordinates
  #"Lvls" names of the CLPs
  #"Z.col", "CLP.col" are the colour specifications for samples and
CLPs
  #"Z.pch", "CLP.pch" are the plotting character specifications for
samples and CLPs
  #"title" title of the figure

#######################################################################
  #Value
  #Returns a biplot.

#######################################################################

#######################################################################
  #dev.new()
  #usr <- par("usr")
  #clip(usr[1], -2, usr[3], usr[4])
  #clip(2, usr[2], usr[3], usr[4])
  par(pty="s")
  temp <<- rbind(CLPs[,1:2],Zs[,1:2])
  rangex <<- range(temp[,1])
  rangey <<- range(temp[,2])
  xminmax <<- c(rangex[1]-exp.factor*(rangex[2]-rangex[1]),
rangex[2]+exp.factor*(rangex[2]-rangex[1]))
  yminmax <<- c(rangey[1]-exp.factor*(rangey[2]-rangey[1]),
rangey[2]+exp.factor*(rangey[2]-rangey[1]))
  plot(x = xminmax, y = yminmax, xlim = xminmax,
       ylim = yminmax, xaxt = "n",
       yaxt = "n", xlab = "", ylab = "", type = "n", xaxs = "i",
       yaxs = "i",  main=title)

#plot(rbind(CLPs[,1:2],Zs[,1:2]),pch="",xaxt="n",yaxt="n",xlab="",yl
ab="",main=title)
  points(Zs,pch=Z.pch,col=Z.col,cex=Z.cex)
  points(CLPs,pch=CLP.pch,col=CLP.col,cex=CLP.cex)


  is.null(Lvls)
  {
    text(CLPs,cex=0.7,label=rownames(CLPs),pos=3,xpd=NA)
  }
```

```
  !is.null(Lvls)
  {
    text(CLPs,cex=0.7,label=Lvls,pos=3,xpd=NA)
  }
}
#MCA Biplot
biplFig(CLPs=mca.sacity_2$colcoord,Zs=mca.sacity_2$rowpcoord,Lvls =
mca.sacity_2$levelnames,CLP.cex=2,Z.col = "gray60",Z.pch = 20,Z.cex
= 1.5,CLP.col = "blue",CLP.pch =
c(9,9,9,16,15,17,25,8,8,8,8,16,16,15,17,17))


#####################################################################
#EMBEDDED WORD MCA BIPLOT#
#####################################################################
library(RColorBrewer)
library(ggrepel)
#Setting colours for each category of NumWords
##Creating index vectors for the Tweets associated with each
category of NumWords
NW1<-which(SACity$NumWords==1)
NW2<-which(SACity$NumWords==2)
NW3<-which(SACity$NumWords==3)
NW4<-which(SACity$NumWords==4)
NW5<-which(SACity$NumWords==5)
##Setting colours
colors.nw<-brewer.pal(n = 5,name = "Set2")


##Figure 3.5 - MCA biplot without embedded words
word.plot.d <-
as.data.frame(rbind(mca.sacity$rowpcoord[,1:2],mca.sacity$colcoord[,
1:2]))
lab1 <- SACity$Words
lab2 <- mca.sacity$levelnames
p.d <- ggplot(data = word.plot.d[1:334,], mapping =
aes(word.plot.d[1:334,1],word.plot.d[1:334,2])) +geom_point(color =
"red", label = lab1)+geom_text(data = word.plot.d[335:350,],mapping
= aes(word.plot.d[335:350,1],word.plot.d[335:350,2],label = lab2,
fontface = 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension
2") + theme_bw()+ theme(aspect.ratio=1)+scale_x_continuous(limits =
c(-4,4))+scale_y_continuous(limits = c(-4,4))+theme(axis.text =
element_blank(),panel.grid = element_blank(),axis.ticks =
element_blank())
```

```
##Figure 3.6 - Embedded word MCA biplot

word.plot.0 <-
as.data.frame(rbind(mca.sacity$rowpcoord[,1:2],mca.sacity$colcoord[,
1:2]))

lab1 <- SACity$Words

lab2 <- mca.sacity$levelnames

p.0 <- ggplot(data = word.plot.0[1:334,], mapping =
aes(word.plot.0[1:334,1],word.plot.0[1:334,2])) +geom_point(color =
"red", label = lab1)+geom_text_repel(color =
"gray60",max.overlaps=300,label = lab1)+geom_text(data =
word.plot.0[335:350,],mapping =
aes(word.plot.0[335:350,1],word.plot.0[335:350,2],label = lab2,
fontface = 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension
2") + theme_bw()+ theme(aspect.ratio=1)+scale_x_continuous(limits =
c(-4,4))+scale_y_continuous(limits = c(-4,4))+theme(axis.text =
element_blank(),panel.grid = element_blank(),axis.ticks =
element_blank())


#Embedded word MCA biplot

word.plot <-
as.data.frame(rbind(mca.sacity$rowpcoord[,1:2],mca.sacity$colcoord[,
1:2]))

lab1 <- SACity$Words

lab2 <- mca.sacity$levelnames

lab2[7:11]<-"" #supressing NumWords

p <- ggplot(data = word.plot[1:334,], mapping =
aes(word.plot[1:334,1],word.plot[1:334,2])) +

  geom_point(color = "red", label = lab1) + geom_text_repel(mapping
= aes(color = as.factor(SACity$NumWords)),max.overlaps=300,label =
lab1) + geom_text(data = word.plot[335:350,],mapping =
aes(word.plot[335:350,1],word.plot[335:350,2],label = lab2, fontface
= 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension 2") +
theme_bw()+ theme(aspect.ratio=1,legend.position =
"bottom")+scale_color_brewer(name = "Number of words",palette =
"Set2",direction = 1)+scale_x_continuous(limits = c(-
2.5,2.5))+scale_y_continuous(limits = c(-3.5,3.5))+theme(axis.text =
element_blank(),panel.grid = element_blank(),axis.ticks =
element_blank())

#Individual embedded word MCA biplots

##One word

word.plot.1 <-
as.data.frame(rbind(mca.sacity$rowpcoord[NW1,1:2],mca.sacity$colcoor
d[,1:2]))

lab1 <- SACity$Words[SACity$NumWords==1]

lab2 <- mca.sacity$levelnames
```

```
lab2[7:11]<-""

p.1 <- ggplot(data = word.plot.1[1:198,], mapping =
aes(word.plot.1[1:198,1],word.plot.1[1:198,2])) +geom_point(color =
"red", label = lab1) + geom_text_repel(max.overlaps=300,label =
lab1,color = colors.nw[1]) + geom_text(data =
word.plot.1[199:214,],mapping =
aes(word.plot.1[199:214,1],word.plot.1[199:214,2],label = lab2,
fontface = 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension
2") + theme_bw()+theme(aspect.ratio = 1)+scale_x_continuous(limits =
c(-2.5,2.5))+scale_y_continuous(limits = c(-
3.5,3.5))+theme(axis.text = element_blank(),panel.grid =
element_blank(),axis.ticks = element_blank())

##Two words

word.plot.2 <-
as.data.frame(rbind(mca.sacity$rowpcoord[NW2,1:2],mca.sacity$colcoor
d[,1:2]))

lab1 <- SACity$Words[SACity$NumWords==2]

lab2 <- mca.sacity$levelnames

lab2[7:11]<-""

p.2 <- ggplot(data = word.plot.2[1:95,], mapping =
aes(word.plot.2[1:95,1],word.plot.2[1:95,2])) +geom_point(color =
"red", label = lab1) + geom_text_repel(max.overlaps=300,label =
lab1,color = colors.nw[2]) + geom_text(data =
word.plot.2[96:111,],mapping =
aes(word.plot.2[96:111,1],word.plot.2[96:111,2],label = lab2,
fontface = 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension
2") + theme_bw()+ theme(aspect.ratio=1)+scale_x_continuous(limits =
c(-2.5,2.5))+scale_y_continuous(limits = c(-
3.5,3.5))+theme(axis.text = element_blank(),panel.grid =
element_blank(),axis.ticks = element_blank())

##Three words

word.plot.3 <-
as.data.frame(rbind(mca.sacity$rowpcoord[NW3,1:2],mca.sacity$colcoor
d[,1:2]))

lab1 <- SACity$Words[SACity$NumWords==3]

lab2 <- mca.sacity$levelnames

lab2[7:11]<-""

p.3 <- ggplot(data = word.plot.3[1:35,], mapping =
aes(word.plot.3[1:35,1],word.plot.3[1:35,2])) +geom_point(color =
"red", label = lab1) + geom_text_repel(max.overlaps=300,label =
lab1,color = colors.nw[3]) + geom_text(data =
word.plot.3[36:51,],mapping =
aes(word.plot.3[36:51,1],word.plot.3[36:51,2],label = lab2, fontface
= 2),color = "blue")+labs(x = "Dimension 1", y = "Dimension 2") +
theme_bw()+ theme(aspect.ratio=1)+scale_x_continuous(limits = c(-
2.5,2.5))+scale_y_continuous(limits = c(-3.5,3.5))+theme(axis.text =
element_blank(),panel.grid = element_blank(),axis.ticks =
element_blank())

##Four and five words
```

```
word.plot.4 <-
as.data.frame(rbind(mca.sacity$rowpcoord[NW4,1:2],mca.sacity$colcoor
d[,1:2]))

word.plot.5 <-
as.data.frame(rbind(mca.sacity$rowpcoord[NW5,1:2],mca.sacity$colcoor
d[,1:2]))

lab1_1 <- SACity$Words[SACity$NumWords==4]

lab1_2 <- SACity$Words[SACity$NumWords==5]

lab2 <- mca.sacity$levelnames

lab2[7:11]<-""

p.4 <- ggplot(data = word.plot.4[1:4,], mapping =
aes(word.plot.4[1:4,1],word.plot.4[1:4,2])) +

  geom_point(color = "red", label = lab1_1) +
geom_text_repel(max.overlaps=300,label = lab1_1,color =
colors.nw[4])+geom_point(data = word.plot.5[1:2,], mapping = aes(x =
word.plot.5[1:2,1],y = word.plot.5[1:2,2]),color = "red", label =
lab1_2) + geom_text_repel(data = word.plot.5[1:2,], mapping = aes(x
= word.plot.5[1:2,1],y = word.plot.5[1:2,2]),max.overlaps=300,label
= lab1_2,color = colors.nw[5])+geom_text(data =
word.plot.4[5:20,],mapping =
aes(word.plot.4[5:20,1],word.plot.4[5:20,2],label = lab2, fontface =
2),color = "blue")+labs(x = "Dimension 1", y = "Dimension 2") +
theme_bw()+ theme(aspect.ratio=1)+scale_x_continuous(limits = c(-
2.5,2.5))+scale_y_continuous(limits = c(-3.5,3.5))+theme(axis.text =
element_blank(),panel.grid = element_blank(),axis.ticks =
element_blank())
```