Financial Sentiment Analysis: an NLP approach towards reputation management

by

Michelle Terblanche

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

December 2020

Publication data:

Michelle Terblanche. Financial Sentiment Analysis: an NLP approach towards reputation management. Masters mini-dissertation, University of Pretoria, Department of Computer Science, Pretoria, South Africa, December 2020.

Sentiment Analysis of Financial Information: an NLP approach towards reputation management

by

Michelle Terblanche

E-mail: michelle.terblanche@gmail.com

Abstract

Sentiment analysis as a sub-field of natural language processing has received increased attention in the past decade enabling organisations to more effectively manage their reputation through online media monitoring. Many drivers impact reputation, however, this thesis focuses only the aspect of financial performance and explores the gap with regards to financial sentiment analysis in a South African context.

Results showed that pre-trained sentiment analysers are least effective for this task and that traditional lexicon-based and machine learning approaches are best suited to predict financial sentiment of news articles. The study contributed to updating an existing sentiment dictionary and developing a full pipeline to filter data for financial topics and predict sentiment. Using a binary logistic regression model and a binary XGBoost classifier on both headlines and article content produced accuracies of >85%. The predicted sentiments correlated quite well with share price and highlighted the potential use of sentiment as an indicator of financial performance.

Model generalisation was less acceptable due to the limited amount of training data used. Future work includes expanding the data set to improve general usability and contribute to an open-source financial sentiment analyser for South African data.

Keywords: financial sentiment analysis, natural language processing, corporate reputation, South Africa, share price

Supervisors : Dr. V. N. Marivate
Department : Department of Computer Science
Degree : Master of Information Technology

"Without data, you're just another person with an opinion."

W. Edwards Deming, American Engineer

"It takes 20 years to build a reputation and five minutes to ruin it. If you think about that, you'll do things differently."

Warren Buffett, CEO of Berkshire Hathaway (1892)

"A brand for a company is like a reputation for a person. You earn reputation by trying to do hard things well."

Jeff Bezos, CEO of Amazon

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Big corporate organisations produce vast amounts of textual information in the form of official financial and non-financial reports, media releases and trading statements. As a result even more information is produced by online news publishers, social media platforms and investors. This deluge of data is available in different formats, from a multitude of sources and produced at varying rates and intervals in time. The nature of the content also vary from financial, to sport to corporate social responsibility. These contribute differently to market perception and with so much information available it is often difficult to contextualise and derive value for improved decision-making.

The communication strategy of an organisation directly impacts it's reputation. It is therefore key to understand the impact of perception on the organisation. One of the industry accepted measures of reputation, *the RepTrak Score*[1], takes into account the following seven drivers of reputation:

- products and services
- innovation
- workplace
- citizenship
- governance

[1] https://www.reptrak.com/reputation-intelligence/what-is-it/

- leadership
- performance

The sentiment (from both customer and media) of these dimensions are continuously tracked to calculate the overall *RepTrak Score*[2]. Last on the list is ***Performance*** which is a measure of the financial health of an organisation.

In the past, historic accounting information formed the basis for financial performance prediction, evolving from statistical models to more sophisticated machine learning models [12]. Subsequent research ventured into the field of qualitative measures such as textual analysis to predict performance [26]. More recent research shows that there is promise in correlating sentiment with financial performance in order to make future predictions [12, 18, 26, 31].

1.1 Motivation

The justification for this study is two-fold and is elaborated on in the sections below.

1.1.1 Financial Value of Reputation

Research has found that there is a financial value linked to the reputation of an organisation [25, 40]. A reputational landscape overview stated that reputation can be seen as an "information signal" that can increase investor confidence [11]. Furthermore, an improvement in reputation can have in the order of a 6% improvement in the company bottom-line [25]. As a result, reputation risk should form a key component of overall corporate strategy [40]

Therefore, understanding and improving reputation can have a significant impact on share value, return on investment and achieving and sustaining a competitive advantage.

1.1.2 Improved Sentiment Analysis of South African Text

Some of the most common methods for sentiment analysis are as follows:

[2]https://www.reptrak.com/reptrak-platform/reptrak-difference/

- A simple dictionary-based approach
- Open-source pre-trained sentiment analysers
- Custom-built models using machine learning

When using for e.g. a dictionary-based approach, general sentiment word lists are often too generic and misclassify words in a financial context [26]. It is hypothesized that a similar misclassification is likely when extending domain-specific dictionaries to different language dialects i.e. American vs. South African English.

Furthermore, even though many sentiment analysers are freely available (some to be discussed in Section 2.2), these models were developed within a given context and relevant to a specific domain. The hypothesis is that these existing sentiment analysers have shortcomings when applied to a new domain or even a similar domain (such as financial) where the use of language may vary for e.g. South Africa vs. USA.

Research has also shown that there is limited South African studies related to financial sentiment analysis using natural language processing (NLP) techniques. This gap in the field of an industry application is the anchor point for the thesis.

1.2 Objectives

Based on the identified problem and motivation, the goal of the dissertation is to answer each of the following research questions and achieve the corresponding objectives:

Main Research Question

> "*What NLP techniques are required to successfully determine the sentiment of financial communication?*"

Objectives:

- Evaluate and compare existing "off-the-shelf" sentiment analysers
- Evaluate and compare alternative methods to predict sentiment
- Recommend the most suitable method for predicting sentiment

- Understand the minimum requirements to effectively predict sentiment for e.g. headlines vs. document content

Sub-Question 1

"*Is there a correlation between the sentiment of financial news and company performance as indicated by share price?*"

Objective: Determine whether a trend can be observed between sentiment of financial-related documents and share price through a graphical representation.

Sub-Question 2

"*How effectively can a narrower sentiment prediction model be applied to a broader scope of finance-related information?*"

Objectives:

- Develop a sentiment prediction model using financial communication for a specific corporate organisation
- Determine how well such a model generalises when using data from a different organisation
- Comment on the application of financial sentiment prediction models and the limitations

In light of the above-mentioned questions and objectives, the thesis of this study is that natural language processing techniques can successfully be used to determine/predict the sentiment from financial documents/articles in a South African context which can then be correlated with company financial performance.

1.3 Contributions

The ultimate goal of this work and the derived future work are as follows:

- A sentiment analysis tool for financial news articles specifically in a South African context.
  - Develop/expand a financial dictionary suitable for financial articles
  - Develop a full pipeline from data collection, to filtering and sentiment prediction
- Setting the foundation for, in future, expanding the work to include a broader sentiment prediction model that takes into account various topics and their contribution to overall sentiment as an indication of company reputation.
- Progress towards an open-source library for financial sentiment analysis developed on South African data.

1.4 Thesis Outline

The remainder of the thesis is organised as follows:

- **Chapter 2** presents a literature review and related work on the various concepts explored to address the research questions and objectives.
- **Chapter 3** covers the research design and detailed methodology for the model development pipeline.
- **Chapter 4** gives a detailed explanation of the data sources, data collection and pre-processing/cleanup requirements.
- **Chapter 5** is related to topic modelling and using it as a method to filter data for specific topics. This Chapter also justifies the thesis and the need for addressing the research objectives.

- **Chapter 6** covers the process of annotating the data set and evaluating various methods in order to propose the most suitable approach for answering the research questions.
- **Chapter 7** briefly illustrates the correlation of sentiment with financial performance.
- **Chapter 8** addresses sentiment prediction at a document level compared with only using headlines (Chapter 6).
- **Chapter 9** extends sentiment prediction to data relating to a different organisation than what was used for model development. It comments on how well it generalises and presents recommendations for improvement.
- **Chapter 10** highlights the main conclusions of the thesis and summarises the most important items for future work.

Chapter 2

Literature Review

2.1 Impact of Reputation on Corporate Performance

There is a growing emphasis on corporate reputation and identifying the main contributing drivers of reputation that have an associated financial value to the organisation [11, 36, 40]. Reputation has been described as a valuable, intangible asset to a company. Consequently the management thereof with regards to traditional accounting and financial reporting is challenging [11, 36].

The field of reputation measurement has been increasingly attracting researchers and creating competition for innovative thinking. Factors that impact reputation which are common to four such systems/studies are ***products and services*** and ***financial performance/soundness*** [6, 10, 36, 40]. Although ***products and services*** are a separate driver, it has a direct impact on financial performance which both impact corporate reputation.

The reputation of an organisation can be linked with a financial value and increasing the perception around the main drivers can have a marked improvement on overall reputation [25, 36]. This emphasizes the importance of understanding and tracking reputation.

A popular approach to measure reputation is through sentiment analysis of online media [7, 12, 18, 20, 31]. A brief overview of the these various approaches are given n Section 2.2.2.

2.2 Sentiment Analysis and Opinion Mining

2.2.1 General

The terms sentiment analysis and opinion mining are often used interchangeably. It involves using natural language processing (NLP) techniques to extract and classify subjective information expressed through opinions or through detecting the intended attitude [28, 34]. It has been one of the fastest developing areas in the last decade, growing from simple online product reviews to analysing the sentiment from various online platforms such as social media and extending the application to predicting stock markets, tracking polls during elections and disaster management [28]. The first mention of public opinion analysis dates back to post-World War II and since 2005, with computer-based sentiment analysis, a tremendous increase in the amount of research in this field was noted [28].

Some of the main challenges in sentiment analysis are as follows [14, 34]:

- Language: Models are language dependent and there is a gap in developing models in languages other than English due to lack of researchers as well as the lack of data in many "less familiar" languages.
- Domain specificity: There is a domain dependence with sentiment classifiers that need to be considered.
- Nature of the topic: Linked with domain specificity and impacts model accuracy.
- Negation: Use of terms such as ***won't*** instead of ***will*** **NOT**. A possible solution is using phrasal sentences that are separately labeled with a sentiment [19].
- Availability of training data: Most available data is unlabeled with only a few pre-labeled data sets freely available which poses challenges when training a model and evaluating accuracy.

Research highlighted three categories of sentiment analysis and are discussed in more detail in the Sections below.

Lexicon-based

This method typically uses a dictionary of words/phrases either manually created or automatically generated. Many research papers are available on this topic. In one of these studies, the authors claimed that in a cross-domain scenario, lexicon-based approaches outperform machine-learning models [39]. Over the years, various approaches were evaluated making use of adjectives, verbs and adverbs (and combinations) and using modifiers to intensify certain words [39].

Pre-trained Lexicon-based and/or Rule-based

The **Python** library, ***TextBlob*** is a simple rule-based sentiment analyser that provides the average sentiment (excluding neutral words) of a text string[1].

VADER[2] is another rule-based sentiment analyser specifically trained on social media texts and generalises quite well across contexts/domains compared with other sentiment analysers [15]. In the development of this analyzer, the authors emphasized the importance of a human-centric approach i.e. making use of human raters/annotators as part of the model development and validation process [15].

A comparison between ***TextBlob*** and ***VADER*** showed the latter is superior with regards to predicting sentiment in social media [22]. Some of the main reasons are that it takes into account emoticons, capitalization, slang and exclamation marks.

Predictive models using machine learning

The main techniques generally used involve either 1) traditional models or 2) deep learning models [8]. These are supervised machine learning models and required data sets to be labeled.

The traditional models are typically Naive Bayes, logistic regression and support vector machines. Deep learning models most often used in text classification are deep neural networks (DNN), recurrent neural networks (RNN) and convolutional neural networks

[1] https://textblob.readthedocs.io/en/dev/

[2] Valence Aware Dictionary for sEntiment Reasoning

(CNN). A study in 2020 on using deep learning for sentiment analysis referenced 32 research papers in this field (since 2016), indicating the interest and increased application to improve sentiment classification tasks [8].

2.2.2 Financial Sentiment Analysis

Exploiting Typical Financial Headline Structure

A potential way to determine the sentiment of a financial title was explored by introducing the concept that ± 30% of such titles follow a hinge structure [41]. The investigation suggested that the hinge, which is typically a word such as ***as, amid, after***, splits the sentence into two parts, both parts carrying the same sentiment. If one therefore determines the sentiment of the first part of the sentence, the overall sentiment is inferred. In Figure 2.1, the top sentence aims to explain this notion. However, the second sentence shows an example where the part of the sentence following the hinge does not carry the same sentiment as the first part.

Furthermore, it was argued that the verbs hold the key as sentiment carrying words. It was identified, however, that using existing, labeled word lists may still fall short since these lists were created using very domain-specific pieces of text. For e.g. a word such as rise may be listed as positive based on prior usage, however, its use in a new application may indicate it to be negative.

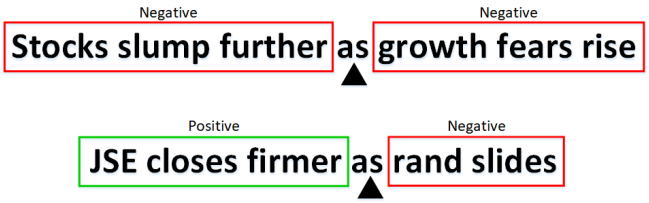

Figure 2.1: Hinge concept of financial headlines.

SemEval-2017: Comparing Approaches for Financial Sentiment Analysis

As part of the 11^{th} workshop on Semantic Evaluation (SemEval-2017)[3], one of the tasks was *"Fine-Grained Sentiment Analysis on Financial Microblogs and News"*[4] of which a sub-task was sentiment analysis on news statements and headlines. It was a regression problem and participants had to predict the sentiment in the range -1 to 1 (representing Negative to Positive). The training data provided was annotated in this same range. Table 2.1 gives a summary of the results and methods for four of the submissions.

Table 2.1: Summary of the performance of the various annotation methods.

Ranking	Score[1]	Modelling Approach
1	0.745	1D convolutional neural network (using word embeddings from GloVe) [27]
4	0.732	Bidirectional Long Short-Term Memory (with early stopping) [30] *Also looked at support vector regression
5	0.711	Ensemble using support vector regression (and gradient boosting regression) [17]
8	0.695	Support vector regression (with word embeddings and lexicon features) [21]

[1] Weighted cosine similarity score

The models used to address the sentiment analysis task range from traditional machine learning to deep learning models with only a ±5% improvement from the latter (Table 2.1). These results indicate that traditional machine learning models can be used quite successfully for this task to set a baseline for further evaluation. The prediction is still far better than random chance of 50%.

For the best performing model, the input to the convolutional neural network was a concatenated vector consisting of word embeddings (from a pre-trained Glove model) as well as lexicon representations from ***DepecheMood*** (a lexicon for emotion analysis).

[3] https://alt.qcri.org/semeval2017/

[4] https://alt.qcri.org/semeval2017/task5/

The authors also added a ***VADER*** score prior to connecting the layers. The results show ∼10% improvement in cosine similarity when using word embeddings (as opposed to excluding it) [27].

From Table 2.1, the 4^{th} ranked submission evaluated traditional and deep learning models. For the support vector regression they introduced a concept of word replacement of domain-specific words to normalise the headings. The authors trained their own Word2Vec model on a corpus of financial texts and used word similarities to replace positive words with the word "excellent" and negative words with the word "poor". The resulting score was ∼10% lower than the neural network. For the LSTM they used their pre-trained Word2Vec model to determine the word representations of the words in the headlines as input to the neural network [30]. An additional comparison would be to use the same inputs to observe the performance difference.

Evaluating various standalone and ensembles of traditional machine learning models also resulted in a good performing model (an ensemble of support vector regression and gradient boost regression) with a score of 0.71, only slightly lower than the 1^{st} ranked [17]. The authors used linguistic features (n-grams, verbs, named entities), sentiment features (counting positive and negative words using a combination of different available sentiment lexicons), domain-specific features (in this case numbers and punctuation that they deemed relevant in financial texts) and word embedding features (concatenated min, max and average of sentence embeddings).

Another similar approach was followed for support vector regression by also using n-grams, sentiment lexicons and word embeddings to determine sentence embeddings [21]. The results show similar performance to the above-mentioned support vector regressor.

From the research above it is seen that various approaches were employed with a number of commonalities. A often used method is pre-trained word embeddings to use as is or to determine sentence representations. Another important feature selection method is using sentiment lexicons to calculate scores for positive and negative words. A few of the authors also commented on the need for domain-specific sentiment lexicons for future work.

As mentioned earlier, training data was annotated with sentiment scores and this was an advantage in the challenge. A large number of supervised learning activities

start with unlabeled data and a substantial amount of time and effort is required to properly annotate data sets.

Existing Popular Financial Sentiment Word Lists

For a lexicon-based approach, a very popular domain-specific (i.e financial) dictionary is the *Loughran-McDonald sentiment word lists* first created in 2009 [26]. The drive for developing these lists stemmed from the authors showing that a more general dictionary, in this case the *H4N* negative wordlist from the *Harvard Psychological Dictionary*, misclassified the sentiment of financial words quite substantially. They found that ∼75% of negative words in the aforementioned list are generally not negative in the financial domain.

2.2.3 Related Work on Financial Sentiment Analysis in the South African Context

Research produced a limited amount of South Africa-related scientific papers on sentiment analysis as a sub-field of natural language processing. Even fewer published results were available on specifically financial sentiment analysis in a South African context.

In 2018, a study on using sentiment analysis to determine alternative indices for tracking consumer confidence (as opposed to making use of surveys) showed high correlation with the traditional consumer confidence indices [32]. These indices are used to better understand current economic conditions as well as to predict future economic activity.

The paper investigated the applicability of such a concept in South Africa as an *emerging market*. The results were promising and motivated the use of large-scale media data sources to monitor this index. They used a combination of general and domain-specific dictionaries and emphasized that a bag-of-words approach is merely "one to one" and does not take into account context. A sentiment score for each document was calculated as the difference between the proportion of positive and negative words (per total count of positive and negative words). This method produces a score on a continuous scale [32]. Lastly, the authors showed that the *Loughran-McDonald sentiment*

word lists performed best and noted that smoother results were obtained for larger data sets. The findings of the study indicated that there were clusters where the proposed sentiment indices corresponded with the traditionally reported consumer confidence index and therefore paves the road for future work to use online media as a means to understand sentiment and its correlation with potential financial performance [32].

Another study, although not necessarily financial sentiment analysis *per se*, was on measuring the online sentiment of the major banks in South Africa [24]. The data source for this analysis was social media only. Machine learning models were used for both detecting topics and analysing the sentiment of user-generated comments relating to those topics. The main contribution the authors made was to highlight the importance of human validation as part of the process to increase accuracy and precision [24].

In summary, even though brand monitoring and reputation management is receiving increased attention, these services are often outsourced to specialist companies at a substantial cost to an organisation. Academic institutions in South Africa are increasing there offerings on programmes, courses and post-graduate degrees in the field of data science and thereby creating opportunities for large-scale upskilling. Organisations can therefore leverage the skills of these individuals to develop tools and systems for in-house reputation management using advanced analytics and sentiment analysis.

Based on the available research, it is deducted that a gap exists for researchers and academics to expand and improve sentiment analysis of online media through natural language processing, especially in the financial domain, in order to increase the knowledge base and pool of technical solutions in the context of South Africa.

2.2.4 Sentiment Correlation with Financial Performance

With the advent of computer-based data analysis including real-time sentiment analysis, more attention has been given to understand whether sentiment analysis of online media can be used to predict share price movement.

The purpose of this Section is to briefly highlight use cases of determining whether sentiment can be used for share price prediction.

A statistical approach to understanding whether stock market prices follow a trend with the sentiment from news articles relating to the stock/company showed promis-

ing results [5]. The method was tested on ∼15 different companies. The study only considered a dictionary-based approach to calculate degrees of positivity, negativity and neutrality. The results showed at 67% correlation between sentiment and share price [5]. The simplistic approach highlighted the existence of correlation and the potential to further explore using sentiment as an indicator of financial performance.

A second paper on predicting market trends using sentiment analysis included a broader context through more diverse data: 1) data sets from the *Financial Times* (for different time periods), 2) news headlines from a worldnews channel hosted by *Reddit*, 3) financial tweets related to the *S&P 500* and 4) share price information for Apple, Google, Hewlett-Packard and JPMorgan Chase & Co.) [31].

The authors evaluated a predictive model using sentiment attitudes (i.e. Positive and Negative), sentiment emotions (such as joy, anger) as well as common technical drivers of share price (for e.g. moving averages, momentum, relative strength index etc.) [31]. Granger-causality was first used to observe whether sentiment attitude and/or sentiment emotions causes stock price changes. The analysis found that only sentiment emotions could potentially be useful indicators.

Predictive machine learning models (a support vector machine and a Long Short-Term Memory recurrent neural network) were developed using the technical drivers as baseline and then including sentiment attitudes and emotions as further input features [31]. The models using only article headlines showed that sentiment negatively impacted performance whereas when using article content, the addition of sentiment enriched model performance in some (but not all) scenarios. The findings highlight the complexity of share price prediction and the fact that it is determined by a number of factors, of which sentiment could potentially add value. The authors highlighted the need to better understand which stocks are impacted by sentiment to determine to applicability of this proposed method [31].

2.3 Topic Modelling

One of the most popular models use to identify topics was introduced in 2003 and uses a Latent Dirichlet Allocation (LDA) method [3]. It is a generative probabilistic model

that is a form of dimensionality reduction. Documents are seen as mixtures of topics each with a topic probability. Each topic in turn is a distribution over words.

Another method that is commonly used in topic modelling is non-negative matrix factorization (NMF). It has been viewed as an alternative to principle component analysis and hence also a dimensionality reduction technique [38]. All the components are non-negative resulting in easier inspection.

A comparison between LDA and NMF on a small data set and a narrow domain (politics, sport and medicine) showed very comparable performance between LDA and NMF with NMF being slightly better [35].

For this thesis, topic modelling will only be used as a high-level method to filter documents according to a dominant topic. Therefore based on the above-mentioned study, non-negative matrix factorisation (NMF) will be used to detect latent topics.

2.4 Logistic Regression for Classification

2.4.1 Justification for use in Text Classification

> "Everything should be made as simple as possible, but not simpler."
>
> Albert Einstein

An informal recommendation (or "rule-of-thumb") is to use a traditional machine learning model to set the baseline for a prediction problem [1]. This allows for better understanding of for e.g. the prediction accuracy for the various classes and the more important features the model learns. Subsequent to this, models with increased complexity requiring more computational resources can be evaluated to determine whether the trade-offs are worthwhile.

One of the big advantages of a logistic regression model over for e.g. a neural network is the ability to interpret the model coefficients and understand the relationships between the inputs and the outputs [38]. Neural networks are more useful when prediction accuracy is of greater importance than interpretation.

A review on published literature surveying the use of neural networks and logistic regression models made mention of the fact that the process of building logistic regression models is easier and allows a reader to reproduce the results more successfully [9]. Also, the architecture of neural networks can be very complex and is generally more prone to overfitting. An important observation was that there is not one specific algorithm that is significantly better than another on a specific data set or a given domain [9].

A high-level comparison of the performance of a Naive Bayes classifier to a logistic regression model on Twitter data indicated that the latter produced superior results. The logistic regression model resulted in both higher accuracy and better precision metrics (with a balanced data set) on a 2-class problem [33].

A separate study comparing additional supervised machine learning models on the *20 newsgroups* data set as well as *IMDb* movie reviews showed that logistic regression and a support vector machine performed the best on these multi-category classification problems [13].

Furthermore, as an example, a study on deep learning for financial sentiment analysis used logistic regression as a baseline and compared various deep learning approaches. The results showed that only a convolutional neural network outperformed the logistic regression model and also justifies the use of a simpler model as a baseline [37].

Various research experiments therefore corroborate the use of simpler models and emphasize the benefits of a logistic regression model due its ease of development and interpretation [13, 33, 37].

2.4.2 Interpretation of a Logistic Regression Model

Logistic regression is generally used to predict a qualitative, binary response variable i.e. Yes/No or Male/Female [16]. From theory, the probability function for multiple logistic regression is given in equation 2.1 with $X = X_1, ..., X_p$,β_1, β_2,. . . ,β_p being the logistic regression model coefficients (slope terms) and β_0 the intercept term [16]. From this function, the logistic regression model is derived and is given by equation 2.2 [16].

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{2.1}$$

$$\log \frac{P(X)}{1 - P(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2.2}$$

The left hand side of Equation 2.2 is referred to as the log-odds or logit. This equation shows that a one-unit change in predictor X_j, changes the log odds by β_j (not a direct, linear relationship with P(X)). However, a positive β_j will increase P(X) for an increasing X_j and vice versa [16]. These can be used to compare the relative importance of one feature ($X = X_1, ..., X_p$) over another.

Multi-class extensions can be implemented but discriminant analysis is preferred generally [16]. The ***scikit-learn***[5] library in **Python**, however, can be used for developing a multi-class logistic regression model.

2.5 XGBoost for Classification

In machine learning, gradient tree boosting is often a popular and effective method used by data scientists to improve model performance [4]. Its wide use in public challenges also indicates that it is frequently the common choice for ensemble techniques.

XGBoost[6] is a scalable implementation of these gradient tree boosting methods that is efficient and flexible and was proven to solve problems with using minimal resources [4]. It provides state-of-the-art results in many machine learning challenges/applications.

2.6 Summary

The purpose of this Chapter was to highlight the research currently available in the various fields of application required to answer the research questions and address the objectives of the thesis. It further justifies the use of the selected techniques/models/approaches for the study and aims to position the relevance of the thesis.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6]https://xgboost.readthedocs.io/en/latest/python/index.html

Chapter 3

Method

3.1 Research Design

In order to test the thesis that natural language processing techniques can be adequately implemented to determine sentiment of financial text a model development pipeline was designed to address the various steps required to answer the research questions and achieve the objectives. This process flow is given in Figure 3.1.

During the first phase of exploratory data analysis, ***Data***, various sources of publicly available textual information was identified and collected. Data is available in different formats, the main ones including portable document format (PDF), comma-separated values (CSV), plain text files and HTML for online news articles. Relevant data from these sources was extracted, cleaned and consolidated.

In the ***Analysis*** phase, topic modelling was used to filter the data for various topics and specifically financial-related documents after which a high-level sentiment analysis can be done. During this phase, existing sentiment analysers were used to observe the sentiment prediction across the different documents and topics. This does not form part of the model pipeline but was required to anchor and justify the thesis objectives.

The next step was ***Data Annotation***. The filtering method using topic modelling was used to extract financial documents. Labeling of the data set was done on document headlines using independent annotators. Further automated methods were then evaluated, compared and the most robust prediction model implemented.

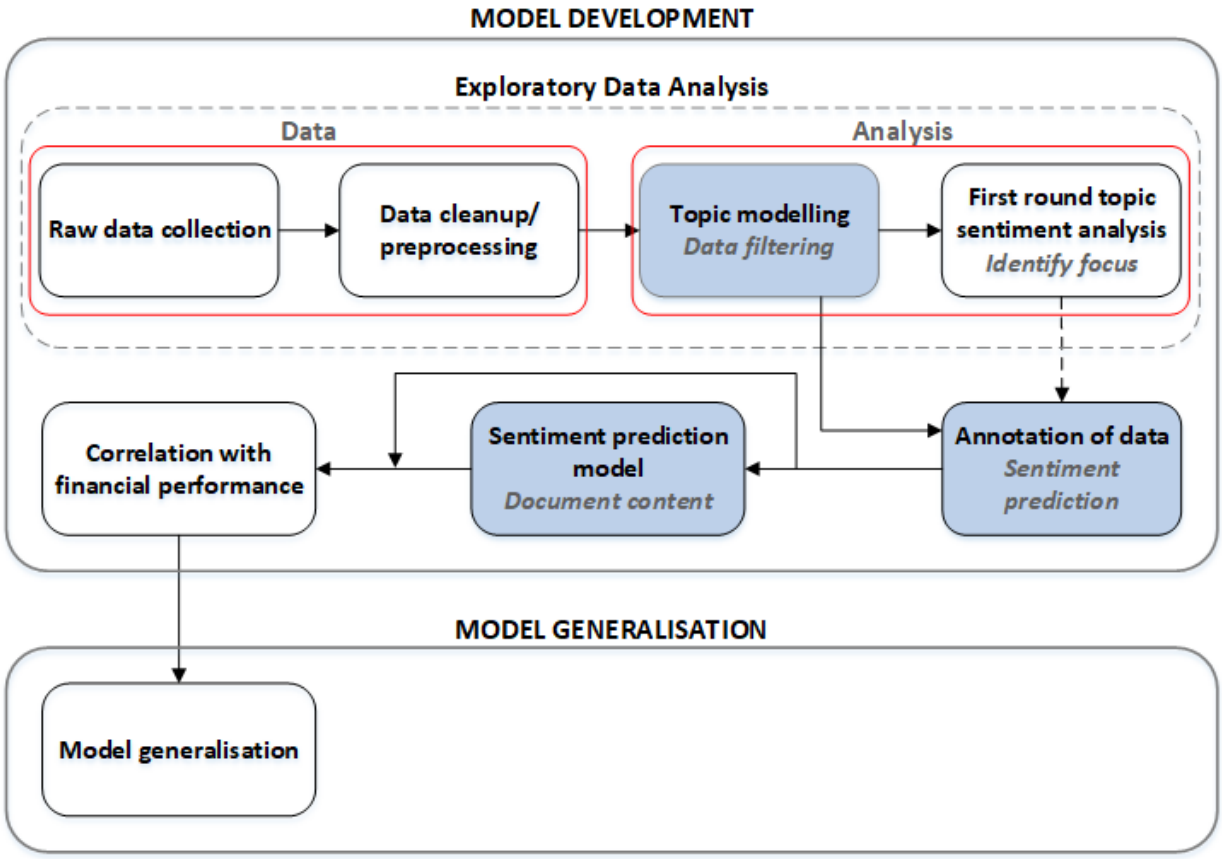

Figure 3.1: Process flow for addressing the research questions and objectives.

The last phase is ***Correlation*** where the predicted sentiments and company financial performance (as indicated by share price), over the same time period, were analysed to observe whether patterns can be recognised.

In parallel with this phase, a sentiment prediction model was developed using the full document/article content. The main purpose is to develop a model that can adequately predict sentiment on financial pieces of text of South African companies.

In order to understand how well the models generalise, it was required to use the developed model pipeline on unseen data from a different organisation. The aim was to determine whether language use in financial articles (mostly online news) follow the same pattern for different organisations. This will inform whether such models can be implemented on a larger scale or whether it is company-specific.

The following sections describe in detail the steps in the model development pipeline. Many permutations of experiments can be envisaged, however, one such construct is discussed.

3.2 Topic Modelling for Data Filtering

Since the main focus of the thesis is to understand financial sentiment analysis, topic modelling was only used as a filtering mechanism to extract financial-related documents.

3.2.1 Data Preparation

Pre-processed data, of which the detailed process is discussed in Chapter 4, was used for this phase of the analysis. The following additional pre-processing steps were performed:

- Tokenized text into words
- Converted words to lower case
- Expanded contractions: replace for e.g. *can't* with *can not*
- Removed English stopwords
- Removed punctuation
- Removed additional words such as ***Sasol, South Africa***
- Lemmatized the words using *NLTK's WordNetLemmatizer*

3.2.2 Topic Modelling with NMF

In order to illustrate the use of unsupervised learning to filter large amounts of data, experiments using a varying number of topics were used to determine an adequate clustering of topics. Term frequency–inverse document frequency (TF-IDF) was used to determine the word vectors as input to the topic model (vocabulary size of 20000 and using unigrams only). Non-Negative Matrix Factorization (NMF) from ***Scikit-learn*** with Kullback-Leibler divergence was used.

In order to assign a unique topic to each document, the NMF output was transformed and normalised for each topic. The maximum value was used to assign a dominant topic.

3.3 High level sentiment analysis

The purpose of this analysis was to evaluate and compare the sentiment per topic from various pre-trained sentiment analysers, with specific focus on financial topics. The following sentiment analysers were used:

- TextBlob (sentiment polarity)
- NLTK's Vader (compound score)
- Stanford CoreNLP (categorical sentiment)

3.3.1 Sentence level to Document level

The premise of this approach was to determine the sentiment for each sentence in a document after which it can be aggregated to a document level sentiment followed by overall topic sentiment. The pre-processed data set (Chapter 4) was tokenized into sentences using ***spaCy***[1] which proved superior to ***NLTK***.

For purposes of this analysis it was decided to use the first 10 sentences of a document for determining sentiment. Where a document had less than 10 sentences, the full document was used. The following two approaches were evaluated for document level sentiment analysis:

- Automated sentiment score from TextBlob and Vader for the 10-sentence paragraph as mentioned above.
- Manually calculated sentiment of the 10-sentence paragraph based on the majority sentiment of individual sentences.

[1] https://spacy.io/api/sentencizer

Since the Stanford CoreNLP sentiment analyzer provides categorical sentiments, these had to be converted as follows: very negative = -1, negative = -0.5, neutral = 0, positive = 0.5 and very positive = 1.

The threshold for a neutral sentiment was assumed to be ±0.05. This is applicable to TextBlob and Vader only since for CoreNLP, neutral = 0.

Algorithm 3.1 outlines the process of calculating the sentiment at document level from sentence level.

The Stanford CoreNLP sentiment analyser interface with **Python** requires sentences as inputs. It is limited to number of characters/words and from experimentation, appears to be inaccurate when multiple sentences separated by punctuation is supplied. It was therefore not used in further evaluations.

3.3.2 Document level to Topic level

The sentiment per topic was calculated in a similar way as was done for document level sentiment. For each unique topic, the majority sentiment across the different documents belonging to the specific topic was used.

Algorithm 3.2 outlines the process of calculating the topic sentiment using the document sentiment calculated with Algorithm 3.1.

3.4 Annotation of data

3.4.1 Ground truth: independent labeling

In order to evaluate whether existing models and/or techniques can be used to determine sentiment, the data first had to be independently labeled.

For this specific pipeline (Figure 3.1) it was decided to use document headlines as a measure of sentiment and since the focus of the thesis is on financial information, the data set was small enough to obtain four independent annotators to label all the relevant documents. The sentiment category options were ***Positive***, ***Negative***, ***Neutral*** and ***None*** (to allow for noisy data that may have been missed during cleanup which for e.g is a single word on nonsensical title). More fine-grained categories can be considered

```
Initialise all variables
Load pre-processed data
for each document d in data do
    Tokenize d into sentences
    for each sentence s do
        Calculate TextBlob, Vader, CoreNLP sentiment and store in alldata
    end for
    if number of sentences > 10 do
        Concatenate sentence 1 - 10
        Calculate TextBlob, Vader sentiment for paragraph and store as
        TextBlob_par, Vader_par in alldata
    else
        Concatenate all sentences
        Calculate TextBlob, Vader sentiment for paragraph and store as
        TextBlob_par, Vader_par in alldata
    end if

Load dataset of all sentences alldata
for each unique document title t in alldata do
    for all sentences in t OR first 10 sentences in t do
        for each sentiment in [TextBlob, Vader, CoreNLP]
            vpos ← count if sentiment >= 0.5
            pos ← count if 0.05 < sentiment < 0.5
            neg ← count if -0.5 < sentiment < -0.05
            vneg ← count if sentiment <= -0.5
            neu ← length(sentences)-vpos-pos-neg-vneg
            labels ← array(vneg,neg,neu,pos,vpos)
            position ← argmax(labels)
            sentiment ← -1 if position = 0 else -0.5 if position = 1 else 0.5
            if position = 3 else 1 if position = 4 else 0
        end for
    end for
end for
```

Algorithm 3.1: Algorithm to calculate document level sentiment.

Load dataset of document sentiments *docsents*
for each *topic t in docsents* **do**
  for all *documents d in t* **do**
    for each *sentiment in [TextBlob, TextBlob_par, Vader, Vader_par, CoreNLP]*
      vpos ← *count if* sentiment >= 0.5
      pos ← *count if* 0.05 < sentiment < 0.5
      neg ← *count if* -0.5 < sentiment < -0.05
      vneg ← *count if* sentiment <= -0.5
      neu ← length(documents)-vpos-pos-neg-vneg
      labels ← *array(vneg,neg,neu,pos,vpos)*
      position ← *argmax(labels)*
      sentiment ← -1 if position = 0 else -0.5 if position = 1 else 0.5
      if position = 3 else 1 if position = 4 else 0
    end for
  end for
end for

Algorithm 3.2: Algorithm to calculate topic level sentiment.

for future (as was looked at in topic modelling) or possibly a continuous scale. The majority label was determined and used as the ground truth sentiment for the relevant documents.

A pre-processing step was required to neaten the appearance of headlines.

On a larger scale, however, only a subset of data can realistically be labeled manually and hence a semi-supervised approach will have to be considered where the labels of the unlabeled data can be inferred from analysing a subset of the data.

The inter-annotator agreement was calculated using the ***AnnotationTask class***[2] from NLTK in **Python**. Fleiss' Kappa was used as the statistical measure of inter-rater reliability of which the interpretation of the statistic is given in Table 3.1. [23].

The following sections describe the various methods of automated sentiment analysis

[2]https://www.nltk.org/_modules/nltk/metrics/agreement.html

Table 3.1: Interpretation of the Kappa statistic.

Kappa	Agreement
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

evaluated which include the following:

- A simplified dictionary approach based on sentiment-carrying words
- Rule based pre-trained models: TextBlob and NLTK's Vader
- Logistic regression model using ground truth labels

3.4.2 A simple lexicon-based approach

For the first experiment using an existing dictionary, the *Loughran and McDonald Sentiment Word Lists* were used as basis [26]. These lists were developed to overcome the fact that more general dictionaries often misclassify financial texts, especially words perceived as negative in a day-to-day context. The sentiment categories are negative, positive, uncertainty, weak modal, strong modal, litigious and constraining.

The above-mentioned word lists were used as is and adapted and the following 3 iterations were evaluated:

1. Experiment 1
   Base dictionary as updated in 2018[3].
2. Experiment 2
   Base dictionary (Loughran and McDonald Sentiment Word Lists) with added syn-

[3]https://sraf.nd.edu/textual-analysis/resources/

onyms (using NLTK's Wordnet Interface[4]). These synonyms are given the same sentiment.

3. Experiment 3

   - Base dictionary (as for experiment 2) but without the addition of synonyms for words in the "modal" lists.
   - Manual addition and deletion of words based on the evaluation of a sample of sentiment predictions from Experiment 2.

The concept of a hinge structure (Section 2.2.2) to annotate article headings was used and the following assumed:

- Headings were split using a ***"comma"*** as a hinge (where no hinge words were present in headings). The two parts of the sentence were split again if hinge words are present in the title (see below). Where either of the splits only consisted of one word, it was not regarded as a hinge point.
- The hinge words included are ***as, but, amid, after, ahead, while*** and ***despite***.
- Individual words in article headings were lemmatized using multiple lemmas i.e. adjectives, verbs and nouns to ensure maximum chance of matching words in the developed dictionary.

In order to assign a sentiment to the headline, it was decided to only use the first part of the sentence (where a hinge or a ***"comma"*** as a hinge was present) alternatively the full sentence was used. This method assumes the sentiment is given by the first part, which could be slightly contradictory to the initial hinge structure proposal (Section 2.2.2).

In the case where multiple sentiment-carrying words are present, the first occurring ***Positive*** or ***Negative*** word was used as the sentiment of the headline (other sentiments were excluded in this round of the evaluation). This approach does not take into account

[4]https://www.nltk.org/howto/wordnet.html

context, however, this simple bag-of-words implementation to detect word sentiments was used for the baseline model development.

The algorithm for determining the sentiment of the article headline (or different parts thereof where relevant) is given by Algorithm 3.3.

3.4.3 Existing rule-based Approaches: TextBlob and Vader

Both TextBlob and Vader were used to calculate the sentiment of headline segments (based on the methodology discussed above in Section 3.4.2) as well as for the full headline. These steps are shown in Algorithm 3.3.

3.4.4 Feature-based Approach: Logistic Regression and XGBoost

A multi-class logistic regression model was developed using a TF-IDF vectoriser as input to the model. Even though more advanced machine learning models have been used for sentiment classification (Section 2.2.1), it was decided to only evaluate a more traditional machine learning model. The main reason being that a significant portion of the thesis is to develop an annotation method in order to set a baseline after which improvements can be investigated. Furthermore, to illustrate incremental improvement as a result of various experiments, TF-IDF was used as a starting point (as opposed to word embeddings, either pre-trained or custom) . Future work would include evaluating word representations as input to the prediction model.

The pre-processing steps for data preparation are the same as for topic modelling (Section 3.2.1) and is applied to the document headlines for this model. In addition, all numeric and non alpha-numeric values very removed from the headline text.

Python's implementation of ***XGBoost*** was also used to develop a binary classifier and the same data pre-processing steps as discussed above were followed. The results were then compared with the logistic regression model.

```
Load developed sentiment dictionary
Load dataset with headlines of financial topics
for each headline h do
    Calculate TextBlob, Vader sentiment for headline and store
    split ← headline.split(',')
    if length(split[0]) = 1 do
        newheadline ← concatenate(split) (excludes comma)
        split ← newheadline.split(',')
    end if
    for each split do
        if any word in split = hinge word do
            split2 ← split.split('hinge word')
            for each split2 do
                Extract bi-grams and match to sentiment dictionary
                Calculate TextBlob, Vader sentiment for split and store
                if match store split2, bi-gram and bi-gram sentiment
                for each word in split2 do
                    Match all lemmatize(word) to sentiment dictionary
                    if match store split2, word and word sentiment
        else if any word in split in sentiment dictionary do
            Extract bi-grams from split and match to sentiment dictionary
            Calculate TextBlob, Vader sentiment for split and store
            if match store split, bi-gram and bi-gram sentiment
            for each word in split do
                Match all lemmatize(word) to sentiment dictionary
                if match store split, word and word sentiment
        else
            Extract bi-grams from split and match to sentiment dictionary
            Calculate TextBlob, Vader sentiment for split and store
            Store split with sentiment = 'Not detected'
```

Algorithm 3.3: Algorithm for calculating the sentiment of headlines (and/or parts thereof).

3.4.5 Programmatic Labeling of Data using Snorkel

Prior to a final decision on the most optimum method/model for annotating the data set, an additional approach using ***Snorkel***[5] was evaluated. This is a system for building training data sets (for supervised learning) without the need for labeling data manually. It uses heuristic rules and/or distant supervision methods through various labeling functions. The following 2 features were used in this evaluation:

- ***MajorityLabelVoter***: Determines a single, probabilistic label for each data point by using a simple majority vote. The disadvantage of this function is that when labeling functions are correlated, certain values may be over-represented in the final result.
- ***LabelModel***: Determines a single label for each data point through probabilities and by learning weights for the labeling functions.

The predicted labels from this approach were compared to the ground truth labels to determine whether the method is useful for annotating the data set.

3.4.6 Final Annotation Model

In order to determine the optimum method for annotating the data, the 4 above-mentioned approaches were evaluated using the following performance metrics:

- Accuracy: the percentage of the total number of correct predictions.
- Recall: the percentage of correct predictions per category/class i.e. positive, negative and neutral.
- F1-score: 2 * (precision * recall) / (precision + recall)[6].

Based on the above model performance comparison and the results from the ***Snorkel*** model (Section 3.4.5), a preferred annotation model was recommended.

[5]https://www.snorkel.org

[6]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

This is an important step in the model development pipeline which is required for model generalisation. When new data is added for evaluation, sentiments will be predicted using headlines based on the most robust prediction model.

3.5 Sentiment Correlation with Financial Performance

For this step in the model development pipeline it was decided to use company share price as the indicator of financial performance. Daily share prices were required for this evaluation (Section 4.3.7).

The following steps were followed to evaluate whether there is a correlation between sentiment (from financial articles) and company performance:

1. The final annotation model (Section 3.4.6), which predicts the sentiment of a document based on title, was used to predict the sentiment of all 808 documents.
2. Unique dates were identified and the majority sentiment for a given date was used.
3. The time-based sentiments and the share price were plotted on one graph to determine whether patterns can be observed.

Since share price prediction is a complex task and impacted by various factors, it was decided to only illustrate whether a directional correlation can be observed. For future work, a statistical correlation can be investigated and potentially include additional drivers known to impact a given stock price.

3.6 Sentiment Prediction at Document Level

There may be a use case for predicting sentiment based on longer text (as opposed to only document titles). To explore the potential of this application, the following process was implemented and evaluated:

1. Used the same data set (consisting of financial documents) as was used for headline sentiment prediction.

2. Determined the distribution of the number of sentences across the document in the data set.

3. Used the statistical mode for the number of sentences as the input paragraph (or used the full article if made up of less than 10 sentences).

4. Used the same pre-processing on words as for the headline prediction model.

5. Developed a new vocabulary from a TF-IDF vectoriser and used as input to a multi-class logistic regression model.

3.7 Case Study

As was discussed in the research design (Section 3.1), the two major components thereof are model development and model generalisation. In order to answer the research questions and address the objectives, data from a big corporate organisation is required. The company identified for developing the financial sentiment model is ***Sasol***[7].

3.8 Summary

The methods and approaches discussed in this chapter are to illustrate the steps required in the model development pipeline. Various other experiments can be proposed at each of the steps, however, for purposes of this thesis, one set of experiments were discussed in detail.

The model pipeline is required to evaluate unseen data and allow for sentiment predictions from data collection to indicating the impact on financial performance.

This detailed approach set the groundwork for understanding financial sentiment analysis for South African documents/articles and achieved the following:

- Highlighted the benefits and/or limitations of existing methods and models for sentiment analysis.

[7] www.sasol.com

- Emphasized the advantages and disadvantages of custom prediction models trained on company-specific data.
- Defined a baseline for sentiment prediction accuracy (and other performance metrics) which can be compared to and improved on by considering more advanced models such as recurrent neural networks or transformers that take into account the sequence of words for improved language understanding.
- Provided insight into the financial dictionary and the impact of size and variety of documents on sentiment prediction.

The results and discussion of the various steps in the model development pipeline are given in Chapters 5 - 8 and Chapter 9 addresses model generalisation. The data collection and handling is discussed in Chapter 4.

Chapter 4

Data

4.1 Introduction

As discussed in Section 3.7, the financial sentiment model was developed using the big corporate organisation, ***Sasol***. This chapter is dedicated to describing the Sasol related textual information.

The first step in exploratory data analysis is data collection and pre-processing (Section 3.1) and comprises a significant portion of time and effort in answering the research questions. The purpose of this section is to provide a detailed explanation of the following data handling steps:

- Identifying the data sources
- Method and criteria of data collection
- Metadata to be included
- Data cleanup

4.2 Data Description

Table 4.1 is a summary of the various data sources collected.

Table 4.1: Data collected relating to Sasol.

Type of data	Format	Period
Financial statements	PDF	Dec 2013 - Dec 2019
Interim: December		
Final: June		
Investor reports	PDF	Nov 2011 - Jan 2020
SENS reports[1]	TXT	May 2015 - Apr 2020
Media releases	TXT	May 2015 - Apr 2020
Social media	CSV	2013 - 2019
Online news articles	CSV	
Links to websites		Jan 2019 - May 2020
Headlines (no content)		Apr 2015 - Dec 2018
Share price	CSV	Jul 2004 - May 2020

[1]Stock Exchange News Service

4.3 Data Collection and Structuring

The following information was extracted from the raw data and included in the structure of the data set:

- Date
- Document title
- Document content
- Label (official or market)
- Publisher (Sasol for official, SENS, online news platform e.g. News24)
- Author (Data description: investor report, Media release, Online news, SENS)
- Link (relevant for online news articles only)

4.3.1 Financial Statements

Company financial statements are publicly available[1] and released bi-annually: interim results (December) and final audited results (June). Publish dates were included in the filenames so that it could be included in the overall data set. These documents are classified as official communication. At the time of the study, it was decided not to include these in the analysis since the reports are factual of nature, however, are available should future work require.

4.3.2 Investor Reports

Investor reports are publicly available documents issued by investment houses where they provide a view on company performance. These are sent to the Sasol Investor Relations Group and is therefore available for use within Sasol (and for purposes of this thesis). Publish dates were included in the filenames so that it could be included in the overall data set.

The ***PyMuPDF*** and ***BeautifulSoup*** libraries in **Python** were used to extract the relevant document content from the PDF files.

4.3.3 Stock Exchange News Reports

These reports are a service provided by the Johannesburg Stock Exchange (JSE)[2] and are company announcements that can have an affect on market movement. These reports are publicly available[3]. Due to the smaller volume of these (168 announcements for the given period), these were manually captured as text files to make data extraction easier. Filenames are the dates in **yyyymmdd** format. Figure 4.1 is an example of a SENS report. The format is standard and was used to identify the body of the announcement.

[1] https://www.sasol.com/investor-centre/overview

[2] https://www.jse.co.za/services/market-data/market-announcements

[3] https://www.sharedata.co.za

```
SOL: SASOL LIMITED - Sasol´s Lake Charles Chemicals Project Ethoxylates (ETO) Unit Achieves Beneficial Operation

SOLBE1: SASOL LIMITED - Sasol´s Lake Charles Chemicals Project Ethoxylates (ETO) Unit Achieves Beneficial Operation
Sasol´s Lake Charles Chemicals Project Ethoxylates (ETO) Unit Achieves Beneficial Operation
Sasol Limited
Sasol Ordinary Share codes: JSE: SOL NYSE: SSL
Sasol Ordinary Share ISIN codes: ZAE000006896 US8038663006
Sasol BEE Ordinary Share code: JSE: SOLBE1
Sasol BEE Ordinary Share ISIN code: ZAE000151817
("Sasol" or "the Company")
SASOL´S LAKE CHARLES CHEMICALS PROJECT ETHOXYLATES (ETO) UNIT
ACHIEVES BENEFICIAL OPERATION
The ETO unit has reached beneficial operation. It is the fourth of the seven Lake
Charles Chemicals Project facilities to come online. The ETO unit has a nameplate
capacity of 100 000 tons per annum and completes the ethylene oxide value chain
which forms part of the Performance Chemicals product range. The achievement of
the ETO unit beneficial operation is well within the previously guided timeline of the
third quarter of the financial year.
30 January 2020
```

Figure 4.1: An example of a SENS announcement to indicate standard format.

4.3.4 Media Releases

A media release an official communication from the company regarding important and/or noteworthy information. By issuing a media release, a company aims to control the narrative of online news by positioning the story they wish publishers to follow. The Sasol media releases are publicly available information[4] and classified as official communication. These documents were also manually captured as text files to ease the extraction of useful data. The format is simple and consistent for e.g. the headline is always in the first line. This standard format ensures that the the content can easily be identified and is shown in Figure 4.2.

4.3.5 Social Media

Sasol uses ***BrandsEye***[5] to collect social media information relating to the company. These are predominantly from *Twitter*, with some mentions on *Facebook*, *YouTube* and *LinkedIn*. Permission was given by Sasol to use these as well as ethical clearance obtained from the EBIT Ethics Committee at the University of Pretoria.

Since the focus of the thesis is on sentiment regarding financial communication and

[4]https://www.sasol.com/media-centre/media-releases/latest-media-releases

[5]https://www.brandseye.com/

Sasol invites bidders for supply of renewable energy to its South African Operations

Date:
14 May 2020
Johannesburg, South Africa – Sasol is inviting bidders to participate in a Request for Information (RFI) process for the supply of renewable energy to its South African operations. An international integrated chemicals and energy company, Sasol's core business is leveraging technologies and the expertise of its people to build and operate world-scale facilities to produce a range of high-value product streams, including liquid fuels, chemicals and low-carbon electricity. The company's largest operations are in South Africa in Secunda, Mpumalanga and Sasolburg in the Free State. In October last year, Sasol released its inaugural Climate Change Report where it committed to reduce its absolute greenhouse gas (GHG) emissions from the South African operations by at least 10% by 2030, off a 2017 baseline. The company has identified renewable energy as a key lever for reducing its GHG emissions and moving it towards producing products in a more sustainable manner. The purpose of the RFI process is to identify partners for the potential deployment of renewable energy projects. It is envisaged that the successful bidder(s) will supply energy as Independent Power Producer(s), in terms of Power Purchase Agreement(s) agreed between the parties.
"We intend procuring, in total, approximately 600 MW of renewable electricity capacity with the aim of reducing our greenhouse gas emissions by approximately 1.6 million tons per annum. This will favourably position Sasol to deliver on our commitment of reducing our South African GHG emissions by at least 10% by 2030," said Hermann Wenhold, Sasol's Chief Sustainability Officer. In aligning with the Integrated Resource Plan (IRP 2019), Wind and Solar Photovoltaic (PV) technologies are favoured at this stage. The projects must show a generation capacity of at least 20 MW to be implemented either as wheeled options from suitable locations across South Africa, or as embedded options close to Sasol's facilities in Sasolburg or Secunda. Interested bidders may apply for access to the RFI by forwarding their company profile together with contact details to: renewable.energy@sasol.com.
The closing date for submissions is Friday, 05 June 2020.

Figure 4.2: Media release example to highlight the simplistic format.

its contribution to overall reputation, it was decided to exclude social media from the analysis. These are typically short text comments and handled differently to longer articles. Also, prediction accuracy decreases with declining text length [29] and when using document content, will impact on overall accuracy. However, the impact of including social media should be considered when expanding the scope of determining a reputation score.

4.3.6 Online News Articles

As is shown in Table 4.1, for articles prior to 2019, only headlines for online news articles were provided (63336 potential articles). Links to websites were provided for 2019 onwards (9833 potential articles) that could be used to access article content.

2015-2018

Information regarding the online news articles were provided as multiple CSV files, one for each month. The most important data extracted from these files are as follows:

- Publish date
- Language (for filtering)

- Headline
- Source name (online news publisher i.e. Engineering News)

Since no links were available to access the articles, an alternative approach was used. The ***googlesearch-python***[6] library in **Python** was used to automate searching of articles to record a URL. The search term used included the Source name and the Headline. However, upon implementation it was found that the process is rate limited to ± 40 searches per 15 minutes. This proved to be very inefficient.

In order to speed up the process, the volume of data was reduced by filtering according to the day of and one day after the Media Release dates (Section 4.3.4). The list of potential articles was reduced to 13771 articles.

The data set was split into data sets consisting of a 1000 articles to be searched so that multiple instances could be run in parallel. Where articles couldn't be found, the URL was left empty.

Once the URLs were obtained, the data was filtered for English articles only and the ***urllib.request***[7] library in **Python** was used to access the URL. The ***BeautifulSoup***[8] library was used to scrape the information from the websites and extract the article content only.

The data set was further reduced by excluding articles where the term 'sasol' does not appear in either the headline or the article content. The data was also checked for duplicate URLs. The final number of articles were 2088.

2019-2020

One CSV file was provided containing the article information relating to Sasol mentioned in online news.

The data was filtered for English articles only after which the ***urllib.request*** and ***BeautifulSoup*** libraries in **Python** were used to extract the article content (the same process as discussed above). Articles of which either the heading or the content did not

[6]https://pypi.org/project/googlesearch-python/

[7]https://pypi.org/project/urllib3/

[8]https://pypi.org/project/beautifulsoup4/

contain the word 'sasol' were removed. Duplicate articles based on repeating URLs were also removed. The final number of articles were 3842.

Additional articles for January 2016 - May 2020

Due to the significant reduction in the amount of articles as a result of filtering, it was decided to supplement the data through obtaining additional news articles. The ***GoogleNews***[9] library in **Python** was used to search Google News articles using the term "sasol".

The additional articles were combined with the above mentioned and checked for duplicates. The final number of online news articles were 7666.

4.3.7 Share Price

Daily, closing share price values were obtained from Sasol, however, this information is also publicly available[10].

4.4 Data Consolidation

The different data sets (excluding share price) were combined i.e. investor reports, SENS reports, media releases and online news articles in preparation for the final data cleanup.

4.5 Final Data Cleanup

The following final data cleanup steps on the document bodies were done as part of data pre-processing:

- Removed ASCII characters
- Ensure there is a space between "th", "st" and "nd" and the next word character. For e.g. **2ndAvenue** needs to be replaced with **2nd Avenue**.

[9] https://pypi.org/project/GoogleNews

[10] https://za.investing.com/equities/sasol-ltd-(j)-historical-data

- Ensure there is a space between "!" and "?" and the next word character. This is to detect sentence boundaries for sentence tokenization For e.g. **ever!The** needs to be replaced with **ever! The**.
- Ensure there is a space between a digit and a "unit of measure. For e.g. **60Hz** will be replaced with **60 Hz**.
- Removed email addresses
- Removed URLs (various possible formats)
- Removed date ranges for e.g. **2019-2020**
- Removed various types of telephone numbers

The above-mentioned list is not necessarily exhaustive and there may be additional cleanup required for future work and improvements.

The final data set is referred to as the pre-processed data set and is the starting point for exploratory data analysis.

4.6 Ethical Considerations

Ethical clearance was required for the use of Sasol social media as well as additional social media data that may be required to address the research objectives. Approval was granted by the Faculty of Egineering and Built Environment (reference: EBIT/38/2020).

The guidelines when using social media data to which will be adhered are as follows:

- Results from the project will be such that any personal identifiable information will not be revealed or inferred.
- Results from the study will not show any affiliation of individuals to organisations of religious and political nature or bias towards specific corporate institutions.
- Where necessary to quote comments from specific users for illustration purposes, it will not be quoted directly but rather paraphrased to avoid identifying specific individuals.

- Images/photographs are not required and will not be included in the dataset.

4.7 Summary

This chapter is a detail explanation of data collection process as well as the challenges experienced. It also gives a detailed overview of the pre-processing and cleanup steps required. In Chapter 5 the various sources of data described are used to highlight emerging topics and perform a high-level sentiment analysis.

Chapter 5

Topic Modelling for High Level Sentiment Analysis

The purpose of this analysis is to utilise an unsupervised machine learning technique to cluster the various articles and documents by topic based on headlines. Since the goal of the thesis is to develop a sentiment prediction tool for financial documents, topic modelling is only used as a high level method to group similar content and extract financial-related articles. It is a rudimentary approach to support the research objectives.

5.1 Emerging Topics

The top 15 words for each of the 25 topics from the NMF topic model are given in Appendix B. From the results, 25 topics may too fine-grained, however, clear topic segregation is noticed for relevant topics. The evaluation to determine the extent of efficient clustering and topic labelling is a manual process based on visual inspection and supported by a good understanding of the various potential topics in the ***Sasol*** domain. The impact of topic choice on sentiment analysis was not evaluated in this analysis. However, this method appears to be a practical solution to filter the data for the purposes of this thesis. Figure 5.1 gives wordclouds for 3 of the topics and show sensible clusters of words.

Furthermore, from Appendix B it appears that topic 10 is also related to financial

(a) Topic 2: Financial (b) Topic 7: Environmental (c) Topic 15: Education

Figure 5.1: Wordclouds to indicate different emerging topics.

documents as shown in Figure 5.2. Most of the investor reports (which are mostly financial related as well) were allocated to Topic 9 (73%) and the remainder to Topic 22. This also indicated a good clustering result using NMF.

Figure 5.2: Top words for Topic 10 indicating financial terms.

5.2 Topic Sentiment Analysis using existing Analysers

Algorithms 3.1 and 3.2 were implemented and the result is given in Figure 5.3. Overall it appears that when using individual sentences and relying on a majority sentiment results in mostly neutral topic sentiments (indicated by *zero* values) and is not very informative (*CoreNLP*, *TextBlob* and *Vader* on x-axis in Figure 5.3).

When looking at topic sentiment using paragraphs as input (*TextBlob_par* and *Vader_par* on x-axis in Figure 5.3)), the majority of sentiments appear positive (values > 0.5).

The following are some of the drawbacks experienced when using CoreNLP in a **Python** environment:

- Length of input is limited to a certain amount of words/characters, therefore sentiment analysis for a paragraph (of for e.g. 10 sentences) can not be performed.
- When multiple sentences, separated by punctuation, is provided as input (within the number of characters limitation), it appears that the sentiment of the first sentence is used as the sentiment of the entire input.

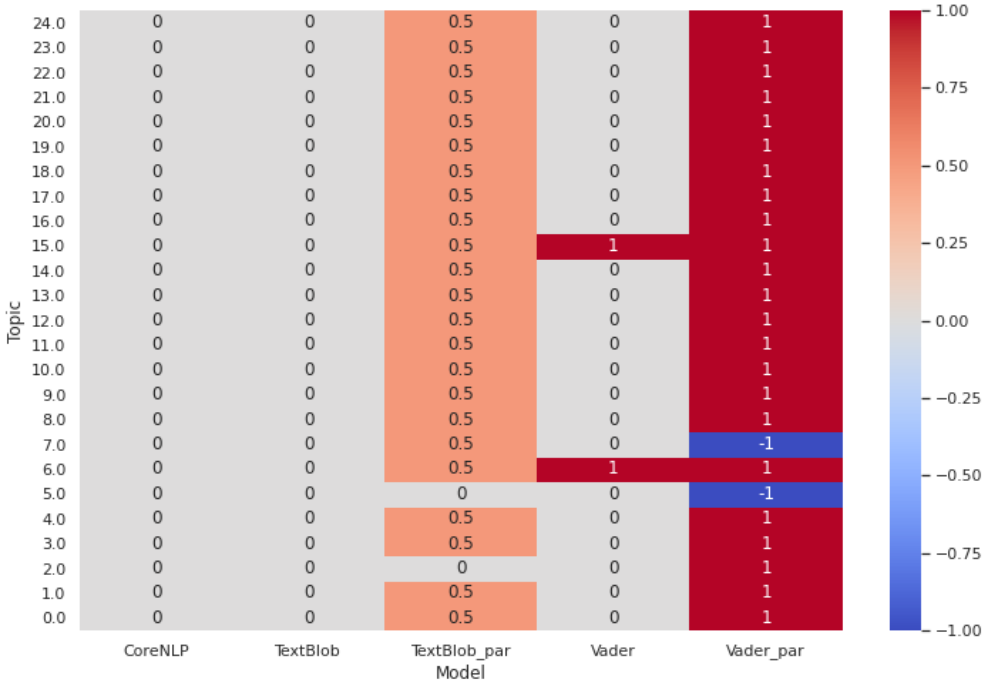

Figure 5.3: Comparison of topic sentiment from various sentiment analysers.

As a result of the above-mentioned disadvantages, CoreNLP was not used for paragraph level.

Furthermore, it can be seen (Figure 5.3) that the topics where there are more significant misalignment between ***TextBlob*** and ***VADER*** (on a paragraph level) are topics 2, 5 and 7 which represent financial, a cluster of Afrikaans words (even though documents were filtered for English only) and environmental respectively. This therefore indicates that off-the-shelf sentiment analysers are inconsistent in predicting sentiment of, in this case, finance-related documents.

The sentiment distribution for topic 2 is shown in Figure 5.4 and it is observed that the two sentiment analysers are predicting quite differently. ***TextBlob*** predicts mostly neutral with the remainder balanced between **Negative** and **Positive**, which is ambiguous. ***VADER*** on the other hand is split between **Very Negative** and **Very Positive**, which is also ambiguous. Both these models are therefore not adequate for this research task.

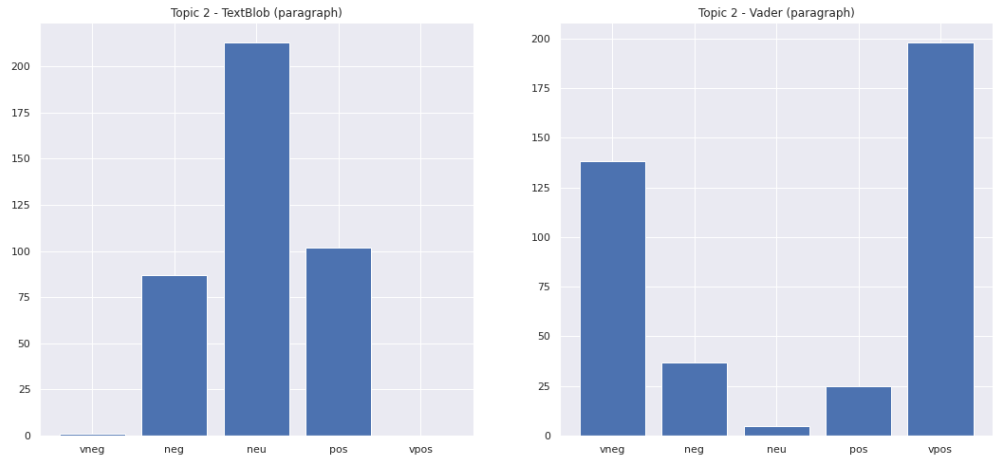

Figure 5.4: Sentiment distribution for topic 2.

As indicated above (Section 5.1), topic 10 is also related to financial. Although Figure 5.3 doesn't show misalignment between ***TextBlob*** and ***VADER*** for topic 10, and a more detailed analysis showed otherwise (Figure 5.5). This justifies the inclusion of both topics as financial information of which more consistent and improved sentiment analysis is required.

The sentiment using a majority sentiment approach based on the first 10 sentences

Figure 5.5: Sentiment distribution for topic 10.

individually is given in Appendix A. The results indicate an overall neutral sentiment for both financial topics (using both ***TextBlob*** and ***VADER***), with ***TextBlob*** appearing to be predominantly neutral and ***VADER*** being more balanced (i.e. more predictions in other categories but on average neutral). Using paragraph-level sentiment predictions provide slightly better information on the extent of misalignment and hence supports the justification for investigating alternative sentiment prediction methods.

As a last step in confirming the observation regarding financial topics, it was decided to look at the first 10 sentences of one document in topic 2 only. This is a case where ***TextBlob*** predicted a **Negative** sentiment and ***VADER*** predicted **Positive**. Figure 5.6 gives the sentence sentiments for the various approaches.

It is clear from Figure 5.6 that ***TextBlob*** and ***VADER*** is fundamentally different. The wording of the sentences are given below and on visual inspection it is noticed that both models have areas where sentiment prediction is opposite to what is expected.

0: The Johannesburg-headquartered company told investors that it decided to hold back on the interim dividend to protect its investment grade.

1: Chief financial officer Paul Victor said the current position of the balance sheet necessitated that the board make the decision in the long-term interest of shareholders.

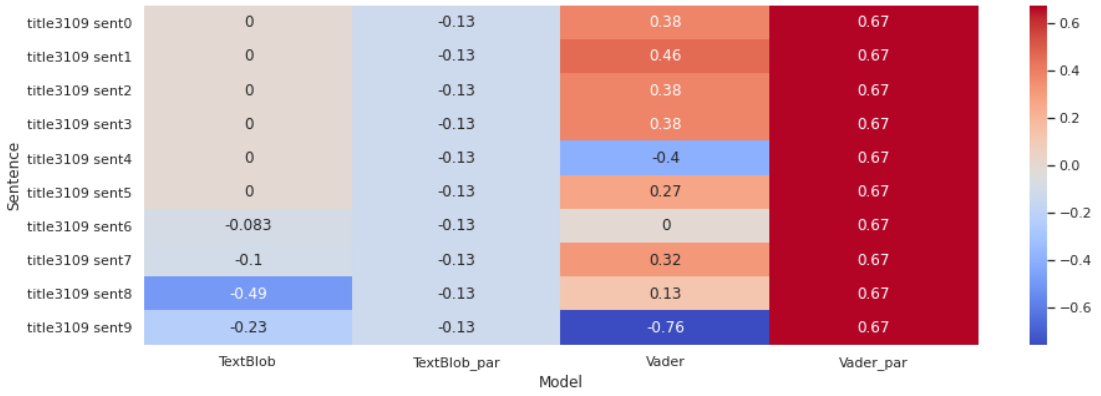

Figure 5.6: Sentence sentiment from various sentiment analysers for a single document.

2: Where we find ourselves in terms of our peak gearing, one of the key priorities for the company is to protect our investment grade, Victor said.

3: Ultimately, in terms of our capital allocation framework, we protect the investment grade and then also allocate resources to sustain operations and the health of the operations.

4: Sasols shares later clawed back their losses to close 3.31 percent lower at R207.

5: Sasols gearing increased to 64.5 percent during the period from 56.3 percent in June 2019.

6: The rise was at the upper end of the previous market guidance of 55 percent to 65 percent.

7: Moodys Investors Service affirmed Sasols Baa3 ratings, with the outlook changing from stable to negative last May. In December 2018, S&P rated the company at a BBB-/A-3 with a stable outlook, which is two notches above the sovereign credit rating.

8: Michael Treherne, a portfolio manager at Vestact Asset Management, said cutting the dividend was a signal that things were tough at the company.

9: Sasol reported that earnings plummeted 74 percent to R4.5 billion during the period, from R23.25bn in the prior period, on the 9 percent weakness in the rand per barrel price of Brent crude oil, softer global chemical prices and refining margins.

Whether using sentence-level or paragraph-level, both ***TextBlob*** and ***VADER*** appear

insufficient for finance-related sentiment prediction. This therefore warrants further investigation to determine a suitable sentiment prediction model for financial information.

5.3 Summary

The results from this high level sentiment analysis confirm that the area of financial topics require improvement and that existing sentiment analysers may not be sufficient. The use of the English language in South Africa as well as financial news reporting is unique and requires an improved method for sentiment prediction.

It was decided to only include Topics 2 and 10 on the assumption from the visual inspection that these account for financial-related content. A more scientific analysis using for e.g. part-of-speech tags, keyword lists or seeded topic modelling could improve the topic assignment/sentiment analysis and should be considered for future work to prove/disprove. Take note, this possible improvement will gain more emphasis where an overall sentiment prediction is required over the range of topics.

Also, the investor reports were excluded since the titles of these are not necessarily sentiment-carrying even though the documents are within the financial domain. These, however, could be considered for sentiment analysis at document level (Chapter 8). The information used for the remainder of the analysis therefore include predominantly online news and some Stock Exchange News Service announcements.

In Chapter 6, the document headlines of the filtered financial data (using this topic modelling approach) are used to annotate the data and develop a sentiment prediction model.

Chapter 6

Annotation of Data

6.1 Manual Annotation

The data set post topic modeling was filtered for the identified financial topics and distributed to the independent annotators. It was noticed that some titles were incomplete, however, was used as is. For future, more attention should be paid to the completeness of the document title.

The data was further filtered to exclude "Official" documents, Investor reports (of which there was only 1) and documents where the ground truth sentiment is "None". The latter category is related to headlines that could not be interpreted and should've been removed prior to distributing to the annotators. These are for e.g. incomplete or one word titles or titles such as "Market Watch" which is non-sensical. The total number of financial articles used was 808 (with only 4% i.e. 33 being SENS reports). It should be noted that there may be duplicate titles originating from different online platforms which was not removed in this analysis and should be considered in future work.

Fleiss' Kappa, which represents the inter-annotator agreement, was calculated as 0.67. According to Table 3.1 (Section 3.4.1) this translates to a ***substantial agreement*** between annotators even though only four raters were used. Future work may require additional annotators and evaluating the impact on model accuracy. Also, cases where an assigned data label was tied 50/50, it was resolved through manually evaluating the annotations to recommend the optimum label (based on domain expertise). Due to the

small data set, this process was possible. However, a more sustainable method for future, larger data sets will need to be investigated and implemented.

Table 6.1 gives the sentiment distribution for the financial data set based on the majority label from the annotators.

Table 6.1: Summary of the sentiment categories of the annotated data.

Sentiment	Count	Percentage
Positive	249	31%
Negative	419	52%
Neutral	141	17%

A future improvement and recommendation is to make used of experience linguists to assist with annotation.

6.2 Lexicon-based Approach

For this approach, the traditional method of having a train and test data set is not used since the method is a simple keyword-based technique used to assign a sentiment to an article headline.

In Experiment 1, the *Loughran and McDonald Sentiment Word Lists* (containing 4140 words) were used as is to determine a sentiment based on key words according to the approach discussed in Section 3.4.2 to observe accuracy (results given in Table 6.2).

The goal of Experiment 2 was to update the word lists and determine whether it improves prediction accuracy. An extract of the sentiment predictions from Experiment 2, based on the base dictionary with added synonyms, is given in Appendix C. Thereafter in Experiment 3, random samples were evaluated to update the dictionary from Experiment 2. It was noticed that some of the synonyms added resulted in incorrect predictions and had to be removed again. Also, the synonyms added in this experiment excluded those for the "modal" word lists. The removed words as well as added words and added bigrams are given in Appendix C. Only 4 words were removed from the original dictionary: break, closed, closing and despite. Due to the small data set, the impact of the updated dictionary was evaluated using the full data set.

The final dictionary contains 9743 words. However, it is recommended that a more robust method be developed to update the dictionary in future since this manual method does not necessarily capture all the required words and may also have redundant words.

The sentiment word lists only consider positive and negative words and if there is no match, the sentiment would be 'Not detected'. For purposes of this analysis, ***Neutral*** ground truth labels were therefore not considered. This results in a somewhat imbalanced data set with 37% of headlines being positive and 63% being negative. For supervised machine learning models, this may lead to slightly more False ***Negative*** predictions.

Furthermore, for evaluating the improvement in the dictionary-based approaches, only ***Positive*** and ***Negative*** headline predictions were compared. After experiment 3, there were 17 headlines not in this group. A recommendation for future work is to refine the sentiment word lists to either eliminate the other categories (e.g. Litigious) or increase the granularity of the sentiment categories.

Table 6.2 gives the results for the 3 experiments and highlights the improvement based on the manual dictionary update. The final headline sentiments were calculated using Algorithm 3.3. After updating the dictionary, the sentiment prediction accuracy improved by 29% compared with the original word lists.

Table 6.2: Summary of the results of the simple dictionary-based approaches.

		Experiment 1		**Experiment 2**		**Experiment 3**	
Sentiment	**Actual Count**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Positive	249	69	28%	123	49%	184	74%
Negative	419	180	43%	323	77%	379	90%
Overall	668	249	37%	446	67%	563	84%

The results from the various experiments highlight the need for not only domain-specific sentiment prediction tools but also region-specific corpora.

The data statement [2] for the updated sentiment word lists is given in Appendix D. The data set is named ***LM-SA-2020*** representing *Loughran and McDonald Sentiment Word Lists* for *South Africa*.

A future improvement is to assess the sentiment for sentences where multiple sentiment-carrying words are present for e.g. the difference between the counts of positive and

negative (Section 2.2.3) and evaluate the impact on sentiment prediction accuracy.

6.3 Rule-based Approaches: TextBlob and Vader

Upon inspection it was noticed that many of the financial document headlines have either of the following formats (publisher name in headline):

> Sasol halts production at inland crude oil refinery as fuel demands drops | Fin24
> Sasol restoring coal stockpiles post strike, but output to be lower - Miningmx

Since a dictionary-based approach only looks for listed sentiment-carrying words, named entities such as the name of the online publisher will not affect the overall result. It is the same for TextBlob and Vader, where named entities will carry a neutral sentiment.

Tables 6.3 and 6.4 show the confusion matrices for using TextBlob on parts of sentences (hinge structure) as well as on the full headline. The overall accuracy ranges from 25% - 28%.

Table 6.3: Summary of the results using TextBlob with the proposed hinge structure.

		Predicted		
		Negative	Neutral	Positive
Actual	Negative	52	326	41
	Neutral	14	101	25
	Positive	14	182	53

Table 6.4: Summary of the results using TextBlob on full headline.

		Predicted		
		Negative	Neutral	Positive
Actual	Negative	68	304	47
	Neutral	15	99	26
	Positive	19	170	60

Tables 6.5 and 6.6 show the confusion matrices for using Vader on parts of sentences (hinge structure) as well as on the full headline. The overall accuracy ranges from 47% - 54%.

Table 6.5: Summary of the results using Vader with the proposed hinge structure.

		Predicted		
		Negative	Neutral	Positive
Actual	Negative	194	151	74
	Neutral	18	102	20
	Positive	26	139	84

Table 6.6: Summary of the results using Vader on full headline.

		Predicted		
		Negative	Neutral	Positive
Actual	Negative	243	90	86
	Neutral	20	96	24
	Positive	53	97	99

From the above results it appears that a simple dictionary based method to annotate the document headlines may prove more accurate than pre-trained sentiment analysers. These analysers were trained on specific corpora and it seems in the context of financial news reporting in South Africa, fall short. The superior performance of ***VADER*** as compared with ***TextBlob*** is consistent with a previous study on their comparison (Section 2.2.1) [15]. Furthermore, since ***VADER*** was trained on social media, the subpar performance on financial headlines is therefore not unexpected.

6.4 Feature-based Approach: Logistic Regression

6.4.1 Model Results

As discussed earlier (Section 6.3), many of the headlines contain the name of the online publisher. Since a TF-IDF vector was used as input to the logistic regression model,

named entities (i.e. publisher names) may impact model performance and needed to be removed. This was not an issue for the lexicon-based approach or using ***TextBlob*** or ***VADER***. In future, these clean-up steps should be performed prior to using any prediction methodology for ease of use in different applications.

The pre-processed data set (from Section 4) was used to create a preliminary list of Publishers. The following steps were implemented to clean the headlines:

- Removed everything to the right of the "|" separator.
- Removed publisher names based on the preliminary list from above.
- Visually inspected the neatened headlines and created an additional list after which these were removed from the headlines as well.

The standard NLTK list of stopwords were adapted by removing words deemed relevant (based on the findings from applying the lexicon-based approach). Hinge words and sentiment-carrying words within a financial context were removed (list given in Appendix E).

TF-IDF vectorisation produced a vocabulary of size 917. This is quite small due to the limited number of documents and only using document titles.

The data was split 80:20 for training and testing. The 80% training data set is used in cross validation and the 10-fold cross validation accuracies are given in Table 6.7. The 20% test data set is the "unseen data set" on which model performance is measured.

Table 6.7: Cross-validation accuracy for headline sentiment using logistic regression.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**	**StDev**	**Test**
0.73	0.8	0.68	0.72	0.69	0.71	0.70	0.77	0.75	0.68	0.72	0.039	0.81

In order to compare this method to the other approaches (and observe overall prediction accuracy compared to ground truth), the logistic regression model (trained on 80% of the data) was used to predict the labels of all 808 document headlines. The accuracy of prediction is 87%. Table 6.8 shows the confusion matrix for this prediction.

Table 6.8: Summary of the results using the logistic regression model on all headlines.

		Predicted		
		Negative	Neutral	Positive
	Negative	406	0	13
Actual	Neutral	52	83	5
	Positive	33	1	215

6.4.2 Model Confidence

Figures 6.1 - 6.3 show the probability distributions for the different categories of sentiment predictions. Negative prediction results seem to have a higher probability of being correct with 83% of predictions having a probability of 50% and higher and 21% a probability of 75% an higher. For positive predictions it is 72% and 0% respectively and for neutral predictions it is 59% and 10%.

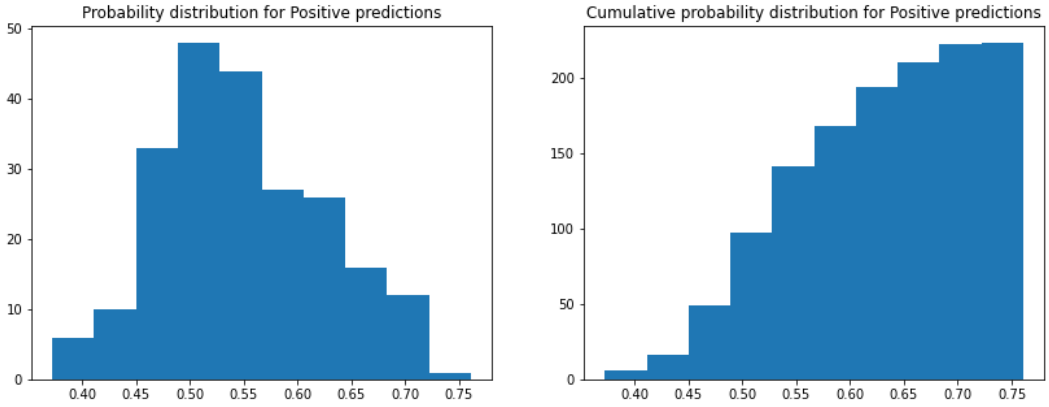

Figure 6.1: Probability distributions for positive sentiment predictions.

A 2-dimensional representation of the TF-IDF vectors further corroborates the observations from the distributions. Figure 6.4 does not show very good clusters, however, the neutral topics are clustered more closely together and towards the right of the illustration, the positive predictions appear to more prominent.

Based on the probabilities of sentiment predictions it is recommended to investigate improving the model to shift the predictions of all categories to include minimum 80%

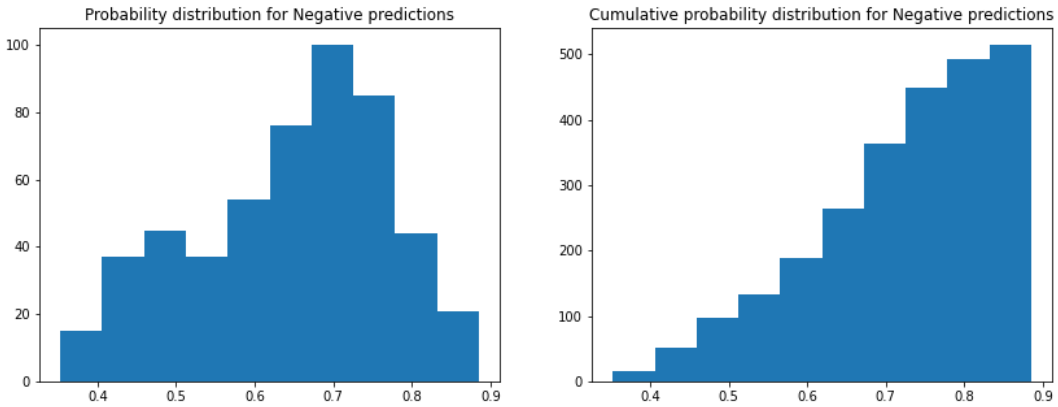

Figure 6.2: Probability distributions for negative sentiment predictions.

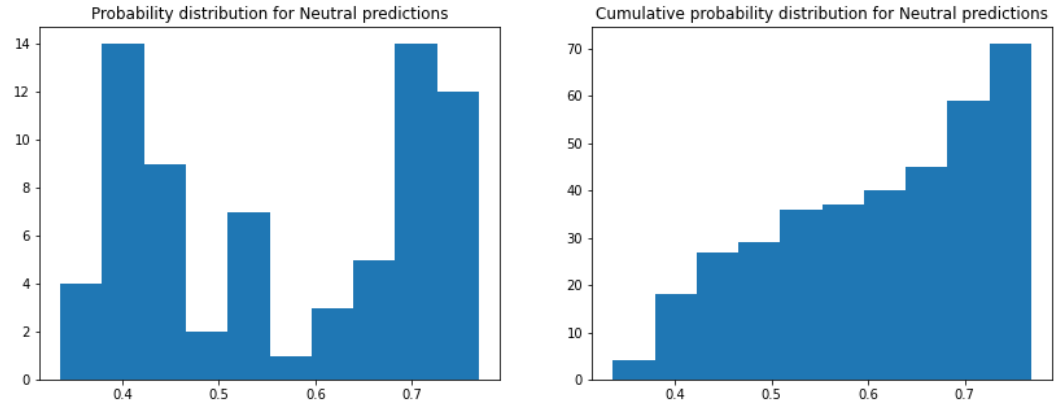

Figure 6.3: Probability distributions for neutral sentiment predictions.

of predictions to have probability of at least 75%. Or alternatively, evaluate additional methods and/or machine learning models (Sections 6.6.2 and 6.6.3).

6.4.3 Model Insights

In order to gain a better understanding of the words that contribute to the sentiment predictions, the following summary of the TF-IDF vectors were evaluated:

- Counted the number of documents in which each word in the vocabulary occurs

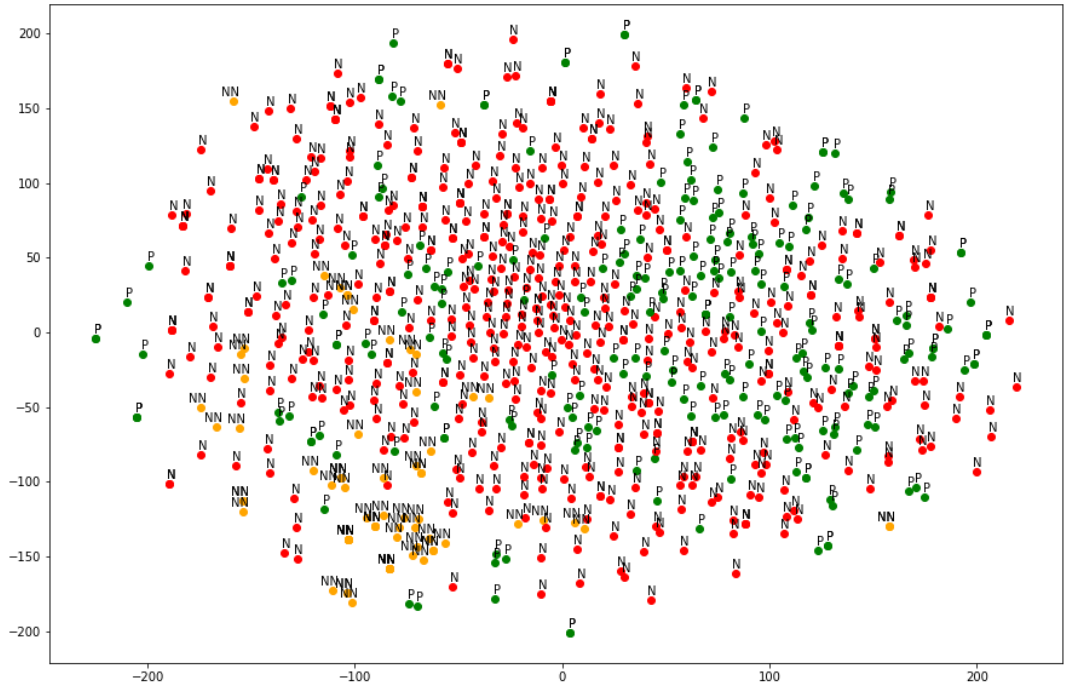

Figure 6.4: 2-dimensional representation of the TF-IDF vectors for the full data set (with sentiment predictions).

(and calculate as a %).

- Calculated the average TF-IDF value for each word (considered only values greater than zero i.e. where the word was present).

Figure 6.5 is a visual representation of the above-mentioned TF-IDF values. As expected, the higher the TF-IDF value of a word, the less the word occurs. However, when looking at the top 20 words with the highest TF-IDF values it is noticed that most of these should not impact sentiment (or are nonsensical) and for future should be looked at and removed during pre-processing of document titles (to allow for the possibility of more frequent occurrence of these words in unseen data that will affect the prediction). Table 6.9 lists these words.

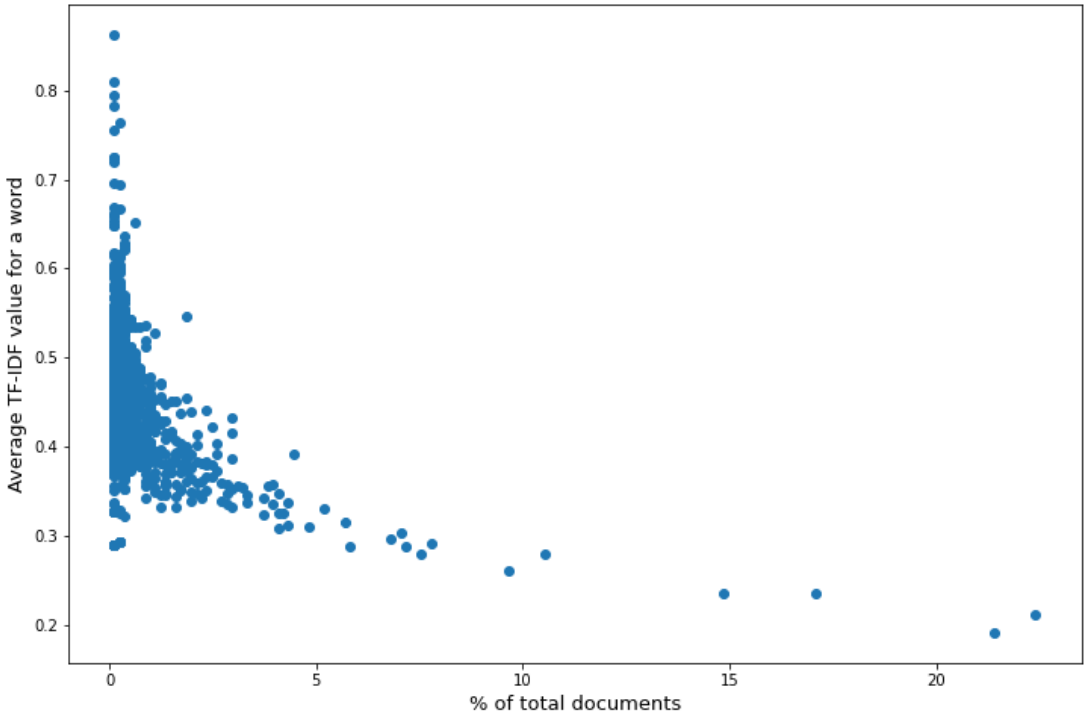

Figure 6.5: TF-IDF values as a function of occurrence.

Although the overall prediction accuracy (on all document headlines) was 87% (Section 6.4.1), it was decided to take a deeper look at the words impacting a specific sentiment prediction. A few random sentiment-carrying words were chosen. Table 6.10 shows the logistic regression model coefficients for the various categories as well as the associated sentiment from the dictionary created in Section 6.2. Recall that the interpretation of these coefficients were discussed in Section 2.4.2.

From just a sample of words it can be seen that there are some words that intuitively should have a higher probability of belonging to a specific category that is the opposite in the logistic regression model. For e.g. rebound, which is ***Positive***, will decrease the log odds and hence the probability of a ***Positive*** prediction and increase the probability of a ***Negative*** prediction and the opposite for crash. The coefficients for the word rally

Table 6.9: Top 20 words with highest TF-IDF values (in increasing order).

1	look	11	small
2	items	12	cushion
3	confluence	13	pique
4	gloomy	14	live
5	thin	15	neck
6	shield	16	ana
7	strength	17	leader
8	finally	18	spar
9	respond	19	eps
10	singularly	20	adrs

Table 6.10: Sentiment comparison between dictionary-based and logistic regression models using a sample of words.

	Logistic regression coefficients			**Dictionary-based**
Word	**Positive**	**Negative**	**Neutral**	**Sentiment**
plummet	-0.29	0.34	-0.08	Negative
higher	-0.18	0.26	-0.08	Not present
retreat	-0.08	-0.28	0.38	Negative
rise	-0.07	0.12	-0.05	Positive
rebound	-0.07	0.12	-0.05	Positive/Constraining
rally	0.05	0.14	-0.17	Positive
resilient	0.34	0.27	-0.82	Positive
up	0.38	-0.32	-0.06	Positive
crash	0.41	-0.28	-0.12	Negative
firmer	0.26	0.11	-0.39	Negative

indicate that there is a higher contribution to a probability of a ***Negative*** prediction than ***Positive***, however, it should be ***Positive***. To illustrate, the following are all sentiment predictions where the logistic regression model did not predict ***Positive*** where the word rally appeared (with the probability given as well):

- Resources rally pulls JSE higher | Fin24 - ***Negative*** - 0.60
- Sasol shares briefly rally on stronger oil price before paring gains | Fin24 - ***Neutral*** - 0.46
- Equities rally as US-Sino trade talks progress | Fin24 - ***Negative*** - 0.50
- Stocks rally on stimulus hopes | Fin24 - ***Negative*** - 0.51
- Stocks rally on the back of lower coronavirus death rate | Fin24 - ***Negative*** - 0.77
- Resources lead JSE rally | Fin24 - ***Positive*** - 0.67

It should be emphasized that the dictionary-based approach is very simplistic and only uses a single word to determine sentiment. The logistic regression model uses the TF-IDF values of all the words in the vocabulary. Sentiment-carrying words may not occur frequently (due to the limited data used) and incorrect words may contribute to predictions that is learned during model training.

6.5 Programmatic Labeling of Data using Snorkel

The 4 approaches evaluated in the preceding sections were used as the labeling functions for the ***Snorkel*** labeling model. Both ***Snorkel's*** *MajorityLabelVoter* and *LabelVoter* were used to assign labels to the data. Table 6.11 gives the recall and F1-score using the new data labels compared with the ground truth labels.

Table 6.11: Summary of the Snorkel model performance.

		Recall			**F1-score**		
	Accuracy	**Pos**	**Neg**	**Neutral**	**Pos**	**Neg**	**Neutral**
MajorityLabelVoter	46%	60%	27%	75%	63%	38%	40%
LabelVoter	46%	84%	11%	84%	69%	20%	43%

From the results in Table 6.11, it can be seen that using this approach yields poor labels when compared with either a dictionary-based or logistic regression model. An informal tutorial commented on the reason for this observation. It was noted that when

strong labeling functions are combined with weaker ones, the degree of disagreement and conflicts between the functions increases which results in overall decreased performance[1]. Both the lexicon-based approach as well as the logistic regression model are strong, good performing models and hence the results from the ***Snorkel*** model. Based on this performance, programmatic labeling will not be used for this study.

6.6 Final Annotation Model

6.6.1 Summary of Annotation Results

Table 6.12 summarises the results of the evaluated annotation methods. The overall accuracy, recall and F1-score for the sentiment categories are given.

Table 6.12: Summary of the performance of the various annotation methods.

		Recall			F1-score		
	Accuracy	**Pos**	**Neg**	**Neutral**	**Pos**	**Neg**	**Neutral**
Lexicon	84%[1]	74%	90%	-	80%	89%	-
TextBlob	28%	24%	16%	71%	31%	26%	28%
Vader	54%	40%	58%	69%	43%	66%	45%
Logistic regression	87%	86%	97%	59%	89%	89%	74%

[1]Only using Positive and Negative, therefore overall accuracy not directly comparable but can be compared with binary prediction models (Section 6.6.2)

From the results it can be seen that the logistic regression model outperformed the other approaches evaluated. However, recall that the probability distributions indicated moderate confidence in predictions. Furthermore the model coefficients from the logistic regression model were sometimes contradictory to expectations. As a result it was decided to investigate the impact of only using two categories i.e. ***Positive*** and ***Negative*** (since these are more informative) to evaluate the impact. Since ***TextBlob*** and ***VADER*** are both pre-trained models, the predictions and accuracies given in Table 6.12 will not be impacted.

[1]https://www.kdnuggets.com/2020/07/labelling-data-using-snorkel.html

6.6.2 A Binary Classification Model using Logistic Regression

The documents with a ground truth sentiment of ***Neutral*** were removed, reducing the data set to 668 documents. Data pre-processing and model training were done exactly the same as for the multi-class prediction model (Section 6.4). The 10-fold cross-validation accuracy was 80% (±4%) with an accuracy of 86% on the unseen test set and an overall accuracy on all document headlines of 90%. The results per category (for the full data set) are given in Table 6.13 (including the lexicon-based model results as reference). There is a slight decrease in recall rate of ***Positive*** predictions but an increase in F1-score (compare with Table 6.12).

Table 6.13: Binary logistic regression model performance metrics on full data set.

	Lexicon		**Logistic regression**	
	Recall	**F1-score**	**Recall**	**F1-score**
Negative	90%	89%	98%	92%
Positive	74%	80%	76%	85%

Recall from Section 6.4.3, sentiment predictions where the word rally featured were evaluated and showed incorrect predictions. Predictions for these sentences using a binary classification model are as follows:

- Resources rally pulls JSE higher | Fin24 - ***Positive*** - 0.74
- Sasol shares briefly rally on stronger oil price before paring gains | Fin24 - ***Positive*** - 0.52
- Equities rally as US-Sino trade talks progress | Fin24 - ***Positive*** - 0.62
- Stocks rally on stimulus hopes | Fin24 - ***Positive*** - 0.72
- Stocks rally on the back of lower coronavirus death rate | Fin24 - ***Negative*** - 0.64
- Resources lead JSE rally | Fin24 - ***Positive*** - 0.70

Only one sentence still had a ***Negative*** prediction and overall there was a slight improvement in the probability of the predictions. It appears a more simplistic approach yielded better results for this application and the size of the data set.

Interpretability

The top 20 words with the highest coefficients from the binary logistic regression model are given in Table 6.14. Most of the important features listed are sensible and can be seen that they are key words for predicting sentiment.

Table 6.14: Highest ranking words for from the logistic regression model (decreasing importance).

1	rise	11	month
2	rally	12	boost
3	rebound	13	recover
4	firm	14	positive
5	up	15	steady
6	higher	16	strengthen
7	firmer	17	gold
8	resilient	18	ease
9	ahead	19	feed
10	despite	20	level

6.6.3 A Binary Classification Model using XGBoost

At this stage in the analysis it was decided to consider an additional traditional machine learning approach i.e. **Python's *XGBoost*** (Extreme Gradient Boosting)[2]. It is a boosting algorithm based on an ensemble of decision trees[3].

In order to compare the performance to that of the above binary logistic regression classifier, the same TF-IDF vectors were used as input (and the same 80:20 train:test data set). Table 6.15 gives the 10-fold cross validation accuracy scores as well as the accuracy on the test set. Table 6.16 gives the recall and F1-score on the full data set. The overall accuracy was 93% using all headlines. The prior method results are also

[2]https://xgboost.readthedocs.io/en/latest/python/index.html

[3]https://www.datacamp.com/community/tutorials/xgboost-in-python

included for reference and it can be seen that the XGBoost classifier outperforms the other models.

Table 6.15: Cross-validation accuracy for headline sentiment using XGBoost.

1	2	3	4	5	6	7	8	9	10	Mean	StDev	Test
0.76	0.74	0.78	0.80	0.85	0.79	0.89	0.89	0.85	0.89	0.83	0.055	0.77

Table 6.16: XGBoost model performance metrics on full data set.

	Lexicon		Logistic regression		XGBoost	
	Recall	**F1-score**	**Recall**	**F1-score**	**Recall**	**F1-score**
Negative	90%	89%	98%	92%	98%	95%
Positive	74%	80%	76%	85%	84%	90%

The following give the sentences evaluated in Section 6.6.2 with the prediction and probability from the XGBoost model. All the predictions are correct when compared with the ground truth labels with a significantly higher probability than the logistic regression model.

- Resources rally pulls JSE higher | Fin24 - ***Positive*** - 0.99
- Sasol shares briefly rally on stronger oil price before paring gains | Fin24 - ***Positive*** - 0.88
- Equities rally as US-Sino trade talks progress | Fin24 - ***Positive*** - 0.92
- Stocks rally on stimulus hopes | Fin24 - ***Positive*** - 0.95
- Stocks rally on the back of lower coronavirus death rate | Fin24 - ***Positive*** - 0.76
- Resources lead JSE rally | Fin24 - ***Positive*** - 0.98

Figure 6.6 shows the top 20 most important features of the XGBoost classifier. The words flagged are interpretable, useful and comparable to the binary logistic regression model's features (from Table 6.14).

The results given in this section leads to the conclusion that an XGBoost classifier performs better than a logistic regression model.

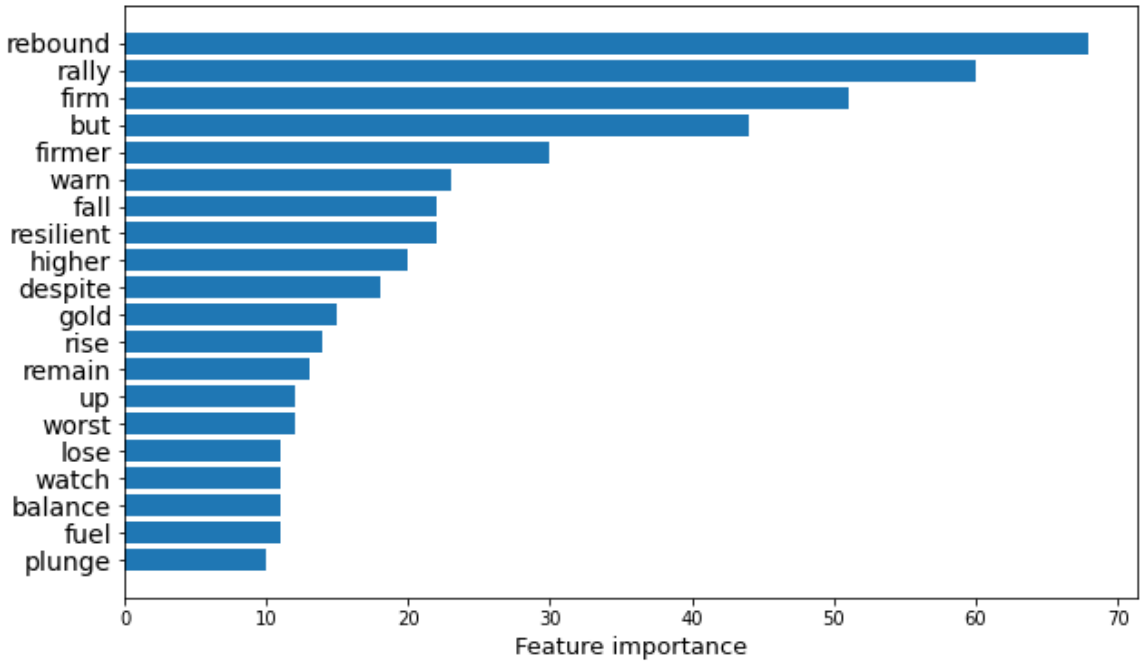

Figure 6.6: TF-IDF values as a function of occurrence.

6.7 Summary

In this Chapter, various annotation methods were explored and compared. A simple dictionary-based approach predicts fairly well for ***Positive*** and ***Negative*** sentiments. Both a multi-class and a binary logistic regression model were evaluated and it was concluded that a binary prediction model suits the specific application better. A further evaluation indicated that model performance can be improved by using **Python's *XGBoost***. The biggest advantage of the binary prediction model over the multi-class prediction model is better overall accuracy when evaluating predictions on headlines as well as improved probability of predictions.

For a potential future improvement, an approach that takes into account the sequence of words in a sentence should be evaluated for e.g. a recurrent neural network (RNN) such as a Long Short Term Memory (LSTM) with attention. These results can then be

compared with the more simpler approaches such as logistic regression and XGBoost in order to determine whether the baseline can be improved.

In Chapter 7, the predicted sentiments based on article headlines are used to determine whether there is a correlation with financial performance (as indicated by share price).

Chapter 7

Sentiment Correlation with Financial Performance

One of the research sub-questions is to evaluate whether there is a correlation between sentiment of financial documents and company performance. As was discussed in Section 3.5, share price was used as the indicator of financial performance.

7.1 Evaluation and Results

In order to observe whether there is a noticeable trend between sentiment and share price, a shorter time period than the extent of the available data (Section 4.2) was used. Figure 7.1 shows this trend for a period of 6 months. The sentiment predictions are from the XGBoost Classifier (Section 6.6.3). Periods A and B are periods where sentiment improved and was reflected by share price. Similarly Period C stands out through a significant amount of negative sentiments and a severe drop in share price.

The graphical representation using the binary logistic regression model is shown in Appendix F.

Degree of correlation

Although on visual inspection it is noted that there are periods where sentiment and share price correlate, to further compare (and justify) the final choice of classifier, a

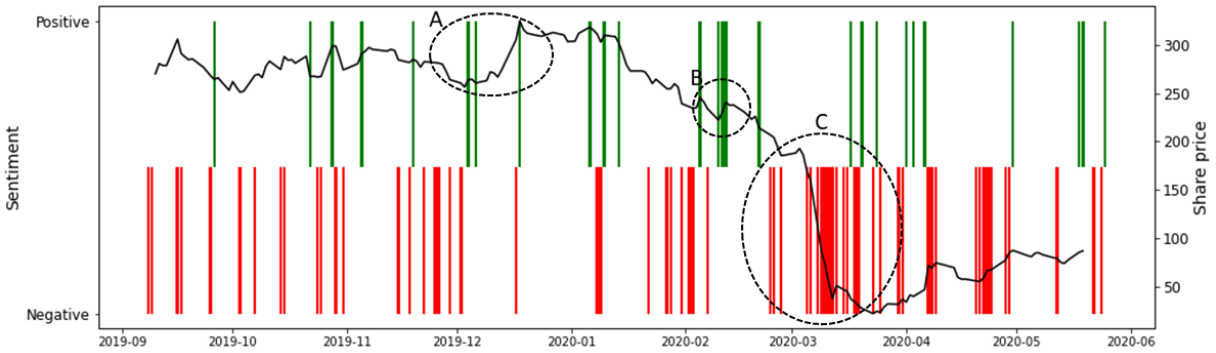

Figure 7.1: Sentiment prediction vs. share price for September 2019 - May 2020 using XGBoost.

correlation coefficient was calculated for both the logistic regression and the XGBoost models.

Table 7.1 is an extract of the data, illustrating how it was converted and used to calculate a correlation coefficient. As described in Section 3.5, the majority sentiment was determined for a specific date. In order to calculate the *Price change* for a date, the share price on the date and the preceding share price (not necessarily one day apart due to stock market closures over weekends) were used. For an increase, a value of 1 was used and for a decrease a value of -1. The correlation coefficient for the XGBoost classifier is 0.09, which is an increase compared to the logistic regression model correlation of 0.05.

Table 7.1: Data sample for calculating correlation between share price change and sentiment.

Date	Share price	Price change	Sentiment
2020-04-09	7449		
2020-04-20	5561	-1	-1
2020-04-21	5502	-1	-1
2020-04-22	5884	1	-1

Even though the statistical correlation appears low, the above results still show promise that there are indeed periods where sentiment (from financial articles/documents) and share price correlate well. There is a question regarding the direction of the correlation or the causal relationship i.e. whether sentiment impacts financial performance or

vice versa.

The use of the correlation coefficient is therefore primarily to assist in choosing the better classifier and not necessarily to evaluate the existence of a causal relationship. This can be evaluated in more detail as part of future work.

An additional factor for consideration is the impact of lag when using share price movement. A more detailed understanding could improve the correlation coefficient and it is recommended to be evaluated in future work.

7.2 Future Improvement of Sentiment

As discussed in Section 1, the measures of reputation are linked to seven key drivers, of which one is financial performance. From the above results (Section 7.1), the sentiment from financial-related communication is correlated with share price. This element can therefore be incorporated into an overall sentiment score i.e. measure of reputation.

It is recommended to expand the sentiment prediction to include additional topics and observe the correlation with share price. These topics can be identified through topic modelling as addressed in Section 5.1. A model can then be developed that determines the optimum weights for the contribution of the additional topics. An alternative is to extract topics according to the seven key drivers that impact reputation (Section 1) and apply weightings to an overall reputation score. Including additional topics will allow for evaluating topic contributions to sentiment as well as the evolution of various topics over time.

Lastly, it is recommended to explore using either more fine-grained categories or a continuous scale for sentiment. This will allow for determining a period-on-period change in sentiment and provide more valuable feedback to an organisation on the effect of communication (as well as adjusted strategies) on market perception.

7.3 Summary

The results presented in this Chapter addresses one of the sub-questions of this thesis namely whether there is a correlation between sentiment and financial performance. The

findings are positive and it is recommended to expand the topics included in sentiment analysis. It is important to note, however, that correlation does not necessarily imply causation and it warrants a more detailed study to determine the true nature of the correlation and whether sentiment is a causal predictor of financial performance.

More granularity will provide an organisation with a better understanding of the impact of various topics on overall reputation as well as highlight areas for improved communication strategies.

Chapter 8 explores a financial sentiment prediction model based on document content to determine whether it is an option for future investigation.

Chapter 8

Sentiment Prediction at Document Level

8.1 Evaluation and Results

The same data set used for predicting sentiment based on document titles (Chapter 6) was used for this analysis. Figure 8.1 shows the distribution of the number of sentences in the documents with the statistical mode shown as 14. This was used as the upper limit for the length of an input paragraph. Therefore, only the first 14 sentences were used otherwise for shorter documents, all the sentences were used.

The same pre-processing steps as were used in the logistic regression and XGBoost classifers for headline sentiment (Sections 6.4 and 6.6) were performed on the input paragraph.

Binary logistic regression and XGBoost classifiers with TF-IDF vectors as input were developed and Table 8.1 gives the 10-fold cross validation accuracy for each. The data set was split 80:20 for training and testing.

The overall accuracy using all 668 documents are 85% and 97% for logistic regression and XGBoost respectively. The prediction model at document/paragraph level using logistic regression is less accurate than the model based on headlines only (Section 6.6.2) with a 80% accuracy on the 20% unseen portion of the data (compared to 86%). The XGBoost classifier on the other performed substantially better than its counterpart de-

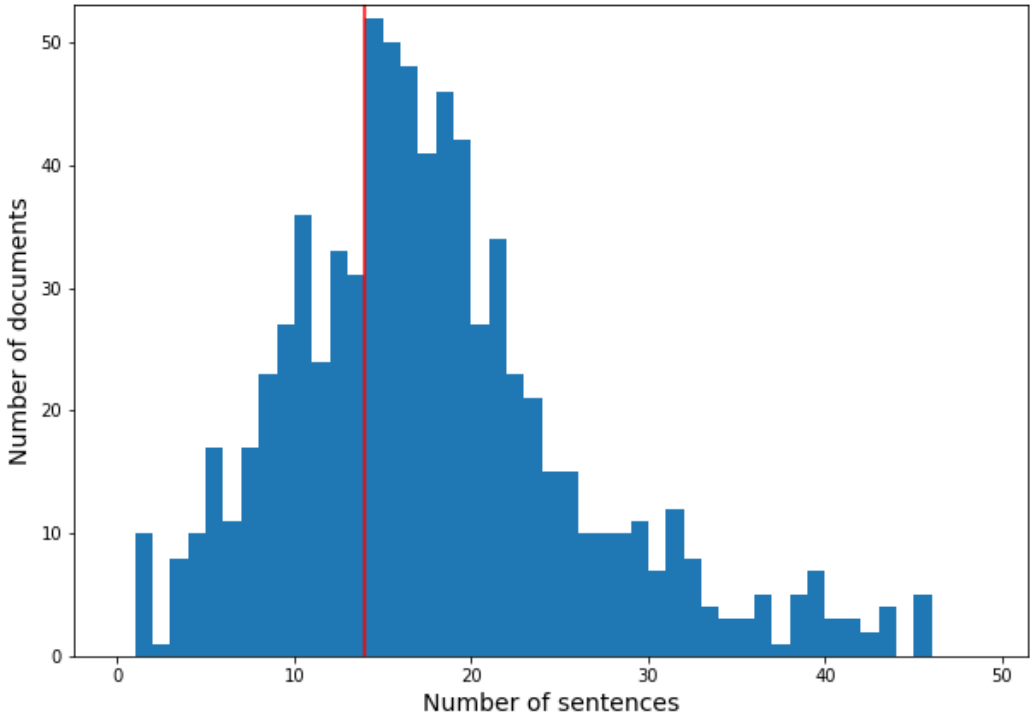

Figure 8.1: Distribution of the number of sentences across the financial documents.

Table 8.1: Cross-validation accuracy for sentiment prediction on document content using logistic regression and XGBoost.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**StDev**	**Test**
LR	0.74	0.80	0.70	0.74	0.74	0.72	0.74	0.72	0.72	0.77	0.028	0.80
XG	0.85	0.80	0.80	0.76	0.85	0.81	0.74	0.79	0.77	0.83	0.038	0.87

veloped on only article headlines (77% vs. 87% on the unseen test data). Table 8.2 gives the recall and F1-score on all the documents when compared to the ground truth.

It is concluded that the XGBoost classifier performs best, whether only on article headlines or on using a paragraph as input. Future work is to focus on paragraph level and consider meaning through using sequences of words. For model generalisation in Chapter 9, however, sentiment predictions are done using document headlines only.

Table 8.2: Final model performance for the binary logistic regression and XGBoost classifiers.

	Logistic regression		**XGBoost**	
	Recall	F1-score	Recall	F1-score
Negative	100%	90%	99%	98%
Positive	62%	76%	96%	97%

8.2 Summary

From the evaluation at document level (assuming a maximum number of sentences), a similar performing prediction model is possible compared to only using document titles. The model using paragraphs as input may contain more information and prove more usable for predicting sentiment on new, unseen data.

In Chapter 9, however, the XGBoost classifier developed using document headlines is used to predict sentiments on another organisation. The bulk of this study focused on document titles with the resulting model performing very well on ***Sasol*** data and is therefore a good baseline to test generalisation.

Chapter 9

Model Generalisation

9.1 Data

From the model development phase, mostly online news articles were used for the financial sentiment model (with 4% being Stock Exchange News Service reports). Therefore, in order to evaluate how well the model generalises, only online news articles were considered.

For purposes of illustrating and testing how well the model extends, data for the corporate organisation, ***Anglo American***, was used.

The ***GoogleNews*** library in **Python** was used to collect data from *Google News* for the period June 2018 - May 2020.

9.2 The Model Pipeline

Recall that the pipeline required (Section 3.1) to predict sentiment (on headline or the first portion of a document) as well as to understand whether sentiment correlates with financial performance, is as follows:

1. Data collection and cleanup/pre-processing (a total of 1758 articles)
2. Filtering of data for financial documents with topic modelling
3. Sentiment prediction using document titles:

- Using the updated dictionary to identify sentiment-carrying keywords (Section 6.2)
- Using the previously developed binary XGBoost classifier based on ***Sasol*** data (without retraining) (Section 6.6.3)

4. Graphically represent daily aggregated sentiments and share price

9.3 Extent of Generalisation

9.3.1 Emerging Topics

The results from the NMF topic model (also using 25 topics) i.e. the top words in each topic are given in Appendix G. The data set is much smaller than that used in model development (Section 5.1) and 25 topics may be too many. However, to illustrate model generalisation, topics 3 and 11 were used as the financial topics and hence the documents to be used for sentiment prediction. Figure 9.1 shows the wordclouds for these 2 topics.

(a) Topic 2: Financial

(b) Topic 7: Environmental

Figure 9.1: Wordclouds to indicate 2 financial related topics.

9.3.2 Sentiment Prediction and Correlation with Share Price

As discussed above, two of the topics were used to filter the data and resulted in a total number of 151 documents in the data set. Table 9.1 gives the predicted sentiments using

Table 9.1: Comparison of sentiment predictions.

	Lexicon-based		**XGBoost**	
	Count	%	Count	%
Positive	49	32%	43	28%
Negative	74	49%	108	72%
Neutral	20	13%	-	-
Other	8	5%	-	-

the dictionary-based and the XGBoost models. The latter is the binary classification model trained on ***Sasol*** data (Section 6.6.3).

From Table 9.1 it seems that the XGBoost classifier is more biased towards negative sentiments whereas the dictionary-based approach appears more balanced. There is only a 52% agreement between the two models. Since there is no ground truth sentiment labels for the data, it was decided to manually evaluate the predicted sentiments (a sample is shown in Appendix H) to provide a more informed view. Herewith three sentences from the ***Anglo American*** data set as an example:

XGBoost incorrect, dictionary correct

"Rand firmer as dollar falls on rate cut bets"

XGBoost: Negative

Dictionary: Positive

XGBoost correct, dictionary incorrect

"JSE rebounds to close firmer | Fin24"

XGBoost: Positive

Dictionary: Negative

XGBoost correct, dictionary correct

"JSE tumbles as global growth fears spread | Fin24"

XGBoost: Negative

Dictionary: Negative

From the manual inspection (Appendix H) it is concluded that the dictionary-based approach predict sentiments more accurately than the XGBoost classifier that was developed using ***Sasol*** data.

Figures 9.2 and 9.3 show the sentiment prediction and share price for both the XGBoost classifier and the model using a simple dictionary to identify sentiment-carrying words.

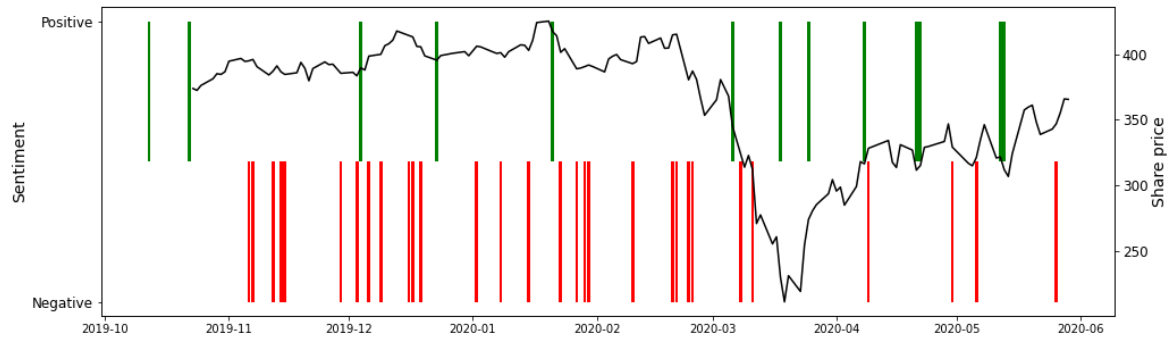

Figure 9.2: Sentiment prediction using a XGBoost classifier compared with share price movement.

Visually, the results look somewhat similar. However, from Figure 9.3 it can be seen that there is an upward movement in share price corresponding to more positive sentiments (post April 2020) which is not as pronounced in Figure 9.2, however, still very comparable. Due to the small amount of data, these results are not conclusive but it is surmised that a XGBoost classifier trained on company-specific document titles may be too specific to extend to other industries.

Furthermore, a larger data set for model training could also improve the generalisation and is recommended to evaluate this for future work.

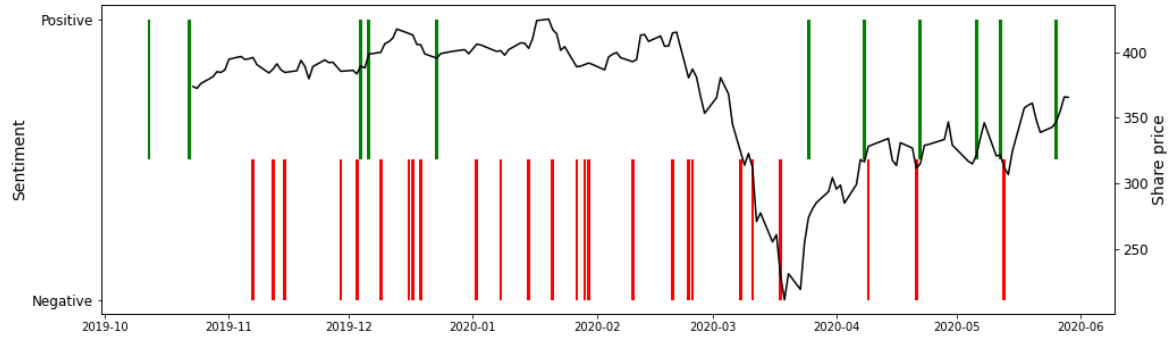

Figure 9.3: Sentiment prediction using a dictionary-based approach compared with share price movement.

It is also recommended to test a model trained on for e.g. both ***Sasol*** and ***Anglo American*** data and observe model performance.

9.4 Vocabulary Shortcomings

In the above section (Section 9.3.2), the XGBoost classifier trained on ***Sasol*** data was used. The input features to this model are the TF-IDF vectors for each document. Recall that the vocabulary size for the previously trained model was 917 (Section 6.4.1).

On first inspection, the XGBoost classifier showed less satisfactory results compared to the dictionary-based approach. A likely explanation is that there is only a small overlap in words between the two companies based on the pre-processing steps implemented. It is recommended to look at removing additional words that is not relevant to sentiment prediction.

For a better understanding, TF-IDF vectors were determined for the ***Anglo American*** document headlines and the vocabulary compared with the pre-trained vocabulary. The vocabulary size using the data was 512, which is smaller than the existing vocabulary. It is not inconceivable to expect a significant portion of overlap of words due to the smaller number of documents in the ***Anglo American*** data set. However, only 29% of the words in this data set exists in the pre-trained vocabulary.

This explains the observation that the XGBoost classifier performs worse than a dictionary-based approach.

9.5 Summary

Since unlabeled data was used to observe how well the sentiment prediction models perform on data related to a different company, measuring the accuracy was done based on observations and was therefore not unbiased. Regardless of this, the approach shows promise in that there were periods where a correlation could be observed.

However, due to the shortcomings already identified in model development (Sections 6.6) regarding misalignment between the sentiment in the dictionary and the predicted sentiments from the logistic regression model and the XGBoost classifier, it is recommended to increase the size of the data set. This should expand the vocabulary as well as the more frequent occurrence of sentiment-carrying words. This can be achieved through either collecting more raw data or through data augmentation using the ***textaugment***[1] library in **Python**. The vocabulary can also be expanded through using paragraphs as inputs.

Based on these results, it should be possible to draw a better conclusion regarding the generalisation of a company-specific prediction model. Lastly, the observed correlation is purely based on a visual analysis. A future improvement is to consider a time-series classification/regression model to improve the correlation between sentiment and share price.

[1] https://github.com/dsfsi/textaugment#mixupaugmentation

Chapter 10

Conclusions

10.1 Summary of Conclusions

Based on the findings it is concluded that natural language processing techniques can be used to derive valuable insights from textual information. It can be used to distinguish various topics and predict the sentiment of financial articles. The study has found that custom models are required in the South African context and is most efficient when developed for a specific company as opposed to generalising too broadly.

10.1.1 Main Research Question

> "What NLP techniques are required to successfully determine the sentiment of financial communication?"

Since unlabeled data was collected, a model development pipeline was designed to filter data for financial documents (using topic modelling), predict the sentiment for each document using headlines (final model is a binary XGBoost classifier) and visually evaluate predicted sentiment and share price to observe whether a correlation exists (Section 3.1).

The following is a summary of the main findings for each of the objectives as given in Section 1.2:

1. Off-the-shelf sentiment analysers

   TextBlob and ***NLTK's Vader*** were evaluated on all documents using the first

10 sentences of a document. The analysis showed that predictions from these models were misaligned mostly for financial-related (and environmental-related) topics which justified the purpose of the thesis.

Furthermore, compared to the ground truth sentiments of the financial articles, these analysers performed poorly with an accuracy of only ∼54% from ***NLTK's Vader***.

2. Alternative sentiment prediction approaches

   (a) Lexicon-based
   The *Loughran and McDonald Sentiment Word Lists* [26] were used as a basis dictionary after which synonyms were added to expand the lists. Words were then manually added and removed based on evaluating a sample of the results from the first dictionary update. Due to a relatively small data set, this manual task was manageable. However, for large data sets and automated process will need to be implemented should this approach be followed. This approach gave an overall accuracy of 84% (only taking into account Positive and Negative predictions).

   (b) Machine learning: Logistic regression and XGBoost
   Both a multi-class and a binary logistic regression model as well as a binary XGBoost classifier were evaluated and the latter gave the better performance with an overall accuracy of 93%. Recall and F1-score on both ***Positive*** and ***Negative*** predictions improved compared to a logistic regression model with the most improvement seen on ***Positive*** predictions.

3. Recommended sentiment prediction model
   From the various approaches to predict sentiment based on document headlines it was found that a binary XGBoost model outperformed the other models and was recommended as the annotation model.

   An additional approach using the first few sentences of each document (rather than the title) was evaluated and the results indicated that the logistic regression model performed worse with regards to ***Positive*** sentiment classification. The XGBoost

classifier, however, showed improved performance for both categories.

The main contributions are as follows:

- Expanded the base financial sentiment dictionary (data statement given in Appendix D).
- Developed a full pipeline to filter data for financial topics and predict the sentiment from the article headline.

10.1.2 Sub-Question 1

> "Is there a correlation between the sentiment of financial news and company performance as indicated by share price?"

A binary XGBoost classifier for predicting the sentiment of financial news articles using headlines showed there are periods where good correlation with financial performance (i.e. share price) can be seen. The approach shows promise, and with refinement, can be used to identify at risk periods for an organisation.

The next step is a sentiment prediction (reputation score) including additional topics to take into account the impact of various areas on overall sentiment. This aggregated score can then be correlated with share price again.

10.1.3 Sub-Question 2

> "How effectively can a narrower sentiment prediction model be applied to a broaden scope of financial related information?"

The sentiment prediction model (based on the model development pipeline) was evaluated using data from a different company to test how well it generalises. Since there were no ground truth data labels for this, a manual evaluation on a sample of the results was done. The dictionary-based approach, binary logistic regression model and the XGBoost classifier were compared and it was concluded that the former was better suited in this case.

Despite these shortcomings, a correlation between predicted sentiment and share price was still observed for certain periods. This substantiates the fact that the method has promise.

Based on these findings it was concluded that the size of the data set used in model development was too small and should either be augmented or additional data collect to increase the vocabulary and occurrence of sentiment-carrying words.

10.2 Future Work

The following is a summary of the main recommendations for future work to improve and expand on the results from this thesis:

- A more sophisticated, streamlined process to update/expand sentiment word lists for a dictionary-based prediction model. Evaluate the potential to leverage of the TF-IDF vocabulary for this process.
- Improve the model generalisation capability. It is recommended to consider the following:
  - Enhanced pre-processing to further remove words that are not relevant to sentiment prediction for e.g. named entities, common words (such as "stocks").
  - Increase the size of the data set to ensure sentiment carrying words occur more frequently and improve prediction accuracy (through data augmentation or increased raw data collection).
  - Use paragraphs as inputs to further expand the vocabulary.
- Evaluate a neural network for sentiment prediction (such as Long Short Term Memory with attention) where the sequence of words are considered to determine whether or by how much it improves the baseline (XGBoost classifier with TF-IDF vectors as input). Based on research (Section 2.2.2), it was shown that there may be performance benefits when considering non-linear relationships captured by neural networks.

- Determine a method to represent sentiment on a continuous scale (or more fine-grained categories). This will allow for tracking a change in sentiment whereas a binary, categorical prediction does not allow for this.
- Investigate the impact of share price movement lag on the correlation with sentiment and enhance the understanding on whether there is a causal relationship between sentiment and financial performance.
- Expand the sentiment prediction model to include additional topics (over and above financial documents). The following should be investigated:
  - Develop a model that determines the relative weights of the added topics required to calculate their contribution to the overall sentiment/reputation score. This process should also consider improved topic modelling based on for e.g. part-of-speech tagging and/or topic keyword lists.
  - Comment on the correlation of the overall sentiment with share price and the impact of various topics over time.

.

Recall from Chapter 1 that the thesis of this study is that natural language processing techniques can be used to predict the sentiment of South African finance-related documents/articles and furthermore be correlated with the movement in company share price (which is used to represent financial performance). Considering the above-mentioned summary of the conclusions regarding the research questions and objectives, the first part of the thesis is proven to be correct with the second part, in principle, proven correct however requires further work to refine. The relationship between sentiment and share price needs further understanding to efficiently distinguish between correlation and causation.

Bibliography

[1] Emmanuel Ameisen. Always start with a stupid model, no exceptions, 2018. Head of AI at Insight Data Science. Available online: [https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa] (Accessed: 13 October 2020).

[2] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 2016.

[5] Spandan Ghose Chowdhury, Soham Routh, and Satyajit Chakrabarti. News Analytics and Sentiment Analysis to Predict Stock Price Trends. *International Journal of Computer Science and Information Technologies*, 5(3):3595–3604, 2014.

[6] Simon Cole. What price reputation?, 2019. AMO Strategic Advisors, Reputation Dividend. Available online: [https://www.amo-global.com/files/media/files/04411587427371e912e14b4b93476f48/AMO_What_Price_Reputation_report.pdf] (Accessed: 20 July 2020).

[7] Elanor Colleoni, Adam Arvidsson, Lars K. Hansen, and Andrea Marchesini. Measuring corporate reputation using sentiment analysis. In *Proceedings of the 15th Inter-*

national Conference on Corporate Reputation: Navigating the Reputation Economy, New Orleans, USA, 2011.

[8] Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics (Switzerland)*, 9(3), 2020.

[9] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6):352–359, 2002.

[10] Korn Ferry. Fortune world's most admired companies, 2019. Available online: [https://www.kornferry.com/insights/articles/fortune-worlds-most-admired-companies-2019] (Accessed: 20 July 2020).

[11] Charles Fombrun and Cees Van Riel. The reputational landscape. *Corporate reputation review*, pages 1–16, 1997.

[12] Petr Hajek, Vladimir Olej, and Renata Myskova. Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738, 2014.

[13] Bi Min Hsu. Comparison of supervised classification models on textual data. *Mathematics*, 8(5), 2020.

[14] Doaa Mohey El Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338, 2018.

[15] C.J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 216–225, 2014.

[16] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2013.

[17] Mengxiao Jiang, Man Lan, and Yuanbin Wu. ECNU at SemEval-2017 Task 5: An Ensemble of Regression Algorithms with Effective Features for Fine-Grained Sentiment Analysis in Financial Domain. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 888–893, 2017.

[18] Kalyani Joshi, Bharathi H. N, and Jyothi Rao. Stock Trend Prediction Using News Sentiment Analysis. *International Journal of Computer Science and Information Technology*, 8(3):67–76, 2016.

[19] Svetlana Kiritchenko and Saif M. Mohammad. The effect of negators, modals, and degree adverbs on sentiment composition. *arXiv*, 2017.

[20] Srikumar Krishnamoorthy. Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, 56(2):373–394, 2018.

[21] Abhishek Kumar, Abhishek Sethi, Shad Akhtar, Asif Ekbal, Chris Biemann, and Pushpak Bhattacharyya. IITPBatSemEval-2017Task5: SentimentPredictioninFinancialText . *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 894–898, 2017.

[22] Nandhini Kumaresh, Venkateswarlu Bonta, and N Janardhan. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6, 2019.

[23] Richard J. Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[24] James Lappeman, Robyn Clark, Jordan Evans, Lara Sierra-Rubia, and Patrick Gordon. Studying social media sentiment using human validated analysis. *MethodsX*, 7:100867, 2020.

[25] Qin Lei. Financial value of reputation: Evidence from the ebay auctions of gmail invitations. *The Journal of Industrial Economics*, 59(3):422–456, 2011.

[26] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66.

[27] Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini, and Jacopo Staiano. Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? Inferring sentiment towards brands from financial news headlines. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 817–822, 2017.

[28] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27(February):16–32, 2018.

[29] Austin McCartney, Svetlana Hensman, and Luca Longo. "How short is a piece of string?" The impact of text length and text augmentation on short-text classification accuracy. *CEUR Workshop Proceedings*, 2086, 2017.

[30] Andrew Moore and Paul Rayson. Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: Predicting sentiment from financial news headlines. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 581–585, 2017.

[31] Andrius Mudinas, Dell Zhang, and Mark Levene. Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward. 2019.

[32] Hanjoodengmailcom Odendaal, Nicolaas Johannes, and Mreidsunacza Reid. Media based sentiment indices as an alternative measure of consumer confidence. 2018. A working paper of the department of economics and the bureau for economic research at the University of Stellenbosch. Available online: [https://towardsdatascience.com/a-new-way-to-sentiment-tag-financial-news-9ac7681836a7] (Accessed: 13 March 2020).

[33] Anjuman Prabhat and Vikas Khullar. Sentiment classification on big data using Naïve bayes and logistic regression. *2017 International Conference on Computer Communication and Informatics, ICCCI 2017*, (January 2017), 2017.

[34] Bilal Saberi and Saidah Saad. Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, 7(5):1660–1666, 2017.

[35] Sheikh M. Saqib, Shakeel Ahmad, Asif H. Syed, Tariq Naeem, and Fahad M. Alotaibi. Analysis of latent Dirichlet allocation and non-negative matrix factorization using latent semantic indexing. *International Journal of Advanced and Applied Sciences*, 6(10):94–102, 2019.

[36] Elliot S. Schreiber. Reputation, 2011. Institute for Public Relations. Available online: [https://instituteforpr.org/reputation/] (Accessed: 7 July 2020).

[37] Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M. Khoshgoftaar. Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, 5(1), 2018.

[38] T. Hastie, R. Tibshirani and J.H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York, 2nd edition, 2009.

[39] Maite Taboada, Julian Brooke, and Kimberly Voll. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[40] Silvija Vig, Ksenija Dumicic, and Igor Klopotan. The impact of reputation on corporate financial performance: Median regression approach. *Business Systems Research*, 8(2):40–58, 2017.

[41] Vered Zimmerman. A new way to sentiment-tag financial news, 2019. Available online: [https://towardsdatascience.com/a-new-way-to-sentiment-tag-financial-news-9ac7681836a7] (Accessed: 13 February 2020).

Appendix A

Calculated sentiment based on sentences

Figures A.1 and A.2 show the distribution of sentiment categories for the two clusters of financial topics. The per topic sentiment was calculated as outlined in Algorithms 3.1 and 3.2 in Section 3.3. These are specifically for sentiments that were calculated using the first 10 sentences of a document and considering the majority sentiment to assign a document sentiment. A majority sentiment approach was also applied to determine the per topic sentiment.

A.1 Summary

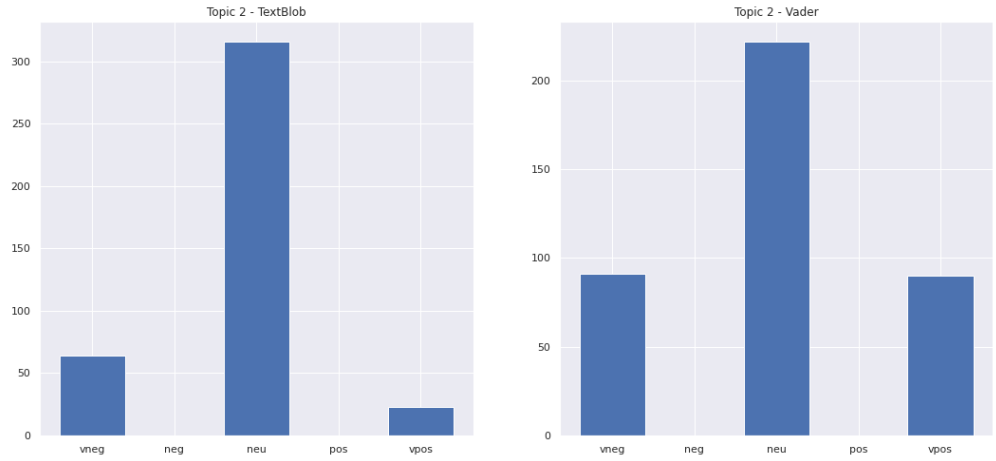

Figure A.1: Sentiment distribution for topic 2 based on individual sentence sentiments.

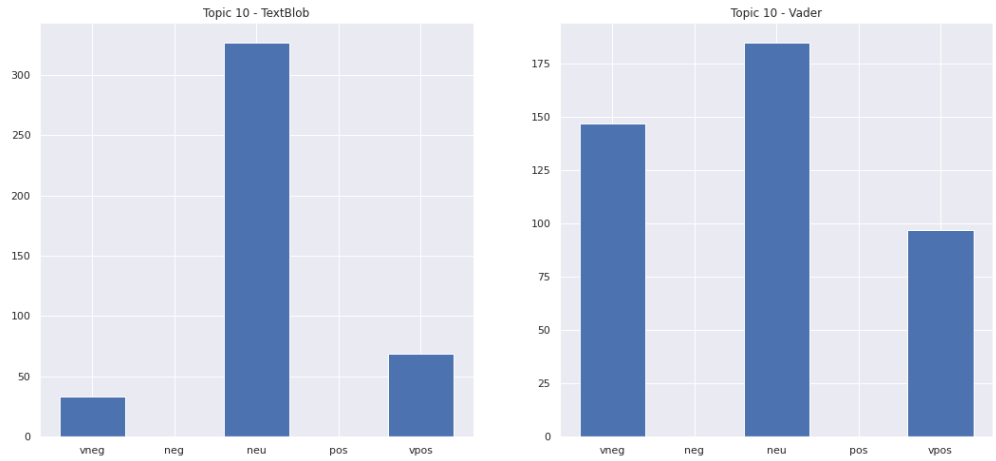

Figure A.2: Sentiment distribution for topic 10 based on individual sentence sentiments.

Appendix B

Model Development: Top Words from NMF Topic Model

Tables B.1 and B.2 give the top words for each topic identified using the ***Sasol*** data.

Table B.1: Top words for topics 0 - 15.

0	say company would also make year one african years time fund take business go government
1	team amaze download 2020 study profile case solar mar branchnews ctp roomadvertisingcopyright printers school jul
2	price share oil percent earn year financial billion lower expect per cost group debt result
3	financial mail every arena thursday except sa transaction january publish distribute month december time acceptance
4	sharenet crypto trade currencies art risk data may market brand website therefore platforms artists technical
5	die het en van n op vir se sy te om wat nie hy meet
6	banyana league womens football alert play players match team ladies game coach cup safa app
7	emissions coal power climate air environmental carbon eskom pollution plant greenhouse dioxide mine emission coal-fired
8	cookies accept say billion rand website cornell project min nqwababa function right petrochemicals edit read
9	research securities macquarie 0 london report capital ltd 1 may product return citi source distribute
10	jse index rand close trade stock market gain jul gold point lose bank dollar platinum
11	learners bursary science engineer programme school technology education study university development graduate young stem skills
12	gas energy oil mozambique project natural fuel supply production development lng pipeline plant use petroleum
13	police station road service suspect vehicle garage petrol park safety car traffic fire area fill
14	water dam vaal river air highveld department week system level quality area mpumalanga gauteng last
15	race gtc championship motorsport marathon second bmw finish car volkswagen round circuit rowe kyalami win

Table B.2: Top words for topics 16 - 24.

16	company business board executive group jul director july ayo service ceo 2019 award mine annual
17	inzalo shop khanyisa scheme share black bee solbe1 shareholders magents empowerment mall relabelled ordinary solidarity
18	eskom ruyter de power ceo nampak utility constable executive appointment gordhan andre electricity eskoms position
19	project lccp charles lake cost chemicals plant company unit knowledge fast cyril context units alec
20	sponsor room advertise branch news parent mr 101 ridge ms advertisers 2020 online copyright vaalweekblad
21	market report research global industry analysis segment growth chemical demand forecast key wax alcohol release
22	morgan stanley investment comment bank investors company stock securities merrill risk lynch shareholders target j.p.
23	iol email share article limit address subscriptions creamer sponsor subscribe media website news via report
24	say agrizzi bosasa court watson commission marine coast drill people environmental appeal bribe officials state

Appendix C

Update of Sentiment Word Lists

C.1 Sample of Sentiment Predictions

The following is an extract of from the results of Experiment 1 (Section 3.4.2) giving the article headlines with the sentiment-carrying words that matched words in the updated dictionary as well as the sentiment associated with that word.

1 Africa's news leader. **- Not detected**
2 Retailers Foschini, Mr Price, Woolworths under pressure - Moneyweb **- Not detected**
3 JSE ***opens*** weaker in lacklustre trade as gold makes gains **- Negative**
4 Sasol dives ***further*** on the JSE over cash-raising plan **- Positive**
5 JSE wrap | ***Reprieve*** for local stocks after volatile session | Fin24 **- Negative**
6 More than R45 billion wiped ***off*** Sasol in a single day | Fin24 **- Negative**
7 South Africa's Sasol half-year ***earnings*** fall 74% as U.S. project weighs | African Mining Market **- Positive**
8 ***WATCH***: Rand drifts weaker overnight **- Negative**
9 The JSE is plunging amid panic selling, and global markets tanking on coronavirus fears **- Not detected**
10 ***Stocks***, oil, and bitcoin plunge as US lawmakers fight over coronavirus rescue package **- Negative**
11 South Africa: Sasol ***Ups*** Dividend Payout Despite Dip in Profit - Footprint to Africa

- Positive

12 Rand ***opens*** stronger ahead of MTBPS and FOMC **- Negative**

13 JSE edges ***higher*** in cautious trade, led by gold miners **- Positive**

14 ANA News | Sasol ***interim*** headline earnings rise 32 percent **- Negative**

15 South African Markets - Factors to ***watch*** on March 7 - Agricultural Commodities - Reuters**- Negative**

16 Beware the ripple ***effects*** of Nene axing**- Negative**

17 Sasol Khanyisa Public – ***Interim*** results release | Company Notices & Announcements **- Negative**

18 JSE edges ***higher*** in cautious trade, led by gold miners**- Positive**

19 JSE ***loses*** nearly 10% on 'worst day in almost 2 decades'**- Negative**

20 Another ***interest*** rate cut could give the rand some much-needed impetus **- Negative**

21 Global ***stocks*** stuck in Sino-US trade war whirlwind **- Negative**

22 ***Positive*** global sentiment lifts the JSE | Fin24 **- Positive**

23 Markets WRAP: Rand ***closes*** at R14.74/$**- Negative**

24 Coronavirus ***collapses*** the global economy - POWER 98.7 **- Negative**

25 Markets in see-saw mode, as rand trade in tight range **- Not detected**

26 South Africa Sasol Seeks ***Partner*** in Lake Charles, Louisiana Plant - Bloomberg **- Positive**

27 Sasol shares plummet, interim dividend put on hold - CHANNELAFRICA **- Not detected**

28 Rand ***slips*** on stronger dollar, stocks up**- Negative**

29 South Africa's ***richest*** people lost R6 billion thanks to the coronavirus **- Positive**

30 UPDATE 1-South Africa's rand ***heads*** for monthly loss, stocks flat - Reuters **- Negative**

C.2 Removal of Words

Table C.1 gives the list of words that were removed after the addition of synonyms during Experiment 2 (exclusion of synonyms for modal words). All variations of words may not have been included since the words in the headlines were lemmatized before matching the dictionary.

C.3 Addition of Words

Table C.2 gives the list of positive and negative words that were added to the dictionary created during Experiment 2. Words in an article title were lemmatized before matched to entries in the sentiment dictionary.

C.4 Addition of Bi-grams

Table C.3 gives a few bi-grams that were added to the sentiment dictionary for cases where one word would be ambiguous. The list is not exhaustive and is to indicate the impact of expanding the sentiment dictionary.

C.5 Summary

The base dictionary, which were the sentiment word lists developed for tagging financial information with an appropriate sentiment, was expanded and updated to better represent the financial articles used for model development. The purpose of this was illustrate the process of building on existing sentiment word lists.

Table C.1: Words removed from sentiment dictionary.

set	sets	say
said	says	saying
interim	results	stock
stocks	closes	closed
close	closing	closes
put	reach	reached
reaches	reaching	price
prices	pricing	bit
bear	bears	bearing
watch	top	black
blue	aim	aims
aimed	sharp	big
biggest	pull	pulls
pulling	pulled	earnings
base	hit	hits
lead	leads	price
small	smallest	smaller
takes	taking	take
capital	rally	rallies
stall	stalls	stalling
stalled	back	track
interest	ahead	despite
soft	commodities	commodity
break	head	heads
headed	heading	richest
profit	see	sees

Table C.2: Words added to the sentiment dictionary.

Negative		Positive	
dip	troubling	momentum	bright
dips	plunge	beneficiary	brightens
dipped	plunges	vogue	brighter
dipping	plunged	breather	weather
blaze	plunging	prioritize	weathers
blazes	plummet	rally	weathered
blazed	plummets	rallies	weathering
blazing	plummeted	robust	piques
sink	dive	climb	pique
sinks	bloodbath	climbs	brighter
sank	retreat	resilient	assuring
sinking	retreats	recovery	assurance
rebounds	retreated	protect	maintain
drops	retreating	protects	maintained
brunt	stall	buoyed	shines
stalls	stalling	maintains	maintaining
stalled	blacklist		
soft	softer		
see-saw	slide		
slides	slumber		
meltdown	rout		

Table C.3: Bi-grams added to the sentiment dictionary.

Negative	Positive
record low	new record
record lows	record high
back foot	record highs
price halves	record production
	on track

Appendix D

Data statement for the LM-SA-2020 Sentiment Word List

Data set name: LM-SA-2020
Citations: N/A
Data set developer(s): Michelle Terblanche
Data statement author: Michelle Terblanche
Collaborators: N/A

A. CURATION RATIONALE

The *Loughran and McDonald Sentiment Word Lists* were developed using corporate 10-K reports between 1994 and 2008 [26]. These reports are relevant to companies in the United States of America and required by the U.S. Securities and Exchange Commission (SEC)[1].

The motivation for building the ***LM-SA-2020*** word list was based on an experiment using the above-mentioned original lists to detect sentiment-carrying words in South

[1]https://www.investopedia.com/financial-term-dictionary-4769738

African financial article headlines. A corpus of 808 financial articles (relating to ***Sasol***[2]) were used and only 37% of headlines had words of which the sentiment matched that of the words in the *Loughran and McDonald Sentiment Word Lists* correctly according to ground truth labels. A gap was therefore identified in developing a method for predicting sentiment of financial articles in a South African context.

Due to the size of data set, it was possible to manually examine the headlines to identify sentiment-carrying words to be included in the original word lists. Furthermore, synonyms were added for the existing words in the *Loughran and McDonald Sentiment Word Lists* using *NLTK's WordNet*[3] interface. The sentiment detection/prediction accuracy improved by 29% using the new word list.

This sentiment word list can be further expanded/improved in future by increasing the size of the data set and/or including data from other companies. It highlights the need for not only domain-specific sentiment prediction tools but also region-specific corpora.

B. LANGUAGE VARIETY

The language of this data set is American English. The original lists were developed using reports from American companies and the expansion was done by adding synonyms using *NLTK's WordNet* interface which was developed by Princeton and hence also American English. The words that were manually added to the list are also considered American English.

C. SPEAKER DEMOGRAPHIC

No specific considerations were made regrading speaker demographic.

The original documents used to develop the sentiment word lists were corporate 10-K reports published by American companies. The specific speaker/author demographic

[2]www.sasol.com/

[3]https://www.nltk.org/howto/wordnet.html

was therefore not available but assumed that the authors are well-versed in the English language.

The financial documents used to expand the sentiment word list were predominantly from South African online news publishers, with 4% of the documents from *Stock Exchange News Reports* published by the ***Johannesburg Stock Exchange***[4]. Again, specific speaker/author demographics are therefore not known but since these articles are forms of formal communication, it is assumed that the authors are fluent in English.

D. ANNOTATOR DEMOGRAPHIC

The original list of words were developed and their sentiment evaluated and annotated by the authors/data set developers, Tim Loughran and Bill McDonald, who are with the University of Notre Dame [26]. No specific information is available regarding their demographics, however from some research, they are white, American males over 50.

Four annotators were used to label the financial articles based on headlines which were then used to expand the word lists. Herewith the demographic information of the annotators:

Table D.1: Annotator demographic

	1	**2**	**3**	**4**
Description	Financial understanding	Financial understanding	Financial understanding	Financial understanding
Age	65-70	50-55	45-50	35-40
Gender	Female	Male	Female	Female
Race/ethnicity	Caucasian	Caucasian	Caucasian	Caucasian
First Language(s)	Afrikaans	English	Afrikaans	Afrikaans English
Linguistics training	No	No	No	No

[4]https://www.jse.co.za/services/market-data/market-announcements

E. SPEECH SITUATION

The original documents used to create the *Loughran and McDonald Sentiment Word Lists* were collected between 1994 and 2008. These are official, corporate reports from American companies and available in written format. Since publishing these is a requirement by the SEC, the documents are formal (i.e. edited/scripted). The intended audience are potential investors.

The financial articles used to develop the ***LM-SA-2020*** were published by South African online news platforms from the period mid 2015 - mid 2020. The news articles are in written format and edited before publishing. The intended audience is the general public.

F. TEXT CHARACTERISTICS

The documents/articles used in generating the original word lists as well as for expanding these to develop the ***LM-SA-2020*** list, are formal communications relating to the financial domain. In the South African context, the news articles used are from official online news platforms. These, however, can be subjected to publisher bias.

G. RECORDING QUALITY

N/A

H. OTHER

The ***LM-SA-2020*** sentiment word list can be made available on request.

I. PROVENANCE APPENDIX

The original word lists, the *Loughran and McDonald Sentiment Word Lists*, were used to developed the ***LM-SA-2020*** data set. The original word lists are publicly available[5] [26].

The data statement for these word lists are available at:
https://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf

References

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66.

[5] https://sraf.nd.edu/textual-analysis/resources/

Appendix E

Update of NLTK Stopword List

Table E.1 is a short list of words removed from the standard NLTK stopword list since these are considered sentiment-carrying.

Table E.1: Words removed from NLTK standard stopword list.

up	down
as	after
while	but
against	between
into	through
during	before
above	below
over	under
again	further
very	

Appendix F

Sentiment Correlation with Financial Performance: Logistic regression

Figure F.1 gives the sentiment prediction from the binary logistic regression model compared with the daily share price. The correlation coefficient for the given period is 0.05.

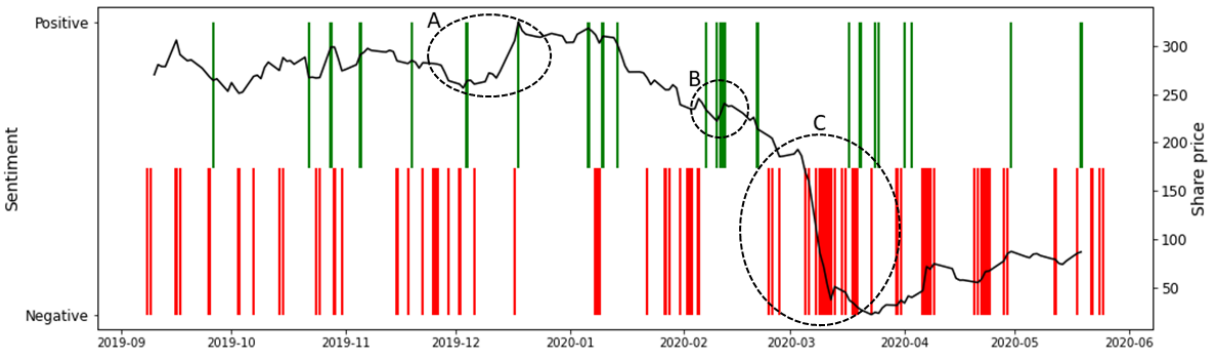

Figure F.1: Sentiment prediction vs. share price for September 2019 - May 2020 logistic regression.

Appendix G

Model Generalisation: Top Words from NMF Topic Model

Tables G.1 and G.2 give the top words for each topic identified using the ***Anglo American*** data.

Table G.1: Top words for topics 0 - 15.

0	mine say company also south years business industry make new one operations take see first
1	regional email separate please variety offer photo store website via password region pdf send subscriptions
2	prank prefer constitute condition remove eu power-mad privacy policy party chance president benefit ownership boost
3	index jse trade stock gold rand market gain weaker week fell close investors well all-share
4	platinum price metal year palladium increase earn say expect amplats share impala higher would pgm
5	diamond diamonds de beers year sales production lower expect 2020 carats market price sell demand
6	like want people job many live one social school create community bring role make go
7	energy technology emissions truck fuel new power electric use need renewable reduce hydrogen target part
8	analysis finance trend join edition cut opportunities international insights risk spot help already expert data
9	see like market look register insider many please feature independent click quality receive great select
10	coal power eskom thermal year station south32 energy south glencore sale would assets comment african
11	world change rise keep get investment need local price share invest fast investors story news
12	cookies accept reserve sign union mine run use include follow ensure statement coal right operations
13	deep oct november friday brief editor analysis ron nov big story fin24 22 dive expert
14	read copper reuters say company construction year demand worlds top min million discover thomson chile
15	biggest include service first latest open minister state reuters government people limit dollar white deputy

Table G.2: Top words for topics 16 - 24.

16	time access site business already day data december simply subscription sign content digital best use
17	safety ore iron kumba provide company operations receive work improve information use brazil number technology
18	article south section legal newspapers say trend iol follow share johannesburg african business group ago
19	company project mine americans include partner development complete venture share digital focus million media plan
20	brazil say company 2018 platinum reuters report tail investors hold zimbabwe mineral fund detail standard
21	remote production online mine copper performance copyright program company innovative ltd. power chile best global
22	gold anglogold resources legal ashanti group right settlement company mine agree form action rainbow party
23	plant water operations communities also facilities operate control well days national financial process impact continue
24	coronavirus minerals covid-19 say share sirius shareholders pandemic use price small lockdown yorkshire fund support

Appendix H

Model Generalisation: Headline Sentiment Prediction

H.1 Sample of Sentiment Predictions

The following is an extract of the headline sentiment predictions using a binary XGBoost classifier as well as a dictionary-based approach.

Table H.1: Comparison of sentiment predictions on ***Anglo American*** data.

Sentence	**XGBoost classifier**	**Lexicon-based**
Aveng execs get R17.7m in bonuses - Moneyweb	Negative	Negative
Sharp (partial) recovery in share prices - Moneyweb	Positive	Positive
JSE tumbles as global growth fears spread | Fin24	Negative	Negative
Anglo American replaces Deloitte with PwC as external auditor after 20 years	Negative	Litigious
Anglo Says S. Africa's Eskom a Major Risk as It Mulls Growth - Bloomberg	Negative	Negative
Another Major Investor Leaves the Pebble Mine | NRDC	Negative	Positive
Mining lobbies and the modern world: new issue of Mine Magazine out now	Positive	Positive
BHP approach to Anglo CEO signals end of Mackenzie era is nearing	Negative	Positive
JSE tracks global markets higher on improved Wall Street data | Fin24	Positive	Positive
Rand firms as dollar, stocks fall	Negative	Positive
Anglo American delivers 3.5-billion USD profit, declares final dividend	Negative	Neutral
Best Mining Stocks to Buy in 2020 | The Motley Fool	Positive	Positive
Rand firmer as dollar falls on rate cut bets	Negative	Positive
The new ministers in charge of the Amazon	Positive	Negative
Anglo American's Cutifani not thinking of retirement as plots coup de grâce - Miningmx	Negative	Negative
Pressure persists for resources stocks | Fin24	Positive	Negative
Markets WRAP: Rand closes at R14.73/$ | Fin24	Negative	Neutral
See the top performers on the JSE in 2018 so far	Positive	Litigious
Anglo American seeks to avert revolt over chief's £14.6m pay | Business News | Sky News	Negative	Negative
Rand, stocks slip as investors await big Trump speech	Negative	Negative