Vrije Universiteit Amsterdam

Bachelor Thesis Information Sciences

Applying Natural Language Processing to the text of financial reports

Author: Wouter van Zeijl (2533591)

Supervisor: Sieuwert van Otterloo

A thesis submitted in fulfillment of the requirements for the bachelor of science degree in Information Sciences

August 3, 2023

Abstract

This study investigates the potential predictive power of CEOs' verbal communications on future stock performance by applying Natural Language Processing (NLP) techniques to textual data derived from annual and quarterly financial reports, as well as CEO quotes. Despite the observed disparities in word frequencies between press releases related to ascending and declining stock prices, the analysis found that these patterns did not strongly predict future stock price changes. The study underscores the complexity and challenges of predicting stock prices based solely on textual data and suggests the need for a multidimensional approach for accurate stock price predictions. The findings highlight the potential utility of NLP in financial discourse analysis and its application in identifying patterns in CEOs' communications. Despite the limitations, the study opens up avenues for further research employing advanced NLP techniques, considering a broader range of data sources, and exploring additional types of corporate communications for better understanding and predicting market behavior..

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

In the financial field, financial analysts are overwhelmed with data. However, most of this data is languishing in data warehouses and repositories because it contains an unstructured form. Unstructured text is easily processed by humans, but machines struggle when they have to process unstructured data. Data mining is a study that deals with analysing this unstructured data and uses statistical techniques to extract information for these large unstructured data collections (1). Annual reports (10-K filings) are an example of data sources that contain numerical and textual data. However, most research and information extraction on these reports have been focused on numerical data. The large chunk of data that is stored in textual form is often left out. The two main reasons not to analyse text data are that of time constraint and subjectivity (2). The amount of text is too abundant to analyse it in a manner that is not too time consuming and each person can interpret a text in a different manner. This introduces the new technique of text mining (3).

Text mining focuses on the task of extracting insightful information from text. It was first introduced by Feldman et al (4) and since then has grown into a wide set of related topics and algorithms for analysing text. Some of these topics are Information retrieval, Information Extraction from text, data mining and Natural Language Processing (NLP) (5). The latter two may seem identical, but they are not. Text mining is about discovering and extracting non-trivial knowledge from unstructured text. This also deals with information retrieval and cluster algorithms. NLP attempts to extract a fuller meaning from unstructured text. NLP also takes the grammatical structure into account, where text mining does not. NLP does this by knowledge representations that represent the grammatical properties of words and their meanings (6). Therefore, NLP can potentially tackle the two problems of time and interpretation when it comes to analysing financial text data.

The interest of data mining has been significant in most fields, however NLP has experienced a dominant increased interest in the field of finance. In finance researchers work on cutting-edge forms of NLP. Financial qualitative market data can be utilised to improve investment decisions. It is suggested that NLP can be used in combination with CNG analysis to help financial managers to more easily read annual reports. However, financial text analysis through NLP has just started (7).

This research aims to apply text mining and NLP methods to financial statements and accompanying press releases. Text mining techniques are deployed on the data sets to generate features. On top of this the Python NLP library spaCy is used to create more insightful features such as sentiment analysis or readability scores. By doing so, this study will contribute to the growing body of literature on NLP in finance and help advance its applications in financial text analysis.

1.2 Problem Definition

The growing complexity and volume of unstructured text data in the financial sector pose significant challenges for financial analysts and decision-makers (8). While recent advancements in NLP offer promising solutions for extracting valuable insights from unstructured financial documents (9), there remains a scarcity of comprehensive evaluation of NLP techniques and tools in the context of financial text analysis. Specifically in the context of the python library spaCy. This gap in the literature has impeded the development of more effective NLP methodologies tailored for the financial sector. Furthermore, it limits the specific needs of financial analysts and their understanding of analysing textual data (10).

The evolving landscape of financial reporting also presents new challenges and opportunities. For instance, the introduction of the Inline XBRL (iXBRL) mandate by the Securities and Exchange Commission (SEC) in the U.S. in 2020 led to a shift in how companies submit their financial reports. The new format blends human-readable disclosures with machine-readable XBRL data tags within the same document, creating an integrated report (11). This shift not only underscores the continuous evolution of financial reporting but also underscores the need for effective text mining tools to process and analyse these integrated reports.

By addressing these challenges, researchers and practitioners can harness the potential of text mining in finance more effectively, leading to better-informed decision-making and improved financial outcomes.

1.3 Research question

This research will answer the following research question:

- **To what extent can Natural Language Processing gather insights from financial annual reports?**

The goal of this research question is to investigate the potential and limitations of Natural Language Processing techniques in extracting valuable insights from financial filings. To answer this primary research question, we take into consideration two sub-questions. The first sub-question is related to word frequency and has been formulated as follows:

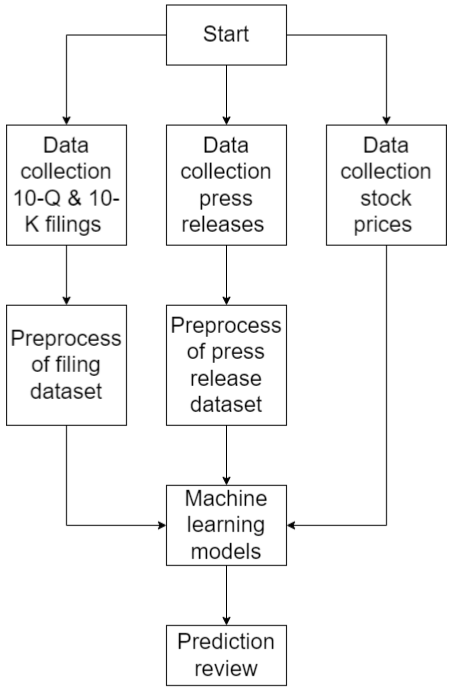
- Does certain word usage correlate with future performance of a company?

This sub-question aims to ascertain whether there is a correlation between word usage in financial reports and the long-term financial well-being of a company, as indicated by its stock price. The second sub-question builds upon the first and is formulated:

- If a correlation exists, can it be used to predict changes in a company's stock market value over the next 12 months?

This sub-question will analyse to what extent the correlation is relevant to the economic health of a company. To answer these research questions we will create two data sets. The first data set will contain annual and quarterly reports and the second data set contains CEO quotes from press releases. Both data sets have been scraped from the SEC government website. The reports are found under the code 10-K and 10-Q. The CEO quotes are found in press releases which are filed under the code exhibit 99.1. The respective stock prices of the companies will be added to each data set respectively. Furthermore, we will process the data to incorporate NLP features. If a correlation is observed between the NLP features and stock prices, we will subsequently employ two machine learning classification models, which are Random Forest and Logistic Regression. This workflow is illustrated in figure 1.1

Figure 1.1: The planned workflow



1.4 Scientific and practical contribution of research

The aim of this research was to investigate to what extent text mining and Natural Language Processing techniques can be applied to gather insights from financial reports. While there has been prior research around stock price prediction with the help of textual data (12, 13), few have delved into the potency of NLP features applied to financial reports. This research provides three main contributions. First, it presents a unique data set which contains the top 50 S&P 500 companies by volume, their respective stock prices, CEO press release quotes and text mining and NLP features. A link to this data set is found in Appendix A1. This data set can serve as a robust starting point for future exploration of data science models for stock price prediction, and it could be further augmented with additional data or refined features to enhance such models. Second, the research offers evidence that certain words from CEO quotes are more prevalent in companies with a rising stock price after a year and vice versa. Third, we introduce a mini Python library that takes as input a CEO quote from a press release and outputs a value indicating the likelihood of the quoted company experiencing a stock price rise after one year.

The practical contribution of this research aims to benefit financial analysts, investors, and decision-makers in companies. For financial analysts and investors, the insights gained from this research can aid them in more accurately predicting stock price changes, thus enabling them to make more informed investment decisions. Moreover, it also takes away time constraints of gathering and reading the data by hand and it prevents different subjective opinions on the data, which can cause conflict between analysts. Finally, the assembled data set can serve as a rich resource for all stakeholders to examine case-by-case instances of company communications and stock price fluctuations.

1.5 Fair data

In order to ensure that the created data set in this research is usable for future reference, FAIR data principles have been applied to make the data set findable, accessible, interoperable, and reusable for future research and practical applications (14, 15). Here's how we've addressed each of these principles:

1.5.1 Findability

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata (defined by R1 below)
- Metadata clearly and explicitly include the identifier of the data they describe
- (Meta)data are registered or indexed in a searchable resource

The data set assembled in this research has been made findable by being available on https://github.com/WoutervanZeijl/Press-releases. Each column in the data set is linked to a detailed description available in a text file named "Explanation".

1.5.2 Accessibility

- (Meta)data are retrievable by their identifier using a standardized communications protocol
- The protocol is open, free, and universally implementable
- The protocol allows for an authentication and authorization procedure, where necessary
- Metadata are accessible, even when the data are no longer available

The data can be accessed and utilized freely by anyone, because it is published publicly on github.

1.5.3 Interoperability

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

The metadata for the data set provides comprehensive descriptions for each feature, facilitating its use in diverse applications. The data set, being in the widely-used CSV format, is interoperable with various types of software, including Python, R, Tableau, and Excel, among others.

1.5.4 Reuse

- Meta(data) are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage license
- (Meta)data are associated with detailed provenance
- (Meta)data meet domain-relevant community standards

The metadata in our data set provides a clear and detailed account of each feature, ensuring the data can be easily understood and reused. The comprehensive descriptions for each column should assist other researchers or practitioners who wish to extend or refine the data set. All data adhere to the standards of the finance domain, enhancing its relevance and utility for future work in the field.

2

Related Literature

Numerous studies related to this topic have already been conducted in this domain. Butler and Keselj attempted to automate the analysis of annual reports. This allowed them to quickly evaluate the textual components of financial filings without the biased opinion of analysts. An n-gram model was created which performed better than the benchmark. The model was not able to capture all the information in the filing, but it captured a significant amount for above average returns. The n-gram model was combined with a bag-of-words approach to obtain better accuracy (12). Falinouss used vector space modelling with term frequency and inverse document frequency to predict stock prices with the help of news articles. The model reached an accuracy of 83%. There is a lot of research on sentiment analysis. Tetlock finds that high pessimism predicts downward momentum in the market. However, Tetlock argues that the changes of pessimistic media are dispersed throughout the trading day (13). Financial reports have also been researched with qualitative analysis. Gupta reviews text-mining in financial statements and discusses the existing literature on text mining in financial applications. Furthermore, text mining methods are discussed in the financial domain. It describes how NLP in the financial sector has developed and what tools are being used for the purpose of data mining on unstructured data (16).

Numerous literature studies show the potential use of NLP in stock market prediction. The most common category of this is sentiment analysis. Chen et al. (2010) and Mukherjee et al. (2013) both perform sentiment analysis but differ in the time interval of the data they use (17, 18). A different flavour of textual sentiment analysis is researched by Schumaker (2009). This research focuses on market trends based on breaking financial news. The authors evaluate the performance of the NLP methodology based on its time-savings effort (19).

3

Theories

This section covers theories related to the conducted NLP techniques on the project. Each theory describes the topic and what it is used for. First the theory for Natural Language Processing is described. This is followed by a description of each technique that was relevant in this project.

3.1 Natural Language Processing

Natural Language Processing in the broadest sense is the technique of enabling computers to understand and analyse natural language much in the same way as humans can. "Natural language" refers to the language that is used for everyday communication by humans. Explicit rules for these natural languages are hard to pin down because they have been passed down from generation to generation, which causes human languages to have an unstructured nature. In contrast, programming languages operate on their binary structure and are a lot faster in their processing. The degree of advancement in NLP ranges from counting words for a frequency analysis to the understanding of complete human phrases. NLP becomes increasingly more widespread. For example, smartphones support the ability to suggest predictive text based on the previous input of the user, search engines are able to retrieve text of unstructured nature and chat bots offer an interactive conversational tool that helps people to perform tasks (20).

NLP faces a challenge in analysing language that is ambiguous and can have more than 1 meaning. Furthermore, it can also be idiomatic and interpret expressions that don't follow grammatical rules.For example, the phrase "Kill two birds with one stone" would be interpreted by a human as someone who has accomplished two things at the same time. However, the NLP algorithm would most likely interpret it as saying that someone killed two birds with one stone. We can overcome these challenges by breaking down language into its constituent parts, this is done with tokenization (see paragraph 3.2). In the early days NLP algorithms faced another challenge of handling digital data. Hard crafted rules had to be written to identify patterns in text. This limited NLP research to applications such as speech recognition and machine translation. However, in the current age NLP algorithms deploy statistical and machine learning approaches to text analysis. An example of this is sentiment analysis on financial statements. Schumaker et al. (2009) used sentiment analysis on financial statements to predict stock prices (19).

3.2 Tokenization

The first step of NLP is tokenization. It is used to split paragraphs, sentences, symbols, words and digits into smaller units called tokens. Tokens form the building blocks of the processed text and simplify the input text for further processing and analysis. This allows extraction of meaningful information and patterns in the given input text. Generally, tokenization is either done by word or per sentence. Furthermore, modern NLP packages allow for more specific rule based methods to tokenize the text. For example, we are able to split the text at punctuation instead of white space. More advanced tokenization methods used machine learning algorithms to understand patterns in the data which are used to handle the tokenization process more accurately (21).

Similar to the challenges discussed in 3.1, tokenization faces challenges in handling multi-word expressions, compound words and language specific nuances. These challenges can be language specific because different languages have varying structures that need to be taken into account.

3.3 Part-of-Speech-Tagging

Part-of-Speech-Tagging (POS) is the process of classifying words into their parts-of-speech and labelling them accordingly. For example, each token is assigned a grammatical category such as verb, noun or adjective in a given text based upon the context and the definition of the token. POS tagging helps NLP tools with question answering parsing, word sense disambiguation and machine translation. There are several approaches to POS tagging (20).

Rule-based POS tagging relies on predetermined rules that consider the morphological and contextual information before assigning a tag to a token. The most common example of this is words ending in "-ing". These words are usually assigned with the tag verb in English. Rule-based methods are able to achieve acceptable accuracy, however, they are difficult to scale across different languages and domains because for each new language the rules have to be reviewed or rewritten.

POS tagging can also be performed with a probabilistic method. This method gathers statistical information from a corpus of text and computes the likelihood of a token having a certain tag. Moreover, this method uses the relationship of adjacent tags and their tokens. This allows for more accurate tagging and can be scaled across different languages. A downside of this is that the calculation of these statistics can be slow and computationally expensive.

Finally, thanks to the help of powerful graphical processor units, Deep Learning and Machine Learning models have recently been applied to POS tagging with success. These models are often able to achieve a very high accuracy. For example, Deshmukh & Kiwelekar (2020) achieved an accuracy of 97% (22).

In paragraph 3.1 we briefly discussed the challenge of language that can be ambiguous. Ambiguous words are the main issue that must be addressed in POS tagging. Words can have multiple meanings and the challenge lies in identifying the correct tag of a word in relation to the sentence it appears in. Another challenge is domain-specific jargon because it has limited exposure in training data. Moreover, jargon can increase the ambiguity in text and complex morphology may increase the ability for POS tagging to recognise the pattern in a sentence.

3.4 Named Entity Recognition

POS tagging proves to be helpful for Named Entity Recognition (NER). NER entities are noun phrases that refer to specific predefined categories, meaning we can classify tokens to real-world entities such as people, organisations, dates etc. NER is great for extracting structured data from unstructured text. This allows for extraction of valuable insights. For example, a financial analyst can use NER to scan a financial statement more quickly. Moreover, NER facilitates text summarization and information retrieval systems. There is also a mismatch between the growth of the amount of digital information and the resources available to manage this data. NER can alleviate this problem by offering the development of tools required for the search and discovery of unstructured data (20). NER contains two approaches that are similar to the first and third POS approaches. The first approach is rule-based. This approach contains hand crafted rules to identify an entity to its correct class. The hand crafted approach generally obtains better precision, but this comes at the cost of a lower recall. Moreover, a hand crafted approach is very time consuming and has to be conducted by computational linguists. The second approach is the use of Machine Learning and Deep Learning models. With the help of statistics, these models can learn relationships within the text data which enables the model to handle a wider range of languages. In contrast to the rule-based approach, this approach is scalable because it doesn't require the time consuming effort that linguists have to put in. NER tools can achieve a high accuracy above 90% and is considered a solved problem. However, Marrero et al. (2013) argue that the current evaluation practices in NER do not allow us to conclude NER as a solved problem. This is because we are not able to say how well the techniques perform with other types of documents that are outside the domain of journalistics. Because there is no sound evaluation method to do so (23).

3.5 Text Classification

Text classification is the task of categorising text into predefined categories. It uses the previously discussed NER methods to structure textual data to make it ready for sentiment analysis and supervised machine learning tasks. It is thoroughly used in the domains of information retrieval, information filtering, sentiment analysis, recommendation systems and in the field of finance law and health.

The first step of text classification is preprocessing. This includes the removal of tokens that are considered stop words such as "the", "of", "and" etc. Furthermore, stemming reduces words to their root form and lemmatization reduces words to their lemma. Stemming is a simpler and faster process than lemmatization, but offers a lower accuracy of classification.

Text classification allows for feature extraction that can be utilised in machine learning models. The most common approach is the Bag of Words model. This model represents text as a matrix of word counts. Another approach is the Term Frequency-Inverse Document Frequency (TF-IDF) model. This model gives higher weights to the words that are more significant in the text in relation to a collection of textual documents. For example a jargon specific word would have more weight and have more focus in the classification process. The formula for TF-iDF is:

$$tfidf_{t,d} = tf_{t,d} \cdot \log_{10}\left(\frac{N}{df_t}\right)$$

Where tf is the relative frequency of t within document d, N the total number of documents and df the document frequency of term t.

After the feature extraction phase, text classification can be used to train machine learning models. The most common models include Naive Bayes, Support Vector Machines and Decision Trees. However, deep learning has recently been more successful in text classification tasks (24).

3.6 Sentiment Analysis

Sentiment Analysis, also known as opinion mining, originated in research as early as 2001. Das & Chen (2001) attempted a research to predict market sentiment from small investors. They found a strong link between market movements and sentiment and achieved 62% accuracy with their classifier (25). From 2001 on wards numerous studies were published with the same phrasing. The term Sentiment Analysis was used to describe the task of classifying reviews into certain sentiment polarity. This started off as a classification as either positive or negative, but later advancements in the literature started deploying statistical methods to classify sentiment numerically. Sentiment Analysis is a specific type of text classification where the goal is to determine the sentiment expressed in text data. It is thoroughly used in finance, specifically in the prediction of stock market prices. Generally speaking there are three levels on which sentiment analysis can be conducted.

On the document level the general sentiment of the text is classified. The goal of this approach is to classify the whole text to either positive, negative or neutral. For example, managerial comments on financial filings can be classified as either positive or negative in terms of financial results. Sentence level sentiment performs classification on individual sentences. Unlike the document level, individual sentences can be classified as a number, often between -1 and 1 to indicate the weight of the sentiment. On the third feature level we not only identify the sentiment in a text but also link this text to specific entities mentioned in the text. NER is especially useful in performing the third level. Moreover, sentiment analysis can either follow a rule-based approach where the sentiment is based on rules or a Machine Learning approach where sentiment is based on statistical models that identify relationships within the text data.

Even though a lot of research has been done in Sentiment Analysis it is not considered a completed problem. A problem that occurs on the document and sentence level is that the entity is not addressed. Therefore, the document level approach is not viable for text that has multiple opinions (26). Similarly to POS Tagging, ambiguity can cause problems for the classification of sentiment. For example, words like “not” and “never” can cause wrong sentiment when used in combination with the word “great”. Ambiguity can also come in the form of sarcasm. This can cause a sentence to turn its sentiment into the opposite of which it actually is.

3.7 Classification

Classification, in the context of machine learning and NLP, is a supervised learning approach where the algorithm learns from the training data, and then uses this learning to classify new observations. Classification involves categorising text into predefined groups. For instance, emails can be classified as "spam" or "not spam", movie reviews can be classified as "positive" or "negative", and so on. The classification task starts with building a model on the training data, and then using this model to classify new data.

This research used three different classification algorithms. First Logistic Regression is used to fit data on a logistic function, which returns the probability that a given input point belongs to a rising stock price or declining stock price. When the probability is above a threshold, the instance is classified into a specific category, otherwise, it’s classified into the other. Second, Random Forest is an ensemble learning method that operates by creating a multitude of decision trees during the training phase. The decision of the majority of the trees is chosen by the random forest as the final decision. Finally, K-Nearest Neighbour is a type of instance-based learning where the function is approximated locally and all computations are deferred until classification. It classifies a new observation based on the K number of training observations nearest to that new observation. It assumes that similar things exist in close proximity. In other words, similar things are near to each other.

4

Research strategies and research methods

4.1 Strategies and Methods

To answer the main research question and the two sub-questions we used a quantitative approach to analyse the text matching the stock prices. Two separate data sets were created. The first data set consists of annual and quarterly reports. These reports are a summary of an organisation's financial performance. They inform potential investors and shareholders about the company's financial status and business activities. These are mandatory reports and are publicly published on national commissions under the filing names 10-K and 10-Q. The second data set contains data from CEO quotes. These CEO quotes can be found in the press releases that come with annual reports, more specifically exhibit 99.1. This exhibit contains additional information, such as CEO quotes and management commentary, that provides insights into the company's current state, performance, or strategic initiatives. Moreover this commentary gives information on various aspects of the company's operations, financial performance, strategic direction, or significant events.

4.1.1 10-K and 10-Q data collection

The first data set that was created consists of the 10-Q and 10-K filings from the top 10 S&P 500 companies by volume, ranging from the year 2015 - 2022. The filings were scraped from the U.S. Securities and Exchange Commission (SEC) government website. The SEC provides an electronic data gathering, analysis and retrieval system (EDGAR), which allows HTTP GET requests. There are paid API's that provide tools to scrape the filings in an efficient manner. However, the filings were scraped with python GET requests in order to keep the cost of the research low. Considering each company has three quarterly reports and one annual report per year, the data quickly rises in size. In this research 309 filings were scraped which exceeded a storage size of 800MB. To complement each company with the relevant stock price, GET requests to Yahoo Finance were made. To adhere to the Center for Research in Security Prices Yahoo Finance automatically adjusts the close price for stock splits and dividend and/or capital gain distributions (27). Finally, the analysis was only focused on the most relevant textual parts of the filings. These parts are Risk Factors (Item 1A), Quantitative and Qualitative Disclosures About Market Risk (Item 7A) and Financial Statements and Supplementary Data (Item 8). It was evaluated that the filings have huge overlap. For example, using a text comparison tool (28), the Risk Factors section of the 10-K filings of 2020 and 2021 of the company Alphabet have 85.2% overlap. The 14.8% distinction comes from the added comments about the Covid-19 pandemic in 2020. The statistics of this overlap instance can be found in Appendix E5. Moreover, the text of these 10-Q and 10-K filings is lawyer-endorsed and shows consistent overlap because of its formal tone. The consistent use of legal language leads to little variation between the content of the filings. Therefore, there is little value in comparing the filings with each other. For these reasons the analysis was primarily focused on the second data set. To show the similarity between the text of different years an example of the first paragraph of the 10-K Risk Factors chapter of Alphabet 2020 can be found in Appendix C3 and an example of the year 2021 can be found in Appendix D4. The highlighted text is the difference in the text between the two years

The second data set consists of the CEO quotes in the annual press releases that come with the annual reports. The press releases considered are the top 50 S&P 500 companies by volume, ranging from the year 2017 - 2022. These filings were also scraped from the SEC website and can be found under Press Release exhibit 99.1. The time span differs from the first data set because the data was significantly harder to find before the year 2017. In a few instances this wasn't the case and the data from the years 2015 and 2016 was also added. An obstacle with the creation of this press release data set in comparison to the filing data set was that the CEO quotes from these press releases had to be manually scraped. Unlike the 10-Q and 10-K filings, the press releases did not have a strict structure, therefore the quotes from the press releases had to be manually scraped instead of with a script. To be precise some companies added more textual context in their Press Releases such as a highlights section from the recent year or recent announcement. To keep the data set consistent only the CEO quotes were taken into account. Furthermore, if there was a quote from the CFO, COO or similar then this quote was also taken into account.

The text below gives an indication of what the CEO quotes in the press releases entail: *"A big thank you to employees across Amazon who overcame another quarter of COVID-related challenges and delivered for customers this holiday season. Given the extraordinary growth we saw in 2020 when customers predominantly stayed home, and the fact that we've continued to grow on top of that in 2021, our Retail teammates have effectively operated in peak mode for almost two years. It's been a tremendous effort, and I'm appreciative and proud of how hard our teams have worked to serve customers," said Andy Jassy, Amazon CEO. "As expected over the holidays, we saw higher costs driven by labor supply shortages and inflationary pressures, and these issues persisted into the first quarter due to Omicron. Despite these short-term challenges, we continue to feel optimistic and excited about the business as we emerge from the pandemic. When you combine how we're staffing and scaling our fulfillment network to bring even faster delivery to more customers, the extraordinary growth of AWS with 40% year-over-year growth (and now a $71 billion revenue run rate), the addition of marquee new entertainment like The Lord of the Rings: The Rings of Power and Thursday Night Football, and a plethora of new capabilities that we're building in areas like Alexa, Ring, Grocery, Pharmacy, Amazon Care, Kuiper, and Zoox, there's a lot to look forward to in the months and years ahead."*

4.2 Preprocessing and construction

The columns that were added to the data set were derived through different methods. This subsection elaborates what each column in the two data sets entail.

Table 4.1: Column explanation

Column	Description
Company	The company to which the row belongs to
Date	The date on which the filing or press release was published
Adj Close	Adjusted Close indicates the stock closing price for that given day.
Volume	The number of stock shares that were traded on this day.
Press release	The CEO quotes that were scraped for a given company and a given date
Price change 12mo	This column presents the adjusted close price change. For example if the stock price in the row is 100 and the price after 12 months is 150, then the 12 month price change for this row will be 50.

Continued on next page

Table 4.1 – continued from previous page

Column	Description
Price change 8mo	Similarly, this price change indicates the same but over a time span of 8 months instead of 12.
Price change 4mo	This price change indicates the change over a time span of 4 months.
Price change pct 12mo	The change in percentage of price change over a 12 month time period
Price change pct 8mo	The change in percentage of price change over a 8 month time period
Price change pct 4mo	The change in percentage of price change over a 4 month time period
Price change sign 12mo	This column contains either a value of -1 for when the price change 12mo is negative or a 1 for when the price change 12mo is positive.
Price change sign 8mo	This column contains either a value of -1 for when the price change 8mo is negative or a 1 for when the price change 8mo is positive
Price change sign 4mo	This column contains either a value of -1 for when the price change 4mo is negative or a 1 for when the price change 4mo is positive.
Compound	The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive). This score has been calculated with the Vader sentiment analysis. Vader is an abbreviation for Valence Aware Dictionary and Sentiment Reasoner. It is a lexicon and rule-based feeling analysis instrument. With a mix of highlighted tokens VADER marks a score to mark each token as either positive and negative with a score between -1 and 1. The next three columns are derived from this.
Negative	A negative score is a score that has a compound score lower than -0.05
Neutral	A neutral score is a score that has a compound score between -0.05 and 0.05
Positive	A positive score is a score that has a compound score high than 0.05
Polarity	The polarity score is a float between 1 and -1. This score indicates a negative sentiment for -1 and a positive sentiment for 1. The library Textblob provides a simple API to conduct polarity and subjectivity scores.
Subjectivity	The subjectivity score is also a float between -1 and 1. According to the Textblob library it refers to the general emotion or opinion.

Continued on next page

Table 4.1 – continued from previous page

Column	Description
Text length	Text length indicates the length of the text that was scraped.
Word count	The word count is the amount of words that occur for the given text. More specifically, it is the text length split on white space.
Word density	The word density is a value that has been calculated by dividing the text length by the word count.
Punctuation count	This column represents the amount of punctuation that occurs per given text in the press release column.
Upper case count	The amount of upper cases that occur
Stop word count	The amount of stop words that occur. The stop words that were taken into account are the stop words that occur in the spaCy library as stop words.
Readability Dale Chall	The readability score is calculated with the Dale Chall readability metric. This metric can be found in the readability library (29). The Dale Chall score is derived from a formula that is based on the use of familiar words, rather than syllable or letter counts. It indicates how easy it is to read a certain text based upon the familiarity of the words that occur in said text
Readability Flesch reading ease	This readability score is based on the Flesch Reading Ease. It is a standard test of readability used for the U.S. Department of Defense and insurance policies.
Noun count	The amount of nouns that occur in the relevant text. The library TextBlob was used to identify each Part of Speech.
Verb count Adj count	The amount of verbs that occur in the relevant text. The amount of adjectives that occur in the relevant text.
Adv count	The amount of adverbs that occur in the relevant text.
Pron count	The amount of pronouns that occur in the relevant text.
Top words	Top words is a list of tuples. Each tuple consists of a word and the amount of occurrences for that word in the press release text
Word calculation 'Continued', 'record', 'cash', 'results', 'customers', 'fiscal', 'billion', 'cloud', 'growth', 'strong', 'fourth', 'performance', 'quarter', 'business', 'sales'	The word calculation is calculated by calculating for each word how likely it is to appear in a text that belongs to a company that has a rising stock price after a year. Then this value per word is multiplied by each other. Every word is a separate column and counts the frequency of occurrences of that specific word in the press release instance.

Named Entity Recognition (NER) can be utilized to exclude words that are irrelevant in certain contexts. For instance, specific references such as company names like "Apple," or years such as "2018," did not contribute meaningful information to our analysis. In our research, we focused on the top 25 most frequently occurring words, because words that appeared less frequently provided insufficient data for reliable interpretation. We manually filtered out irrelevant words, prioritising those that conveyed sentiment. However, in the case of a larger data set, NER could be employed to automatically exclude words with minimal or no sentimental value.

4.3 Data analysis

For the analysis part there was a specific focus on the word frequency that appeared for companies with a rising or declining stock price. Moreover, the analysis investigated if certain words were more common for companies with a rising stock price than a declining stock price and vice versa.

Furthermore, the analysis deployed two classification algorithms, namely, Random Forest and Logistic Regression. The random forest decision tree classifier uses a five cross validation to test the criterion hyper parameter Entropy and Gini. Entropy uses information gain as the criterion for splitting. Information gain is a measure to estimate the reduction in entropy after the split. A high entropy indicates that the data has a high disorder, meaning that the data points are distributed more evenly among the classes. Gini is a hyper parameter that uses impurity as the criterion for splitting. A Gini impurity measures the probability that a randomly chosen element is misclassified. Both the hyper parameter Entropy and Gini were trained on two, three, five and ten leaves. To be clear, a leaf node of 2 indicates that the Random Forest will be limited to having a maximum of 2 leaf nodes, meaning the three can only make one split.

The Logistic Regression model also deployed a cross validation of five. Furthermore, four hyper parameters were tested. The first hyper parameter is Limited-memory Broyden-Fletcher-Goldfarb-Shanno. This is an optimization algorithm that finds the optimal weights, which works well for smaller data sets. The second hyper parameter is the Library for Large Linear Classification. This works well for high-dimensional data sets with a small number of samples. Newton Conjugate Gradient is also an optimization algorithm which uses the gradient to solve the linear system of equations. The final hyper parameter is Stochastic Average Gradient and optimises the model with the stochastic gradient.

The primary metric used to evaluate the two models was an F1-score, which was derived from the corresponding confusion matrix gained from the test set. Furthermore, we examine the random forest model more in-depth to understand which features were critical for the classification process. Likewise, we look at the coefficients of the Logistic Regression to see which features gained the most weight.

Finally, to eliminate any correlation between the stock prices of companies we produced table 1 in the appendix. This table illustrates the number of positive, negative and no changes of the stock price. In this context a positive change is a price change above 5% after 12 months. A negative change is a change below 5% and no change is the boundary between ±5%. Because the majority of companies have a more than zero negative changes it can be excluded that the companies correlate with each other and a prediction purely based on the company name is not possible.

5

Results

The purpose of the analyses was to draw conclusions from the constructed data set, specifically focusing on the textual data extracted from CEO quotes. The following sections detail the obtained results.

5.1 Class imbalance

The press release data set is composed of 248 entries. Among these, 168 instances represent a positive change in the corresponding company's stock price after one year, while the remaining 80 denote a negative shift within the same time frame. 5.1 illustrates the distribution of stock price changes. They are categorised into stock increases, decreases and neutral. In this context neutral is defined as a stock price that has not moved more than $\pm 5\%$. Furthermore, it can be observed that the three time periods have a similar distribution to each other. Logically it follows that the 12 month time period has the lowest amount of neutral instances because it had the longest time for a stock to move out of the $\pm 5\%$ boundary. The 12 month time period also experiences the highest amount of positive stock increase. The analysis was focused on the 12 month time period, because this time period has the lowest neutral occurrences and is therefore the least vulnerable to random fluctuations.

5.2 reveals that there is a high correlation between the price change over a 12-month period and that of 8 and 4 months. Therefore, the analysis over only the 12 month period is sufficient and does not have to be recalculated for other time periods.

Figure 5.1: The distribution of companies with a rising vs declining stock price

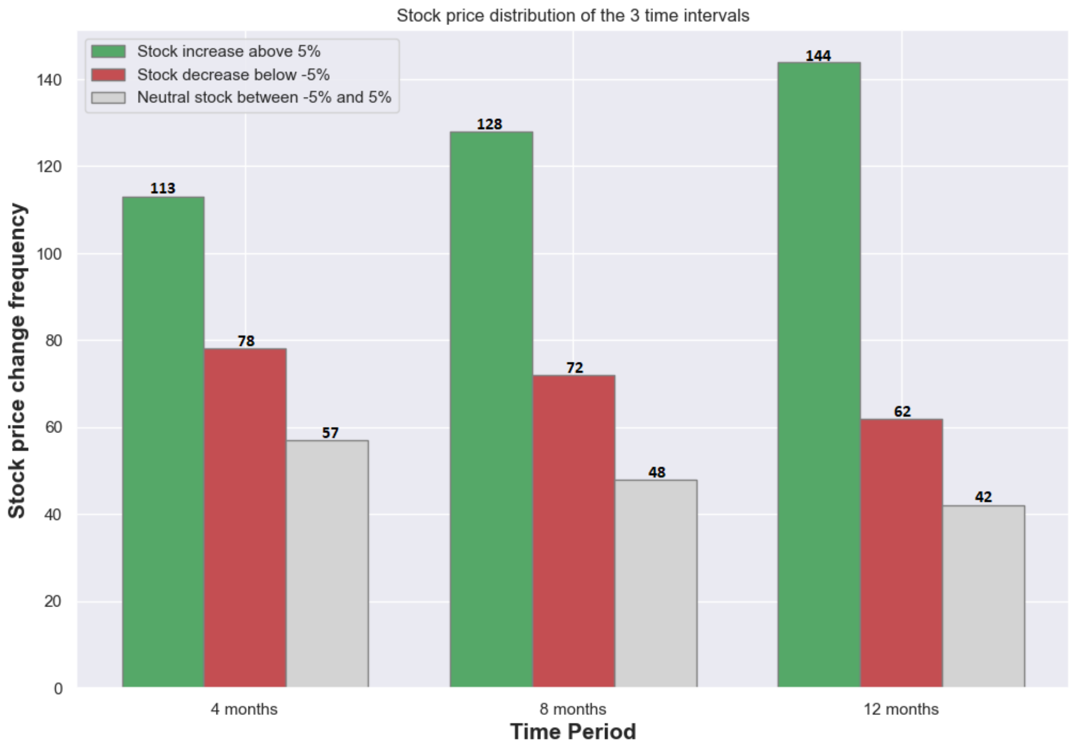
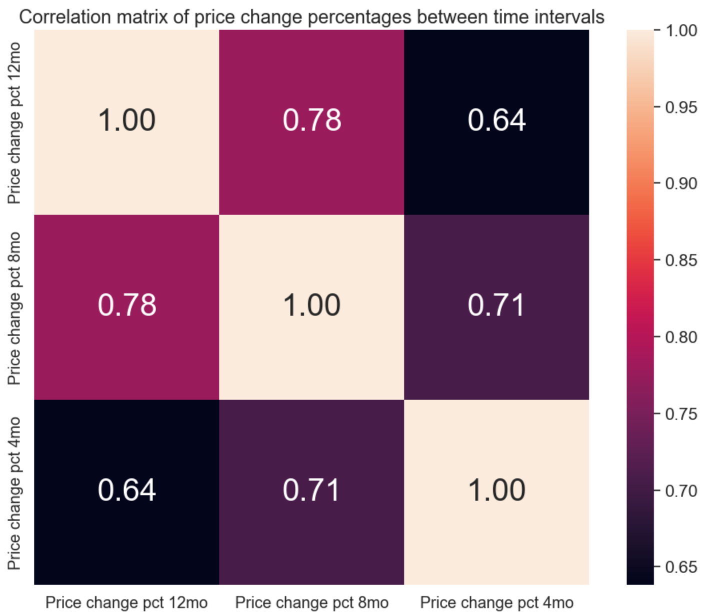


Figure 5.2: The correlation matrix of price changes between time intervals



5.2 Word occurrence

This section focuses on the words that are observed more frequently in the CEO quotes from press releases of companies that experience a rise in stock price over a year, as compared to those where the stock price descends within the same period.

Figure 5.3: Occurrence of word for a positive stock

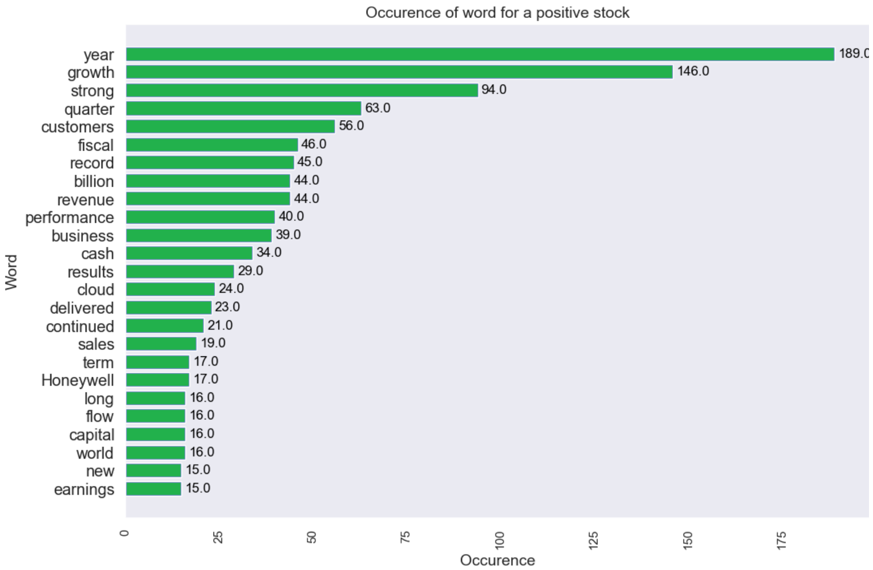
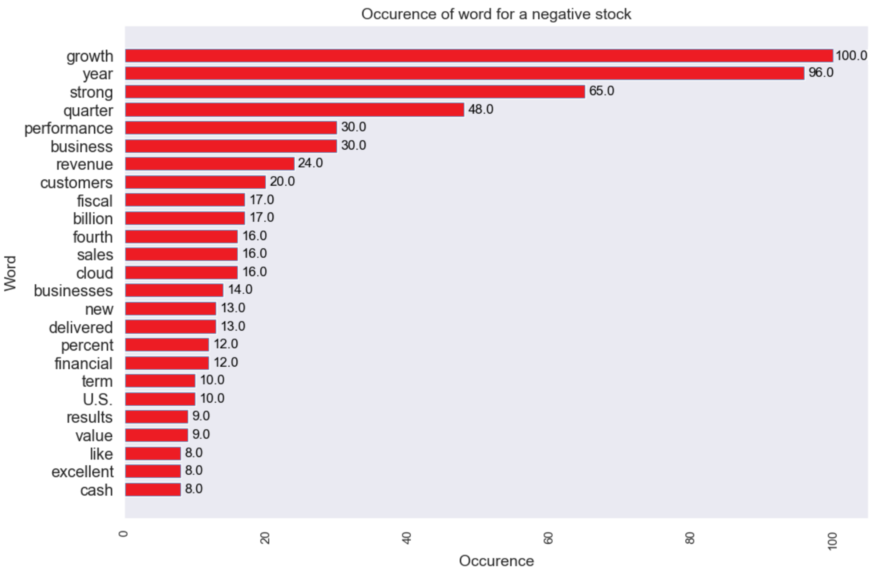
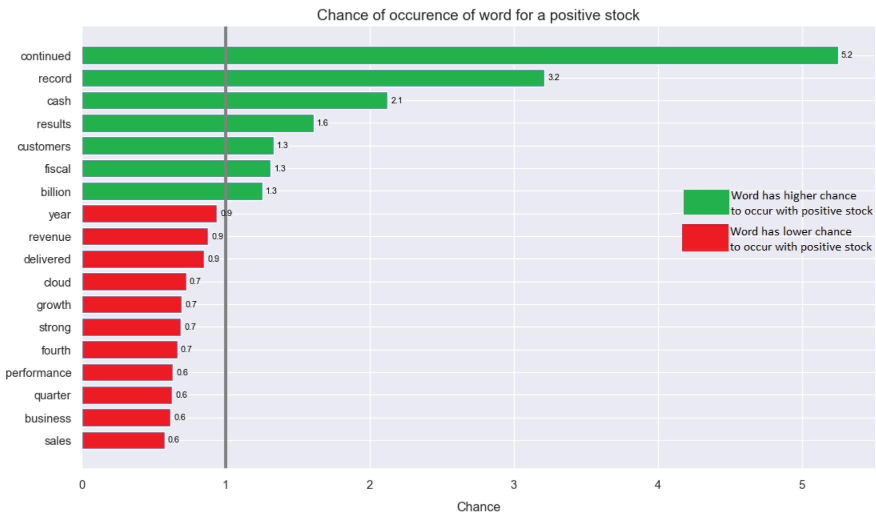


Figure 5.4: Occurrence of word for a negative stock



Figures 5.3 and 5.4 show that certain words occur more often when they are related to a rising stock price and vice versa. For instance, the word 'growth' occurs 46 more times in cases of a rising stock price. However, to ensure a meaningful comparison between the two graphs, it was crucial to account for class imbalance in the word occurrence data. This was achieved by normalising the word frequency data based on the class distribution, as presented in figure 5.5. This normalization reveals that the word 'continued' is five times more likely to appear in texts related to an ascending stock price. On the other hand, the word 'strong' is 0.6 times as likely to be found in text pertaining to a rising stock price, implying it is 1.66 times more frequent in text associated with a declining stock price.

Figure 5.5: Chance of occurrence of word for a positive stock



5.2.1 Feature selection

The following features were selected as independent variables for the machine learning models.

Negative	Neutral	Positive	polarity
subjectivity text	length	word count	word density
punctuation count	upper case word count	stop word count	readability Dale Chall
readability Flesch reading ease	noun count	verb count	adj count
adv count	pron count	continued	record
cash	results	customers	fiscal
billion	cloud	growth	strong
fourth	performance	quarter	business
sales			

Table 5.1: Chosen features for the models

5.2.2 Outcomes of the random forest model

The decision tree model reaches the highest mean F1 score on the training data with the hyper parameter Gini and 10 nodes. Using these hyper parameters on our test set we obtain the following values:

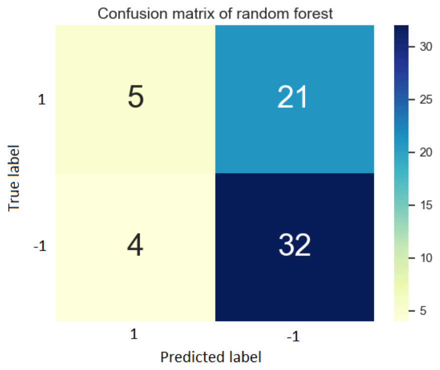
F1-score	Precision	Recall	Accuracy
0.286	0.556	0.192	0.597

Table 5.2: Metric table random forest.

The F1 score highlights that there is room for improvement. A better balance between precision and recall can be found. Considering there is a class imbalance in the data set where 68% of the price changes are positive, an accuracy of .597 does not outperform the baseline and shows insufficient results. These results indicate that the selected features may not be strongly predictive of the price change. The confusion matrix in figure 5.6 represents a lot of false negatives, which causes a low F1 score.

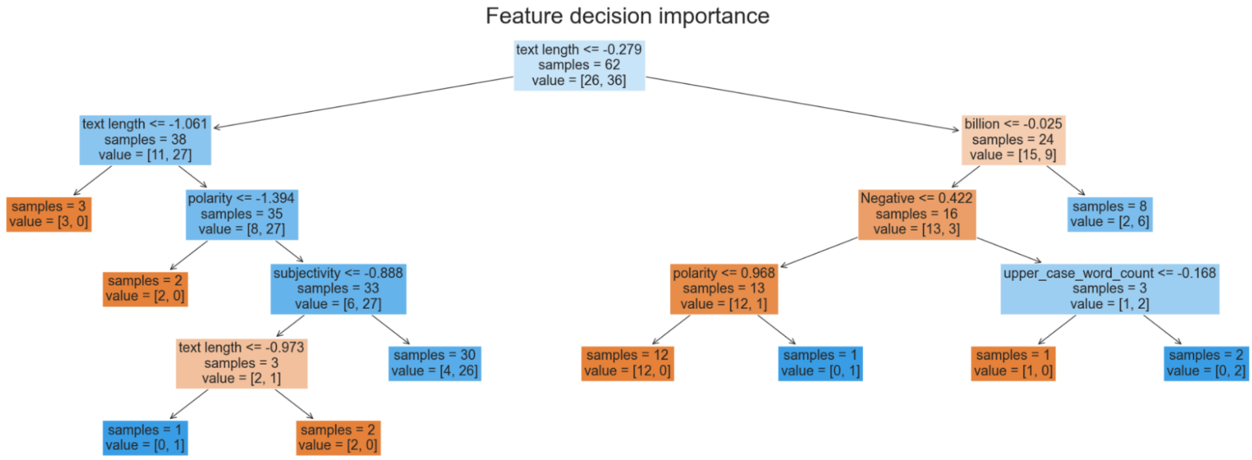
Figure 5.7 explains the feature importance of the decision tree. This model shows that text length is used as the first node. Out of the 62 samples in the test set, 38 have a text length smaller than -0.279. An important observation is that even though we showed

Figure 5.6: Confusion matrix of random forest



evidence that certain words occur more often for certain stock movements in figure 5.5, the use of these word features are only taken into account once in this model. The feature 'billion' occurs in the second depth on the right side, but no other word feature is used by the model.

Figure 5.7: Feature decision importance

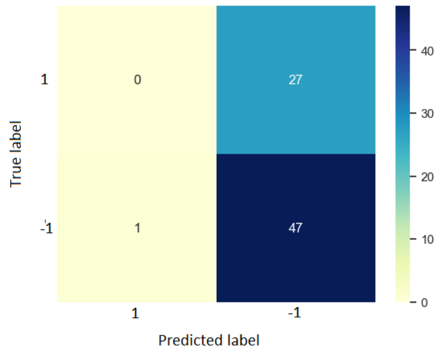


5.2.3 Outcomes of the Linear Regression model

The Logistic Regression model reaches the highest F1 score on the training data with the Library for Large Linear Classification hyper parameter. Deploying this on our test set gives us the confusion matrix in figure 5.8

We observe in this matrix that zero values were predicted to be true positives. Adjusting

Figure 5.8: Confusion matrix of Linear Regression



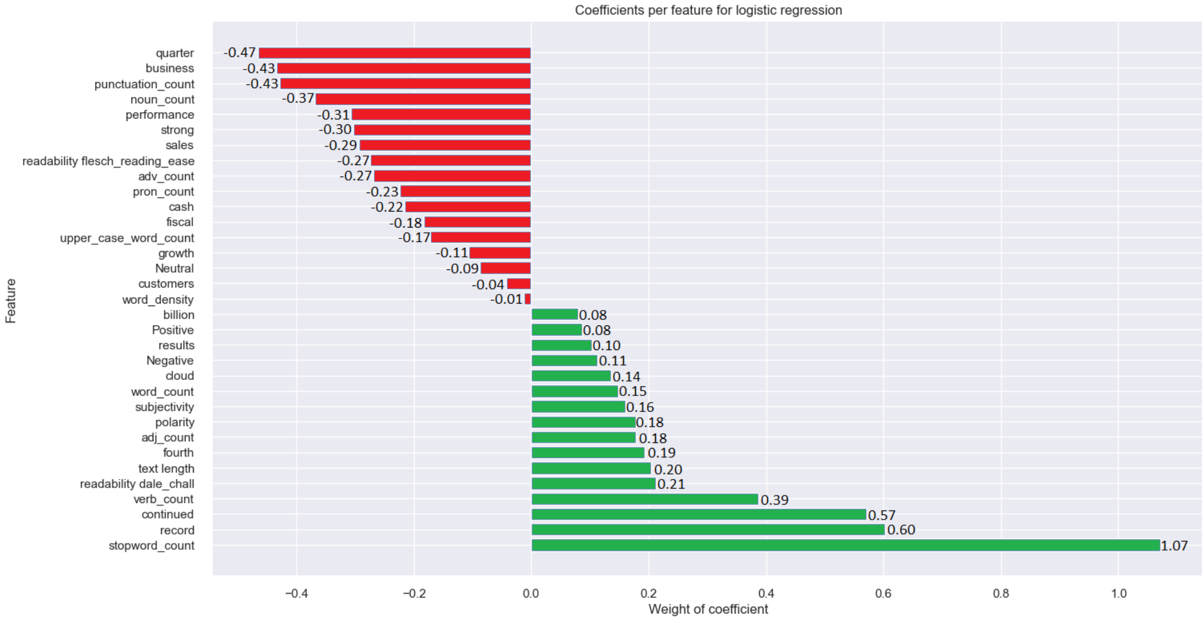
for this zero by adding 1 to it we obtain the following metrics:

F1-score	Precision	Recall	Accuracy
0.067	0.500	0.036	0.632

Table 5.3: Metric table linear regression.

The low F1 score indicates that the model was biased towards one class and does not predict the target label well. Furthermore, the random forest model with an F1-score of 0.286 performed better than the linear regression model. Finally looking at the weights of the coefficients displayed in figure 5.9 we can see that the feature 'stopword_count' had the most importance with the highest weight of 1.07. Moreover, it can be observed that the individual word features do not receive higher weights than the NLP features.

Figure 5.9: Weights of coefficients Linear Regression model



6

Discussion

6.1 Answering research questions

This research aimed to answer several questions relating to the correlation of specific word usage with a company's future performance. Text mining techniques and Natural Language Processing (NLP) were used to create features to extract insights from financial annual reports and CEO quotes from press releases. Our first sub-question, "Does certain word usage correlate with future performance of a company?" found a positive answer. As presented in Figure 5.5, there are noteworthy disparities in word frequencies between texts associated with rising and declining stock prices, indicating a noticeable pattern in word usage. However, our second sub-question, "Can this correlation be used to predict the change in stock market value in the long run?" remained unaddressed to a satisfactory extent. Despite applying two different machine learning algorithms, we found that the derived features couldn't outperform the baseline determined by the class imbalance. This result suggests that while specific word usage patterns exist, they might not directly translate into predictive models for long-term stock price changes. This leads us to the primary research question, "To what extent can Natural Language Processing gather insights from financial annual reports?" Our findings indicate that NLP provides limited value in analyzing 10-Q and 10-K reports due to their formal language and consistent overlap shown in appendix E5. However, CEO quotes within press releases proved more enlightening. Significant differences were noted in word occurrence within quotes relating to companies with rising and declining stock prices, showcasing the potential of text mining and NLP in extracting meaningful insights from certain types of corporate communications. Even though these differences in word occurrence were noted, the random forest and linear regression algorithms were not able to predict future companies performances based on the features created. Moreover, the two models did not seem to prefer the significant occurrence of certain words over the other created features.

Finally, this research was able to create a useful data set which contains CEO quotes, stock prices and features created through text mining and Natural Language Processing. This data set is named "Press release dataset with stock prices and features" and found by following the link in Appendix A1. This data set can serve as a robust starting point for future exploration of data science models for stock price prediction, and it could be further augmented with additional data or refined features to enhance such models.

6.2 Obstacles and limitations

While our research yielded some insightful results, it's important to acknowledge the obstacles and limitations encountered during the study.

An unexpected challenge we encountered was the absence of unique text between the 10-Q and 10-K filings. Year-to-year variations in these filings were primarily due to remarks related to the COVID-19 pandemic. This research utilized an online text comparison tool to reach these conclusions. However, to automatically determine if different texts warrant comparison, spaCy offers a text comparison method (30). Moreover, to decide whether a text belongs to a 10-K filing or a press releases statement a text classification model could be developed to assess whether the texts should be used for comparison. Although this was beyond the scope of our current research, the spaCy library offers a pipeline to build such a model (31).

Additionally, the extraction of CEO quotes from press releases presented a substantial hurdle. These had to be manually scraped from Exhibit 99.1, which was a time-consuming process. This limitation restricted our analysis to the period from 2017 to 2022, as press releases before 2017 were difficult to find, thereby reducing the size of the data set.

6.3 Future Work

Our study reveals an intriguing field where further research could be beneficial. Exploring additional language features, investigating more advanced NLP techniques, or integrating other data sources might improve the predictive capabilities of the models. It might also be worthwhile to examine different corporate communications like earnings call transcripts or social media postings, which may provide different insights compared to 10-Q and 10-K reports and press releases.

In conclusion, while our study has highlighted the complexities of using text mining and NLP for stock market predictions, it also underscores the potential that lies in further investigating this topic. By refining the techniques and expanding the scope of analysis, future research could potentially unlock more of the predictive power inherent in corporate textual communications.

References

[1] NIDA TÜREGÜN. **Text mining in financial information**. *Current analysis on economics & finance*, **1**:18–26, 2019. 1

[2] CRAIG LEWIS AND STEVEN YOUNG. **Fad or future? Automated analysis of financial text and its implications for corporate reporting**. *Accounting and Business Research*, **49**(5):587–615, 2019. 1

[3] DANIEL T LAROSE AND CHANTAL D LAROSE. *Discovering knowledge in data: an introduction to data mining*, **4**. John Wiley & Sons, 2014. 1

[4] RONEN FELDMAN AND IDO DAGAN. **Knowledge Discovery in Textual Databases (KDT).** In *KDD*, **95**, pages 112–117, 1995. 1

[5] MEHDI ALLAHYARI, SEYEDAMIN POURIYEH, MEHDI ASSEFI, SAIED SAFAEI, ELIZABETH D TRIPPE, JUAN B GUTIERREZ, AND KRYS KOCHUT. **A brief survey of text mining: Classification, clustering and extraction techniques**. *arXiv preprint arXiv:1707.02919*, 2017. 1

[6] ANNE KAO AND STEVE R POTEET. *Natural language processing and text mining.* Springer Science & Business Media, 2007. 1

[7] INGRID E FISHER, MARGARET R GARNSEY, AND MARK E HUGHES. **Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research**. *Intelligent Systems in Accounting, Finance and Management*, **23**(3):157–214, 2016. 2

[8] ITAY GOLDSTEIN, CHESTER S SPATT, AND MAO YE. **Big data in finance**. *The Review of Financial Studies*, **34**(7):3213–3225, 2021. 2

[9] KEITH CORTIS, ANDRÉ FREITAS, TOBIAS DAUDERT, MANUELA HUERLIMANN, MANEL ZARROUK, SIEGFRIED HANDSCHUH, AND BRIAN DAVIS. **Semeval-2017**

task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535, 2017. 2

[10] Carlos Pinho, Mara Madaleno, and Helder Santos. **The usefulness of financial analysts' reports: a content analysis**. *International Journal of Management*, **30**(2):631, 2013. 2

[11] **Inline XBRL**. 2

[12] Matthew Butler and Vlado Kešelj. **Financial forecasting using character n-gram analysis and readability scores of annual reports**. In *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*, pages 39–51. Springer, 2009. 4, 7

[13] Paul C Tetlock. **Giving content to investor sentiment: The role of media in the stock market**. *The Journal of finance*, **62**(3):1139–1168, 2007. 4, 7

[14] **FAIR data: what is it and why is it important**. 5

[15] **FAIR Principles**. 5

[16] Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. **Comprehensive review of text-mining applications in finance**. *Financial Innovation*, **6**(1):1–25, 2020. 7

[17] Keke Cai, Scott Spangler, Ying Chen, and Li Zhang. **Leveraging sentiment analysis for topic detection**. *Web Intelligence and Agent Systems: An International Journal*, **8**(3):291–302, 2010. 7

[18] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. **Exploiting topic based twitter sentiment for stock prediction**. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, 2013. 7

[19] Robert P Schumaker, Yulei Zhang, and Chun-Neng Huang. **Sentiment analysis of financial news articles**. In *20th Annual Conference of International Information Management Association*, 2009. 7, 10

[20] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009. 9, 10, 11

[21] Murhaf Fares, Stephan Oepen, and Yi Zhang. **Machine learning for high-quality tokenization replicating variable tokenization schemes**. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I 14*, pages 231–244. Springer, 2013. 10

[22] Rushali Dhumal Deshmukh and Arvind Kiwelekar. **Deep learning techniques for part of speech tagging by natural language processing**. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 76–81. IEEE, 2020. 11

[23] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. **Named entity recognition: fallacies, challenges and opportunities**. *Computer Standards & Interfaces*, **35**(5):482–489, 2013. 12

[24] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. **Deep learning–based text classification: a comprehensive review**. *ACM computing surveys (CSUR)*, **54**(3):1–40, 2021. 13

[25] Sanjiv Ranjan Das and Mike Y Chen. **Yahoo! for Amazon: Sentiment parsing from small talk on the web**. *For Amazon: Sentiment Parsing from Small Talk on the Web (August 5, 2001). EFA*, 2001. 13

[26] Muhammad Taimoor Khan, Mehr Durrani, Armughan Ali, Irum Inayat, Shehzad Khalid, and Kamran Habib Khan. **Sentiment analysis and the complex natural language**. *Complex Adaptive Systems Modeling*, **4**(1):1–19, 2016. 13

[27] **yahoo finance**. 16

[28] **text comparison tool**. 16

[29] **py-readability-metrics**. 19

[30] **spacy text comparison**. 32

[31] **text categorising**. 32

Appendix

A1 Link to the dataset and mini python library

https://github.com/WoutervanZeijl/Press-release-dataset

B2 Company price correlation

Table 1: Company correlation

Company	Nr of positive changes	Nr of negative changes	Nr of no changes
Abbott	3	1	1
Accenture	4	1	0
Adobe	4	1	0
Alphabet	2	1	0
Amazon	2	1	2
AMD	4	1	0
Amgen	4	0	1
Analog	3	1	1
Apple	4	1	1
ATT	1	3	1
Boeing	0	3	2
Booking	3	2	0
Broadcom	4	0	1
Caterpillar	4	1	0
Chevron	3	2	0
Cisco	3	1	1
Cola	3	0	2
Deere	4	1	0
Disney	1	3	0
Electric	2	3	0
Elevance	4	0	1
Exxon	2	2	1
Home	4	0	1

Continued on next page

Table 1 – continued from previous page

Company	Nr of positive changes	Nr of negative changes	Nr of no changes
Honeywell	4	1	0
IBM	2	2	1
Intel	0	2	1
Lockhead	3	2	1
Materials	2	1	2
McDonalds	3	0	2
Meta	2	1	0
Medtronic	2	2	0
Microsoft	5	0	1
Nike	3	0	2
Nvidia	3	2	0
Oracle	3	1	1
Pfize	3	0	2
Philips	1	1	3
Prologis	3	2	0
Salesforce	4	1	0
Service	4	1	0
Starbucks	4	3	1
Stryker	3	0	2
Thermo	3	1	1
Union	4	1	0
United	5	0	0
Ups	3	2	0
Verizon	1	1	3
Visa	5	0	1
Wallmart	1	6	1

C3 Risk Factors alphabet 2020 10-K filing

ITEM 1A. RISK FACTORS Our operations and financial results are subject to various risks and uncertainties, including but not limited to those described below, which could harm our business, reputation, financial condition, and operating results. Risks Specific to our Company We generate a significant portion of our revenues from advertising, and reduced spending by advertisers, a loss of partners, or new and existing technologies that block ads online and/or affect our ability to customize ads could harm our business. We generated over 83% of total revenues from the display of ads online in 2019. Many of our advertisers, companies that distribute our products and services, digital publishers, and content providers can terminate their contracts with us at any time. These partners may not continue to do business with us if we do not create more value (such as increased numbers of users or customers, new sales leads, increased brand awareness, or more effective monetization) than their available alternatives. Changes to our advertising policies and data privacy practices, as well as changes to other companies' advertising policies or practices may affect the advertising that we are able to provide, which could harm our business. In addition, technologies have been developed that make customized ads more difficult or that block the display of ads altogether and some providers of online services have integrated technologies that could potentially impair the availability and functionality of third-party digital advertising. Failing to provide superior value or deliver advertisements effectively and competitively could harm our reputation, financial condition, and operating results. In addition, expenditures by advertisers tend to be cyclical, reflecting overall economic conditions and budgeting and buying patterns. Adverse macroeconomic conditions can also have a material negative effect on the demand for advertising and cause our advertisers to reduce the amounts they spend on advertising, which could harm our financial condition and operating results.

D4 Risk Factors alphabet 2021 10-K filing

ITEM 1A.RISK FACTORS Our operations and financial results are subject to various risks and uncertainties, including but not limited to those described below, which could harm our business, reputation, financial condition, and operating results. Risks Specific to our Company We generate a significant portion of our revenues from advertising, and reduced spending by advertisers, a loss of partners, or new and existing technologies that block ads online and/or affect our ability to customize ads could harm our business. We generated over 80% of total revenues from the display of ads online in 2020. Many of our advertisers, companies that distribute our products and services, digital publishers, and content providers can terminate their contracts with us at any time. These partners may not continue to do business with us if we do not create more value (such as increased numbers of users or customers, new sales leads, increased brand awareness, or more effective monetization) than their available alternatives. Changes to our advertising policies and data privacy practices, as well as changes to other companies' advertising and/or data privacy practices may affect the advertising that we are able to provide, which could harm our business. In addition, technologies have been developed that make customized ads more difficult or that block the display of ads altogether and some providers of online services have integrated technologies that could potentially impair the availability and functionality of third-party digital advertising. Failing to provide superior value or deliver advertisements effectively and competitively could harm our reputation, financial condition, and operating results. In addition, expenditures by advertisers tend to be cyclical, reflecting overall economic conditions and budgeting and buying patterns. Adverse macroeconomic conditions, including COVID-19 and its effects on the global economy (as discussed in greater detail in our COVID-19 risk factor under ral below), have impacted the demand for advertising and resulted in fluctuations in the amounts our advertisers spend on advertising, and could have an adverse impact on such demand and spend, which could harm our financial condition and operating results.

E5 Overlap of the risk factors of 2020 and 2021 10-K filing of Alphabet

Figure 1: Overlap between the Alphabet risk section 2020 and 2021