Twitter Sentiment Analysis

Supporting Auditor Judgment

Alexander van Oostveen

(2564530)

March 30, 2021

Thesis MSc Business Analytics

VU Supervisors:
Rob van der Mei
Sandjai Bhulai

KPMG Supervisor:
Alexander Boer

Management Summary

A large part of a company's financial statements can no longer be explained by physical assets. Instead, nowadays, intangible assets can make up for more than 70% of a company's value. Due to the lack of physical evidence indicating the value of a company's intangibles, auditors have to rely more and more on their professional judgement when valuing such intangible assets. This is especially the case with unidentifiable intangibles, such as goodwill. However, many sources of information are still untapped. One example is the sentiment of Twitter data.

It has been shown that Twitter sentiment can be an early indication of abnormal economic results. Especially in stock market movement this has been heavily researched. However, only little research has been devoted to the relationship between other economic factors and Twitter sentiment. This thesis takes a closer look at the sentiment of Tweets about a company as an additional resource for the identification of goodwill impairment triggers. In doing so, this thesis is the first investigation into the relationship between goodwill impairment and Twitter sentiment.

Problems

This research aims to answer the question: *How can the sentiment of Tweets about a company be used as an additional source of information in assessing whether a company's goodwill is impaired*? In order to answer this question, three sub-problems are tackled:

1. A comparative analysis between three NLP-based sentiment analysis models: the VADER sentiment intensity analyzer, an artificial neural network (ANN) and BERT.
2. A qualitative analysis with the goal of determining the perspective of auditors on the usefulness of Twitter sentiment as an additional resource in identifying goodwill impairment indicators.
3. A qualitative analysis with the goal of discovering what sources of information from Twitter could be most useful for identifying goodwill impairment triggers and how this information could be transferred to auditors who lack programming experience.

Results

The results obtained from the sub-problems tackled by this research are the following:

1. The findings of the comparative analysis imply that BERT significantly outperforms the other models in terms of predictive power. However, the ANN has the best ratio between speed and performance.
2. The findings of the qualitative analysis show that auditors believe that Twitter sentiment could be a useful additional source of information in identifying goodwill impairment triggers.
3. Furthermore, from the interviews it has become clear that, in order for Twitter sentiment to become a viable source of input in financial reporting, auditors must have access to as much detailed information as possible in an easy-to-use environment such as an app or dashboard.

Recommendations

Based on the results presented in this thesis, NLP-based sentiment analysis on social media data is found to be an interesting additional source of information when assessing whether a company's goodwill is impaired and for the general risk assessment around a company. Therefore, it is advised that companies use social media sentiment as an additional source of information. However, this information should be highly detailed in order to be used in financial reporting. Furthermore, since auditors generally lack a programming background, it is advised that this information is presented in an easy-to-use app or dashboard.

Keywords: Goodwill, impairment, NLP, sentiment analysis, BERT, Recurrent Neural Network, VADER.

Contents

1 Introduction

A large part of a company's financial statements can no longer be explained by physical assets. Instead, intangible assets can make up for more than 70% of a company's value nowadays (Duff and Phelps, 2018; Klingbeil, 2010; EFRAG, 2016). Due to the lack of physical evidence indicating the value of a company's intangibles, auditors have to rely more and more on their professional judgement when valuing such intangible assets (Tsai et al., 2012). This is especially the case with unidentifiable intangibles, such as *goodwill*, since unidentifiable intangibles remain hidden until some transaction gives rise to their identification.

According to (among others Higson (1998); Churyk (2005)), goodwill is defined as the difference between the acquisition price of a business and the fair value of its net identifiable assets. Once goodwill has been determined as the result of an acquisition, its value can only change by being impaired. According to Tsai et al. (2016), the valuation of goodwill can be significantly improved with the help of big data analytics and deep learning. With the rise of technologies, more sources of non-financial evidence are becoming available. A question that arises is, what sources of non-financial data can help auditors in goodwill accounting and how can deep learning be used to support auditor's judgement in this process. Therefore, this research intends to take a closer look at the relationship between Twitter sentiment and goodwill impairment. Specifically, this research aims to answer the question: *How can the sentiment of Tweets about a company be used as an additional source of information in assessing whether a company's goodwill is impaired?*

This rise in technology has also caused investing in the stock market to shift from being reserved to financial investors, to a medium available to the public (Kalda et al., 2021; Lyócsa et al., 2021; Eaton et al., 2021; Pagano et al., 2021; van der Beck and Jaunin, 2021). Especially during the COVID-19 pandemic, retail trading activity has soared (van der Beck and Jaunin, 2021). According to Eaton et al. (2021); Pagano et al. (2021); van der Beck and Jaunin (2021), these traders are generally uninformed traders, behaving like noise traders. Despite their relatively small market share, their impact on the cross-sectional variation in stock returns during the second quarter of 2020 has been significant (van der Beck and Jaunin, 2021).

According to Lyócsa et al. (2021), this surge in retail trader activity has been accompanied by a rise in social media activity related to these investments. This trading-related social media activity has resulted in so-called "hype-stocks". These are stocks, such as GameStop, AMC Entertainment Holdings, Blackberry and Nokia, that were subject to a decentralized short squeeze that exploited the short positions of institutional investors. In the stock market, a short squeeze is a quick surge in a company's stock price as the result of an increase in short selling of this company's stock rather than underlying fundamentals. According to Lyócsa et al. (2021), these specific examples were likely initiated by retail investors concentrated around the social media platform Reddit. Another example of social media encouraged trading behavior has been demonstrated by Ante (2021), who shows a significant relation between Elon Musk's cryptocurrency-related Tweets and the corresponding trading volumes. Another example of Elon Musk's Tweets influencing the stock market is shown in the figure below:

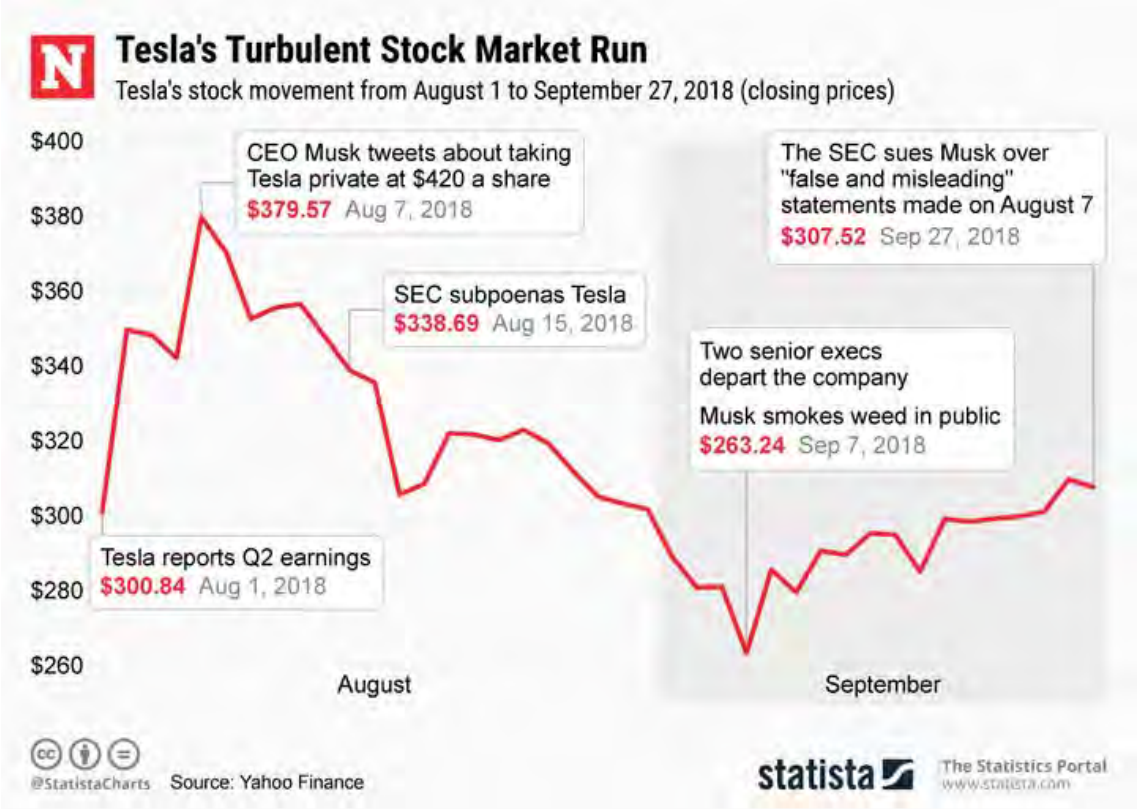

Figure 1: Tesla stock following Tweet from Elon Musk. (Murdock, 2018)

These are some examples of a causational relationship between the stock market and social media platforms such as Reddit and Twitter. Even though this causational relationship is not generally the case, Bollen et al. (2011); Zhang et al. (2011a); Nguyen et al. (2012); Rao and Srivastava (2016); Pagolu et al. (2016) and many others have shown that a significant relationship between stock market movements and Twitter sentiment exists. Bollen et al. (2011) argue that very early indicators can be extracted from Twitter to predict changes in various economic and commercial factors. However, besides the relation between Twitter sentiment and the stock market, not much research has been done into the relationship between social media sentiment and other financial aspects.

Examples of literature that shows a relation between financial aspects and Twitter sentiment can be found in the following works: Gruhl et al. (2005) show that book sales can be predicted by looking at online chat activity. Mishne et al. (2006) predict movie sales with the help of blog sentiment. Albert and Barabási (2002) look at different product sales and is able to predict them using a Probabilistic Latent Semantic Analysis (PLSA) model to extract indicators of sentiment from blogs. Schumaker and Chen (2009) have analyzed the relationship between breaking financial news and stock movement. Furthermore, Google search data has proven to give early indications of buyer behavior and of flue contamination rates (Choi and Varian, 2012). Asur and Huberman (2010) were able to accurately predict box office receipts of premiering movies with the help of the Twitter sentiment about these movies. Recently, Albrecht et al. (2019) have shown that higher search volume, positive sentiment and the increased use of emotive language on Twitter are linked to a high capitalization of block chain startups and their initial coin offerings (ICO).

The financial links between public sentiment and financial indicators have been shown in many fields. In the fields of psychology and behavioral economics, the ground-breaking works by (among others Damasio (1994); Dolan (2002); Kahneman and Tversky (2013)) have shown that emotions, together with information, play an important role in human descision-making. Furthermore, Nofsinger (2005) has provided further proof that financial decisions are significantly driven by emotion and mood.

Most of the predictive analyses that have been performed up until now, based on Twitter sentiment or other sources of public sentiment, have been efforts to predict some kind of economic quantity. However, no research has yet attempted to use Twitter sentiment in relation to financial risk assessment, while this is currently more crucial than ever. The devastating impact on financial markets of the rapid spread of COVID-19 has generated unprecedented levels of risk and made the stock market more volatile than ever (Zhang et al., 2011b; Baker et al., 2020; Ramelli and Wagner, 2020). This makes it crucial for companies to utilize every available resource in the process of performing financial risk assessments. Therefore, this research aims to answer the question: *How can the sentiment of Tweets about a company be used as an additional source of information in assessing whether a company's goodwill is impaired?*

In order to take a closer look at this relationship between the indicators of goodwill impairment and Twitter sentiment, this problem is divided into three sub-questions. First, a comparative analysis is performed, comparing three state-of-the-art natural language processing (NLP) models and their ability to predict Twitter sentiment. These models are the VADER sentiment intensity analyzer, the artificial neural network and the BERT model. Next, a qualitative investigation is performed to answer the second and third sub-question. This qualitative assessment is in the form of interviews with KPMG auditors specialized in the field of goodwill accounting. The first goal of these interviews is to determine auditors' perception of the potential relevance of having Twitter sentiment about a company as an additional source of information assessing whether a company's goodwill is impaired. The third sub-question, answered through these interviews tackles the problem of accessing and assessing data from Twitter. This information is not very easily retrieved, let alone transformed into insights, especially for those without a background in data science, which is often the case with auditors. This final part of the research investigates what pieces of information could be most valuable to auditors in relation to goodwill impairment and a company's general risk assessment, and how this information could be made available in order to make it accessible to auditors.

The structure of this thesis is as follows: Section 2 gives a review of the literature about all topics discussed in this research, together with an extensive theoretical background. Section 3 discusses the methods used in this research. Section 4 gives the results found in this research. Section 5 provides a discussion of the findings from this research and finally, section 6 concludes this research.

2 Literature Review

In this section the necessary theoretical background is provided for the different topics presented throughout this thesis. This background is fully based on academic literature and cited accordingly. The topics brought forward in this thesis can be roughly divided into three sections:

1. Goodwill accounting.
2. The economic impact of Twitter.
3. Deep learning in audit.

Therefore, this section is also divided in such a manner. First, some background information will be given about goodwill accounting together with some of the most important literature about the developments surrounding the impairment of goodwill and related accounting practices. Next, a report of some important literature about the influence of social media on economic factors and the use of deep learning in social media settings is given. Finally, the literature surrounding the use of deep learning in auditing and in combination with social media is described.

2.1 Goodwill accounting

In this section, relevant literature about goodwill accounting and the financial risk assessment with regards to goodwill impairment is provided. First, a background is given about goodwill and related financial regulations and their history. Then, the focus will be shifted towards goodwill impairment and the identification of impairment indicators.

2.1.1 International financial reporting standards

In order to ensure consistent and reliable financial standards across the entire world, international regulations have been put in place to hold entities accountable in this regard. The most commonly followed standards are the *International Financial Reporting Standards (IFRS)*. The IFRS Foundation is a not-for-profit, public interest organisation established to develop a single set of high-quality, understandable, enforceable and globally accepted accounting standards and to promote and facilitate adoption of the standards (IFRS, 2021).

In 1973 the official financial regulatory instances of Australia, Canada, France, Germany, Japan, Mexico, Netherlands, United Kingdom/Ireland and the United States formed the *International Accounting Standards Committee (IASC)* and agreed to adopt *International Accounting Standards (IAS)* for cross-border listings (IFRS, 2021).

In the year 2000, the IASC agreed to restructure itself into a full-time International Accounting Standards Board, overseen by independent trustees. Subsequently, the IFRS Foundation is established, with Paul Volcker appointed Chairman of the Trustees and Sir David Tweedie as Chairman of the Board. In 2001 the IASB holds its first meeting and adopts the IASC Standards (IFRS, 2021). The current chairman of the IASB is Hans Hoogervorst, the Netherlands' former minister of health and minister of finance.

According to the IFRS (2021), currently 144 of the 166 profiled jurisdictions around the world follow the IFRS Standards. These standards are set by the IFRS Foundation's standard-setting body, *the International Accounting Standards Board (IASB)*. Some countries follow other regulations, such as the United States, who follow the *Generally Accepted Accounting Practices (GAAP)*

(Sartore, 2020). In this thesis, however, the focus will be on the standards as they are set by the IASB.

2.1.2 Goodwill

Traditionally, a firm's value was determined by assessing the value of their physical assets, such as land, capital, and labor. Recently, this landscape has been changing. Due to the rise of digital technologies there has been a change in the factors determining whether a company is profitable or not (Tsai et al., 2012). This has also caused a shift in the valuation process of companies. Specifically, many assets that determine a company's value are no longer physical, such as brand name, trademarks or goodwill. These non-physical aspects of a company are called intangible assets.

Two types of intangible assets can be distinguished, *identifiable intangible assets* and *unidentifiable intangible assets*. Identifiable intangibles must be identifiable and separable (Churyk, 2005). Generally, these intangibles fall under intellectual property. Unidentifiable intangibles remain hidden until some transaction gives rise to their identification. A well known unidentifiable intangible asset is *goodwill*.

According to Higson (1998), goodwill is defined as the difference between the acquisition price of a business and the fair value of its net identifiable assets. While this approach is intended to represent the excess value created by a going concern, it is possible that the amount of goodwill recorded reflects an over payment for the acquired firm (Churyk, 2005).

According to Boennen and Glaum (2014), goodwill is intended to capture the expected future economic benefits from intangible assets that are not individually identifiable and therefore cannot be recognized separately in companies' balance sheets. They continue by arguing that goodwill can be created through internal factors or because of *business combinations* when the acquisition price exceeds the fair value of the net identifiable assets. Traditionally, goodwill can only be generated through acquisitions. As a result, internally generated goodwill may not be recognized. This is due to the notion that unidentifiable intangible assets are considered to be too difficult to identify and to measure (Boennen and Glaum, 2014).

2.1.3 Goodwill impairment

According to Vergoossen (2004), the impairment of fixed assets is more and more prevalent in auditing. This is not solely due to poor economic conditions but has mainly been caused by a change in auditing standards. These standards have been developing increasingly towards the practice of *fair value accounting*, instead of *historical* or *nominal value accounting*.

Historical value accounting reports assets and liabilities at the price that was reported when the original transaction took place. On the other hand, fair value accounting reports assets and liabilities at their current market value. Jaijairam (2013) states that, even though both methods of financial reporting have an effect on auditing statements, the fair value method affects the balance sheet the most because it is more volatile. However, the fair value method is believed to be a more accurate representation, since it represents the present-day market value. This, as opposed to the

historical method, which stems from the past value. On top of that, the fair value method provides more transparency and actual financial information about a company.

As a result of this transition towards fair value accounting, the IASB made an important change in the standards of accounting for goodwill (Boennen and Glaum, 2014). Following the the U.S. *Financial Accounting Standards Board (FASB)*, who introduced SFAS 141 "Business Combinations" and SFAS 142 "Goodwill and Other Intangible Assets" in 2001, the IASB introduced IFRS 3 "Business Combinations" and IAS 36 "Impairment of Assets" by the IASB in 2004. These new standards mandate that goodwill is no longer amortized over its expected useful live. Instead it must be tested at least annually for impairment *("impairment-only approach")*. This approach states that, once an acquisition has taken place and the value of goodwill has been determined, this value can only be changed through *goodwill impairment* (IFRS, 2020a). According to Späth and Trampler (2018) the goal of this measure was to provide investors more information about management's investment decisions.

IFRS 3 focuses on the financial reporting standards concerning the acquiring party in a business combination and the way they should report for goodwill. According to IFRS (2020b), the core principles in IFRS 3 are as follows: "An acquirer measures the cost of the acquisition at the fair value of the consideration paid; allocates that cost to the acquired identifiable assets and liabilities on the basis of their fair values; allocates the rest of the cost to goodwill; and recognises any excess of acquired assets and liabilities over the consideration paid (a 'bargain purchase') in profit or loss immediately. The acquirer discloses information that enables users to evaluate the nature and financial effects of the acquisition."

According to (IFRS, 2020a), the core principle in IAS 36 are as follows: "An asset must not be carried in the financial statements at more than the highest amount to be recovered through its use or sale. If the carrying amount exceeds the recoverable amount, the asset is described as impaired. The entity must reduce the carrying amount of the asset to its recoverable amount, and recognise an impairment loss. IAS 36 also applies to groups of assets that do not generate cash flows individually (known as cash-generating units)."

2.1.4 Indications of impairment [IAS 36.12]

According to IFRS (2020a), the recoverable amount of goodwill acquired in business combination must be assessed each year. Here, the recoverable amount is defined as the higher of (a) fair value less costs to sell and (b) value in use. Furthermore, the recoverable amount of all assets must be assessed when there is an indication that the asset may be impaired. Such an indication is also referred to as an *indicator* or a *trigger*. IAS 36.12 contains a list of external and internal impairment triggers.

External sources:

- Market value declines.
- Negative changes in technology, markets, economy, or laws.
- Increases in market interest rates.

- Net assets of the company higher than market capitalisation.

Internal sources:

- Obsolescence or physical damage.
- Asset is idle, part of a restructuring or held for disposal.
- Worse economic performance than expected.
- For investments in subsidiaries, joint ventures or associates, the carrying amount is higher than the carrying amount of the investee's assets, or a dividend exceeds the total comprehensive income of the investee.

IAS 36.13 states that these lists merely form a guideline and are not mean to be exhaustive. Furthermore, IAS 36.17 indicates that an impairment trigger may also indicate that the asset's useful life, depreciation method, or residual value may need to be reviewed and adjusted.

2.2 The economic impact of Twitter

As mentioned in the introduction, Bollen et al. (2011); Zhang et al. (2011a); Nguyen et al. (2012); Rao and Srivastava (2016); Pagolu et al. (2016) and many others have shown that a significant relationship between stock market movements and Twitter sentiment exists. Bollen et al. (2011) even go as far as arguing that very early indicators can be extracted from Twitter to predict changes in various economic and commercial factors. However, the focus in current research has mainly been on the relation between Twitter sentiment and the stock market. In comparison, relatively little research has been done into the relationship between social media sentiment and other financial aspects. Some of the most important literature regarding the economic impact of Twitter is stated below.

Bollen et al. (2011) show that public mood can be an indicator for collective decision-making by predicting stock market movements based on Twitter sentiment. In doing so, they extended the notion from behavioral economics that emotions affect human behavior. In further research, Bollen et al. (2011) perform a sentiment analysis of all Tweets published on Twitter during the second half of 2008. They find that social, political, cultural and economic events have a significant, immediate and highly specific effect on the various dimensions of public mood. They speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators.

Asur and Huberman (2010) examine the relationship between Twitter sentiment about a premiering movie and the resulting box office receipts. As a result, they were able to accurately predict box office revenues of these movies using Twitter sentiment analysis.

Rao and Srivastava (2016) investigate the complex relationship between Tweet board literature (like bullishness, volume, agreement, etc) with the financial market instruments (like volatility, trading volume and stock prices). Their results show high correlation between stock prices and twitter sentiments.

Luo et al. (2013) show that there is a predictive relationship between social media and firm equity value. Interestingly, they find that standard online behavioral metrics such as Google searches and web traffic are found to have a significant yet substantially weaker predictive relationship with firm equity value than social media metrics. They also find that social media has a faster predictive value, i.e., shorter "wear-in" time, than conventional online media.

Zhang et al. (2011b) attempt to predict stock market indicators by using Twitter sentiment. They found that the percentage of emotional Tweets is significantly negatively correlated with Dow Jones, NASDAQ and S&P 500, but displayed significant positive correlation to VIX. They conclude that checking on Twitter for emotional outbursts can serve as a predictor for the stock market returns of the next day.

According to Pagolu et al. (2016), nowadays social media is perfectly representing the public sentiment and opinion about current events. They continue by arguing that especially Twitter has attracted a lot of attention from researchers for studying the public sentiments. In their results, it is shown that a strong correlation exists between the rise and falls in stock prices and the public sentiments in Tweets.

Nisar and Yeung (2018) explore the relationship between politics-related sentiment and FTSE 100 movements by conducting a short-window event study of a UK based political event. Their findings suggest that there is proof of a relationship between the public sentiment and trading volumes. Furthermore, their research even shows a causational relationship between public sentiment and stock market movements. In conclusion, the results from their research are promising regarding the use of Twitter sentiment analysis for forecasting stock movements.

Rather than focusing on the general public's opinion, Hales et al. (2018) examine a public platform designed to convey insider information - Glassdoor.com, where employees voluntarily share their opinions on a number of issues, including the company's near-term business outlook. In particular, the authors of this research find evidence that employee opinions are useful in predicting growth in key income statement information, transitory reporting items (e.g. restructuring charges), earnings surprises, and management forecast news.

Ruan et al. (2018) look at the predictive power of Twitter on abnormal stock returns. They built a trust network among Twitter users as an additional filtering and amplifying mechanism for the Twitter sentiment data, to increase its correlation with stock movements. Their results showed that by using the trust network to weigh Tweets, Twitter sentiment scores are able to enhance their ability to predict abnormal stock returns.

More recently, Albrecht et al. (2019) have shown that higher search volume, positive sentiment and the increased use of emotive language on Twitter are linked to a high capitalization of block chain startups and their initial coin offerings (ICO).

2.3 Deep learning in audit

According to Tsai et al. (2012, 2016), a shortage of standards regarding intangible assets makes it hard for investors and auditors to assess a company's intangible assets. They continue by arguing that machine learning and deep learning models can be used effectively for assessing the value of intangible assets evaluation due to their ability to use large amounts of data to find complex and

non-linear patterns.

Sun and Vasarhelyi (2017) show that deep learning techniques can be applied in financial environments to create additional audit evidence. They argue that these applications have the potential to improve the effectiveness and efficiency of audit decision making and automation.

The potential impact of deep learning applications, such as *natural language processing (NLP)*, visual recognition, and structured data analysis on the audit environment, is illustrated in Sun (2019). The author argues that these applications of deep learning serve two major functions in supporting audit decision-making: *information identification* and *judgment support*.

In Azimi and Agrawal (2021), a state-of-the-art text classification approach from deep learning is proposed, with the goal of more accurately measuring sentiment in 10-Ks. They find that both positive and negative sentiment in these documents predicts abnormal returns and trading volumes around the 10-K filing date and future firm fundamentals and policies. Sun and Vasarhelyi (2018) also insist that textual data exploration and the use of deep learning techniques can provide new resources of auditing evidence. They show different applications and provide a guide for auditors to start implementing novel ways of creating financial reporting evidence with the help of deep learning and NLP.

2.3.1 Natural language processing methods

Deriving the sentiment of Twitter data, is done with the help of *natural language processing (NLP)*. In this thesis, three NLP models are considered for the task of sentiment analysis of Twitter data about companies. The proposed models are the *VADER Sentiment Intensity Analyzer*, an *Artificial Neural Network* and the *BERT* model.

2.3.2 VADER: sentiment intensity analyzer

Hutto and Gilbert (2014) propose a novel, heuristics-based, natural language processing model, designed to cope with the language encountered in social media. The model presented by these authors is the *Valence Aware Dictionary for sEntiment Reasoner (VADER)*. It is based on several heuristics, specific to the inherent way of communication encountered on social media platforms. Along with these heuristics, they expanded the models dictionary to include typical "microblogging" jargon. The model was trained to determine the sentiment of Tweets using a *valence score*, which was developed for this specific model, and was found to outperform human raters in this task. According to Hutto and Gilbert (2014), the valence score is calculated on a scale from [-4], representing the most negative score, to [+4], most positive, with a score of 0 representing a neutral sentiment. The valence score is given to the word that is dealt with by taking into account a certain context such as observations and experiences rather than word-for-word scoring. In the process of calculating the valence score, VADER depends on a dictionary that contains a mapping of numerous lexical features that are customary to Twitter language. Some of these features are listed below:

- A full list of Western-style emoticons (for example - :D and :P);

- Sentiment-related acronyms (for example - LOL and ROFL);
- Commonly used slang with sentiment value (for example - Nah and meh).

Furthermore, as mentioned before, the VADER sentiment intensity analyzer is a heuristics based model. As the result of a qualitative analysis of micro-blogging language behaviour, Hutto and Gilbert (2014) derived five main heuristics for their model. These heuristics capture word-order sensitive relationships between terms and are also specific to the language used in environments such as Twitter. VADER uses these heuristics to determine the valence score of an input sentence and create some form of context awareness. The heuristics use the following indicators:

1. Punctuation
2. Capitalization
3. Degree modifiers
4. Conjunction
5. Negation

Punctuation, such as an exclamation mark, may increases the intensity of a sentence without specifically changing the words contained in this sentence. For example: "The weather is hot!!!" is more intense than "The weather is hot." Capitalization, specifically using ALL-CAPS can have a similar effect.

Degree modifiers are adverbs that have an impact on the subsequent word by either increasing or decreasing the intensity. For example: "The weather is extremely hot." is more intense than "The weather is hot.", whereas "The weather is slightly hot." reduces the intensity.

Polarity shift due to conjunctions, is based on the idea that the conjunctive words such as the word "but" signal a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. For example, "The weather is hot, but it is bearable." has mixed sentiment, with the latter half dictating the overall rating.

Finally, the negation heuristic looks at polarity shifts caused by negation. An example of a negated sentence would be "The weather is not really that hot." According to Hutto and Gilbert (2014), by examining the 3 words preceding high sentiment scoring words, almost 90% of the cases where negation would flip the polarity of the text are identified.

2.3.3 Artificial neural network

One of the most famous forms of deep learning is the *artificial neural network (ANN)*. The mathematical modeling of Artificial Neural Networks (ANN) originated as a result of a series of scientific breakthroughs. The main idea was that the brain's structure and function could be altered, even throughout adulthood. This phenomenon is called *neuroplasticity*. Until then, the structure and functioning of our nervous system was believed to become fixed as one reached adulthood.

This new wave of thinking was brought about by great scientists in the field of psychology and neuroscience. As early as 1890, James (1890) used the term plasticity in combination with (adult) human behavior, describing its' ability to change over time. Not long thereafter, Ramón y Cajal (1894) described the inner workings of a neuron. However, it was not until 1948 that Konorski (1948) coined the term *neuro plasticity*, in a continuation of the famous work of Pavlov and Gantt (1928). In their research, they showed that through training a neural network had the ability to change. These were some of the pioneers of our modern way of thinking about the behavior and physiology of a human neural network.

As first described by Ramón y Cajal (1894), and fine-tuned over the years, a neural network roughly works as follows: The human nervous system consists of a system of neurons, called a neural network. Each neuron consists of three parts; the *soma*, the *axon* and the *dendritic tree*. Impulses from other neurons arrive through the dendritic tree, where they are transmitted to the soma and on to the axon. The transfer from the soma to the axon is of particular interest for the case of artificial neural networks. This is where the complex processing takes place, deciding whether a signal is transmitted on to a next neuron. If the total excitation, brought about by the incoming impulse, exceeds a certain threshold, an output signal is emitted. This output signal is then propagated along the axon and its branches to other neurons, through a junction, referred to as a *synapse* (among others, Ramón y Cajal (1894); Abbott and Kepler (1990); Gerstner (1998); Albert and Barabási (2002)). The neuron sending the impulse is commonly referred to as the *presynaptic* neuron and the receiving neuron as the *postsynaptic* neuron.

Hebb (1949) claimed neuroplasticity to be the process of enhanced synaptic potency as a result of presynaptic neuron's recurrent and continuing stimulation of a postsynaptic neuron. The working paper of the above-mentioned book was translated into a mathematical model by McCulloch and Pitts (1943). This paper is widely considered to be the origin of modern-day artificial neural networks and inspired Rosenblatt (1958)'s *perceptron* model. The mathematical logic derived from neural network is also known as *threshold logic*, where a certain threshold must be surpassed in order for a signal to be propagated to a following layer. This threshold is often defined by an *activation function*. In order to explain the way this works, consider the following node:

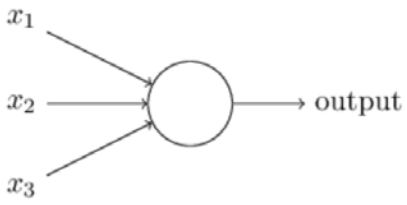

Figure 2: Example of a node in a neural network (Nielsen, 2015).

The node in our example takes three binary inputs, say x_1, x_2 and x_3, and one binary output. Rosenblatt (1958) proposed a simple way to evaluate the output of such a perceptron node. He introduced weights, $w_1, w_2, \ldots$, real numbers expressing the importance of the respective inputs to the output. The neuron's output, 0 or 1, is determined by whether the sum of these weights, $\sum_j w_j x_j$, is larger or smaller than some threshold value. Just like the weights, the threshold is a real number which is a parameter of the neuron. In mathematical terms this gives:

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

The problem with networks containing perceptrons, however, is that a small change in weights or bias can completely flip the output of a model when it is exactly enough to change the output of the perceptron from 0 to 1. This is not desired. In order to optimally learn from data, small changes in weights should have related effects on the output. Therefore, instead of perceptron nodes, often one of the following nodes are used:

- Sigmoid
- Softmax
- ReLU

Sigmoid neurons are similar to perceptrons, but modified so that small changes in their weights and bias cause only a small change in their output. That's the crucial fact which will allow a network of sigmoid neurons to learn. The formula for for assessing the ouput sigmoid node is the following:

$$\sigma(z) \equiv \frac{1}{1+e^{-z}}$$

To put it all more explicitly, the output of a sigmoid neuron with inputs $x_1, x_2, \ldots$, weights $w_1, w_2, \ldots$, and bias b is

$$\frac{1}{1+\exp\left(-\sum_j w_j x_j - b\right)}$$

The softmax activation function is an alteration of the sigmoid function and looks as follows:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

This function is more favorable when considering multiclass classification, as opposed to the sigmoid function which works well with binary classification. Finally, the rectified linear unit or ReLU function is a linearized version of the sigmoid function. This function is very simple;

$$y(x) = max(0, x)$$

This feature makes it quite desirable due to lower computational cost. Together with some other desirable features, such as sparse activation and fast convergence make this one of the most popular activation functions.

The general structure of an artifical neural network consists of three types of layers. An *input layer*, one or more *hidden layers* and an *output layer*. The input layer contains the models input, for

example text in the case of an NLP model. The output layer contains a prediction, so for example the label "Positive", "Neutral" or "Negative" in the case of sentiment analysis. The hidden layer(s) are not different from other layers, they are just not on the outside of the model, but "hidden" in between the input and output layers. Each of these layers takes input from previous nodes and passes a transforms this input based on its' activation function and then passes this on to the next node. An example of the layout of a simple neural network architecture can be seen below:

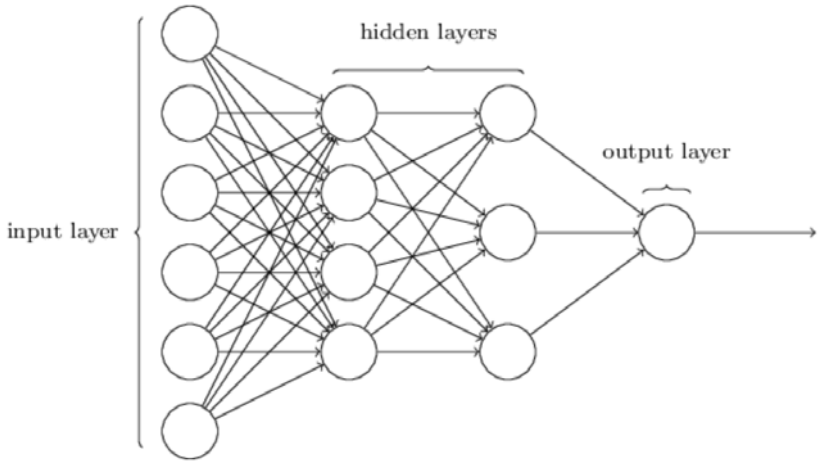

Figure 3: Example of a neural network (Nielsen, 2015).

Models where output only travels in the direction of the output are called *feedforward neural networks*. More complex neural networks also feed output back into previous layers, causing a node's input to depend on their own output. This makes it possible to update weights into even more depth. These method of feeding output back to previous layers is known as *back-propagation*.

With the introduction of back-propagation, the process of distributing the error term back up through the layers of a neural network by modifying the weights at each node (Werbos, 1974), and *connectionism*, the simultaneous consideration of multiple pieces of information (Rumelhart et al., 1986), artificial neural networks significantly improved in terms of explanatory power. Later on, with the rise of computing power, deep neural networks such as *recurrent neural networks (RNN)* and *convolutional neural networks (CNN)* were introduced. These models have significantly enhanced performance in the fields of speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics (LeCun et al., 2015). Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech (Socher et al., 2011, 2013).

A popular model for so-called "sequence prediction" tasks is the *Long Short Term Memory (LSTM)* network introduced by Hochreiter and Schmidhuber (1997). This model has been shown to outperform conventional feed-forward models and RNN's (Sundermeyer et al., 2012). This is due to its' ability of being able to remember patterns for multiple iterations.

RNN's have one major drawback and that is their short term memory. This is due to the *vanishing gradient problem*. This is a phenomenon that occurs as data travels through the nodes in a neural network. As the weights are updated on each node, this weight is transformed into

a multiple of the learning rate, the error term from the previous layer and the input to that layer. Thus, the error term contains the product of all previous layers' errors. With an activation function such as sigmoid, this error which is determined using the gradient of the activation function, the small values of its derivatives gets multiplied multiple times as we move towards the starting layers. As a result of this, the gradient almost vanishes as we move towards the starting layers, and it becomes difficult to train these layers.

In LSTMs, information is kept in cell states. This way, LSTMs can selectively remember or forget things. The information at a particular cell state has three different dependencies:

- The previous cell state
- The previous hidden state
- The input at the current time step

The previous cell state contains the information that was present in the memory after the previous time step. The previous hidden state contains the information of the previous cell. An the input at the current time step contains the information about the new information that is being fed in at that moment. These dependencies to different cell states gives the LSTM the power of being able to remember data structures and makes this model strong in terms of predictive power.

In the famous paper by Mikolov et al. (2013) the word embedding model word2vec was introduced. Word embedding is basically the function of assigning a context-based value to a word. Word2vec was a revolutionary in comparison to other word embedding models. It consists of two shallow neural networks which map words to the target variable. These models are the *Continuous bag of words (CBOW)* model and the *Skip-gram* model.

The CBOW predicts the probability of a word given a context. This context can either be single word or multiple words. The Skip-gram model does exactly the opposite. It predicts the context based on a word. The combination of these two models have made it possible for the word2vec word embedding model to significantly outperform other word embedding models.

2.3.4 BERT

In the revolutionary paper Vaswani et al. (2017), the Google team proposes a completely new way of text classification. The best performing models up until then were based on sequence models that implement complex recurrent or convolutional neural networks, including encoder-decoder architecture. The best performing models also include an attention mechanism in the encoder and the decoder stacks. In contrast, the *Transformer* model is based completely on this attention mechanism, eliminating recurrence and convolutions entirely. The original transformer architecture is a stack of 6 encoder-decoder networks that uses self-attention on the encoder side and attention on the decoder side. The authors show that the model significantly outperforms other current models in the task of language translation.

The basic example from Alammar (2018) can help with understanding the encoder-decoder architecture used in transformer models. In this example, the task at hand is to translate a French sentence to English, see below:

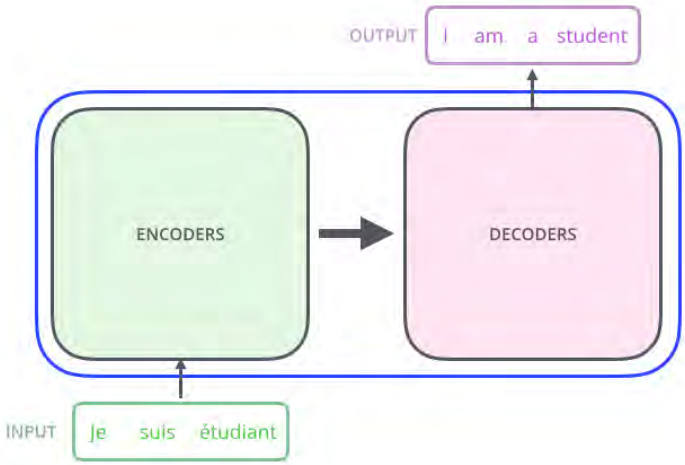

Figure 4: Example of encoder-decoder architecture (Alammar, 2018).

By zooming into the encoder-decoder structure, the following general architecture can be found:

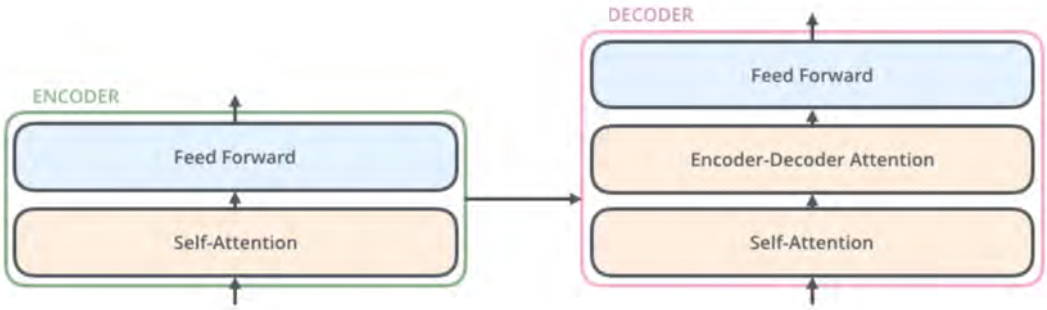

Figure 5: Example of encoder-decoder architecture with some basic information about the inner workings (Alammar, 2018).

Here we see two aspects which are key to the encoder part of a transformer model. The first is the self-attention mechanism. This is a layer that makes it possible for the encoder to look at other words in the input sentence as it encodes a specific word. The output of this layer is fed into a feed-forward neural network. Each encoder layer in a transformer model has the exact same feed-forward neural network structure. Then, the decoder part of the transformer has both layers, but between them is an attention layer that helps the decoder focus on relevant parts of the input sentence.

Devlin et al. (2018) have implemented the encoder part of the Transformer in their novel language representation model *BERT*, which stands for *Bidirectional Encoder Representations from Transformers*. BERT is trained with the entire Wikipedia corpus and an additional 10000+ books from BookCorpus, an online library, totalling to almost 3.5 billion words of training data. This already accounts for a large part of its performance.

In contrast to previous language representation models, which are trained to read textual input in a certain direction, BERT is trained to read textual input bidirectionally. The authors argue that, as a result, the pretrained BERT model can be fine-tuned, using only one supplementary output layer to create state-of-the-art models for a wide range of NLP related tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT has shown impressive results in a large number of NLP related general scoring metrics, outperforming its peers significantly.

So in fact, BERT is just a stack of encoders. BERT$_{BASE}$ consists of 12 layers of encoders while BERT$_{LARGE}$ has 24 layers of encoders. These are more than the Transformer architecture described in the original paper, which consists of 6 encoder layers. The base and large BERT architectures are also built with larger feedforward-networks, consisting of 768 and 1024 hidden units respectively, and more attention heads, 12 and 16 respectively, than the Transformer architecture suggested in the original paper. This model contained 512 hidden units and 8 attention heads. BERT$_{BASE}$ contains 110M parameters while BERT$_{LARGE}$ has 340M parameters.

The bidirectional nature of BERT has posed some novel challenges in terms of training. Classical language understanding training strategies were all devised for directional models. Therefore, the training of this model required new strategies which ensure that the model considers an entire sentence in one time, rather than making directional assumptions. This was done through two main strategies:

1. Masked Language Modelling (MLM)
2. Next Sentence Prediction (NSP)

The idea behind MLM is that a certain percentage of the words fed into the model as training input were replaced by a [MASK] token. However, Devlin et al. (2018) also included some specifications to this model in order to further improve this technique.

The process of replacing words by [MASK] tokens was done by randomly masking 15% of the words, in order to prevent the model from focusing too much of on a particular masked position or token. Furthermore, the masked words were not always replaced by [MASK] since these tokens never appear during the fine-tuning phase. Therefore, Devlin et al. (2018) used the technique listed below:

- 80% of the time the words were replaced with the masked token [MASK]
- 10% of the time the words were replaced with random words
- 10% of the time the words were left unchanged

While in MLM BERT is taught to recognize the relationships between words, in NSP BERT is trained for tasks that require an understanding of the relationship between sentences. Since this is a binary classification task, the data can be easily generated from any corpus by splitting it into sentence pairs. In 50% of the cases, a pair will consist of a sentence and the corresponding next sentence. However, in the other 50% of the cases, the second sentence in a pair will be a

random sentence. This is roughly how BERT is trained. This combination of MLM and NSP is why it is outperforming all of its' peers in the entire range of NLP tasks. This is shown in BERT's performance in the General Language Understanding Evaluation (GLUE) benchmarking test.

3 Methods

In this section, the methods used to perform this research are discussed. As mentioned in the introduction, the goal of this research is to identify how the sentiment of Tweets about a company can be used as an additional source of information in assessing whether a company's goodwill is impaired. In order to address this question properly, it has been divided into three sub-questions. The first part of this question is the sentiment analysis of Twitter data. For this part, a comparative analysis is performed between several NLP models that were discussed in depth in section 2; the VADER Sentiment Intensity Analyzer, a pretrained heuristics-based model developed specifically for Twitter data, the Artificial Neural Network, and BERT. For the second and third sub-questions a qualitative analysis is performed in the form of interviews with auditors who are specialized in goodwill accounting. In the second sub-question, the goal is to identify how auditors perceive the potential relevance of having Twitter sentiment about a company as an additional source of information in relation to the process of identifying whether a company's goodwill is impaired. The final sub-question is meant to assess what information from Twitter could be useful and how this information should be transferred in order to make it easily accessible for financial auditors who lack programming experience.

In order to properly display the different methods used to answer these problems, this section has been divided into six parts, describing the main components of this research. The first part of this section briefly mentions all packages used in this research and their particular relevance. The second part is a detailed description of the methods used to scrape data from Twitter. The third part of this section shows the preprocessing steps undertaken to prepare the scraped Tweets for modelling. The fourth part of this section is a report of the three models that were developed for this research and their specifications. In this segment, the data used for training, testing and validation are discussed as well as the evaluation metrics. In the fifth part of this section, the process of building a dashboard to display the information from Twitter data is described. Finally, the last part of this section discusses the interviews. In particular, it touches on the structure of these interviews and how they help answer the sub-questions and eventually, the main research question.

3.1 Packages

In this section, a short description is given of the different packages used for the modelling part of this research. All programming in this research was performed using Python. As a compiler, Jupyter Notebook was used. Python contains many libraries or packages that can be very useful for data analysis. A list of the main packages that were used and how these were implemented is given below.

First of all, for the data manipulation, reading and some of the analysis, the packages *pandas* and *NumPy* were used. Next, for scraping Tweets from Twitter, the package *Tweepy* was used.

This library allows users easy access to the Twitter API. However, to access the Twitter API, a user needs obtain Twitter credentials. In order to get these credentials, one needs to apply for a Twitter developer account. Once this application is accepted and the necessary credentials have been obtained, Tweepy can be used to access the Twitter API. In the free version of the Twitter API it is possible to fetch Tweets from up to 7 days in the past, based on usernames, hashtags, words contained in a Tweet and some other options. The Twitter Enterprise API has access to the full archive of Twitter, making it possible to look up historical Tweets from any date in the Twitter database and enables researchers to easily access historic data for validation purposes.

Next, since the raw returns from the Twitter API, and Tweets in general, contain a lot of noise, several libraries were used for preprocessing the Tweets. The most important package from this collection is the *Natural Language Toolkit (NLTK)* package. NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities. It also contains useful dictionaries such as stopwords. The VADER Sentiment Intensity Analyzer model is also contained in the NLTK library, including the VADER lexicon, which contains frequently used "Twitter language", such as western style emojis, i.e. ":-)", and slang.

Then, in the process of understanding and analyzing the data, several visualization tools were used. Some examples are; Matplotlib, Seaborn, WordCloud and Plotly. However, these visualizations were not only used for the data understanding part of this research but also for transferring the information that was gained from Twitter data to auditors. For this part of the research, a dashboard was built using Plotly's Dash app-building framework.

Finally, the deep learning models were built using scikit-learn, TensorFlow, Keras and KTrain. Scikit-learn was mainly used for preprocessing, model evaluation and cross validation. The Keras framework runs on top of the open source library TensorFlow and was used to build the Artificial Neural Network. Finally, the KTrain library was used to load and train the BERT model.

3.2 Twitter Scraping

In this section, a more detailed description is given of the process of Twitter scraping. Where in the previous section the relevant packages were discussed, this section will deal with a more in-depth analysis of how the search was performed, what filtering techniques were used to eliminate noise and the preprocessing steps that were undertaken in order to arrive at the final dataset which could be used for the modelling and visualizations in the dashboard. This section will contain a description of the possibilities available in the free version of the Twitter API, which was used for this research, and for the Twitter Enterprise API, which has significantly more add-ons.

As mentioned earlier, the goal of this research is to identify how the sentiment of Tweets about a company can be used as an additional source of information in assessing whether a company's goodwill is impaired. In order to achieve this goal it is necessary to extract data from Twitter about a company. The first step in tackling this problem was discussed in the previous section. This step consists of getting the credentials for a Twitter developer account and using these credentials together with Tweepy for calling the Twitter API.

Once these steps have been taken, the next challenge is to get the right data from Twitter. Since the goal is to determine the sentiment of Tweets about a target company, it is important to get Tweets that are actually about this company. Generally, when a search for a specific subject or

target is performed, the results contain a lot of noise. Therefore, a search needs to be fine-tuned in order to enhance the results. There are a couple of main aspects of a search which can be filtered. Below, these aspects are mentioned. Note that some of these options are only available in the Twitter Enterprise API.

1. Contents
2. Accounts
3. Attributes
4. Geo-localization
5. Date

In the contents aspect one can specify the content of a Tweet by searching for specific keywords, an exact phrase, hashtags, mentions, a url, and some more functions. The accounts aspect makes it possible to specify the user from-, or to which the Tweet or retweet was sent. The attributes aspect is only available for enterprise API and contains metrics about a Tweet, such as the number of retweets, likes and views. These could be used to measure the importance of a Tweet. The geo-localization aspect makes it possible to search for Tweets within a certain longitude, lattitude and radius. For the enterprise API, it is also possible to filter on country, region, locality and some other interesting location related information. Finally, it is possible to search for Tweets between certain dates. Note, however, that the standard API has a historical limit of 7 days. Therefore, searches with dates outside of the 7 day limit will come up empty.

The filtering of the searches in this research was mainly based on adjusting the contents and the account aspects of a search. For the contents, the specifications were that a Tweet must be in English, contain the name of a company and some specifics about the number of Tweets returned per loop, to avoid getting timed out of the Twitter API server. In some cases this still leads to a lot of unwanted Tweets. Therefore, company specific alterations were made, to return more accurate results per company. For the account, the focus was on retweets. Since Tweets are often retweeted multiple times, one search can contain a lot of identical Tweets. Even though this information can be useful when looking at the importance of a Tweet, it does not provide additional information about a company and can blur the results in terms of unique sources of information. Therefore, the retweets were removed from the search.

The lack of options in the standard version of the Twitter API has made it fairly hard to properly filter out much of the noise. The Twitter Enterprise API gives room for many filtering techniques and since it comes with full archive access, it does not cause a filtered search to come up practically empty, which is often the case when putting too many filters on a standard Twitter API search.

3.3 Preprocessing

Now that Twitter has been scraped, the returned Tweets must be prepared for modelling. This procedure is called preprocessing. In order to determine the sentiment of a body of text, a model must be provided with as little noise as possible. In order to do this, much of the contents of

a Tweet have to be removed. The specific preprocessing steps differ per model, but the general methods applied will be discussed here.

First of all, each Tweet is transformed to lowercase and stripped of url's, RT's, punctuation marks and numbers. Next, we replace elongated words by their normal version, i.e. "Hellooooo" becomes "Hello". Then, stopwords are removed. Stopwords are words such as "the", "he", "have" etc. These words do not generally add much meaning to a sentence. Furthermore, we perform stemming. This is the practice of reducing each word to their stem.

The next step is handling negation. An example of negation is; "His grades aren't (are not) good". This sentence is clearly negative, but due to the word "good", a simple model could score the overall sentence as positive. Handling negation, searches for the word "not" and / or a word containing "n't" and replaces the following word in the sentence by its antonym. So in our example this becomes; "His grades are bad".

Finally, once the Tweets are cleaned, they are also tokenized. This is the process of dividing a body of input text into words. These tokens help create a context-based understanding for the NLP model. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

3.4 Models

In this section, the different models presented in this research are discussed. In section 2 a detailed description of the different models is provided. In this section, the specifications per model are given. This includes some special preprocessing steps, implementation methods and the general structure and hyper parameters chosen for each model. Then, the training, testing and validation datasets are discussed. Finally, the choice of evaluation metrics are given.

3.4.1 VADER

As described in the literature review, the VADER sentiment intensity analyzer model is a heuristics-based sentiment analysis model. It is pretrained using a Twitter-based lexicon. The implementation of this model is fairly straightforward. It can be loaded straight from the NLTK library and requires no further training. Therefore, VADER consumes fewer resources as compared to the ANN and BERT, since there is no need for large amounts of training data. Additionally, VADER does not suffer from a speed-performance trade-off.

3.4.2 Artificial neural network

For this research, the second model considered is an artificial neural network. As described in section 2, the artificial neural network's design allows it to identify complex non-linear relationships in data. Therefore, this model is perfectly suited for the task of NLP-based sentiment analysis. However, within the domain of artificial neural networks there are many different types of architectures. Depending on the design, the functionalities of this model, and its' ability to understand

the structure of the underlying data, change.

The model developed in this research was built using Keras' sequential model. A Sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.

In our case, the neural network consists of four layers. These layers are noted below:

- Embedding layer
- Dropout layer
- LSTM layer
- Dense layer

The embedding layer was created using a word2vec model. As explained in section 2, this layer makes it possible for the model to derive the value of words based their context. The dropout layer drops out 50% of the data randomly as to avoid overfitting. The LSTM layer, as explained in section 2, generally performs well in sequential tasks due to its' ability to remember data structures through after iterations. The final layer is a dense layer. A dense layer is a layer where all nodes are connected to all nodes from the previous layer. This dense layer's activation function is the sigmoid function, since it appropriate for performing binary classification.

3.4.3 BERT

Finally we have the BERT model. As explained in section 2, this model is based on the transformer models, introduced by a team from Google in Vaswani et al. (2017). This model roughly consists of two components, the encoder and the decoder. While the encoder focuses on language representation, the decoder is a task-specific add-on which does the desired prediction. The BERT model, introduced by Devlin et al. (2018), only consists of the encoder part. This part is trained with the entire Wikipedia corpus and an additional 10000+ books from an online library source. Needless to say, that the language understanding part of this model is state-of-the-art.

In contrast to previous language representation models, which are trained to read textual input in a certain direction, BERT is trained to read textual input bidirectional. Devlin et al. (2018) argue that, as a result, the pretrained BERT model can be fine-tuned, using only one supplementary output layer to create state-of-the-art models for a wide range of NLP related tasks, such as question answering and language inference, without substantial task-specific architecture modifications. In this research, BERT was loaded using KTrain and then trained and fit on a dataset containing labeled Twitter data. KTrain contains a built-in preprocessing model which processes the Tweets for the BERT model.

3.4.4 Data

In this section, a description of the data used for the training testing and validation of the models is given. Two datasets were used in this research. For training, testing and validation of the ANN

and BERT a dataset containing 1.6M labeled Tweets was used (Michailidis, 2018). This dataset was created using the methods described in Go et al. (2009). To double-check the performance of the models, with respect to the specific purpose of this research, a dataset containing 3000 Tweets, was manually labeled and used as a second validation set. This second validation set contains scraped Tweets using the Twitter scraping model built for this research. This validation set was meant as a final check, to test the reliability of the results obtained during training, testing and validation from the pre-labeled dataset from Michailidis (2018).

3.4.5 Evaluation metrics

Since sentiment analysis is basically a classification problem, appropriate evaluation metrics are the *precision*, *recall*, *F-score*, *accuracy* and *specificity*. The formulas for these metrics are given below. Here, TP stands for true positive, TN for true negative, FP for false positive and FN for false negative.

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\
specificity &= \frac{TN}{TN + FP}
\end{aligned}
$$

Figure 6: List of relevant metrics for classification tasks.

Precision is a measure that returns the proportion of positive identifications that was actually correct. Recall, or sensitivity, is a measure that returns the proportion of actual positives that were identified correctly. The F-score or F_1-score is the harmonic mean of precision and recall. Accuracy is the degree of closeness to the true value and specificity represents the proportion of correctly classified negative identifications.

Furthermore, the models are also compared in terms of their size and speed. For size, the models are compared based on their training size and model size. For speed, the comparison is based on training speed and predictive speed.

3.5 Dashboard

As mentioned earlier, in order to provide auditors with information about the sentiment of Tweets about a company in real time, a dashboard was built. This dashboard was built in Plotly's Dash-app environment. The main goal of this app was to provide auditors with the sentiment of Tweets about a company in an environment that is easy-to-use.

The idea here is that an auditor chooses a company for which they want to know the Twitter sentiment and that the dashboard returns this information. The first step is the search of Tweets. This was done as described in section 3.2. Once the search is completed, the Tweets were preprocessed, as described in section 3.3. Then, the sentiment analysis is performed by one of the models described in section 3.4. Once the sentiment analysis has been performed, all information needed to return the appropriate visualizations is available. Based on coordinated efforts with multiple auditors, the following visualizations were deemed most interesting:

- Time series sentiment data
- Pie chart of positive, negative and neutral Tweets
- Histogram of most common words
- Most negative & positive Tweets

In the following section a qualitative analysis will be described in which the features from this dashboard are put to the test. The results section will also look at potential additions to this dashboard, based on interviews with auditors.

3.6 Interviews

Finally, a qualitative analysis was performed in the form of interviews with auditors from three different teams responsible for three different companies. The goal of these interviews was to help answer the second and third sub-questions. These were the following:

1. What do auditors think about the relevance of Twitter sentiment as an additional source of information for assessing whether a company's goodwill is impaired?
2. What sources of information from Twitter could be most useful to help auditors in the process of identifying whether a company's goodwill is impaired and how should this information be transferred?

The interviews were generally divided into three parts;

1. Current process
2. Relationship between Twitter sentiment and goodwill
3. Usefulness of the dashboard

That being said, each interview took its natural course. However, the main structure of each interview was the same. First an introduction of both parties was initiated. Then, the motivation for the interview was provided. Then the interview started. The first part was focused on the auditor's current process of identifying goodwill impairment triggers. Here the main questions were:

- What does the current process of identifying goodwill impairment triggers look like?
- At what occasions do you have to perform the task of identifying goodwill impairment triggers?
- Are the steps taken and resources used in the process of identifying goodwill impairment triggers different for internal vs external triggers?
- What resources do you use in the process of identifying goodwill impairment triggers?
- Are all of your resources public information?
- Are there sources containing non-public information?
- Are there p2p (peer to peer) sources?
- Are there social media sources?
- Are there news sources?
- Is any part of this process automated through software?
- Is there, in your opinion, room for automation?

Once a clear view of the auditor's current practices was acquired, the the conversation was steered towards the relevance of Twitter sentiment as an additional source of information for assessing whether a company's goodwill is impaired. This was initiated by reminding the auditors of all goodwill impairment triggers, as stated by the IFRS. Then, the auditor was asked to answer the following type of questions:

- How do you become aware of these triggers?
- Do you think the activation of one of these goodwill impairment indicators would trigger a response on Twitter?
- Could social media contain some of the information listed in these triggers?
- How do you think you would respond if you discovered that one of this events had occurred?
- Do you think the occurrence of such a trigger could lead to a negative response from financial analysts or other experts such as yourself.
- Based on red flags / current sources of information; do you think social media contains "not-yet-public"-information?
- Do you use Twitter?
- If so: Do you follow financial Twitter accounts?
- If so: Have you experienced these accounts to disclose information about financial topics of which you were not yet aware?
- Based on recent events, such as game stop, doge coin, bitcoin and anything retweeted by Elon Musk; Do you think these events have any impact on the goodwill of these companies?

- Following these events the hedge fund Melvin Capital lost more than 50%, as a result the hedge funds Citadel and Point72 infused around $3 billion into the hedge fund, resulting in drops in their positions as well. Do you think these events could cause a goodwill impairment and how would you determine this to be true or not?

Finally, once an auditor had provided a clear view of their current process, the gaps of information in this process and the relation between the different goodwill impairment triggers and Twitter sentiment, the auditor was given a demonstration of the tool. Following this demonstration the following information was asked of the auditor being interviewed:

- Then the auditor was asked to give pro's and cons about each feature of the tool
- Could you access this data if it were not on an easy-to-use app or dashboard?
- How could the current aspects of the tool aid in the process of finding impairment triggers?
- What would you really want to see in this tool, that is missing?
- What do you think is an addition to the current information you have regarding goodwill impairment trigger detection?
- What information from social media do you think, could be valuable in the process of identifying goodwill impairment?
- Final remarks?

4 Results

In this section the results from this research are given. First, the results from sub-question 1, the comparative analysis of the NLP-based sentiment analysis models, will be given. Then, the results from sub-questions 2 & 3, the qualitative analysis in the form of an interview, will be given.

4.1 Model performance

This section deals with the results from the comparative analysis which was performed to answer sub-question 1: *What NLP-based sentiment analysis model is most suitable for Twitter sentiment analysis?* For this comparative analysis, three models are compared: VADER, Artificial Neural Network (ANN) and BERT. These models are compared based on two labeled datasets: a subset containing 10000 instances from the Kaggle dataset (Michailidis, 2018) and the manually scraped and labeled dataset containing 3000 instances. In this part, the models are compared based on 5 well known classification metrics: Precision, Recall, F-Score, Accuracy and Specificity. These results can be found in tables 1 and 2 respecctively. Furthermore, the models are also compared in terms of model size, training size, training speed and prediction speed. These results can be found in table 3.

Table 1: Validation performance per model based on 5 classification metrics Kaggle dataset.

Model	Metrics				
	Precision	Recall	F-Score	Accuracy	Specificity
VADER	0.64	0.84	0.74	0.70	0.52
ANN	0.78	0.79	0.79	0.78	0.77
BERT	0.86	0.84	0.85	0.85	0.84

Table 2: Performance per model based on 5 classification metrics manually labeled dataset.

Model	Metrics				
	Precision	Recall	F-Score	Accuracy	Specificity
VADER	0.66	0.86	0.75	0.71	0.55
ANN	0.79	0.78	0.78	0.78	0.78
BERT	0.87	0.86	0.86	0.86	0.85

The two models show great similarity. This shows that the training data did a good job of representing the scraped data and shows robustness of the models. In terms of predictive power, BERT wins on all metrics. Furthermore, the ANN scores pretty well and shows great consistency. For a plot of the training and validation loss and accuracy per epoch of the training process of the ANN, see the appendix. In contrast, the VADER model is not consistent at all. In some areas it

performs quite well whereas in others it performs poorly. This is probably due to the fact that the Kaggle dataset and the manually labeled dataset only contain the labels "Positive" and "Negative", whereas the VADER model produces three predictions: "Positive", "Negative" and "Neutral". In the calculation of the metrics of the VADER model, all rows containing "Neutral" predictions were not taken into account. Therefore, these results are slightly optimistic, but this might also explain the inconsistencies in the metrics' results.

Table 3: Results per model in terms of size and speed.

Model	Metrics			
	Model Size	Training Size	Training Speed	Prediction Speed
VADER	-	-	-	Fast
ANN	405MB	1.28M	15hr	Fast
BERT	1.3GB	64K	25hr	Slow

In terms of model size, training size and training speed the VADER model clearly outperforms both the ANN and BERT. Furthermore, BERT requires far less training data than the ANN, however, note that BERT has already seen around 3.5 billion words during pre-training. Even though BERT only used 5% of the data that was used by the ANN for training, it still needs significantly more training time than the ANN. This is most likely due to the huge amount of features, approximately 110M, incorporated in this model. Also, for the prediction speed it takes about 40 times as long as the VADER model and the ANN, which both predict at about the same speed.

4.2 Interviews

As mentioned earlier, a qualitative analysis was performed in the form of interviews with auditors from three different teams responsible for three different companies to help answer the second and third sub-questions. These were the following questions:

1. What do auditors think about the relevance of Twitter sentiment as an additional source of information for assessing whether a company's goodwill is impaired?
2. What sources of information from Twitter could be most useful to help auditors in the process of identifying whether a company's goodwill is impaired and how should this information be transferred?

The interviews were generally divided into three parts;

1. Current process
2. Relationship between Twitter sentiment and goodwill
3. Usefulness of the dashboard

In the following sections the main findings in each of these sections is given. Finally these findings are used to assess the results from sub-question 2 and 3.

4.2.1 Current process

The first part of the interview was focused on the auditor's current process of identifying goodwill impairment triggers. Here the main questions were:

- What does the current process of identifying goodwill impairment triggers look like?
- At what occasions do you have to perform the task of identifying goodwill impairment triggers?
- Are the steps taken and resources used in the process of identifying goodwill impairment triggers different for internal vs external triggers?
- What resources do you use in the process of identifying goodwill impairment triggers?
- Are all of your resources public information?
- Are there sources containing non-public information?
- Are there p2p (peer to peer) sources?
- Are there social media sources?
- Are there news sources?
- Is any part of this process automated through software?
- Is there, in your opinion, room for automation?

The main findings are stated below:

In the current process of goodwill impairment trigger detection, auditors stated that they generally only perform goodwill impairment tests when this is requested by a client. When assessing a company's goodwill impairment they generally use public financial sources and combine this with professional judgment. However, occasionally other sources such as news sites, peer to peer communication and even reddit are being used, but never as an official source. There are slight differences in the approach of investigating internal and external triggers, however, this is mostly based on their economic quantities.

Auditors generally seem to believe that there is room for automation in this field. Some auditors even said that their client has software in place which automatically tracks the levels of all goodwill impairment indicators, raising red flags when these reach a certain threshold. However, none currently implement any kind of sentiment.

4.2.2 Relationship between Twitter sentiment and goodwill

Once a clear view of the auditor's current practices was acquired, the conversation was steered towards the relevance of Twitter sentiment as an additional source of information for assessing whether a company's goodwill is impaired. This was initiated by reminding the auditors of all goodwill impairment triggers, as stated by the IFRS. Then, the auditor was asked to answer the following type of questions:

- How do you become aware of these triggers?
- Do you think the activation of one of these goodwill impairment indicators would trigger a response on Twitter?
- Could social media contain some of the information listed in these triggers?
- How do you think you would respond if you discovered that one of this events had occurred?
- Do you think the occurrence of such a trigger could lead to a negative response from financial analysts or other experts such as yourself.
- Based on red flags / current sources of information; do you think social media contains "not-yet-public"-information?
- Do you use Twitter?
- If so: Do you follow financial Twitter accounts?
- If so: Have you experienced these accounts to disclose information about financial topics of which you were not yet aware?
- Based on recent events, such as game stop, doge coin, bitcoin and anything retweeted by Elon Musk; Do you think these events have any impact on the goodwill of these companies?

The main findings are stated below:

Auditors were generally pretty optimistic about the relationship between Twitter sentiment and goodwill impairment. All believed the two to be correlated somehow. However, none believed there to be a causational relationship. Furthermore, only one auditor seemed to use Twitter and indicated that this was not with any analytical intention but solely for personal use. The recent events surrounding "hype stocks" had reached all auditors and they all believed this to have some form of impact on the goodwill of those companies, but not all of them.

4.2.3 Usefulness of the dashboard

Finally, once an auditor had provided a clear view of their current process, the gaps of information in this process and the relation between the different goodwill impairment triggers and Twitter sentiment, the auditor was given a demonstration of the tool. After this demonstration the following information was asked of the auditor being interviewed:

- What are the pros and cons about each feature of the tool?
- Could you access this data if it were not on an easy-to-use app or dashboard?
- How could the current aspects of the tool aid in the process of finding impairment triggers?
- What would you really want to see in this tool, that is missing?
- What do you think is an addition to the current information you have regarding goodwill impairment trigger detection?
- What information from social media do you think, could be valuable in the process of identifying goodwill impairment?
- Final remarks?

The main findings are stated below:

In this final part of the interview, the auditors all stated that this information could be very useful as an additional resource in the general risk assessment of a company. Most indicated that if they were to use such a tool, a very negative sentiment would definitely put them on alert and make them do further research into the cause of this negative sentiment and any possible economical reactions. However, most auditors stated that the information provided in this tool would need a lot more in-depth information about the source of the general sentiment and specific Tweets. They also indicated that in order for these results to be usable for a company's risk assessment, they should be more adequately filtered, whereas currently the results still contained a lot of noise. Furthermore, some other interesting suggestions were to add information about the industry and to add corresponding data from the stock market. Finally, to the question whether the auditors would be able to access the data if it were not on an easy-to-use app, they all answered no. None of the auditors had any real experience programming in python, let alone scraping Twitter for data and performing a sentiment analysis.

Based on the findings from the interviews, the sub-questions have been answered as follows: The first sub-question was: "What do auditors think about the relevance of Twitter sentiment as an additional source of information for assessing whether a company's goodwill is impaired?" This question can be answered by saying that the auditors were all rather optimistic. They all thought a correlation between the two factors exists. However, they also noted that they did not believe that a causational result would exist. Furthermore, they thought this information could be a useful early indication of economic shift, and perhaps also goodwill impairment.

The second sub-question was: "What sources of information from Twitter could be most useful to help auditors in the process of identifying whether a company's goodwill is impaired and how should this information be transferred?" The auditors responded by indicating that all presented features were interesting, but detailed information about for example the origin of the data was missing. Furthermore, they would also add industry information and stock market information to the dashboard. Finally, the auditors did not have a background in programming and would not be able to access this information if it would not be in an easy-to-use dashboard environment.

5 Discussion

With the rise of technology, it has become possible for everyone to invest in the stock market, get news from across the globe and communicate with anyone, instantly from their phone. Through social media people post whatever it is that they are thinking about or experiencing. If one pays close attention to the information coming from these platforms, one might find indicators of economic change. This has brought many opportunities, but has also created a lot of challenges for companies. Companies have become more vulnerable to public opinion. However, at the same time there are more available sources of information that can be used to build a wall of knowledge to minimize risk.

In auditing, companies are currently barely tapping into all available sources of information. Financial reporters mainly focus on publicly available financial data, when performing audit procedures. In these tasks, auditors generally react and almost never act preventively. This is largely due to the gap in technological advancements in this field.

It has been shown that very early indicators for stock market movement and other economic changes can be found by analyzing the sentiment from social media platforms such as Twitter. However, except for research about stock market prediction, not much is currently being done with this information. For other economic factors, there is some research regarding their relation to social media sentiment, but not many practical implementations are yet in place, as discussed in section 2. In the field of risk assessment, however, no research has been currently been done regarding the implementation of social media sentiment as an additional source of information.

This research has been an early attempt at investigating the relationship and practical relevance of tapping into this source of information and combining it into one form of financial risk assessment: the identification of goodwill impairment triggers. In order to investigate this relationship, three sub-questions were devised. The first sub-question is about what NLP-based sentiment analysis model is most appropriate for the task of uncovering the current sentiment of Twitter data about a company. The second and third sub-question are qualitative assessments. The second subquestion is about identifying the perceived relevance of using Twitter sentiment as an additional source of information in the process of goodwill impairment trigger detection, according to financial auditors. Finally, the third sub-question continues this qualitative research by investigating what information auditors would find interesting to obtain from Twitter and how this information could be transferred.

The first sub-question resulted in a comparative analysis between three models; the VADER sentiment intensity analyzer, the Artificial Neural Network and BERT. These models were tested based on different aspects, but mainly on predictive power, speed and size. In terms of predictive power, BERT was the overall winner, followed by the Artificial Neural Network, leaving the VADER model last. However, in terms of predictive speed the VADER model and the Artificial Neural Network were able to produce quick results, about 40 times faster than BERT. In terms of business implications, this could mean two things. First of all, in order to get current information from Twitter on a large scale, a model must be quick. Therefore, the ANN model would be preferred in this regard since it outperforms BERT in speed and VADER in predictive power. However, a business may have access to a cloud computing service which generates a larger amount of processing power. In this case all three models will be able to perform on a large scale in adequate speed. Then, the only relevant metric would be the performance and BERT is by far the strongest

in this regard.

The second sub-question is about the perceived relationship between goodwill impairment and Twitter sentiment, from the point of view of an auditor. The results from the interviews performed in this research imply that auditors are convinced that some form of correlation exists between the two, however, they do not think it's a causational relationship. Besides this relationship, auditors noted that their current measures often heavily rely on domain knowledge and other judgmental aspects. These are generally prone to human error. The auditors all said that social media sentiment could be a useful additional resource of information. In the future, auditors might have access to real time data about the social media sentiment about a company and have automated alarms or red flags going off when the information returned by these models shows unusual behavior. This could possibly change their position from reacting to economic movement to acting preventive in this regard.

Finally, the third sub-question was about what information from Twitter could be useful and how this should be transferred in order to make it easily accessible. In this case, the analysis was performed based on a demo of the tool that was built for this research. This tool contains four features; a line-graph showing the sentiment per day for the past week. A pie-graph showing the number of positive, neutral and negative Tweets. A histogram containing the most frequent words and a list of the most negative and most positive Tweets based on their sentiment scores. Auditors said all of these features could be interesting as additional data points, but they would need more information about the sources of the Tweets and their impact. This information could be retrieved with an enterprise Twitter API. Combining this with a trust-like network analysis, much like the trust framework proposed by Ruan et al. (2018), could make Twitter sentiment analysis a reliable source of information. Furthermore, the auditors stated that the dashboard was easy to use and comprehensive. They also stated that programming was out of the question for them. Therefore, in order to make this kind of information accessible to auditors, it should be provided in an environment which is friendly for people without programming experience.

6 Conclusion

In conclusion, this thesis is the first approach at incorporating Twitter sentiment analysis in a company's financial risk assessment. This has been done by answering the question; How can the sentiment of Tweets about a company be used as an additional source of information in assessing whether a company's goodwill is impaired? This can be seen as a step in the direction of cognitive automation in the field of financial reporting and could in the future protect companies from unnecessary risks through automated monitoring of Twitter sentiment related measures.

Furthermore, in future research, this relationship could be tested quantitatively by comparing historical Twitter data to historical goodwill impairment data. In order to perform this research, one would need access to a full archive Twitter API and a database containing historical goodwill impairment data. Showing a statistically significant correlation between the two could be a step towards the implementation of social media sentiment analysis as an additional source of information in the fields of goodwill accounting and financial risk assessment.

Auditors responsible for the financial reporting regarding other intangible and tangible assets could also benefit from this type of information. Investigative research into the relation between

other financial aspects in these fields and social media sentiment could prove to be fruitful. Furthermore, this analysis need not be limited to Twitter as a source of social media sentiment. For example, Reddit has recently proven to be a qualified contender in terms of explanatory power of economic factors.

Besides other financial aspects and other sources of social media data, it might also be interesting to look into more types of NLP-based methods which could help identify risks and extract other useful sources of information from social media. This information could be more detailed than just extracting the sentiment. As mentioned before, BERT performs well on a wide range of language understanding tasks, such as the identification of topics discussed in Tweets about a company. Another application is that an NLP-based model could be be used to improve a company's understanding of consumer needs by analyzing customers' comments, posts and online conversations.

Finally, many financial auditors lack a programming background. This, however, should not prevent them from getting all the insights available in order to perform their financial reporting. On the contrary, these are the positions where all possible resources should be available and are most crucial. Therefore, these streams of information should be made more accessible in the form of easy-to-use apps or dashboards. This could save companies enormous amounts of money and increase their level of risk awareness significantly. It would change the field of financial reporting from a reactive into a proactive practice.

Appendix

Artificial Neural Network training and validation

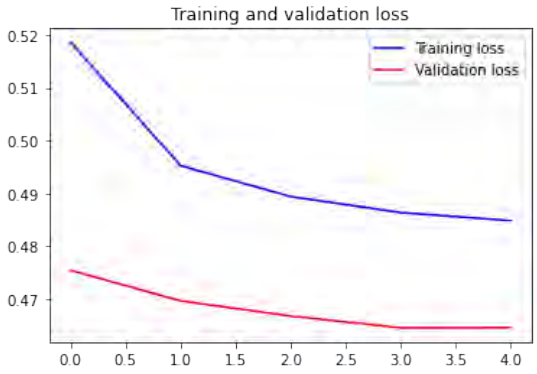

Figure 7: Training and validation loss of the neural network per epoch.

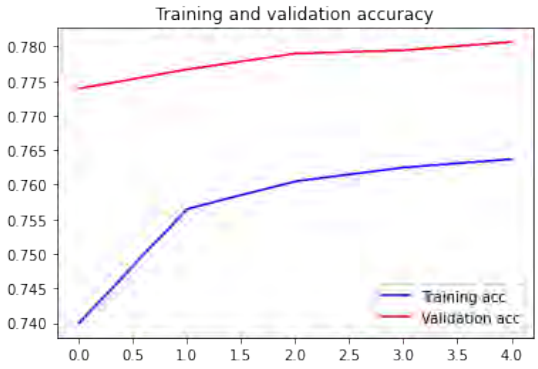

Figure 8: Training and validation accuracy of the neural network per epoch.

References

Abbott, L. and T. B. Kepler (1990). Model neurons: from hodgkin-huxley to hopfield. In *Statistical mechanics of neural networks*, pp. 5–18. Springer.

Alammar, J. (2018). The illustrated transformer. http://jalammar.github.io/illustrated-transformer/.

Albert, R. and A.-L. Barabási (2002). Statistical mechanics of complex networks. *Reviews of modern physics 74*(1), 47.

Albrecht, S., B. Lutz, and D. Neumann (2019). How sentiment impacts the success of blockchain startups–an analysis of social media data and initial coin offerings.

Ante, L. (2021). How elon musk's twitter activity moves cryptocurrency markets. *Available at SSRN 3778844*.

Asur, S. and B. A. Huberman (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, Volume 1, pp. 492–499. IEEE.

Azimi, M. and A. Agrawal (January 5, 2021). Is positive sentiment in corporate annual reports informative? evidence from deep learning. *Forthcoming, Review of Asset Pricing Studies*.

Baker, S. R., N. Bloom, S. J. Davis, K. J. Kost, M. C. Sammon, and T. Viratyosin (2020). The unprecedented stock market impact of covid-19. Technical report, National Bureau of Economic Research.

Boennen, S. and M. Glaum (July 4, 2014). Goodwill accounting: A review of the literature. *Forthcoming, Review of Asset Pricing Studies*.

Bollen, J., H. Mao, and A. Pepe (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 5.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of computational science 2*(1), 1–8.

Choi, H. and H. Varian (2012). Predicting the present with google trends. *Economic record 88*, 2–9.

Churyk, N. T. (2005). Reporting goodwill: are the new accounting standards consistent with market valuations? *Journal of Business Research 58*(10), 1353 – 1361. La Londe Seminar 2003.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *science 298*(5596), 1191–1194.

Duff and Phelps (2018). 2018 us goodwill impairment. https://www.duffandphelps.com/GWIstudies.

Eaton, G. W., T. C. Green, B. Roseman, and Y. Wu (2021). Zero-commission individual investors, high frequency traders, and stock market quality. *High Frequency Traders, and Stock Market Quality (January 2021)*.

EFRAG (2016). What do we really know about goodwill and impairment: A quantitative study. https://www.efrag.org/Publications.

Gerstner, W. (1998). Spiking neurons. Technical report, MIT-press.

Go, A., R. Bhayani, and L. Huang (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford 1*(12), 2009.

Gruhl, D., R. Guha, R. Kumar, J. Novak, and A. Tomkins (2005). The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 78–87.

Hales, J., J. R. Moon, and L. A. Swenson (2018). A new era of voluntary disclosure? empirical evidence on how employee postings on social media relate to future corporate disclosures. *Accounting, Organizations and Society 68-69*, 88–108. New Corporate Disclosures and New Methods.

Hebb, D. O. (1949). The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology 62*, 78.

Higson, C. (1998). Goodwill. *Financial Accounting eJournal 30*, 141–158.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Hutto, C. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 8.

IFRS (2020a). Ias 36. https://www.ifrs.org/issued-standards/list-of-standards/ias-36-impairment-of-assets/.

IFRS (2020b). Ifrs 3. https://www.ifrs.org/issued-standards/list-of-standards/ifrs-3-business-combinations/.

IFRS (2021). International financial reporting standards. https://www.ifrs.org.

Jaijairam, P. (2013, 01). Fair value accounting vs. historical cost accounting. *Review of Business Information Systems 17*, 1–6.

James, W. (1890). The principles of psychology. chapter iv. habits. *New York, NY: Henry Holt and Company*.

Kahneman, D. and A. Tversky (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific.

Kalda, A., B. Loos, A. Previtero, and A. Hackethal (2021). Smart (phone) investing? a within investor-time analysis of new technologies and trading behavior. Technical report, National Bureau of Economic Research.

Klingbeil, M. C. C. (2010). Intangible assets and goodwill in the context of business combinations: An industry study. KPMG AG Wirtschaftsprüfungsgesellschaft.

Konorski, J. (1948). Conditioned reflexes and neuron organization.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *nature 521*(7553), 436–444.

Luo, X., J. Zhang, and W. Duan (2013). Social media and firm equity value. *Information Systems Research 24*(1), 146–163.

Lyócsa, Š., E. Baumöhl, and T. Vŷrost (2021). Yolo trading: Riding with the herd during the gamestop episode.

McCulloch, W. S. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 5*(4), 115–133.

Michailidis, M. (2018). Sentiment140 dataset with 1.6 million tweets. https://www.kaggle.com/kazanova/sentiment140.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mishne, G., N. S. Glance, et al. (2006). Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pp. 155–158.

Murdock, J. (2018). Is tesla better off without elon musk? fraud charge brings uncertain future. https://www.newsweek.com/tesla-better-without-elon-musk-sec-charges-fraud-1143837.

Nguyen, L. T., P. Wu, W. Chan, W. Peng, and Y. Zhang (2012). Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, pp. 1–8.

Nielsen, M. A. (2015). *Neural networks and deep learning*, Volume 25. Determination press San Francisco, CA.

Nisar, T. M. and M. Yeung (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science 4*(2), 101–119.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance 6*(3), 144–160.

Pagano, M. S., J. Sedunov, and R. Velthuis (2021). How did retail investors respond to the covid-19 pandemic? the effect of robinhood brokerage customers on market quality. *Finance Research Letters*, 101946.

Pagolu, V. S., K. N. Reddy, G. Panda, and B. Majhi (2016). Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1345–1350.

Pavlov, I. P. and W. Gantt (1928). Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous activity (behaviour) of animals.

Ramelli, S. and A. F. Wagner (2020). Feverish stock price reactions to covid-19. *The Review of Corporate Finance Studies 9*(3), 622–655.

Ramón y Cajal, S. (1894). The croonian lecture.—la fine structure des centres nerveux. *Proceedings of the Royal Society of London 55*(331-335), 444–468.

Rao, T. and S. Srivastava (2016). Analyzing stock market movements using twitter sentiment analysis. In *2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review 65*(6), 386.

Ruan, Y., A. Durresi, and L. Alfantoukh (2018). Using twitter trust network for stock market analysis. *Knowledge-Based Systems 145*, 207–218.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *nature 323*(6088), 533–536.

Sartore, M. (2020). What is gaap? https://www.accounting.com/resources/gaap/.

Schumaker, R. P. and H. Chen (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS) 27*(2), 1–19.

Socher, R., C. C.-Y. Lin, A. Y. Ng, and C. D. Manning (2011). Parsing natural scenes and natural language with recursive neural networks. In *ICML*.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Späth, G. and R. Trampler (2018). Goodwill impairment: Predicting goodwill impairment with the market reaction to acquisitions.

Sun, T. and M. Vasarhelyi (2018, 01). Embracing textual data analytics in auditing with deep learning. *The International Journal of Digital Accounting Research 18*, 49–67.

Sun, T. and M. A. Vasarhelyi (2017). Deep learning and the future of auditing: How an evolving technology could transform analysis and improve judgment. *CPA Journal 87*, 24–29.

Sun, T. S. (2019, 05). Applying Deep Learning to Audit Procedures: An Illustrative Framework. *Accounting Horizons 33*(3), 89–109.

Sundermeyer, M., R. Schlüter, and H. Ney (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Tsai, C.-F., Y.-H. Lu, Y.-C. Hung, and D. C. Yen (2016). Intangible assets evaluation: The machine learning perspective. *Neurocomputing 175*, 110 – 120.

Tsai, C.-F., Y.-H. Lu, and D. C. Yen (2012). Determinants of intangible assets value: The data mining approach. *Knowledge-Based Systems 31*, 67 – 77.

van der Beck, P. and C. Jaunin (2021). The equity market implications of the retail investment boom. *Available at SSRN 3776421*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vergoossen, R. (2004, 06). Bijzondere waardevermindering van vaste activa in de regelgeving. *Maandblad Voor Accountancy en Bedrijfseconomie 78*.

Werbos, P. (1974). Beyond regression:" new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*.

Zhang, X., H. Fuehres, and P. A. Gloor (2011a). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences 26*, 55–62.

Zhang, X., H. Fuehres, and P. A. Gloor (2011b). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences 26*, 55–62. The 2nd Collaborative Innovation Networks Conference - COINs2010.