5-5-2019

An exploration of a financial lexicon-based approach to sentiment analysis and its application to financial news and reports

Avery N. Kirchner

NORTHERN ILLINOIS UNIVERSITY

An exploration of a financial lexicon-based approach to sentiment analysis and its application to financial news and reports

A Capstone Submitted to the

University Honors Program

In Partial Fulfillment of the

Requirements of the Baccalaureate Degree

With Honors

Department Of

Accountancy

By

Avery Kirchner

DeKalb, Illinois

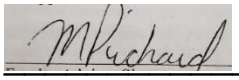
May 5, 2019

University Honors Program
Capstone Approval Page

Capstone Title (print or type)
An exploration of a financial lexicon-based approach to sentiment analysis and its application to financial news and reports

Student Name (print or type) Avery Kirchner

Faculty Supervisor (print or type) Dr. Matt Pickard

Faculty Approval Signature 

Department of (print or type) Accountancy

Date of Approval (print or type) 05/04/2019

Check if any of the following apply, and please tell us where and how it was published:

☐ Capstone has been published (Journal/Outlet):

_____

☐ Capstone has been submitted for publication (Journal/Outlet):

_____

☐ Capstone has been presented (Conference):

_____

☐ Capstone has been submitted for presentation (Conference):

_____

Completed Honors Capstone projects may be used for student reference purposes, both electronically and in the Honors Capstone Library (CLB 110).

If you would like to opt out and not have this student's completed capstone used for reference purposes, please initial here: _____ (Faculty Supervisor)

HONORS CAPSTONE ABSTRACT

For as long as the stock market, financial news, and financial reports have been around, people have been trying to gain an edge on the market and identify trends in advance, allowing for more intelligent financial decision-making. This has been quite difficult, but in recent years a branch of Natural Language Processing known as sentiment analysis has made this more realistic. By using machine learning or lexicon-based methods to sentiment analysis to analyze the emotion behind text, research has shown potential for investors and companies to identify financial trends in advance that can help them guide strategy or investment decisions.

In this project, by analyzing sentiment analysis research over a variety of domains, a lexicon-based approach to sentiment analysis was examined. Its methodologies, benefits, and drawbacks were studied, and through this research, a potential model was proposed that could improve upon the current lexicon-based model and be specifically applied to finance specific works such as financial news articles or financial reports.

This paper identifies a potential improved way of doing sentiment analysis for financial news and reports using a lexicon-based approach. By incorporating increased polarity aspects, there is potential to be able to delve to deeper granularity levels when it comes to analyzing sentiment. Typically, a lexicon-based approach to sentiment analysis will depict text as being either positive, negative, or neutral, but the model proposed in this paper describes a way to do this analysis while determining to what extent text is positive or negative. This allows for greater understanding of the sentiment behind financial news and reports and this information can be used to better understand the health of the market, future company earnings potential, a company's risk levels, and a company's future ROA numbers among other things.

Introduction

Predicting the future when it comes to both business and the world is seen as an impossible task, but with recent technological innovation, it is becoming more of a possibility every day. Through harnessing the power of big data and sentiment analysis, it is possible to identify trends and patterns and use this information to aid in decision-making in order to stay ahead of the curve. Sentiment analysis allows users to analyze and understand text strings to identify underlying sentiment or emotions from the text. By applying this to the business world, especially the fields of finance and accounting, it is possible to discover trends in market price movement, risk hidden within financial statements, and gauge how public opinion changes in real time in relation to financial news, thus allowing an investor to make decisions faster and more intelligently than an investor who does not use sentiment analysis.

What is the stock market?

Stock markets are a financing channel that allow for the optimal allocation of capital (Wang et al., 2019). Stock markets not only help in maintaining business values, but also assist in protecting investors by allowing them to invest in companies without being held liable for problems that those companies might run into. Stock markets also offer an outlet for people to invest in commodities and currencies.

The efficient markets theory states that a market is acting under known information, so stock prices reflect the information available in the market. As a result, the only ways to generate returns different than what the market should provide would be to invest in riskier products or try to trade when market prices deviate from their fundamental values. In reality, markets are not efficient, with humans focusing too much on the short-term and not enough on the long-term, causing errors in market analysis to get amplified which results in market crashes (Bouchaud, 2008). Economists do not have models to make sense of "wild" markets since their beliefs are based off of economic axioms like the rationality of economic agents where agents always act to maximize their profits, the invisible hand where agents acting to maximize their profits are led to do what is best for society as a whole, and market efficiency where market prices reflect all known information about assets (Bouchaud, 2008). Sentiment analysis is one method that would allow investors to make use of data that most investors do not normally take into account, allowing investors or companies to make investment decisions as changes in the market are identified through changes in sentiment (Sung, Cho, & Ryu, 2019).

What is Natural Language Processing?

Natural language refers to the way that humans talk to each other through speech and text. Natural Language Processing is a field of study concerned with the automatic manipulation of natural language in the forms of speech and text through software. Natural Language Processing provides humans tools that can be used to understand natural language and its forms. Working with natural language is difficult due to the ambiguity of its structure and uses. Different words can have different meanings depending on the context of a situation and there are some rules that dictate how things are written and spoken, but there are not enough rules in place to make it clear in every situation how words are used (Liu, 2012). As a result, the study of Natural Language Processing combines aspects of both linguistics and statistics to make sense of the unstructured data that natural language creates.

Overall, Natural Language Processing seeks to make sense of linguistic observations and creations in the world, make statistical inferences from data created from natural language, and it encompasses any computer manipulation of natural language data including the automatic computational processing of human languages covering algorithms that take human-produced text as input as well as algorithms that produce natural looking text as outputs (Liu, 2012). Due to the rise of digitally based opinion text through social media posts, sentiment analysis has become one of the most active research areas within the field of Natural Language Processing (Liu, 2012).

What is sentiment analysis?

Sentiment analysis is a growing research and analysis technique in the business arena. According to Bing Liu, a prominent sentiment analysis research professor at the University of Illinois at Chicago, "sentiment analysis, also called opinion mining, is a field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2012, p.7). Sentiment analysis is a relatively new and unexplored field with little research having been done regarding people's opinions and sentiments prior to the year 2000. Using sentiment analysis and the opinion-related information it contains can provide great value and competitive insights for people in the business world.

Opinions have influenced human behaviors for millennia and the rise of social media has given people access to more opinions than ever before. No longer are people limited to opinions just from friends and family, but now through social media websites are able to access opinions on political candidates, products, and business from anyone on the social media site of their choice. As a result, more and more people and organizations are using the content from social media to guide their decision-making (Liu, 2012). Due to the massive amounts of information available through social media, it can be difficult to find relevant information and make decisions using the right information. Automated sentiment analysis systems can assist users in identifying relevant sites and extracting and summarizing the opinions housed within (Liu, 2012).

Typically, there are three levels that sentiment analysis is performed at. Sentiment analysis has been investigated at the document level where an entire document is analyzed to determine whether or not a document expresses a positive or negative sentiment (Liu, 2012). An example of this would be a product review of a single product. Since the review is focusing on one and only one product, sentiment analysis can be performed at the document level to determine what the overall sentiment of the review was in relation to the product (Liu, 2012). This technique is known as document-level sentiment classification.

Sentiment analysis can also be performed at the sentence level. At the sentence level, each sentence is analyzed to determine whether each sentence expresses a positive, negative, or neutral opinion (Liu, 2012).

The last and most complex level is the entity and aspect level. This level does not focus on sentiment analysis for sentences, documents, or paragraphs, but rather, focuses on positive or negative sentiment of an opinion and how it relates to a target (Liu, 2012). This comes into effect when looking at an example of a quote relating to an iPhone's performance, "The iPhone's call quality is good, but its battery life is short" (Liu, 2012, p.11). The call quality sentiment of the

quote is positive, but the battery life sentiment of the quote is negative. By separating this quote into two separate aspects: call quality and battery life, the aspect for each entity can be separately analyzed turning unstructured text into structured data (BL Book). Separating text information like this into two separate aspects makes it possible to run qualitative and quantitative analysis on the newly create structured data. This type of analysis becomes even more complex when comparing two entities like Coke and Pepsi. If someone were to say, "Coke tastes better than Pepsi," now they are comparing two separate entities on a shared aspect (taste) that expresses a preference for Coke (Liu, 2012).

There are two primary approaches to sentiment analysis: supervised and unsupervised (or lexicon-based). A lexicon-based approach to sentiment analysis measures the sentiment of a sentence or piece of text by assigning a sentiment score to each individual word in the sentence or text and then summing up the sentiment score from each individual word to determine whether the text's sentiment is positive or negative (Domanska, 2018). A lexicon-based approach needs an external lexical resource that has already mapped words to a categorical class of positive, negative, or neutral or a numerical score that provides the same insight. Due to this, a lexicon-based approach's effectiveness is limited by how effective its lexical resource is. A lexicon-based approach saves time compared to a supervised model because it skips the step where people have to label training data, but it has some drawbacks such as a word in a lexical resource having different meanings (and as a result different sentiments) depending on its use that a lexicon-based approach cannot account for (Domanska, 2018). Related to this concept, sentences with a sarcastic tone can warp the degree of sentiment and a sentence with sentiment words not in the lexical resource will fly under the radar. According to research done by a CoreValue data science engineer, a lexicon-based approach works best when it relates to binary classification for sentiment (is it positive or negative) with classification accuracy reaching 69%, competitive with a set of experiments from Mechanical Turk indicating that humans only agree 79% of the time (Domanska, 2018). When the lexicon-based approach dealt with classifying scores in 10 classes it was only 0.65% accurate and when it came to classifying scores in 3 classes, it was 19% accurate (Domanska, 2018).

A supervised approach to sentiment analysis works with training data to do the analysis. By using training samples and entering their output values into an algorithm before applying it to an actual data set, the algorithm can handle new unknown data in the future, providing more accurate sentiment classification in specific domains for which it is trained (Domanska, 2018). Some of the most common algorithms for sentiment analysis are Naïve Bayes classification and Support Vector Machines (SVM). The same research done by the CoreValue data science engineer found that a supervised approach to sentiment analysis was 63% accurate when it came to classifying scores in 10 classes, 99% accurate when it came to classifying scores in 3 classes, and 100% accurate when it came to classifying scores in a binary class (Domanska, 2018). As expected, machine learning methodologies outperformed the lexicon-based method in all categories, with only the binary classification attempts being somewhat competitive.

The value of the Loughran and McDonald finance-specific lexicon

A sentiment lexicon is vital to conducting sentiment analysis over text, but a general-purpose sentiment lexicon will often be inadequate when applied to information coming from financial sources. Researchers Loughran and McDonald noticed that while the Harvard

Psychosociological Dictionary was a valuable tool for word classifications, English words can have many meanings depending on the context, and as a result, a word categorization scheme that works well in one discipline might not work as well in another discipline like finance or accounting (Loughran & McDonald, 2011**)**. After analyzing 50,115 firm-year 10Ks produced between 1994 and 2008 and the 2.5 billion words within those reports, Loughran and McDonald found that the Harvard Psychosociological Dictionary often misclassified words when analyzing tone in a financial setting with almost 75% of the negative word counts resulting from the Harvard Psychosociological Dictionary list not being considered negative in a financial setting (Loughran & McDonald, 2011**)**. The Harvard Psychosociological Dictionary considers words that appear frequently in most 10-Ks like "tax", "cost", "capital", "board", "liability", "foreign", and "vice" to be negative words (Loughran & McDonald, 2011**)**. In the case of 10-Ks, these words are naming parts of the company and bear no negative meaning (i.e. "Board" of Directors, "Vice"-president). Also, there are many other common, industry specific words like "mine", "cancer", "crude" (oil), "tire", or "capital", that would be considered negative on the Harvard list, but from a financial perspective, are part of doing business (or a specific industry) and hold no negative meaning in a financial setting (Loughran & McDonald, 2011**)**. This means that a lexicon needs to be tailored to be specific to financial contexts in order to properly classify the sentiment behind words found in financial reports.

Loughran and McDonald saw the need for a finance-specific lexicon and created a list of 2,337 words that have negative meanings in a financial sense (Loughran & McDonald, 2011**)**. Due to the many polystemes in the English language, it is impossible to map every single word in the language to a financial sentiment, but the list that was created by the two researchers is a promising start (Loughran & McDonald, 2011**)**.

Loughran and McDonald's list uses term weighting to account for differences in the number of times that words typically show up in a financial report. Doing their analysis, the pair found that the word "loss" showed up 1.79 million times amongst the reports analyzed while the word "aggravates" showed up 10 times within the same sample (Loughran & McDonald, 2011). Since the collective use of loss is not 179,000 times more important than aggravate, term weighting helps to provide a more balanced weighting to the impact that each word has when it shows up in the financial reports. This ensures that even if a word is typically used less frequently than another word that it's still able to have an impact when it comes to determining financial sentiment. The Loughran and McDonald list also accounts for inflections in words (Loughran & McDonald, 2011). An example of this is that if accident is considered a negative word, then "accidental", "accidentally", and "accidents" are all considered negative as well. Accounting for inflections in their list brings the number of words in their list up from 2,337 words to 4,187 words compared to the 2,005 words found in the Harvard Psychosociological Dictionary (Loughran & McDonald, 2011). Though it would be easier (and make for a shorter list) to only use a list that details the most impactful words based off of the 10-Ks that the researchers analyzed, since the lists are publicly available, it would allow financial managers to manipulate their financial reports so that those impactful words are excluded (Loughran & McDonald, 2011). As a result, the list that Loughran and McDonald created is quite exhaustive to account for any manipulation that creators of the financial reports might attempt to do.

Figure 1

Variable	Full 10-K Document ($N = 50,115$) Mean	Median	Standard Deviation	MD&A Section ($N = 37,287$) Mean	Median	Standard Deviation
Word Lists						
H4N-Inf (H4N w/ inflections)	3.79%	3.84%	0.76%	4.83%	4.79%	0.89%
Fin-Neg (negative)	1.39%	1.36%	0.55%	1.51%	1.43%	0.67%
Fin-Pos (positive)	0.75%	0.74%	0.21%	0.83%	0.79%	0.32%
Fin-Unc (uncertainty)	1.20%	1.20%	0.32%	1.56%	1.48%	0.62%
Fin-Lit (litigious)	1.10%	0.95%	0.53%	0.60%	0.51%	0.43%
MW-Strong (strong modal words)	0.26%	0.24%	0.11%	0.30%	0.27%	0.17%
MW-Weak (weak modal words)	0.43%	0.39%	0.21%	0.43%	0.34%	0.32%
Other Variables						
Event period [0,3] excess return	−0.12%	−0.19%	6.82%	−0.23%	−0.28%	7.26%
Size ($billions)	$3.09	$0.33	$14.94	$2.12	$0.30	$9.62
Book-to-market	0.613	0.512	0.459	0.611	0.501	0.477
Turnover	1.519	0.947	2.295	1.695	1.104	2.508
One-year preevent FF alpha	0.07%	0.04%	0.20%	0.07%	0.05%	0.21%
Institutional ownership	48.34%	48.07%	28.66%	49.23%	48.52%	29.33%
NASDAQ dummy	56.15%	100.00%	49.62%	60.12%	100.00%	48.97%
Standardized unexpected earnings	−0.02%	0.03%	0.76%	−0.03%	0.03%	0.82%
Analysts' earnings forecast dispersion	0.17%	0.07%	0.33%	0.19%	0.08%	0.36%
Analysts' earnings revisions	−0.21%	−0.04%	0.69%	−0.24%	−0.05%	0.74%

The above chart provides an example of the analysis that Loughran and McDonald were able to do over the 50,115 10Ks and 37,287 MD&A sections that they analyzed while also making some comparisons to the Harvard list (H4N) to show the difference when using a finance-specific list (2011).

Research from Loughran and McDonald (2011) created a finance-specific lexicon that is comprised of 6-word lists that can be used to break apart sentiment found in financial reports. These word lists include financially negative terminologies such as "deficit or default"; financially positive terminologies such as "achieve or profit"; financially unclear words such as "appear or doubt"; words indicating a potential for future litigation such as "amend or forbear"; words expressing strong confidence levels such as "always or must"; and words expressing weak confidence levels such as "could or might" (Wang, Tsai, Liu, & Chang, 2013). These lists break apart the different types of words that might be found in financial reports and allow for a finance-specific lexicon to better classify the sentiment behind words in a financial context.

Figure 2

Negative	Positive	Uncertainty	Litigous	Strong Modal	Weak Modal
ABANDON	ABLE	ABEYANCE	ABOVEMENTIONED	ALWAYS	ALMOST
ABANDONED	ABUNDANCE	ABEYANCES	ABROGATE	BEST	APPARENTLY
ABANDONING	ABUNDANT	ALMOST	ABROGATED	CLEARLY	APPEARED
ABANDONMENT	ACCLAIMED	ALTERATION	ABROGATES	DEFINITELY	APPEARING
ABANDONMENTS	ACCOMPLISH	ALTERATIONS	ABROGATING	DEFINITIVELY	APPEARS
ABANDONS	ACCOMPLISHED	AMBIGUITIES	ABROGATION	HIGHEST	CONCEIVABLE
ABDICATED	ACCOMPLISHES	AMBIGUITY	ABROGATIONS	LOWEST	COULD
ABDICATES	ACCOMPLISHING	AMBIGUOUS	ABSOLVE	MUST	DEPEND
ABDICATING	ACCOMPLISHMENT	ANOMALIES	ABSOLVED	NEVER	DEPENDED
ABDICATION	ACCOMPLISHMENTS	ANOMALOUS	ABSOLVES	STRONGLY	DEPENDING
ABDICATIONS	ACHIEVE	ANOMALOUSLY	ABSOLVING	UNAMBIGUOUSLY	DEPENDS
ABERRANT	ACHIEVED	ANOMALY	ACCESSION	UNCOMPROMISING	MAY
ABERRATION	ACHIEVEMENT	ANTICIPATE	ACCESSIONS	UNDISPUTED	MAYBE
ABERRATIONAL	ACHIEVEMENTS	ANTICIPATED	ACQUIREES	UNDOUBTEDLY	MIGHT
ABERRATIONS	ACHIEVES	ANTICIPATES	ACQUIRORS	UNEQUIVOCAL	NEARLY
ABETTING	ACHIEVING	ANTICIPATING	ACQUIT	UNEQUIVOCALLY	OCCASIONALLY
ABNORMAL	ADEQUATELY	ANTICIPATION	ACQUITS	UNPARALLELED	PERHAPS
ABNORMALITIES	ADVANCEMENT	ANTICIPATIONS	ACQUITTAL	UNSURPASSED	POSSIBLE
ABNORMALITY	ADVANCEMENTS	APPARENT	ACQUITTALS	WILL	POSSIBLY
ABNORMALLY	ADVANCES	APPARENTLY	ACQUITTANCE		SELDOM
ABOLISH	ADVANCING	APPEAR	ACQUITTANCES		SELDOMLY
ABOLISHED	ADVANTAGE	APPEARED	ACQUITTED		SOMETIMES
ABOLISHES	ADVANTAGED	APPEARING	ACQUITTING		SOMEWHAT

The above chart provides a sample of some of the different types of sentiment words and their inflections from the Loughran and McDonald list (2011). The above chart also demonstrates how the Loughran and McDonald list has been broken into 6 different sentiment classifications (indicating positivity, negativity, uncertain as to what type of sentiment, words indicating litigation in the future, and words that amplify other words).

Incorporating polarity aspects into the financial-lexicon model

This financial-lexicon model can then be improved by incorporating polarity aspects. Thus far, not much work has been incorporating polarity into lexicon models used for analyzing sentiment in a financial context. Social media analytics researchers have been able to incorporate polarity aspects into their lexicon-based approach to sentiment analysis. This allows analysis to be done at a more granular level. Rather than focusing on whether a string is positive or negative, analysis can be focused on to what extent something is positive or negative. In order to do sentiment analysis with a lexicon, a dictionary of positive or negative words is needed. As mentioned previously, a general lexicon will not suffice since it does not consider the changes in sentiment that occur when words are used in a financial context. Loughran and McDonald's dictionary is one example of a lexicon that is more finance specific. By integrating polarity into this dictionary, it is possible to have a clearer understanding of the underlying sentiment behind text.

To incorporate polarity into the McDonald and Loughran dataset, the dictionary would need to have a positive or negative sentiment value assigned to each word (Jurek, Mulvenna, & Bi, 2015). Lexicon-based approaches to sentiment analysis work by considering each word in a text string as a word from a bag of words. As a result, each individual word has its own positive

or negative value coming from the lexicon dictionary. Sentiment values from the dictionary are assigned to each word, and then the sum or average of the total sentiment behind the text is taken to determine the overall sentiment for the message, positive or negative (Jurek et al., 2015). The lexicon-based approach has an advantage over machine-learning models in that it does not need to generate a labelled training set which is difficult, time-consuming, and expensive to ensure a dataset has been sufficiently and correctly labelled (Jurek et al., 2015). It is also easier to understand how a lexicon-based approach works (compared to a machine learning model) and it is easier to generate a lexicon and make changes to it than it is to do the same with a machine learning model.

The proposed lexicon-based sentiment model would use McDonald and Loughran's financial dictionary, but further segment the words in each list by assigning them values to represent sentiment between -100 (most negative) and 100 (most positive) (Jurek et al., 2015). This model would also incorporate negation and intensity aspects in order to provide the most possible polarity data in the text. To do this, a negating function can be used with a lexicon of negating words from Jurek, Mulvenna, and Bi's research (2015) that considers how a negating work could affect the polarity of a text string.

Figure 3

$$F_N(S) = \begin{cases} max\left\{\frac{S+100}{2},\ 10\right\} & if\ S < 0 \\ min\left\{\frac{S-100}{2},\ -10\right\} & if\ S > 0 \end{cases}$$

*****F_N represents final negation value

*S represents sentiment value from financial lexicon

By creating the above function, researchers were able to find negation in a sentence and then find the first non-neutral word occurring in the following three positions after the negator was found to determine the new sentiment value of the sentence. This differs from the typical approach to handling negation in a lexicon-based approach by reversing polarity of the lexicon item next to the negator (ex. "good": 100, "not good": -100), by resulting in a more accurate assignment of sentiment to negating words (Jurek et al., 2015). Using this function, a sentence like "I don't hate this stock", would have a sentiment value of 10 rather than 100 ("hate" has a value of -100 in researcher's lexicon, don't would reverse the sentiment completely under typical approach), meaning that the use of a negating function ensures that neither a very high or low sentiment can be created through negation (Jurek et al., 2015).

In a similar vein, this model would also be able to handle intensifiers like "very", "most", and "quite", as well as amplifiers like "slightly" that increase or decrease the sentiment behind text. The researchers were able to divide intensifiers into 3 categories based off polarity: downtoners, weak amplifiers, and strong amplifiers (Jurek et al., 2015). Downtowners represented intensifiers that decreased sentiment value by 50% while weak and strong amplifiers

increased sentiment by 50% and 100% respectively (Jurek et al., 2015). Negators and intensifiers would be considered neutral words if in a sentence surrounded by neutral words, but in a group of positive or negative words, they would be considered as non-neutral since they influence the sentiment found in the sentence (Jurek et al., 2015).

Once each word in the text string has been classified, a combining function can be used to determine the polarity behind the sentiment within the string. Since each word within the financial lexicon has been assigned a value between -100 and 100, by using a combining function, it is possible to determine to what extent the sentiment is positive or negative. The combining function should also be able to model the relation between sentences depending on the number of non-neutral words and the value of the sentiment contained (Jurek et al., 2015). Researchers Jurek, Mulvenna, and Bi proposed a normalization function that combines the best of both worlds between sum and average functions. The normalization function that they developed considers the difference between the overall positive and negative sentiments expressed within the text string as well as the number of positive and negative words within the message (Jurek et al., 2015). As a result of this, the overall positive or negative sentiment in a message should be represented as a product of the average sentiment and a coefficient that's value depends on the number of positive or negative words within the message (Jurek et al., 2015). This normalization formula can be depicted as:

Figure 4

$$F_P = min\left\{\frac{A_P}{2 - log(3.5 \times W_P + I_P)}, 100\right\}$$

$$F_N = max\left\{\frac{A_N}{2 - log(3.5 \times W_N + I_N)}, -100\right\}$$

*F_P/F_N calculate overall positive or negative sentiment within a message

**W_P/W_N represent the number of positive or negative words within a message

***A_P/A_N represents the average of positive or negative sentiment

****p represents a parameter determining the shape of the log function. The larger that p is, the faster the value of FP/FN will increase as the number of non-neutral words change. p can be determined by analyzing a large sample of tweets or financial reports in order to determine the distribution of non-neutral words' distribution across the tweets or financial reports (Jurek et al., 2015). Using a logarithmic function provides the ability to model the relationship between the number of positive or negative words in a text string compared to the average positive or negative sentiment within the sentence (Jurek et al., 2015).

*****I_P/I_N stand for number of intensifiers that refer respectively to positive or negative words in a sentence. They act by decreasing or increasing the number of words by values of 0.5 (for I_P) and 1 for (I_N).

Once this function has been used to run analysis on a text string, it has now been segmented into a negative sentiment value and a positive sentiment value. The value from 0-100 represents the total positive sentiment of the text string and the value from -100 to 0 represents the total negative sentiment of the text string. Based on the larger value that is returned, the text can be classified as either positive or negative as well as the intensity of the text (Jurek et al., 2015). In the case of a text string that has a mixed sentiment, a combination formula was developed to determine if the text string is more positive or more negative.

Figure 5

$$e_P = min\left\{\frac{A_P}{2 - log(3.5 \times W_P)},\ 1\right\}$$

$$e_N = max\left\{\frac{A_N}{2 - log(3.5 \times W_N)},\ -1\right\}$$

*W_P/W_N indicates the number of positive or negative words in the text string

**e_p/e_N indicates the number overall positive or negative evidence in a sentence.

Based on the final sentiment function that returns the value of F_P/F_N and the evidence function that returns the value of ep/eN, then it is possible to determine the overall sentiment in a text string. If there are only positive words in a text string under analysis, then the final sentiment value is based on F_P/e_p only (Jurek et al., 2015). If there are only negative words in a text string under analysis, then the final sentiment value is based on F_N/e_N only. In the case of a sentence or text string that has both positive and negative words, based on the larger evidence value (positive or negative) returned using the evidence function to determine the larger value (Jurek et al., 2015). By taking the difference between the positive and negative evidence values, if the evidence difference is much higher than the other (difference greater than 0.1) then return positive or negative sentiment, but when there is not evidence available or there is not much difference, then the final determination of sentiment in a text string is made based off of the F_P/F_N function, taking the higher returned value to determine the sentiment behind a text string as well as the intensity (Jurek et al., 2015).

This sentiment can best be logically displayed through the following code (Jurek et al., 2015)**:**

IF (W_N==0)

RETURN finalSentiment(F_P,e_p)

ELSE IF (W_P==0)

RETURN finalSentiment(F_N,e_N)

ELSE {

IF(F_P-F_N>0.1)

RETURN finalSentiment(F_P,e_p)

ELSE IF (F_N - F_P>0.1)

RETURN finalSentiment (F_N,e_N)

ELSE {

IF (F_P + F_N > 0)

RETURN finalSentiment (F_P,e_p)

ELSE IF (F_P + F_N < 0)

RETURN finalSentiment (F_N,e_N)

ELSE

RETURN 0

}}

finalSentiment (F,e) {

if(|F|>25 || |e| > 0.5)

RETURN F

ELSE

RETURN 0

}

The above code puts this classification process into a mathematical logic form that breaks down how the above functions are used at different points based on the sentiment values that are returned. At times the functions build on each other, and at times one or two functions are enough to determine the sentiment behind a text string. At times, the sentiment behind a text string will simply be 0, or neutral.

Relating the improved lexicon model to accounting and finance

As mentioned, this model was originally designed for social media analytic analysis, specifically for tweets, but by tailoring it as mentioned, it should apply to financial reports and financial-related tweets. By using the McDonald and Loughran dictionary as a financial-lexicon foundation, it is possible to further classify the words within this dictionary into not just "positive" or "negative" words, but also to what degree the words are positive or negative to incorporate polarity functionality into the model. Now, instead of saying "This sentence is positive…" or "This sentence is negative…", it is possible to say, "This sentence has a positive

sentiment value of 65, this sentence has a positive sentiment value of 80, and this sentence has a positive sentiment value of 20." Likewise, it is possible to say, "This sentence has a negative sentiment value of 65, this sentence has a negative sentiment value of 80, and this sentence has a negative sentiment value of 20." Thus, if this improved lexicon model were to be applied to financial reports or financial news, it would be possible to break down the underlying sentiment in the financial reports into more granular details, which would provide investors, auditors, and financial analysts more insight into what is being said within financial news and reports.

Using the Loughran and McDonald financial lists serves as a strong starting point for developing the lexicon for this model, but more work needs to be done in terms of figuring out how to tag words within the list within the -100 to 100 range to demonstrate their weight on the sentiment behind text. More research needs to be done in terms of ascribing weights to words within the positive and negative ranges to depict how they affect sentiment. What word is given a score of positive 80? What word is given a score of -20? Also, currently this list of financial terms is only for the English language. Many companies around the world produce financial reports that might not be in English. Developing similar lists for financial words in these languages would be helpful when using this type of model to take a lexicon-based approach to sentiment analysis for those companies and their reports. These are questions that still need to be answered before this model can be rolled out to its fullest potential.

Sentiment analysis and financial reporting

The fields of finance and accounting have unique terms associated with them which will require sentiment analysis models to be adapted to their linguistic and semantic differences. The following model depicts a generic model for how sentiment analysis can be used in conjunction with scoring sentiment in financial statements (Sousa Maia, 2018). In finance, words like "amend" mean "to change by some formal process" and for a company that frequently has that word in their financial reports it indicates more risk (Wang, Tsai, Liu, & Chang, 2013). The word "deficit" means an excess of some type of liability over assets and is often indicative of a company with higher risk (Wang, Tsai, Liu, & Chang, 2013). By using a financial polarity-lexicon based approach, investors and auditors could score financial statements based on the sentiment found within and investigate how financial sentiments from the financial statements can affect future company performance.

This type of lexicon-based approach to financial reports can also be used by auditors or readers of the financial reports. Narratives are a key information source for financial analysts in determining the health of a company, with most financial analysts surveyed in 2000 by the AIMR indicating that management discussion is very or extremely important part of evaluating firm value (Balakrishnan, Qiu, & Srinivasan, 2010). A research study found that the MD&A section of a 10-K represents the largest proportion of information cited by analysts when doing their analysis on a firm's value (Balakrishnan et al., 2010). As a result, knowing how to apply tools to analyze the MD&A section as well as other sections of a financial report is key to being able to extract the most data possible from the report.

Using a finance-specific lexicon can assist auditors and investors in identifying relationships between risk and sentiment within financial reports as well as help them understand a firm's value (Wang, Tsai, Liu, & Chang, 2013). Using their sentiment list for finance terms, Loughran and McDonald were able to identify a positive and significant effect through regression analysis based off of their list that showed a strong relationship that when there was a higher proportion of negative words within a financial report that there was a greater likelihood of the firm reporting material weaknesses within their internal accounting controls (Loughran & McDonald, 2011). This means that when there are many negative words in a financial report, investors and auditors should be more wary of relying on that firm's accounting measures. A similar study found that within the MD&A section of 10-Ks for firms that had fraudulent disclosures, that those firms tended to use more activation words and incorporated less lexical variety (Ren, Wu, Wang, W., Ge, Wang, G., & Liao, 2013). Also, when a company's risk disclosure section begins to change, it is associated with a change in the stock market (Ren et al., 2013).

Figure 6

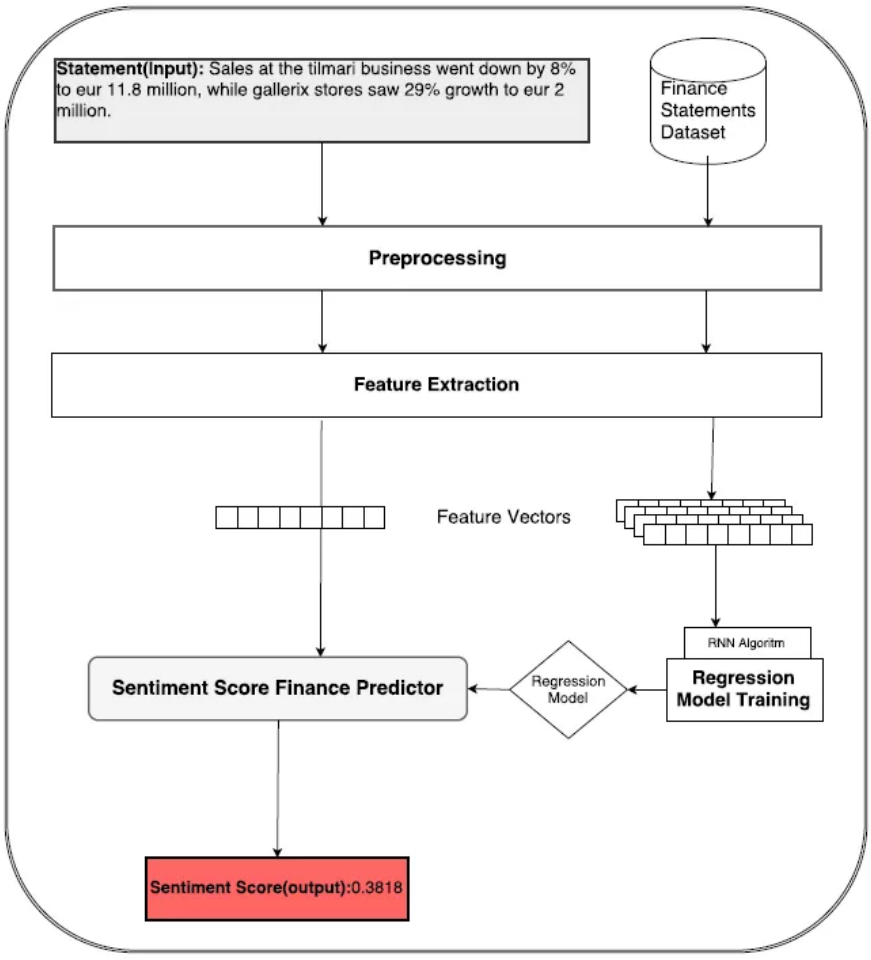


Figure 1: Distant Supervision Model Training Diagram

Using the aforementioned financial lexicon-based polarity model in conjunction with this above diagram (Sousa Maia, 2018), first the text to be analyzed is located. Then the financial lexicon is referenced, and any relevant financial sentiment words or amplifiers are assigned the values they have been assigned within the lexical resource and the other words are assigned neutral values. Then analysis is run on the words using the functions represented in the lexicon model section and based on the results from using the functions and logic of the functions, the final sentiment score is output, showing not only if text is positive or negative, but to what extent the text is positive or negative (-100 most negative to 100 most positive).

This is a powerful tool because it can help investors can use sentiment analysis when it comes to making sense of financial reporting. A rudimentary form of sentiment analysis done in 2006 found that changes in textual data usually precede changes in the financial performance by about one fiscal quarter (Li, 2006). By analyzing 34,180 10-K filings between 1994 and 2005 and counting the frequency of words relating to risk or uncertainty and using it as a risk sentiment proxy, Li was able to find a statistical association between an increase in risk sentiment and lower future earnings. Another study found that when examining the MD&A sections of Form 10-Ks, that the average tone of the MD&A section is positively related to the company's future earnings (Ren et al., 2013). Tone was also important in determining the future ROA of a firm, with the proportion of positive and negative words within an annual report being linked to the future prediction of ROA in a different research study (Davis, Piger, & Sedor, 2006). Lastly, a different group of researchers were able to apply sentiment analysis techniques to analyze corporate financial reports in making a predictive stock portfolio that yielded an average size-adjusted return of 12.16% per year (Balakrishnan et al., 2010).

By using a financial lexicon model in order to do sentiment analysis on financial reports, investors, financial analysts, and auditors can pick up on information not explicitly displayed within the financial reports. Based off of the model, investors, analysts, and auditors can choose text from within the MD&A section of the 10-K or text found throughout the entirety of the 10-K, analyze it to find the sentiment: positive, negative, or neutral; and then determine to what extent the sentiment is positive or negative: score of 100 to -100. Using this information, auditors can begin to piece together how sentiment is changing in certain parts of a financial report and home in on higher risk audit areas. For investors and financial analysts, they can use the information from running sentiment analysis to change their investment strategies. For a company that demonstrates high levels of negativity within their financial reports, investors can begin backing off on their position in those companies as they observe high levels of risk, poor internal controls, and morose future ROA numbers. Or vice versa, if they identify high levels of positivity within their financial reports, investors can increase their position as they identify low risk levels within the company and good future ROA numbers from running sentiment analysis on the company.

Sentiment analysis and financial news

This financial-lexicon model can be used to not only classify words within financial reports, but also in financial news. Financial markets are complex and trading decisions are usually based on a variety of socioeconomic, political, and societal events. Stock markets

provide valuable information on the interplay of many decisions, but often serve as the final record of the actions that agents in the market take. There are not many ways to gain insight into earlier stages of the process, but sentiment analysis is one way to gain insight into how different trends within a market form and develop (Curme, Preis, Stanley, & Moat, 2014).

The efficient market hypothesis asserts that financial market valuations reflect all existing, new, and hidden information with investors acting as rational agents to maximize their profits. This means that stock market prices are largely driven by new information and cannot be predicted since the news is unpredictable and as a result, stock prices will follow a random walk pattern and cannot be predicted with more than 50 percent accuracy (Bollen, Mao & Zeng, 2011). However, numerous studies show that stock market prices do not follow a random walk pattern and other studies show that while news may be unpredictable, it is possible to find early indicators from online sources like social media or the news media (Bollen et al., 2011). The growing field of behavioral finance challenges this idea of the efficient market hypothesis by promoting the importance of behavioral and emotional factors as well as social mood in determining market valuations (Mao, Counts, & Bollen, 2011). News media content has been shown to be important in shaping investor sentiment with high levels of pessimism on Wall Street serving as an indicator of lower market returns the next day (Mao et al., 2011). For individual firms, high negative sentiment has been found to forecast lower firm earnings (Mao et al., 2011).

Behavioral finance incorporating sentiment analysis to analyze the role of human emotion, sentiment, and mood in financial decision-making challenges the efficient market hypothesis and shows signs of being a tool that can be used as a financial market predictor. Researchers analyzing the sentiment behind news article headlines were able to use a Granger causality tests between log returns and sentiment indicators found that in the case of negative news sentiment, there was a statistically significant relationship (both positive and negative) between the log returns and the sentiment behind news headlines (Mao et al., 2011). This means that when sentiment in a collection of news articles is found to be positive or negative, the market will move in a similar direction.

By using a financial lexicon model in order to do sentiment analysis on financial news, investors, financial analysts, and auditors can pick up on information not explicitly displayed within the financial reports. Based off the model, investors and analysts can choose text from within financial news articles, analyze it to find the sentiment: positive, negative, or neutral; and then determine to what extent the sentiment is positive or negative: a score of 100 to -100. Using this information, investors and financial analysts can use the information from running sentiment analysis to change their investment strategies. For a company that has negative financial news related to it from headlines or news articles, analysts can identify the relationship between that and determine that the company's future earnings will probably decline in the future and back off on their position. Or vice versa, analysts can identify positivity within financial news articles for a company or a headline related to a company and identify a strong future earnings trend and increase their position within the company. Lastly, analysts and investors can use this sentiment

analysis model to analyze news articles and headlines to determine the overall health of the market and begin to make decisions from that.

Conclusion

Sentiment analysis is a branch of Natural Language Processing that has a wide variety of applications to the fields of finance and accounting. A lexicon-based approach to sentiment analysis entails creating or using a lexicon of terms, assigning sentiment values to each of the terms, and then scoring the text that those terms appear in and evaluating them to determine the overall sentiment of the text. Using the financial lexicon-based approach model described within this paper provides the potential to drill down to even more granular sentiment levels within text. When it comes to analyzing financial news or reports, researchers or other model users would be able to analyze sentiment not just in terms of positive or negative, but identify to what extent text is positive or neutral. This information could then be applied to financial news to determine the health of the overall market or companies' future earnings potential. This information could be applied to financial reports to determine high risk areas within a company to spend more time auditing, the value of a company's stock, future ROA numbers, and internal controls risks.

To continue moving this project along, more research needs to be done to look into how to best assign scores on a scale to depict different levels of sentiment for the words found in the financial lexicon used with this model. Some words typically indicate greater levels of negative or positive sentiment than others, but this is not always the case. Sometimes a word might be extremely positive in one case, while it is only slightly positive in another case. As a result, many financial reports would need to be analyzed to determine the typical type of sentiment score that could be associated with a word so that when it is used in this model, a fair value is assigned to it. Also, it would be useful to continue exploring the effects of sentiment analysis and its application to financial news, reports, and its application towards social media in determining the changes that can take place in a market and how it might be possible to identify these changes early on. Lastly, it would be helpful to further analyze how something like sarcasm might affect the results of a lexicon-based approach to sentiment analysis.

Reference Page

Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3), 789–801. https://doi.org/10.1016/j.ejor.2009.06.023

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

Bouchaud, J.-P. (2008). Economics needs a scientific revolution. *Nature*, 455, 1181.

Curme, C., Preis, T., Stanley, HE., & Moat, HS. (2014). Quantifying the semantics of search behavior before stock market moves. *Proc National Academy of Sciences*, 111(32), 11600–11605.

Davis, A., Piger, J., & Sedor, L. (2006). Beyond the numbers: an analysis of optimistic and pessimistic language in earnings press releases. Working paper, Federal Reserve Bank of St. Louis, 1-42.

Domanska, O. (2018). Using sentiment analysis for gaining actionable insights. CoreValue, 1-11.

Jurek, A., Mulvenna, M., & Bi, Y (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics* 4.1, 1-13.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan, 1-54.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv.org, *Quantitative Finance Papers*, 1112.1051.

Ren, J., Wu, X., Wang, W., Ge, H., Wang, G., Liao, S. (2013). Effective sentiment analysis of corporate financial reports. Thirty Fourth International Conference on Information Systems, Milan, 1-9.

Sousa Maia, M. (2018). Sentiment classification about finance statements. Retrieved from https://ssix-project.eu/sentiment-classification-about-finance-statements/

Sung, S., Cho, H., & Ryu, D. (2019). The behavior of an institutional investor with arbitrage opportunities and liquidity risk. *Emerging Markets Finance & Trade*, 55(1), 1–12. https://doi.org/10.1080/1540496X.2018.1498333

Wang, C.J., Tsai, M.F., Liu, T., Chang, C.T. (2013). Financial sentiment analysis for risk prediction. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, 802-808.

Wang, J., Xie, Z., Li, Q., Tan, J., Xing, R., Chen, Y., & Wu, F. (2019). Effect of digitalized rumor clarification on stock markets. *Emerging Markets Finance & Trade,* 55(2), 450–474. https://doi.org/10.1080/1540496X.2018.1534683

Supplemental Reference Page

These sources were not cited within this paper, but were valuable in building a knowledge base over the subjects contained within and could prove a valuable jumping off point for future researchers in addition to the "Reference Page".

Agaian, S. & Kolm, P. (2017). Financial sentiment analysis using machine learning techniques. *International Journal of Investment Management and Financial Innovations*, 3(1), 1-9.

Devitt, A. & Ahmad, K. (2007). Sentiment analysis in financial news: A cohesionbased approach. *Proceedings of the Association for Computational Linguistics*, 984–991.

Kolchyna, O., Souza, T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: lexicon method, machine learning method and their combination. [Online]. Available: http://arxiv.org/abs/1507.00955.

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

Moreno-Ortiz, A & Fernandez-Cruz, J. (2015). Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. *Social and Behavioral Sciences*, 198, 330-338.

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

Tetlock, P. C. (2007). Giving content to investor sentiment: the role of media in the stock market. *The Journal of Finance*, 62, 1139-1168. .