USING NLP TO QUANTIFY THE EFFECTS OF NON-GAAP MEASURES TO PREDICT THE OUTCOME OF SECURITIES LAWSUITS

by

Stacey Dianne Taylor

Submitted in partial fulfillment of the requirements
for the degree of Master of Electronic Commerce

at

Dalhousie University
Halifax, Nova Scotia
December 2019

Table of Contents

List of Tables

List of Figures

Abstract

The Management Discussion and Analysis (MD&A) is arguably the most tonal section of the reports provided to the U.S. Securities and Exchange Commission. As part of that dialogue, companies use non-standardized financial metrics known as non-GAAP measures that do not conform with Generally Accepted Accounting Principles. Our research presents a novel extractive approach using Sentiment Analysis to measure the impact that non-GAAP measures have on the common investor versus those who are financially savvy. We find that sentiment declines once the non-GAAP sentences have been extracted with a statistical significance at the p=0.01 level. Building on this, our second research question investigated if we could use a similar approach with machine learning to predict the outcome of securities class action lawsuits. We find that we are able to predict the aggregate outcome of the lawsuits with a recall of 0.9142 using the Random Forest classifier.

Glossary

10-K/A A form required by the U.S. Securities and Exchange Commission for all amendments made to a company's 10-K form.

10-K405 Annual report on the performance and health of a registered company required by the U.S. Securities and Exchange Commission where one (or more) of its directors failed to file Form 4 on time that is no longer in use.

10-KSB Annual report on the performance and health of a registered small company required by the U.S. Securities and Exchange Commission that is no longer in use.

10e Under Regulation SK, 10(e) prohibits the smoothing of items through the adjustments to non-GAAP measures.

10-K Annual report on the performance and health of the company required by the U.S. Securities and Exchange Commission.

10-Q Quarterly report on the performance and health of the company required by the U.S. Securities and Exchange Commission.

Accuracy A machine learning measure calculated as the summation of true positives and true negatives divided by the total number of samples.

Adjusted Earnings Per Share Earnings Per Share, a specified calculation in accordance with the Generally Accepted Accounting Principles, which is then further adjusted by Management, rendering it a non-GAAP measure.

Annual Report Report legally required to be provided to shareholders detailing a public company's financial health and activities.

Bag of Semantic Orientation Similar to the bag-of-words model, it is a "bag" of semantic orientation (i.e. positive, negative, or neutral) with no consideration of representative order or placement in the text.

Balance Sheet One of the four main financial statements which details assets, liabilities, and owners' (shareholders') equity at a specified date during an accounting period.

Bank Stabilization Act The 2008 bank bailout in the United States.

Class period The alleged damage period.

Core Earnings A non-GAAP measure of earnings that also considers all non-recurring earnings and expenses.

Court of Cassation Supreme Court of France.

Deep Learning Machine Learning that uses algorithms capable of learning independently of labelled data.

Diction A lexical library used in machine learning.

Discovery An exchange of information by all parties involved in legal proceedings.

Earnings before Interest A non-GAAP measure that (usually) adds back interest, tax, depreciation, amortization, and restructuring or rent costs to net income in order to provide a proxy for operating profitability.

Earnings before Interest and Tax A non-GAAP measure that provides a proxy for a firm's profitability before interest and taxes have been considered.

Earnings Before Interest A non-GAAP measure that (usually) adds back interest, tax, depreciation and amortization to net income in order to provide a proxy for operating profitability.

Electronic Document Gathering A system used by the U.S. Securities and Exchange Commission to receive, catalogue, and maintain required filings by registered companies.

F1 Measure A machine learning measure calculated as two times the product of precision and recall divided by the summation of precision and recall.

Financial Accounting Standards Board Independent board that is responsible for the development and maintenance of the U.S. Generally Accepted Accounting Standards.

Free Cash Flow A non-GAAP measures that (usually) is the cash remaining after supporting operations and maintaining capital assets.

Funds From Operations A non-GAAP measure that provides the cash flow generated by operations, most often used by real estate companies or real estate investment trusts.

General Inquirer Dictionary A psychological dictionary that lists positive and negative words.

Hyperplane A machine learning tool that separates data in n-dimensional space.

Income Statement One of the four main financial statements which details revenues and expenses, and provides Net Income, over an accounting period.

Institutional Brokers Estimate System (I/B/E/S) A database maintained by Thomson Reuters of historical earnings estimates for most publicly traded companies.

International Accounting Standards Board (IASB) Independent board that is responsible for the development and maintenance of the International Financial Reporting Standards (IFRS).

International Financial Reporting Standards (IFRS) The reporting framework used by foreign registered companies that conforms with Generally Accepted Accounting Principles.

Jargon Domain specific terminology.

K-Fold Cross-validation technique where "K" defines the number of folds to be used.

K-Means A machine learning technique that clusters data using unsupervised learning.

Latent Dirchlet Allocation An algorithm that identified topics of documents.

Lexicon Vocabulary of words which can be general or domain specific.

Libraries Pre-compiled resources that are available to programmers to be called in programs.

Loughran-McDonald Dictionary The most comprehensive dictionary created for the financial domain that uses semantic orientations beyond positive, negative, and neutral, to include other categories such as litigious.

Multivariate Regression Used to determine the degree at which independent and dependent variables are linearly related.

Naive Bayes A machine learning algorithm based on Bayes' Theorem which uses the "naive" assumption of variable independence.

Neural Network A deep learning algorithm modelled after the human brain.

P-Value Statistcal hypothesis test to determine if results are achieved by chance alone.

Precision A machine learning measure calculated as true positives divided by the summation of true positives and false positives.

Quantitative Discourse Analysis Package (QDAP) A collection of dictionaries in R.

Recall A machine learning measure calculated as true positives divided by the summation of true positives and false negatives.

Re-Sampling A technique that continually re-draws samples from data.

Regulation SK A rule mandated by the U.S. Securities and Exchange Commission which, in part, governs the use of non-GAAP measures.

Regulation G A rule mandated by the U.S. Securities and Exchange Commission relating to the presentation of non-GAAP measures as a result of the Sarbanes-Oxley Act (SOX) of 2002.

Return on Capital Employed A non-GAAP measure that provides the ratio of operating profit and capital employed.

Revised Net Income Net Income, a specified calculation in accordance with the Generally Accepted Accounting Principles, which is then further adjusted by Management, rendering it a non-GAAP measure.

Rule 10b(5) A U.S. Securities and Exchange Commission rule which addresses deception and making false statements, among other things.

Sentiment Analysis A Natural Language Processing Task that determines the semantic orientation of the text as either positive, negative, or neutral..

Statement of Shareholders' Equity One of the four main financial statements which shows the changes in Shareholders' Equity over an accounting period.

Statement of Cash Flows One of the four main financial statements which details how changes in the Balance Sheet and Income Statement affect cash over an accounting period.

Statistical Machine Learning Machine Learning that uses algorithms developed using statistical analysis.

Summary Judgment A judgment entered in by the court, either in favour of the plaintiff or the defendant, without conducting a full trial.

Support Vector Machines Algorithm that finds the best separation between classes using a hyperplane.

Tokenization Separating textual documents down into "tokens" which can be done on the word level or sentence level.

Tukey A statistical method used to create confidence intervals for paired differences.

Unbilled Revenue A non-GAAP measure that represents accrued revenue that has not been billed to customers, and is, therefore, not reflected in the financial statements.

US GAAP The reporting framework used by registered domestic companies (or foreign registered companies that elect to use this framework) that conforms with Generally Accepted Accounting Principles.

Wilcoxon Signed Rank Test Statistical test used for paired differences where the distribution is non-parametric.

Word Co-Occurrence The likelihood that two words will appear in order together in a document.

List of Abbreviations and Symbols

Abbreviations

EBIT Earnings Before Interest and Tax

EDITDA Earnings Before Interest, Tax, Depreciation, and Amortization

EBITDAR Earnings Before Interest, Tax, Depreciation, Amortization and Rent/Restructuring

EPS Earnings Per Share

FASB Financial Accounting Standards Board

FCF Free Cash Flow

GAAP Generally Accepted Accounting Principles

GI Harvard-IV General Inquirer Dictionary

IASB International Accounting Standards Board

I/B/E/S Institutional Brokers Estimate System

IFRS International Financial Reporting Standards

MD&A Management Discussion and Analysis

Symbols

$\geq$ Greater than or equal to

$\leq$ Less than or equal to

$>$ Greater than

$\overline{X}$ X Bar (for sample mean)

μ Mu (for population mean)

Acknowledgements

Thank you to my family, particularly my sister Kylie for her unwavering support.

I would also like to thank the DNLP lab — your friendship, support and feedback mean more to me than you will ever know.

To Dijana Kosmajac and Colin Conrad — thank you for the laughs and the craziness. I could not have made it here without you two!

Without Dr. Jacek Wołkowicz, I would never have discovered Natural Language Processing, and subsequently changed the direction of my career. Thank you for your guidance and inspiration.

Finally, I wish to sincerely thank my supervisor Dr. Vlado Kešelj for everything. You are an unparalleled mentor.

Chapter 1

Introduction

Applying sentiment analysis to the financial domain is not new — in fact, there is an existing body of research that has done just this, evidenced by numerous articles and The Handbook of Sentiment Analysis in Finance [65]. In most cases, however, the overriding focus of the research has had the end goal of improving stock prediction, through an analysis of the available textual resources such as financial filings, annual reports to shareholders, earnings call transcripts, press releases, press coverage, or analyst predictions. Being able to reasonably predict stock movement and pricing has important ramifications such as shareholder wealth, which is the main goal of every for-profit company, yet, this should not be the main focal application of sentiment analysis. While there will always be sentiment analysis research interested in tying text to stock prediction, the uses and interest of sentiment analysis are spreading to other domains. In recent years, applying sentiment analysis to the legal domain has surged. It has been applied to the areas such as contracts, to better understand case law, and to predict case judgments.

In the area of finance, traditionally, the bag-of-words (BOW) approach has been used, where the textual resource is chopped up into individual words, with the sentiment analysis done using pre-formed dictionaries that have already identified words as negative, positive, or neutral. Scoring can vary between tools, as some tools will compare positive (negative) words against each other in each category, to determine if one word is more positive (negative) than the other, and thereby give it a higher or lower score than the other. However, the general premise of the BOW approach is to add up all of the positive and negative words and average those over the number of words in the bag. The resulting score provides how positive, negative, or neutral the document is. The range of scores usually ranges from -1 to 1, where -1 represents a document of entirely negative words, and similarly, 1 is a document of entirely positive words. Generally, 'extreme" documents do not exist under normal conditions,

meaning that, typically, scores are not -1 or 1, but usually values in-between.

Natural language — human language — is very complex and has evolved to include hidden elements such as nuance, ambiguity, sarcasm, and misdirection, to name a few. Humans often struggle with these elements of language, which makes it doubly challenging to teach a computer to understand these elements. In all of these examples, research is still working on finding compatible solutions to help the computer to *learn* these elements. Although the phrase that "Computers are incredible fast, accurate, and stupid." is usually attributed to Albert Einstein, who actually never said that, the phrase is, nevertheless, true. Computers are excellent processing machines, but we, as programmers, struggle on how to instruct the computer how to process these hidden elements. Language is very fluid and changes over time, and at varying speeds. If programmable language rules never changed, creating code to instruct the computer would be a hefty, but relatively straightforward task.

Take the colloquial use of the word "wicked" for example. Traditionally, this word means *bad*, but the Merriam-Webster dictionary also indicates that it was also used to describe clever, conniving, extremely, and cool [64]. Already, it can be seen that the different uses of this word presents a problem, as its usage will need to be contextualized in the text to understand which version is being used. This shows just how important contextualization is, which strongly suggests that a BOW approach is not optimal. Prose is structured in such as way that there are main words and supporting words. Using the phrase "John rode his bicycle to school today" means nothing if we take *John* out. This means that the words "rode his bicycle to school today" are all connected, and provide contextualization for John. Yet, if we were to take a traditional BOW approach to this phrase, we would chop it up into individual words, unconcerned with contextualization, main, or supporting words.

In finance, contextualization is extremely important. Conducting sentiment analysis without proper context will lead to poor and misleading sentiment scoring. We illustrate this point using the phrase "In the past year, we have increased our debt by $2 million in order to make investments in equipment." — the main focus is the debt. It is important that users of this financial information are provided the context to this debt, which, in this case is that the increase in the debt is $2 million and that

it is being used to invest in equipment. This provides valuable information to the financial user to understand what the company has done, and why it has done it.

Companies registered with the United States (U.S.) Securities and Exchange Commission (SEC) are required to submit quarterly and annual filings, both of which contain a section called *Management's Discussion and Analysis of Financial Condition and Results of Operation* (MD&A). It is arguably the most tonal section of the filing as it is where the company will speak directly to shareholders and stakeholders alike on the past performance of the company, and what it expects to occur in the near future. While the SEC has provided rules of information that must be disclosed [89], disclosure beyond that is at the discretion of the company, which affords considerably licence to discuss other information. It is these additional disclosures that our research is interested in.

One common form for these additional disclosures is non-GAAP measures (NGM), where companies discuss measures that do not conform to Generally Accepted Accounting Principles (GAAP). We have conducted extensive research on these non-GAAP measures included in the MD&A, by extracting a pre-defined list of NGMs to see how that changes the sentiment of SEC filings. Our method is straightforward in that we measure the sentiment of the original document as filed with the SEC, we extract the non-GAAP measures, along with the rest of the text contained in the sentences that contain the NGM, and then re-measure the sentiment after extraction. We performed this experiment in several different contexts. The results indicate that extracting the non-GAAP measures creates a significant drop in our sample set, and that drop can be used to predict the outcome of our sample of Securities Class Action lawsuits.

1.1 Motivation

Language is very important; it conveys more than simply the words on the page. Like actors in many domains, companies choose their words very carefully when communicating with stakeholders, or potential stakeholders. Each year, companies registered with the SEC are required to submit regulatory filings that convey the financial and operational health of their company. In doing so, companies often follow

an accepted industry practice of discussing financial measures that are outside of the accounting rules set out in the Generally Accepted Accounting Principles. This made us question why, given that the calculation of these measures are not regulated, nor are they auditable. For many years, there has been significant concern over the use of these measures, and has divided the literature and researchers alike into two camps: those in favour of the use of the unregulated measures who indicate that they provide pertinent additional information, and those who have demonstrated that these types of measures which are outside of the rules are predatory. Our motivation here was to determine, using sentiment analysis, if we could quantitatively determine the effect that these unregulated rules have on financial reports filed with the SEC.

In line with understanding the quantitative effect, we also wanted to determine how financial domain-specific dictionaries, versus all-purpose dictionaries, would handle these unregulated measures. The goal here was to use the dictionaries (two of each) as proxies for how investors (those with significant financial experience versus the average investor) would interpret the tone of reports with unregulated rules versus those without. We see this as an important step to learning how to better protect the average investor from making poor decision based on measures that can easily obfuscate the information presented.

Finally, along similar lines, we wanted to see if this average investor versus seasoned investor approach could be used for Securities Class Action Lawsuits, which are usually driven by the gap between actual performance and corporate language hype. The focus here was, again, to help protect the average investor by helping companies understand what language could prompt a Securities Class Action lawsuit. Regardless of outcome of the lawsuit, investors always lose. Not only is the company having to spend money on defending the lawsuit when it could have been using that money to build shareholder wealth, even if investors win, they are rarely fully compensated for the loss they have suffered in the share price [71].

Undoubtedly, regulations have improved significantly since the first stock market crash in 1929, but there is still work to be done to protect the average investor. Our work aims to contribute to this ongoing effort.

1.2 Contributions

We believe that our research provides three main contributions:

1. Our research brings together four different, but highly related fields: accounting, finance, law, and artificial intelligence.

2. To our knowledge, this use of Natural Language Processing, in particular the extractive approach to the change in the Sentiment Analysis of the non-GAAP measures in financial reports has not been done before. Again, to our knowledge, using this approach in potential lawsuit classification has also not been done before. As such, our approach and findings have tremendous cross-domain implications and potential, and open up a new area of research in each of these domains.

3. Although the various inputs for our dataset, with the exception of the actual sentiment scoring, are publicly available, they have not been collated into one dataset. Therefore, we see the dataset itself as a contribution to the research in this field.

1.3 Practical Benefits

There are two main groups that this research benefits: investors who are not considered finance professionals (which we term the *common* investor and companies who prepare the financial filings. For clarity, we define the term *finance professional* in our research broadly to include professional investors, investment and financial analysts, designated accounts (under various designations around the world), and Chartered Financial Analysts. We also include those with no financial training but who have significant experience in finance and the market, as we recognize that experience and knowledge can be commensurate with training in certain cases.

The common investor needs protection. Accounting and finance rules, investing regulations, and regulatory frameworks are complex — so much so that the SEC has had to establish a "Plain English Initiative" to help the general public understand disclosures made in all of its documents, which includes the financial filings submitted

by corporations [76]. Even with that rule in place, the regulations and laws that underpin investing are still a challenge — the non-GAAP measures are a perfect example of this. As will be seen, in certain cases certain non-GAAP measures are considered to be non-GAAP whereas in other situations, those same measures are considered to be **not** non-GAAP measures, to borrow the wording of the SEC [79]. In reviewing the current research available, it does not seem that these differences have been considered in the research thus far. So, in this particular case, the common investor would have to wade through all of the regulations, tie it back to the accounting rules, understand the implications, and then make their financial decisions.

On the other hand, it could be argued that because investing is so complex and has such large financial consequences if things go wrong, that investing should not be for the common investor to undertake themselves, but rather only professionals should be involved. Fundamentally, this goes against the spirit of the market, which is open to everyone — win, lose, or draw — and has been this way in the United States when the very first U.S. stock exchange, the Philadelphia Stock Exchange (PHLX) [68], was established. While the average investor will (likely) not be as financially sophisticated as a professional investor, that should not mean that investing is closed off to those who are not savvy.

The first stock market crash of 1929 highlighted the fact that oversight and investor (and stakeholder) protection was critical to economic stability [75]. Leading up to 1929, the market saw approximately $50 billion in investments in new securities, half of which were deemed to be worthless after the crash. Needless to say, this was a significant contributing factor to the *Great Depression* [75]. That $50 billion in 2008 dollars (the year that Lehman Brothers collapsed) equates to approximately $629 billion. Comparatively, the Bank Stabilization Act of 2008 for the bank bailout was $700 billion. This means that the *Great Crash* was only $70 million short of the 2008 amount, but with no Federal Government bailout. Each crash, scandal, and bailout tests the economic stability of the stock market, and has real-life consequences on those who have invested in the market — either on their own, or through a professional investor. As such, it is important to protect the common investor, as that will help protect all investors, thereby helping to shore up the stability of the markets.

Companies who prepare these filing will also benefit from our research. As will

be seen, Securities Class Action lawsuits are extremely expensive. In our sample set alone, the settlements amounted to in excess of $1.6 billion dollars over forty eight companies (the half of our sample set whose lawsuits were settled). While the settlement amounts were, by no means, the same among the companies which settled, if the average is taken, it amounts to approximately $33 million per company. The highest settlement amount recorded was $410 million dollars. These, however, are just the settlement cost, and does not include the legal fees for the defendants (i.e., the companies who settled). Normally, the legal fees for the plaintiffs are included in the settlement. It should also be pointed out that the investors are rarely, if ever, fully compensated for the damage to their share value [71]. This also means that in the full amount of $1.6 billion, the legal fees for both the plaintiffs and defendants of the companies whose lawsuits were dismissed have not been included, as that information was not publicly available.

We know from the previous research [96] that investors are affected by the sentiment and tone of words. This is also further supported by the research of Rogers *et al.* [71], as they point to specific wording from financial disclosures that does not agree with the company's performance, which has instigated the litigation. It would be of great value for companies to know whether the language that they are using in their financial disclosures are putting them at greater risk of litigation or not.

1.4 Outline

This rest of the thesis is structured in the following way:

Chapter 2: First, we present an overview of the accounting rules and framework. We then discuss the accounting and regulatory rules related to the financial filings submitted to the SEC, and then finish the background with a discussion of non-GAAP measures. We then present the related work for sentiment analysis in the financial and machine learning domains.

Chapter 3: In this chapter, we present our research methodology where we will discuss the data collection, pre-processing, the non-GAAP measures selected for extraction, as well as explain in more detail how the extractive process is done. We will also discuss the dictionaries used to conduct the sentiment analysis, and how the

sentiment analysis itself was conducted.

To address our first research question regarding the overall sentiment for, we compiled a dataset of 10,000 company reports, comprising annual and quarterly filings over a period of 25 years. Using that dataset, we conducted sentiment analysis using four dictionaries from the *Sentiment Analysis* library in R: General Inquirer, QDAP, Henry and Loughran-McDonald. The first two dictionaries are general purpose which were used as proxies for the average investor. The latter two are finance domain specific which were used as proxies for the financially sophisticated investor. To determine the aggregate sentiment for each dictionary, we summed all of the sentiment scores calculated for each individual filing document. The aggregate results for each dictionary was negative, ranging (in aggregate) from -27.13332 to -0.81217.

For our second research question which focuses on using the sentiment measured for each filing document, we determined that we are able to use Natural Language Processing methods to predict the outcome of Securities Class Action lawsuits. Using data in the Stanford Class Action Clearinghouse (SCAC), we selected 96 random lawsuits promulgated under Rule 10b — three from each of the top and bottom three sectors as listed on the SCAC heat map, classified from most to least sued.

Chapter 4: Here, we will provide a discussion on the experiments conducted for each of the following research questions:

1. Given a sample of 10,000 company filings, what is the overall sentiment under each dictionary used?

2. Are we able to use the change in sentiment to predict the outcome of a sample of lawsuits in the top three sectors, in Stanford Class Action Clearinghouse heat map?

3. Similar to the second question, are we able to predict the outcome of a sample of the lawsuits in the bottom three sectors in Stanford's Class Action Clearinghouse heat map?

Chapter 5: In this chapter, we will discuss the results of our experiments. Overall, we find that the aggregate sentiment of all dictionaries decrease over our sample

of 10,000. We also find that, using Recall as our main measure, we are able to predict the outcome of Securities lawsuits to a recall of 0.9142 using Random Forest.

Chapter 6: Here, we will provide our concluding remarks, and briefly discuss future work brought about from our research.

Chapter 2

Background and Related Works

2.1 Generally Accepted Accounting Principles

In an effort to restore investor confidence and to help recover from the *Great Depression*, the United States Congress (Congress) enacted a law in 1933 that created the United States (U.S.) Securities and Exchange Commission (SEC). This law gave the SEC broad regulatory powers over the U.S.' stock exchanges. Public companies wishing to trade in the U.S. market must register with the SEC [81]. In addition, *Section 12* of the *Exchange Act* also indicates that under certain conditions, private companies must also register [82].

Registered companies must use Generally Accepted Accounting Principles (GAAP) to maintain their accounting records and prepare their financial statements. Currently, there are two forms of acceptable GAAP in the United States: U.S. GAAP and International Financial Reporting Standards (IFRS). Domestic companies must use U.S. GAAP as their framework, whereas international companies can either use IFRS or U.S. GAAP [34, 52, 89]. There is also a requirement that, annually, financial statements must be audited and signed off by a qualified Certified Public Accountant (CPA). As part of that, auditors also provide an opinion which indicates if the financial statements are fair representations (or not) of the financial activities of the entity during the year, and that the accounting records and resulting financial statements were prepared in accordance with Generally Accepted Accounting Principles [13].

GAAP exists for a reason. It is the rules on which accounting is built, providing guidance on how a company should record its *transactional* life. Everything that flows through the business will eventually become an input to the financial statements, disclosures in the notes to the financial statements, or both. The main financial statements are the Balance Sheet, Income Statement, Statement of Cash Flows, and the Statement of Shareholders' Equity [77]. The Balance Sheet provides information

on a company's assets, liabilities, and shareholders' equity; the Income Statement indicates how much revenue and expenses the company made during their fiscal year, and also shows the Earnings Per Share; the Statement of Cash Flows details the inflows and outflows of cash; and finally, the Statement of Shareholders' Equity details the changes in shareholders' equity over the fiscal year.

The accounting rules serve to provide a common and accepted approach to maintaining the financial records, and, ultimately, the creation of the financial statements. This framework is very important as it is one of the main conduits that companies have to communicate to stakeholders [36]. These rules enable companies to communicate financial information in a logical, organized, and regulated way. GAAP also makes a company's statements comparable from one fiscal year to the next, as well as comparable from one company to the next. This is useful to investors in a number of ways such as determining if one company is a better investment than another, or if one company has competitive advantage over another. This information is also useful for the company, as it can benchmark against its own prior performance, as well as compare itself to competitors in the industry, to determine if the company is achieving its financial and strategic objectives (or not).

Using this common approach also makes certain financial information auditable because GAAP stipulates how the inputs are to be recorded and how the lines of the financial statements are to be calculated. One of the most, if the not *the* most, focal points of the financial statements is the figure *Net Income*. It is the bottom line number on every Income Statement, and, at the end of the year, represents the income that every company has made — regardless of industry, sector, geography, or size. And because it is a GAAP figure, it can be relied on to mean and represent the same calculation, giving comfort to financial statement users that, when they examine and compare *Net Income*, they are using comparable figures.

2.2 SEC Filings

At minimum, registered companies are required to submit four filings per year to the SEC — three quarterly filing and one annual filing. As previously stated, domestic companies are required to use U.S. GAAP as their accounting framework. As such

they will use form 10-Q for their quarterly filings and form 10-K for their annual filings. Foreign companies who elect to use U.S. GAAP as their framework will also use the 10-Q and 10-K forms [80].

One of the main sections of the Q and K filings is called *Management's Discussion and Analysis of Financial Condition and Results of Operation* (although it is also sometimes referred to as *Management's Discussion and Analysis or Plan of Operation*, among other names). Often, it is shortened in Industry to "MD&A".

This is considered one of the main textual components of the SEC filing as it is management's narrative and opportunity to speak directly to the stakeholder on how the company has performed financially in the last three years and how the company intends to remedy any issues and/or concerns going forward. The SEC provides guidance as to what topics must be addressed in the MD&A such as Liquidity and Capital Resources, Results of Operations, and Off-Balance Sheet Arrangements, for example [78]. The SEC also indicates that the discussion should be based on the financial statements [88], which also implies that it should be a GAAP-based discussion.

Yet, the SEC does not preclude a company from discussing other topics it deems necessary to, in its own words, fully address the company's results. This means that, beyond addressing the required topics, any further discussion is voluntary, and can provide valuable information (interpreted as both *informative* and *predatory* — both of which are discussed later) on how management views the company. This latitude in disclosure also affords the company the ability to discuss items that fall outside of the GAAP framework. Known as "non-GAAP measures" (NGM), these items start with a GAAP item on the financial statements and is then adjusted by management [83], by either adding back or subtracting amounts. Two examples of common Non-GAAP Measures are Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) and Free Cash Flow (FCF). Each is discussed below along with an example for clarification.

EBITDA starts with Net Income on the Income Statement (a GAAP measure) and (usually) adds back interest, taxes, depreciation, and amortization, to arrive at Earnings before the named deductions were taken. In its 10-K filing, Medco Health Solutions, Inc. (Medco) indicates that it calculates EBITDA as "earnings before

interest income/expense, taxes, depreciation, and amortization" [63], which is in-line with the usual approach.

The Consolidated Statement of Income for Medco Health Solutions, Inc has been presented below in Figure 2.1. This is one the four main financial statements, which culminates in the GAAP bottom line figure of *Net Income*.

As of and for Fiscal Years Ended	December 31, 2011(1)(2)(3)
Consolidated statement of income data:	
Total product net revenues(4)	$ 68,563.3
Total service net revenues	1,500.0
Total net revenues(4)	70,063.3
Cost of operations:	
Cost of product net revenues(4)	64,919.0
Cost of service revenues	522.1
Total cost of revenues(4)	65,441.1
Selling, general and administrative expenses	1,744.7
Amortization of intangibles	291.9
Interest expense	208.5
Interest (income) and other (income) expense, net	1.3
Total costs and expenses	67,687.5
Income before provision for income taxes	2,375.8
Provision for income taxes(9) (d)	920.1
Net income	$ 1,455.7

Figure 2.1: Medco Health Solutions, Inc 2011 Consolidated Income Statement

From there, based on Medco's definition of EBITDA, *Net Income* is adjusted to add back interest, taxes, depreciation, and amortization, as seen below in Figure 2.2. Medco's calculation is the most common approach to EBITDA.

Free Cash Flow (usually) starts with Cash from Operations on the Cash Flow Statement (which is another of the four mandatory financial statements) and subtracts Capital Expenditures [30]. *Net cash provided by operating activities* is one of the major line items on the Cash Flow Statement and is a GAAP figure.

As per the notes in the 10-K, Facebook defines Free Cash Flow as "net cash

For Fiscal Years Ended	December 31, 2011(a) (b)
Net income	$ 1,455.7
Add:	
Interest expense	208.5
Interest (income) and other (income) expense, net	1.3
Provision for income taxes	920.1 (d)
Depreciation expense	213.1
Amortization expense	291.9
EBITDA	$ 3,090.6

Figure 2.2: Medco Health Solutions, Inc 2011 EBITDA

FACEBOOK, INC.
CONSOLIDATED STATEMENTS OF CASH FLOWS
(In millions)

	2018
Cash flows from operating activities	
Net income	$ 22,112
Adjustments to reconcile net income to net cash provided by operating activities:	
Depreciation and amortization	4,315
Share-based compensation	4,152
Deferred income taxes	286
Other	(64)
Changes in assets and liabilities:	
Accounts receivable	(1,892)
Prepaid expenses and other current assets	(690)
Other assets	(159)
Accounts payable	221
Partners payable	157
Accrued expenses and other current liabilities	1,417
Deferred revenue and deposits	53
Other liabilities	(634)
Net cash provided by operating activities	29,274

Figure 2.3: Facebook 2018 Cash Flow Statement (Partial)

provided by operating activities reduced by net purchases of property and equipment" [30]. This means that Facebook is starting with the GAAP figure of *Net cash provided by operating activities* (boxed in red) and then adjusting it by reducing the *Net cash* by net purchases of property and equipment, seen in Figures 2.3 and 2.4. Once management adjusted the GAAP figure, it becomes a non-GAAP measure.

	2018
Net cash provided by operating activities	$ 29,274
Purchases of property and equipment, net	(13,915)
Free cash flow	$ 15,359

Figure 2.4: Facebook 2018 Free Cash Flow

We have qualified both of these examples using the word "usually". Although these calculations generally follow a common approach, there is nothing preventing management from taking a different approach. Unlike GAAP measures, NGM are not regulated as vigorously. The only requirements that the SEC has for using non-GAAP measures (under Regulation G and Regulation S-K 10(e)) is that [53]:

- they cannot appear on the financial statements themselves, or in the notes that accompany the financial statements;
- a reconciliation must be provided that presents it in comparison to the closest GAAP measure (this can either be done through a textual formula or discussion, or through a side-by-side reconciliation);
- disclosure on why management has included the particular NGM in its discussion and how it will add value to investor's understanding of the company;
- that the non-GAAP measure not be misleading; AND
- it cannot be shown in greater prominence than its GAAP counterpart.

As such, there are no rules to limit the discussion of NGMs to a certain percentage of the discussion, prevent management from making their own further adjustments to the measures, or stop the company from changing the calculations from one year to the next — even if there is a (mostly) accepted approach to a majority of NGMs

that fall outside of the GAAP framework. The ramifications of this mean that it is quite possible that the NGMs will not be comparable from one year to the next.

This also means that there is nothing preventing management from creating their own non-GAAP measures either. As long as companies follow the presentation rules, they are free to discuss *anything*. As the extant literature has noted, NGMs have become "commonplace" [9], and are the norm now, rather than the exception. As will be addressed later in the literature review, while this is not surprising given the fact that the majority of companies use these measures, it does raise some very important questions that stakeholders must seriously consider as to the quality of information that they are being provided.

Some non-GAAP measures also pose a particular problem in that in some cases they are considered by the SEC to be non-GAAP, but in others, they are considered to be **not** non-GAAP (to borrow the wording of the SEC), thus requiring contextualization of the non-GAAP measure in question. In reviewing the industry literature by the top accounting firms, we have noted two major points regarding NGM that has a direct effect on our research. Firstly, under rule Accounting Standards Codification (ASC) 280, companies are required to disclose information on items such total assets, revenue, as well as profit and loss [11]. Whether disclosed separately, or as part of a single measure, these figures could be discussed as part of the MD&A if considered integral to the understanding of the business [84]. From our research, it appears that companies who are required to report segment information are using *Adjusted EBITDA*. Outside of segment reporting, however, *Adjusted EBITDA* would be considered a non-GAAP measure. Secondly, measures can also be considered **not** non-GAAP if they are metrics, used for, or required by the Government [84]. While this does not *absolve* every non-GAAP measure that is used as possibly being of **not** non-GAAP, it does suggest, that in many cases, the NGM being used require contextualization to understand if they are being used in accordance with the rules that require the use of certain non-GAAP measures, or not. Given this, we discuss in further detail in Section 3.3 how we have accounted for these constraints in our research.

Given the amount of direction that is provided by the accounting firms and the Securities and Exchange Commission, it is unlikely the average investor would be

sufficiently familiar with either the implications of non-GAAP measures, or know under which circumstances non-GAAP measures are permitted, and which non-GAAP measures those are. The SEC has, for some time, expressed serious concern regarding the use of non-GAAP measures, which prompted the issuance of Regulation G and Regulation S-K, Item 10(e) in 2003 — all of which to regulate (to some degree) the use of non-GAAP measures [84]. Since 2003, the SEC has issued more guidance on the use and presentation of NGM, in an effort to better protect average investors [84].

2.3 Related Work

2.3.1 Introduction

There is considerable disagreement in the finance and regulatory community over the use of non-GAAP measures. The dichotomy is that one side believes that non-GAAP measures provide relevant, additional information to stakeholders and interested parties, whereas the other side believes that non-GAAP measures are used mainly for opportunistic purposes, mislead investors, and obfuscate true performance [32]. While it is not our primary objective to comment on managements' intention for their use of non-GAAP measures, our results themselves will bring up relevant questions as to intention, and will also provide indirect commentary on their uses. We also believe that in order to fully appreciate the significance of our results that a firm understanding of both points of view of the uses of non-GAAP measures is needed.

Therefore, we begin our discussion examining the related work in accounting and finance, also presenting the arguments in the extant literature *for* and *against* the use of non-GAAP measures in communications with shareholders and stakeholders alike. We then finish our literature review by discussing Machine Learning in the financial and legal domains.

2.3.2 Accounting and Finance Literature

As previously mentioned, there are two allowable GAAP frameworks in the United States — U.S. GAAP and the International Financial Reporting Standards (IFRS). The Financial Accounting Standards Board (FASB) is responsible for U.S. GAAP

while the International Accounting Standards Board (IASB) is responsible for IFRS. The FASB has expressed concerns for some time that using measures that are outside of the Generally Accepted Accounting Principles goes against its mission of having established *Standards* [11]. While our research is focused only on companies who report under U.S. GAAP, and therefore, under the perview of the FASB, is it worth noting that the IASB has also expressed grave concern that using non-GAAP measures undermines the integrity of accepted financial reporting standards [11]. Former FASB chairman Russell Golden has acknowledged that the prolific use of non-GAAP measures signifies that GAAP, in its current iteration, may need to be improved to better meet the needs of users [11]. The SEC has also expressed concerns over the use of non-GAAP measures, and, as cited earlier, created Regulations G and S-K in order to better reign in the way that NGM were being used [84]

In the 1990s, the use of non-GAAP measures in reporting rose to prominence and has increased significantly ever since [11]. In their sample between 2007 and 2017, [7] show that managements' use of non-GAAP measures has risen almost 86%. As Isidro and Marques [43] point out, while it is not possible to ever know what managements' true intentions are when it uses non-GAAP measures, there is incentive to use NGM as investors and analysts react to this type of information.

In a further effort to increase investor protection, the SEC introduced a "plain English" rule in 1998. The purpose of this rule was to make SEC filings more accessible (and readable) for the general public [59, 39, 76]. While the SEC wanted (and still wants) to reduce the amount of "jargon" and "legalese" [76] in their filings, there is, ultimately, a lexicon challenge. The Cambridge Dictionary defines *Jargon*, in American English, as "Words and phrases used by particular groups of people, esp. in their work, that are not generally understood" [26]. Given that, it is likely that a good number of the accounting and finance terms will be seen by the general public as jargon — particularly the non-GAAP measures where terms such as *EBITDA* and *FCF* are common.

Yet, this focus on the general public also seems contradictory to the FASB's view on this matter, as it indicated that text should be understandable to those who have a practical understanding of business and are driven to understand this information [47]. This highlights the fact that there are two standard setting bodies

(the SEC and the FASB) who appear to have differing goals on who the intended audience of the financial reports are — the *true* general public or the general public with financial savvy. It could be argued that it is unlikely that those of the general public who are engaging in trading on the stock market would only do so if they had a reasonable understanding of business and the market itself. Yet, even with a functional understanding of business, the complexities of regulation and the agendas of financial reporting may not be fully understood by those without financial training or expertise.

The use of non-GAAP measures is far more complex than at first glance — there are situations where some measures are considered non-GAAP in certain cases, but those same measures are deemed to be **not** non-GAAP in others. In his research, Young [100] makes some very keen observations. Those involved in the market are a very diverse group, which raises the distinct possibility that investors will have different responses to NGM, depending on their level of financial experience. He points out that it is very unclear if (all) investors truly understand exclusion adjustments made to measurements [100], which Black *et al.* [9] point out are increasing over time. Along similar lines, Elliott [29] notes that there is evidence that less financially experienced investors look to the non-GAAP measures more than GAAP, whereas professional investors look for the (required) reconciliation between non-GAAP and GAAP before determining the reliability of the NGM [29].

There are also (often hidden) agendas behind financial reporting. Nobel Laureate Daniel Kahneman has also indicated that cognitive bias is an important factor in managements' decision to disclose information, and that it is unrealistic to think that management will act as *rational agents* when making their disclosure decisions [46]. Rather, management creates a strategy when determining what information to disclose [25] by conducting cost-benefit analyses [28], and looking at the costs of disclosure [7]. This idea is further entrenched by the fact that disclosure of non-GAAP measures is entirely voluntary; in fact, in reviewing a sample set of 10-Q and 10-K reports, some companies have elected not to use non-GAAP measures at all.

Kang *et al.* [47] point out that, unlike quantitative analysis, text affords plasticity of message and tone, allowing management the ability to minimize poor results using positive language [47]. In their research, Loughran-McDonald [58] also indicate that

managers are hesitant to use language that could be seen as red flags to investors, which also points to the fact that managements' choice of words in discussions and disclosures is extremely important for a variety of reasons [56, 49]. This is also inline with Davis and Tama-Sweet [25] who indicate that language that management uses provided insight into the company's disclosure strategy and choices of management.

2.3.3 Research FOR and AGAINST the Use of Non-GAAP Measures

The recurring theme of research that supports the use of non-GAAP measures is altruism — that they provide additional, relevant information that the Generally Accepted Accounting Principles cannot [9]. Black *et al.* also point out that non-GAAP measures provide another angle of evaluation other than Net Income as calculated under GAAP [9]. The measures are everywhere in the financial ecosphere and have become accepted as part of the *fundamental* financial narrative. One main argument for the use of non-GAAP measures is based in forecasting where the "normal" or "recurring" income is needed [14]. Boyer points to the fact that GAAP does not provide guidance or regulation on how to determine normal or recurring income, which is something that non-GAAP measures can help to provide. Bradshaw and Sloan, Bhattacharya *et al.*, and Frankel and Roychowdhury indicate that non-GAAP measures provide very important inforamtion, particularly in the area of valuation [15, 8, 35, 41].

Bloomfield suggested in his research that having to report negative corporate results may be the underlying reason for certain word choices, rather than deliberately making the explanation more abstruse [10]. Asay, Libby, and Rennekamp designed an experiment which, among other things, tested Bloomfield's assertion. As part of the study, two hundred and five experienced portfolio managers assumed the role of investor relations for a fictitious company, and were then asked to provide disclosures regarding firm performance [5]. Their findings indicated that there was no "intentional obfuscation" in the disclosure reports prepared [5]. While the intentions of the study were good, there is a major flaw in the design: none of the "experienced managers" had anything to lose — they had no real stake in the performance or results of the "company". Although fraud and intentionally clouding performance results are, for the most part, very different, they still share a major common element: pressure.

While the use of non-GAAP measures has its supporters, there are many more detractors who cite evidence that strongly suggests that the motives are opportunistic rather than altruistic. Earnings targets are a fundamental part of measuring corporate goals. Companies set these objectives to help the company grow, but also demonstrably communicate to investors that the company is worth investing in. Given the latter importance, companies do not take missing earnings targets lightly, which opens companies up to financial and reporting manipulation. Researchers have found that there is a higher percentage of companies that are meeting or beating their earnings targets relative to those that do not. This strongly suggests that there is some degree of financial "management" [18, 17, 37, 72, 55, 8, 25, 28, 9] and one of the tools available to do that is non-GAAP measures, given the freedom afforded by non-regulated measures where companies are free to make any desired adjustments, as long as there is a reconciliation to the *closest* GAAP measure.

Research has also found non-GAAP measures, even as supplementary measures, are misleading given their persuasive nature [32, 5] as the company is essentially implying, through the adjustments that they make, that its *actual* performance is different (and in some cases starkly different) from its *audited* performance. Alee *et al.* also raises the concern that non-GAAP earnings, in particular, may confuse and mislead the average investor [3], particularly when, non-GAAP profits are created through adjustments from an *originally* non-GAAP loss [100].

Corporate language has also been closely scrutinized by researchers and professionals alike, exposing concerning issues. Kang *et al.* found that when management discloses information to stakeholders, it tends to use "flexibility" in the tone used in order to limit the damage by framing the negativity in positive ways [47, 49]. Every reader of financial reports, and in our case, specifically the MD&A and Market risks are either a stakeholder or a potential stakeholder, and is using the document to evaluate the company. This fact is tremendously important, because it speaks to corporate motivation of disclosure. Loughran-McDonald found that this motive entices writers to reframe negativity into positivity because the impact of negative words on stakeholders (or potential stakeholders) is inexorable [58]. Therefore, carefully use of word constructs can help to avoid, or at least, significantly limit the pervasive affect brought on by negative wording. This idea is also echoed by Rogers *et al.* [71]

who indicate that overly optimistic tones can be catalysts for Securities Class Action Lawsuits. This will be discussed in more detail in the literature review for the case studies.

Finally, the lack of consistency and comparability of non-GAAP measures has been found to distort analysis year-over-year for the same company (if they have been inconsistent in their own use of NGM), and between companies [14, 48]. From a stakeholder's point of view, it is important to know how a firm is performing over time, and if the firm is growing or declining. Using inconsistent non-GAAP measures make understanding this difficult. Furthermore, when investors are looking at the market, the ability to compare one company against another becomes a determining factor in investment decisions. Again, using unregulated measures challenges the consistency and comparability [41] assumptions, making it difficult for the investment community to develop reasonable metrics by which to evaluate company (and ultimately industry). performance.

2.3.4 Conclusion for Accounting and Finance

The lack of consensus on the use of non-GAAP measures in the financial community, as well as in academia and industry strongly supports that the cessation of the use of NMG is unlikely. It also points to the fact that GAAP, in its current iteration, is not sufficient to meet the needs of all of the stakeholders and users [12]. Given the arguments *For* and *Against* the use of non-GAAP measures, it is understandable that the accounting profession and the SEC are wary. While these NGM can (and do, for some) provide additional supplementary information, used or interpreted incorrectly can have (un)intended negative consequences or lead to sub-optimal decision-making on the part of the investor, particulary those who are not financially savvy.

2.4 Machine Learning in the Financial Domain

2.4.1 Introduction

From our research, it appears that the majority of the research and literature has been focused on prediction in the financial markets. Both statistical machine learning approaches and deep learning approaches have been used, and in some cases, researchers have developed their own tools to better evaluate text in the financial domain.

2.4.2 Machine Learning for Finance

One of the main approaches in finance has been the use of dictionaries that have developed word lists of positive, negative, and neutral words, as well as other categories such as uncertainty, litigious, and modal. Loughran-McDonald's dictionary is, by far the most cited word list in the research, and in their seminal paper "When is a liability not a liability?"(2011), they use a bag-of-words approach to look at the 10-K SEC filings to evaluate whether or not the Harvard-IV dictionary (which is also referred to as the General Inquirer) is appropriate to be applied to business or not. In their research findings, Loughran-McDonald demonstrate that, when applied to a purely financial domain, that the Harvard-IV is a poor choice due to the miscategorization of words such as debt and taxes [58]. Loughran-McDonald, however, are not the only researchers who have developed word lists, nor were they the first. As pointed out in the Accounting and Finance related works above, Elaine Henry was the first to develop a word list. While her list is significantly shorter and focuses solely on positive and negative (and by omission, neutral) words [22], it was the first word list of its kind, and, as our research demonstrates, is robust bridge between highly financially sophisticated and general purpose.

As part of their research, Chan and Chong developed a *Sentiment Analysis Engine* that goes beyond word tokens and digs down to the phrase level [19]. Although their *Engine* is general purpose, they did look specifically at finance, using words that have been aligned with the stock market.

Kang *et al.* [47] studied the relationship between firm performance and the tone of

the 10-K. Their research focused on two important aspects —- a possible relationship between the frequency of sentiment and firm performance, as well as determining if there was an "overtone" (inflated positivity of the narrative in relation to its earnings) or an "undertone" (a less robust positivity in relation to its earnings). To do this, the authors used the ordinary least squares regression model as well as a firm cluster-robust regression model. The results showed that there does appear to be a correlation between the sentiment and performance, but more importantly that companies that overstate positivity in their financial narratives relative to performance are less able to deliver on the company's expected future performance. To further this, Kang *et al.* [47] also looked at the effects of overtone in the stock price in the short term and found that they were positively correlated. This is important as it suggests empirically that investors either do not understand or struggle with fully comprehending the underlying overtone and its true meaning [47].

Jegadeesh *et al.* [44] identified that in a significant amount of previous research, words —- positive and negative —- were considered equal, meaning that the weighting for each word was the same. But, using the idea of term frequency-inverse document frequency put forth by Manning and Schütze [60], Jegadeesh *et al.*'s [44] research used the market's reaction to annual reports to determine the weighting that was assigned to each word. Doing this, the authors indicated, would provide a more realistic weighting for each word, thereby providing a much more accurate sentiment evaluation. To determine the weightings, they used multivariate regression, which they also indicated could be easily adapted to be used with Naïve Bayes [44]. By using Naïve Bayes, however, the authors cited the major drawback of assuming that the words are all independent, which is rarely the case [44]. They adapted their multivariate regression to take into account word co-occurrences (i.e. the occurrence of two words appearing together that is greater than chance alone) and magnitude to better weight each word.

Another challenge for machine learning is negation, which humans can recognize far more easily than machines. This challenge can also mean that sentiment can be incorrect if negation is not properly handled, as Pröllochs *et al.* have pointed out [67]. To handle this, the authors formulated ten rules to help contain the negation, in conjunction with the Brill tagger [67]. While the rule-based model proved helpful,

it did create certain impediments that stem from the writing styles of different domains [67]. To get around this, Pröllochs *et al.* also implemented a Hidden Markov Model to overcome these limitations. Both methods were used before calculating the sentiment. When they measured the predictive performance of a manually labelled dataset against their model, they found that correct handling of the negatition was far more impactful than relying on the manual labels [67].

Chan and Chong proposed a Sentiment Analysis Engine that leverages grammar as its main linguistic analysis method [20]. The authors indicate that by taking this approach, they are able to conduct sentiment analysis at both the word and phrase level. Once the sentiment analysis is done at the foundational sentence layers, the common sentiment of the text is then determined heuristically [20]. Chan and Chong's *Engine* find that when assessing stocks in a time series, that the tenor continues over time [20], which is an important factor to consider when determining the influencers on stock prices.

The most recent published work is from the European Language Resource Association's conference, also known LREC, where the first financial narrative workshop was held. Sarderlich *et al.*'s [73] work focused on building a novel financial lexicon for sentiment analysis based on Yahoo Message Stock Boards to determine new weightings for financial terms. They found that there is a strong bias towards positive words — either due to wishful thinking or overconfidence on the part of message board participants. Sardelich *et al.* [73] used a sparse vector space model, which considers each term in a separate dimension, to develop what they have called a "bag of Semantic Orientation" that is specific to market terminology (long, short, put, call, etc). In taking this approach, Sardelich *et al.* [73] were able to extend existing dictionaries, but also capture the formal and informal language used in stock trading and better classify the document tone.

Chapter 3

Research Methodology

This research focuses on domestic U.S. companies which are required to use U.S. GAAP as their reporting framework [33] as well as foreign companies who have elected to report under U.S. GAAP [86] In both cases, companies are required to use form 10-Q (for quarterly filings) and 10-K (for annual filings). We will be focusing on two sections of the SEC filing: the Management Discussion and Analysis of Financial Position and Results of Operation (known as the MD&A) and the Quantitative and Qualitative Disclosures About Market Risk. Originally, we were only going to look specifically at the MD&A, but we decided to include the Market Risks as Loughran-McDonald [58] indicate that these sections are usually analyzed together.

3.1 Data Collection

Figure 3.1 represents a high level overview of what steps were taken in the creation of the dataset we used for our research. Each step has been discussed in greater detail below.

Bill McDonald [58] maintains a repository of SEC filings dating back to 1993, linked through his webpage at the University of Notre Dame, and hosted on Google Drive. We have used the "Stage One 10-X" parse data for our research which have already been "cleaned" by stripping out the tables, any ACSII-encoded graphics, as well as HTML tags [62]. This information, while useful in other contexts and areas of research, is not relevant to ours.

Although McDonald's repository includes other types of forms such as the 10-KSB (which was a 10-K form for small businesses) and 10-K405 (which indicated that that the a company's officer, director, or beneficial owner of more than 10% of a class of stock had failed to file its statement of ownership on time) [42, 87]. Both of these forms (along with others) are no longer used by the SEC. There is

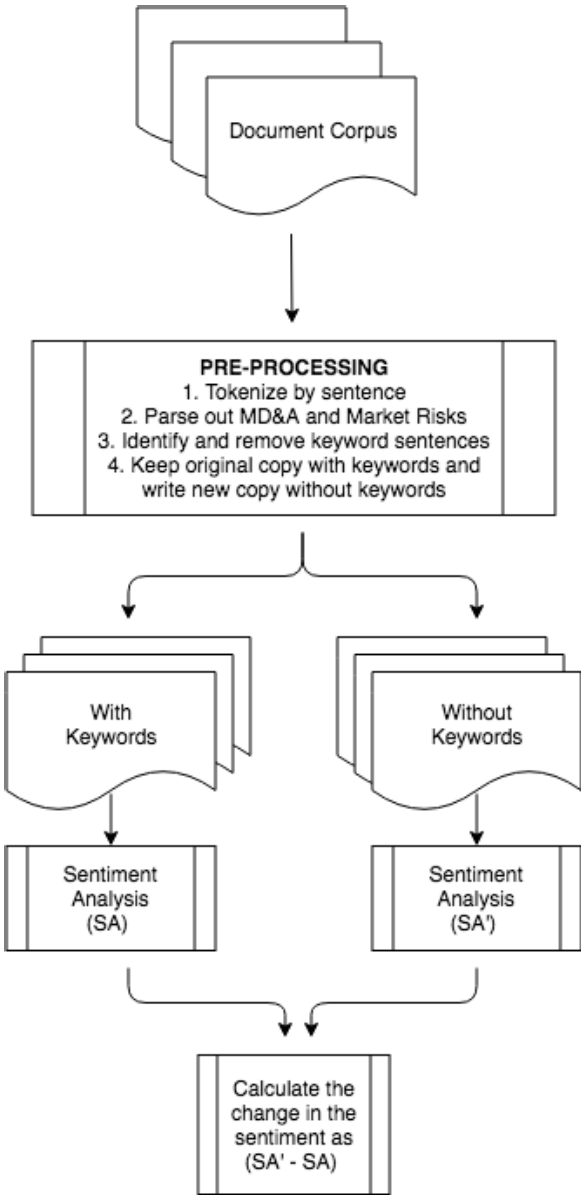

Figure 3.1: Methodology

also an"amendment" form, which is denoted by an "/A" at the end of the file, and is used to amend the original filing. Therefore, a 10-K/A (the amendment) filing should be read in conjunction with the original 10-K filing in order to get the complete and correct information of the report. There is no standardization to the amendment form — it is made available to companies to correct any missing or incorrect information (which may pertain to any section in the 10-Q or 10-K, not just the MD&A and/or Market Risks) found in the original filing. As such, the decision was made to use only 10-Q and 10-K forms. This ensured that we were comparing forms that have been in continual use through the years of the dataset, and that the information was consistently applicable to the MD&A and Market Risks.

3.2 Pre-Processing

As our research has never been done before, and we could not use the entire dataset due to constraints described above, we had to determine an appropriate sample size to use for our experiments. Borrowing from the field of biomedical informatics, we followed the rationale laid out by David Juckett [45], which used a sample size of 10,000 in their research on the number of documents needed to create a gold standard corpus. We then randomly chose 10,000 reports (using 10-K and 10-Q filings only). We divided that 10,000 into experiment sizes of 100 reports, thereby capturing reports from each quarter, from the 4th quarter of 1993 (the beginning of McDonald's dataset) to the 2nd quarter of 2018 (the end of McDonald's dataset when we downloaded it from Google Drive).

The 10-Q and 10-K reports are relatively "standardized" in how they are named and assembled. The SEC provides blank 10-Q and 10-K forms that companies can (should) follow for their filings. Although these are readily available to registered companies, there was variation (sometimes considerable) in the assembly and naming conventions of the reports. The MD&A and Market Risks should be found in Part I (for the 10-Q) and Part II (for the 10-K). We found this was the case. Within the 10-Q, the MD&A and Market Risks should be found under Items 2 and 3, respectively. Similarly, in the 10-K, these two sections should be found under Items 7 and 7a, respectively. We found that this was not always the case. This made extracting the

MD&A and Market Risks challenging.

The naming convention also presented significant challenges in that, officially, it is called the Management Discussion and Analysis of Financial Position and Results of Operations, but was also found to have names such as "Management's Discussion and Analysis or Plan of Operation", "The Registrant's Discussion and Analysis of Financial Condition and Results of Operation", "Trustee's Discussion and Analysis and Plan of Operation", "Financial Review", "Quantitative and Qualitative Analysis of Financial Condition and Results of Operation", "MD&A" , "Discussion", or had no title. This variation made it difficult to correctly parse out the MD&A from the full filing. We found, however, that Market Risks had no variation in the name. This is inline with the challenges faced by Davis and Tama-Sweet [25].

Another challenge we encountered was that the name of the section appears in the Table of Contents, as well as in other sections as a reference where readers are told to either read that section or particular points in conjunction with the MD&A, or are referred to the MD&A for more information. To overcome this, during the parsing procedure, we used the Python module *File-Read-Backwards* that allowed us to read the 10-Q or 10-K from the bottom up, so that the first time that Python encountered the MD&A, it would more likely than not be the actual section itself, and not a reference, or the table of contents. Once Python found this section, the code then reversed the reading direction and wrote the MD&A and Market Risks, each to a new file, line by line, top to bottom, stopping at the beginning of Controls and Procedures (Item 4.) in the 10-Q or the Financial Statements and Supplementary Data (Item 8.) in the 10-K. We found this parsed the MD&A and Market Risks more accurately, on average, than reading in the file top down.

As Python parsed out the MD&A and Market Risks, it was also checking the list of the non-GAAP words (discussed below). It created two copies of the MD&A and Market Risks, each tokenized by sentence, — one *with keywords* (i.e. the original MD&A and Market Risks as filed with the SEC) and one *without key words* where the sentences containing the non-GAAP words were removed. This allowed us to have what we have termed a "before" and "after" copy available for the sentiment analysis portion of our research. We also put in an "if" statement for the non-GAAP measure *adjusted EBITDA*, discussed below. If the key word found during parsing

was *EBITDA*, then it would check to see if it was preceded by the word "adjusted" to be the non-GAAP measure *adjusted EBITA*. If the complete measure was *adjusted EBITDA*, then the sentence was left in. If it was just *EBITDA* on its own, the sentence was taken out.

Companies are allowed to incorporate their MD&A into the 10-Q and 10-K filings through a reference to their Annual Report, which is (now) typically made available on their website. Also following Loughran-McDonald [58], we required at least 250 words in the MD&A (excluding the Market Risks) to be included in the dataset for analysis. This helped to avoid inadvertently including filings where text was included simply as a placeholder to tell the reader that the report was being incorporated by reference.

3.3 Non-GAAP Measures Selected

Three pervasive issues discussed earlier had significant influence over which non-GAAP measures were selected for our research:

- There can be a signficant amount of fluidity and creativity with how companies define their non-GAAP measures;
- Companies have the ability to create their own non-GAAP measures within limits discussed in the background; AND
- In certain circumstances, measures which are normally non-GAAP may be **not** non-GAAP, thus requiring contextualization to determine its status.

To determine which non-GAAP measures we would use, we started with a list of common NGMs published by the accounting firm Deloitte as our starting point [84]. We then reviewed the SEC's rules as well as the accounting principles for each of the non-GAAP measures on the Deloitte's list, removing any NGM that could, under certain circumstances, be judged to be **not** non-GAAP. This allowed us to create a final list of non-GAAP measures which, based on our research and interpretation of the both accounting and the SEC's dislcosure rules, are always judged to be non-GAAP and do not require any contextualization. The measures we selected are as follows:

- Revised Net Income
- Earnings Before Interest and Taxes (EBIT)
- Earnings Before Interest, Taxes, and Depreciation (EBITDA)
- Earnings Before Interest, Taxes, Depreciation, Amortization, and Rent/Restructuring (EBITDAR)
- Adjusted Earnings Per Share
- Free Cash Flow (FCF)
- Core Earnings
- Funds From Operations (FFO)
- Unbilled Revenue
- Return on Capital Employed (ROCE)
- non-GAAP
- Reconciliation

Note: "Revised" or "Adjusted" variants of measures, such as "Adjusted EBIT" were also included, as were commonly accepted variations of naming of the non-GAAP measures such as "debt-free cash flow" and "unlevered free cash flow". Also, we added the word "reconciliation" into our short list. Non-GAAP measures that, from the company's point of view, are NGMs will be accompanied with a reconciliation, as per regulations. We believe that, in that case, no further contextualization is required to determine if the measure is GAAP or non-GAAP, as the company has self identified the measure as non-GAAP. Similarly, we used the term *non-GAAP* to identify any other measures outside of this list that the company had, again, self-identified as non-GAAP, as any measures labelled as such would not require further contextualization.

We then extracted the entire sentence that the non-GAAP measure appeared in. Our rationale for taking this approach is that the non-GAAP measure is the focus of

the sentence, and therefore, the words in that sentence exist only for discussing that measure. To illustrate that point, we offer the following:

"Our EBITDA decreased 2% for the first quarter of fiscal 2012 compared to the first quarter of fiscal 2011, due to a slight decrease in net revenues and a slight increase in operating expenses." (Taken from TD Ameritrade's 10-Q filing made on 2012-02-08.)

If we take a bag-of-words approach to this sentence and only remove the non-GAAP measure — in this case EBITDA — that leaves the rest of the words in the sentence. Yet, without the non-GAAP measure, the sentence no longer makes sense:

"Our decreased 2% for the first quarter of fiscal 2012 compared to the first quarter of fiscal 2011, due to a slight decrease in net revenues and a slight increase in operating expenses."

In this second iteration, we have removed the non-GAAP measure EBITDA. As can be seen, the sentence no longer makes sense, and the reader is left questioning what decreased in order to contextualize the rest of the sentence. Using the bag-of-words approach, all of these words would be left in, only removing the non-GAAP measure, when, in reality, all of the words left in the sentence exist only to discuss and contextualize the non-GAAP measure removed, EBITDA. Therefore, the entire sentence needs to be removed, not just the keyword. By taking this approach, we are able to quantitatively measure the full effect of the non-GAAP measure on the tone of the document. To ensure that this process worked as expected, we chose a small sample to manually review after extraction to ensure that the non-GAAP measures were removed. In the sample selection, we found that there were no issues with the extraction process.

3.4 Dictionaries Used For Analysis

The financial lexicon and jargon used by professionals, which subsequently appears in reports, financial statements and filings (such as the 10-K and 10-Q reports we examined for our research), can be quirky and nuanced. As noted by Loughran-McDonald [58], there are a lot of words which, out of the financial context, elicit emotional responses that may not be warranted. The word "debt" (which is a financial liability) is a good example. When used in a *business* context, the word itself is

neutral; it is expected that businesses will have debt and, until that debt has been contextualized by taking into account the rest of the facts, figures, and discussions, it is not appropriate to assign it a tonal label. If, for instance, the business takes on too much debt or cannot pay its debt, then the sentiment is justifiably negative. If, however, the business is trying to maximize its capital structure by including or increasing debt to lower its cost of capital (as interest on the debt is tax deductible [66], and that the increase is to the optimal level so that the business is able to sustain that debt load), then a positive tonal label would be appropriate.

This has presented a challenge to researchers who have developed financially oriented dictionaries. As will be discussed in more detail below, Henry [22] addressed this challenge by focusing on descriptive words such as "deteriorate" (negative) or "improved" (positive) to characterize the financial terms. Loughran-McDonald's word list is much more comprehensive [58], but unlike Henry [22], includes negative financial terms that require little contextualization in order to glean the correct understanding such as "bankruptcy" or "defrauded".

We used R to conduct our sentiment analysis, as it has four built-in dictionary libraries: Harvard IV-General Inquirer, Quantitative Discourse Analysis Package (QDAP), Henry, and Loughran-McDonald, each of which is discussed below.

1. *Harvard-IV:* Developed by Harvard University, this is a multi-classification psychological dictionary that include categories such as positive, negative, weak, strong, affiliation, and hostile, for example [101]. The implementation of this dictionary in R is strictly a binary classification. There are 1,316 positive words and 1,746 negative words. Words such as *debt*, *interest* and *taxes* are negative words in this dictionary, and are assigned a score of -1 [31] This dictionary does take into account the challenges that financial words present, and therefore, we believe, fairly represents the average investor with no finance training.

2. *QDAP:* The target of this dictionary is, as the name suggests, quantitative discourse analysis to bring together qualitative discussion and statistics [69]. The unique thing about QDAP is that it is a collection of dictionaries and includes subsets of the Harvard-IV, the Hu-Liu word list from their paper *"Mining Opinion Features in Customer Reviews"* [40], Dolch's 220 most common words by

reading level [27], census data collected by the U.S. Government, among others [31]. The R implementation of this dictionary is a binary classification and 1,208 positive words and 2,952 negative words. Words such as *debt*, *interest*, and *taxes* are neutral words in this dictionary, and are assigned a score of -1 [31]. As this dictionary is a collection, it is not targeted to the domain of finance, and therefore, we believe is another fair representation of the average investor.

3. *Henry:* Elaine Henry was the first researcher to develop a financially oriented dictionary that took the challenges of financial words into consideration. The dictionary was formed in conjunction with her article "Are Investors Influenced By How Earnings Press Releases Are Written" [39]. This dictionary has a binary classification with 53 positive words and 44 negative words. Words such as *debt*, *interest*, and *taxes* are neutral words in this dictionary, and are assigned a score of 0. As this dictionary takes into account the nuances of financial words, we believe that this represents the point of view of finance professionals.

4. *Loughran-McDonald:* As part of their seminal work, [58], Loughran-McDonald created their master dictionary, which is based on the *2of12inf* dictionary released by SourceForge [91]. They found this dictionary useful because it uses inflections rather than stems, and is therefore less prone to errors for the purpose of tone. Loughran-McDonald then extended this dictionary using the 10-X filings that they collected from the SEC's online filings repository, EDGAR into a multi-classification dictionary that include categories such as positive, negative, uncertainty, and litigious [57]. As the authors have noted, *"Language is dynamic"* and to keep up with that dynamism, they update this dictionary on an annual basis. Since 2012, no words have been deleted from their dictionary, but 343,606 words have been added and 265 words have been reclassified [57].

   The R implementation of this dictionary is a binary classification only, with 145 positive and 885 negative words. Words such as *debt*, *interest*, and *taxes* are neutral words in this dictionary, and are assigned a score of 0 [31]. This dictionary comprehensively addresses the challenges of financially oriented words. As such, we believe that this fairly represents the analytical approach that finance professionals would take.

3.5 Sentiment Analysis

We wrote a script that leveraged the *Sentiment Analysis* library in R and returned us with the sentiment from all of the dictionaries discussed above. The script first read and returned the sentiments for the files in the "with keywords" folder, and then moved over to the "without keywords" folder, returning the same for those files. Each results file contained the position number of the file in the folder, the overall sentiment score, the negativity and positivity scores, the uncertainty score (for the Loughran-McDonald dictionary only), the file number assigned by EDGAR at the time of filing, and the date filed — all of which are standardly provided using the *Sentiment Analysis* library in R. We kept the overall sentiment scores for each dictionary, the file number assigned by EDGAR, and the date. We discarded the other sentiment scores, as they were all captured in the overall sentiment score in that the negativity score (and the uncertainty score in the case of Loughran-McDonald) was subtracted from the positivity score to provide the overall sentiment score. We also discarded the position number of the file in the folder, as it provided no additional value.

The EDGAR file number is very long and can be challenging to read. The following information is contained in the EDGAR file number, which we found very useful , as it includes the date file, the conformed submission type (such as 10-K, 10-Q etc), the central index key, and accession number (a unique identifier that is assigned when the submission is accepted) [85]. Below, is an example, with labels to each number in the EDGAR file number.

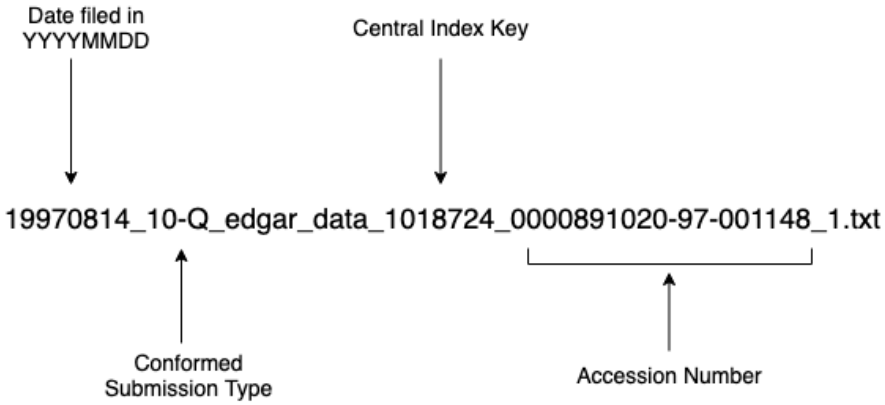

Figure 3.2: EDGAR file number

Chapter 4

Extractive Sentiment Analysis

In our research, we introduce a novel extractive methodology to quantitatively measure the impact of non-GAAP measures in financial reports filed with the SEC. There has long been concern that including the non-GAAP measures is *inflating* the sentiment and obfuscating the actual tone of the reports, creating a positivity that may not be supported by the company's performance. As the name indicates, non-GAAP measures are not within the GAAP rules, and are not auditable. Due to this, companies are free to create their own non-GAAP measures, though the vast majority use those that have been "accepted" by the industry. The corollary is, however, that there is no standardization that companies need to adhere to, nor is there a requirement for companies to compute their non-GAAP measures the same from year-to-year. While there has been guidance given by the SEC on non-GAAP measures, it is on presentation only and does not address the quantitative aspect of these non-GAAP measures. Therefore, our research question compares paired samples, one containing the non-GAAP measures (i.e. as filed with the SEC) which we denote as X in the experiments and one that does not contain the non-GAAP measures, denoted as X′. The equation (X′ - X) gives the difference between the extracted report and the report as filed, quantitatively supporting the tone change once the non-GAAP measures have been extracted.

4.1 Background

As has been discussed, prior research in the finance domain has used the Bag-of-Words (BOW) approach, which chops the text into single words, and is unconcerned with order. Yet, in financial documents, the order of the text matters significantly. Words are selectively chosen and arranged for a particular purpose [25] and the words in each sentence are carefully crafted around the central idea of that sentence. Therefore,

the BOW approach is not appropriate. If we were to take this approach in our research, only the non-GAAP words themselves (such as Earnings before Interest, Tax, Depreciation and Amortization, commonly abbreviated as EBITDA or Free Cash Flow, also commonly abbreviated as FCF) would be extracted, leaving the rest of the words in that particular sentence which only exist to support the non-GAAP measure(s) in the "after" version. This would be incorrect and could possibly skew the results of the sentiment analysis as words that would not normally be in the document if the non-GAAP words were not included would be counted in the sentiment, and potentially (depending on the score assigned), add to the sentiment score.

To the best of our knowledge, our approach of extracting the non-GAAP measures from the reports in order to do comparative research has not been done before, and is therefore a new area of research.

4.2 Methodology for the Extractive Sentiment

The Loughran-McDonald dataset comprises 1,057,957 filings during the period of 1998 to 2018, broken up into quarters. As stated above, we have only used quarterly (Q) and annual (K) reports to give consistency over our sample. As we chose our filings randomly, we did not achieve a dataset that comprised 50% K and 50%Q. The final breakdown of our dataset is, by chance, relatively even, with 5,514 K reports and 4,486 Q reports, totalling 10,000. Using 100 filings per quarter, we conducted one hundred experiments, preparing our files and conducting the sentiment analysis using the overall methodology discussed above. We approached our research in this manner to evaluate two important hypotheses, discussed below. We also broke it up this way so that we could see what the incremental changes were over time, and then determine possible causes for any noticeable differences in the change in tone during those periods.

The results of the sentiment analysis range from -1 to 1, and each parsed MD&A and Market Risks, whether before or after extraction the non-GAAP sentences can take on any score in that range, including -1 and 1. A positive tone (or sentiment as we use these terms interchangeably) is one that is above zero; a negative tone is one that is below zero; and a neutral tone is one that is at zero. Similarly, a positive

tone change is one that increases after the non-GAAP sentences have been removed, and a negative tone change is one that decreases after the non-GAAP sentences have been removed.

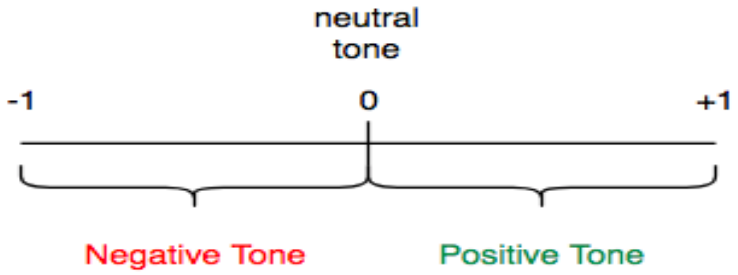

Figure 4.1: Measure of Tone

We also use the term *finance professionals* in our research. We use this term broadly for those who have financial sophistication. This would include professional investors, investing and financial analysts, designated accountants (under various designations around the world), Chartered Financial Analysts, as well as those with no financial training but who have significant experience in finance and the market. We believe that in including those with significant financial experience in this group is appropriate, given knowledge and exposure can be commensurate with training in certain cases.

4.3 Assembled Dataset Prior to Experimentation

The table below outline the composition of the dataset.

Type	# of Files	Word Count (Before)	Avg. Word Count (Before)	Word Count (After)	Average Word Count (After)	Total Words Extracted
K	5,251	18,011,601	3,430	16,716,747	3,184	1,294,854
Q	4,749	13,750,937	2,896	12,553,833	2,643	1,197,054
Total	10,000	31,762,538	6,326	29,270,630	5,827	2,491,908

Table 4.1: Dataset

4.4 Hypotheses

The use of non-GAAP measures have been shown to improve the tone of documents. Due to this, we developed two hypotheses in relation to our aggregate dataset:

Hypothesis 1. *Overall Aggregated Tone*

Given that there is a significant body of existing research that indicates that non-GAAP measures are used opportunistically and thereby present a rosier view of a company's financial situation, we have postulated that when the tone changes for a given dictionary have been aggregated for all 10,000 reports, that the tone will decrease:

Null Hypothesis: The aggregate tone of the dictionary under evaluation is ≥ 0
Alternative Hypothesis: The aggregate tone of the dictionary under evaluation is <0

Hypothesis 2. *Statistical Significance*

Our novel method compares the report two versions of each report, thereby effectively providing a way for us to compare the sentiment that each dictionary provides under each scenario — with (or, as filed with the SEC) and without non-GAAP measures.

Null Hypothesis: After extraction, the mean (μ) of the tone change for the dictionary under evaluation = 0
Alternative Hypothesis: After extraction, the mean (μ) of the tone change for the dictionary under evaluation <0

4.5 Experiments

Once the datasets were created, we conducted our statistical experiments. It is important to note that as we are using "before" (X) and "after" (X′) from the sentiment analyses as our basis, our data is therefore paired, which must be kept in mind as test chosen must be be for matched samples [4].

Our first hypothesis sought to answer the aggregate tone for each dictionary for our sample of 10,000 filings. To determine this, we performed a sentiment analysis for

each record in the dataset using each of the dictionaries described above and summed the results.

Our second hypothesis examined the change in the mean of each dictionary between the report as filed (denoted as X) and the report where the non-GAAP measures had been extracted (denoted as X′). We hypothesized that this change would be negative for each dictionary.

To evaluate this, we first looked at the distribution of the data for each dictionary to determine if it was parametric or non-parametric, by evaluating the histograms of each.

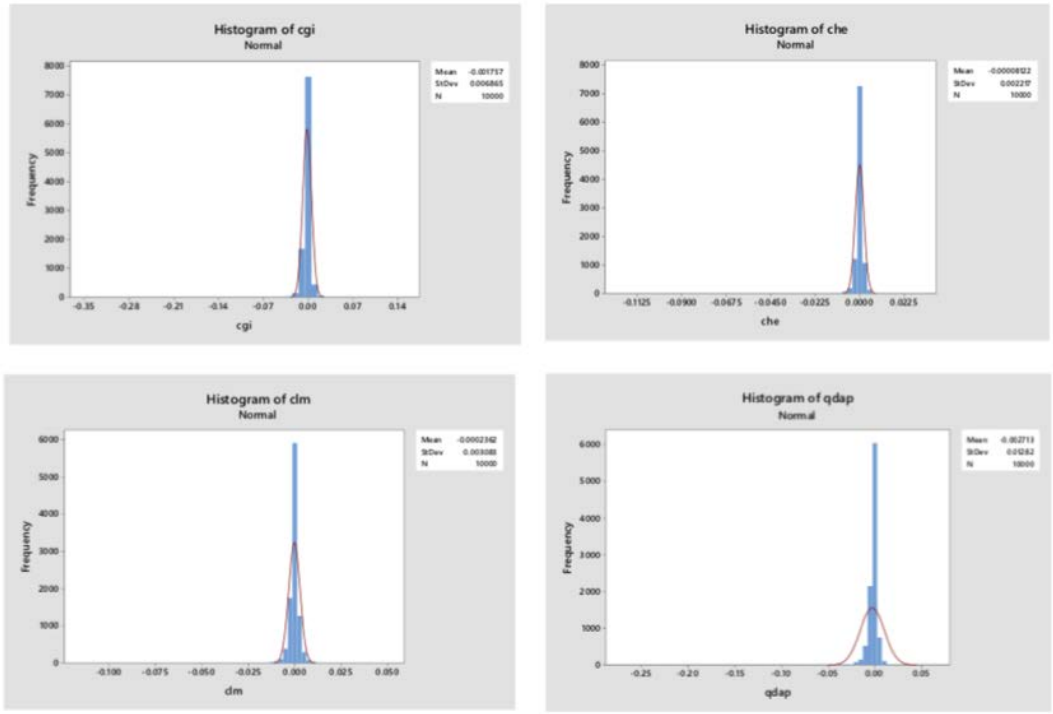

Figure 4.2: Histograms for the aggregated dataset

As the histograms showed normal distribution of the data for each dictionary, we then used the paired t-test to determine if there was statistical significance for any of the dictionaries. In conducting these tests, we used a 95% confidence interval to evaluate our hypothesis.

4.6 Results

Hypothesis #1: Overall Aggregated Tone

We used four dictionaries in our experiments. The General Inquirer and QDAP dictionaries were used as we believe that these are good proxies for the average investor, as they are not domain specific. Loughran-McDonald and Henry were used to represent the financially savvy investors, as these two dictionaries were created with specific orientations to finance.

Figure 4.3 outlines the results for each dictionary, where "cgi" denotes the sentiment change in the General Inquirer; "che" denotes the sentiment change in the Henry; "clm" denotes the sentiment change in the Loughran-McDonald; and "cqdap" denotes the sentiment change in the QDAP. As seen below in Figure 4.3, the aggregate tone change for each dictionary is negative, meaning that, overall, the sentiment decreased in tone once the non-GAAP measures (and the supporting words) were extracted. The most pronounced negative results are for the two dictionaries that were used as proxies for the average investor: the General Inquirer and the QDAP scored -17.57297 and -27.13332 respectively. As these two dictionaries are not domain specific, the results strongly suggest that the average investor is much more sensitive to non-GAAP measures than seasoned investors, which is in line with Black *et al.* and Marques [9, 61]. As well, it is interesting that the results from the Henry and Loughran-McDonald dictionaries are also negative. As both of these dictionaries are financially oriented, the negative result draws credence from the evidence of researchers that non-GAAP measures inflate the positivity of the text [41, 9, 28, 100, 6, 24, 97, 50, 55, 5, 48]

Reporting Quarter	File Name	K (1) Q (0)	Date_Filed	cgi	che	clm	qdap
Q3-18	20181010_10-Q_edga	0	20181010	-0.0101182	-0.000153	0.002228726	-0.0074967
Q3-18	20181011_10-Q_edga	0	20181011	-0.00148	-2.52E-05	-0.00411334	0.00104623
Q3-18	20181011_10-Q_edga	0	20181011	0.00401739	0.0005106	0.000758538	0.00011235
Q3-18	20181011_10-Q_edga	0	20181011	-0.0025215	-0.00172	0.002472894	0.00049252
Q3-18	20181011_10-Q_edga	0	20181011	0.00270776	-0.001678	0.000373329	0.00243339
Q3-18	20181011_10-Q_edga	0	20181011	0.00246886	0.000188	-0.00127438	-5.053E-05
Q3-18	20181011_10-Q_edga	0	20181011	-0.0048782	-0.000422	0.000396483	-0.0050087
Q3-18	20181011_10-Q_edga	0	20181011	-0.0080319	0.000152	0.001665008	-0.0061931
	AGGREGATE TOTALS			-17.57297	-0.81217	-2.36182	-27.13332

Figure 4.3: Aggregate Tone And Word Change for Each Dictionary

Given that the Henry dictionary is barely negative, this may raise questions as to if the inflationary assertion still holds for the dictionary; we believe it does. The Henry

dictionary's focus is on descriptive words that are used in finance such as "growth", "opportunity", "declining", and "deteriorated" [31], not on the financial words themselves such as "debt" or "interest". Based on the evidence of the experiments, these descriptive words have been used as supporting words for non-GAAP measures. We can also infer that, based on the results, that sufficient positive descriptive words have been used with the non-GAAP measures that, when removed, have returned an overall decrease in the sentiment, thereby further entrenching the postulation that even though the result is minimally negative, the inflationary assertion still holds.

We also looked at the distribution of the non-GAAP measures over the 100 experiments performed. As seen in Figure 4.4 below, the results show an increasing trend, which is consistent with what industry professionals are reporting [54]. In fact, PricewaterhouseCoopers LLP (PWC), one of the "Big Five" accounting firms, has indicated that there has been a substantial increase in the usage of non-GAAP measures when comparing today's reporting with that of twenty years ago [54]. PWC also indicates that nearly all of the companies listed on the Standard & Poor 500 (better known as the S&P 500) use at least one non-GAAP measure. All considered, it has raised concerns with the SEC to ensure that the non-GAAP measures are not misleading and that there is transparency in the information being provided [54]. Yet, again, the SEC focuses on the presentation of the information, rather than regulation of the measures themselves.

We also looked at the aggregated results to see which measure or measures were the most used over the period of the data set (4th quarter of 1993 to 3rd quarter of 2018). As there were many keywords that we used to pinpoint the non-GAAP measures selected such as "EBITDA", "revised EBITDA", and "adjusted EBITDA", for example, we have rolled these variations all up into EBITDA, as this is the root non-GAAP measure. We have done this with the remainder of the non-GAAP measures where we used variants to capture different iterations of the NGM, in order to understand the full extent of the usage of the base measure.

Based on the results from the graph showing the distribution over time, we compared two time periods to get an understanding of how the non-GAAP measure distribution had changed. We first looked at the period of Q4-1993 to Q4-2005. This looked at the first 49 experiments, which can be seen in Figure 4.5 below.

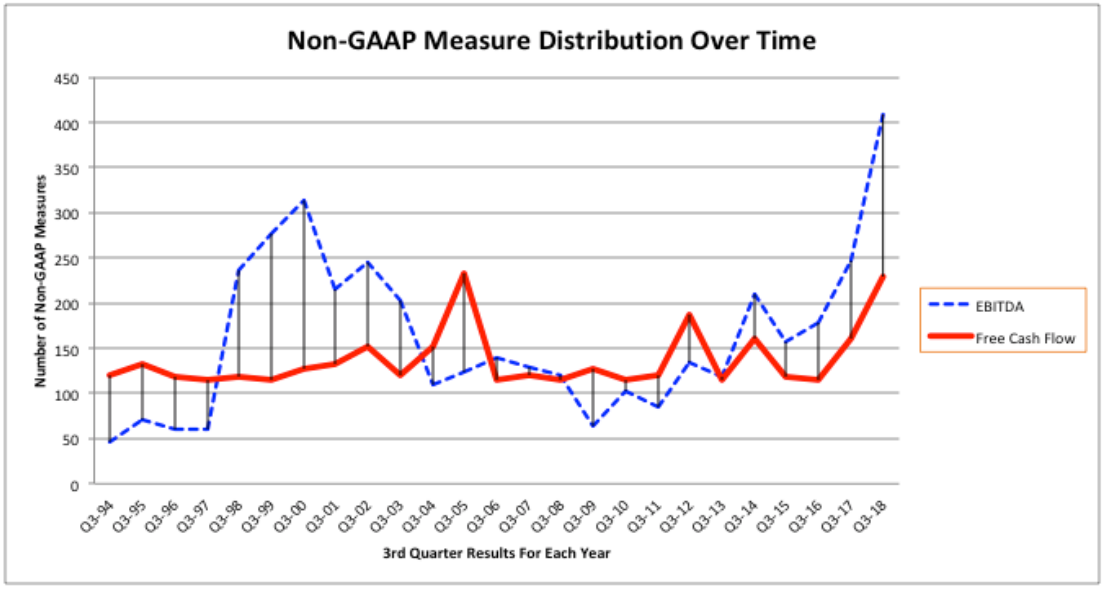

Figure 4.4: Non-GAAP Measure Distribution Over Time

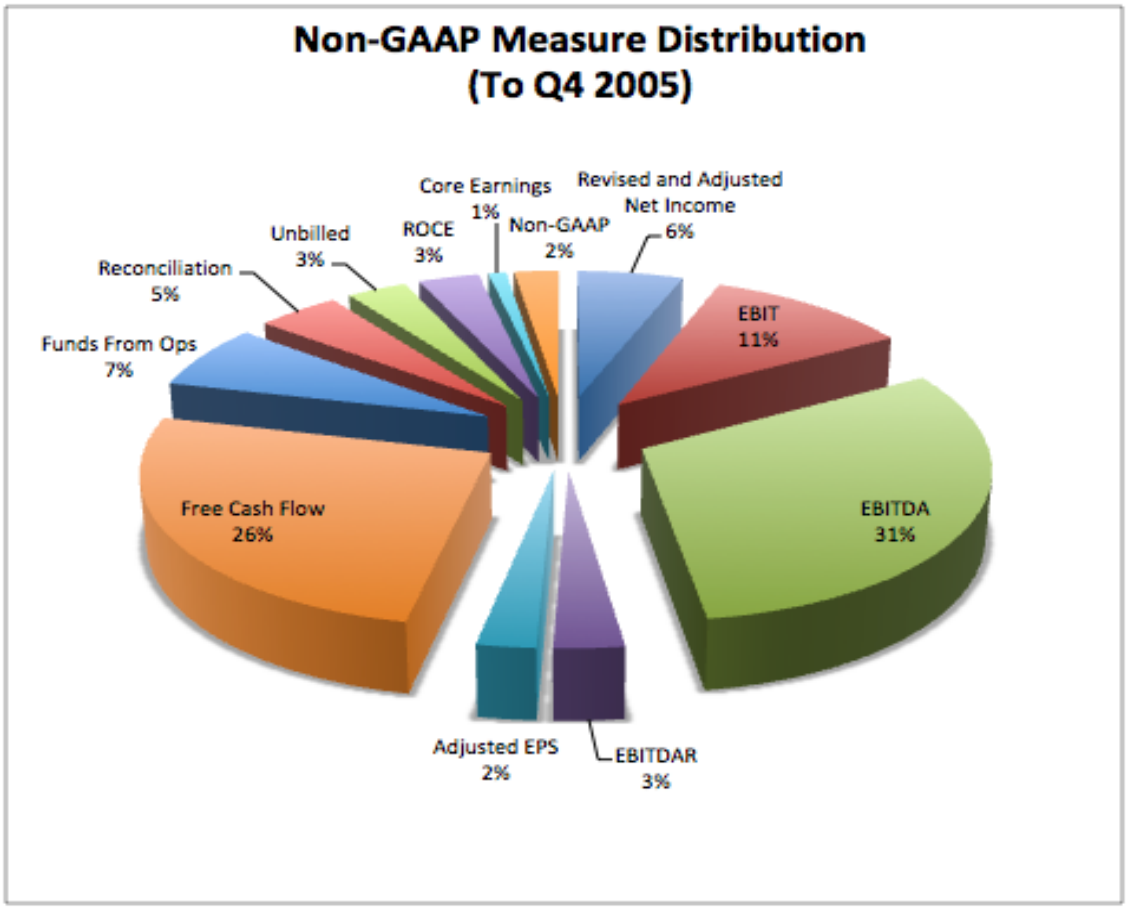

Figure 4.5: Non-GAAP Measure Distribution to Q4 2005

The two most used measures are EBITDA at 31% and Free Cash Flow at 26%. While these are *relatively* even, there is still a difference of 5% between the two. The third most used non-GAAP measure is EBIT, which is, essentially, the pre-cursor to EBITDA is at 11%. Yet, these change quite substantially when we looked at the aggregate over all of the years, seen in Figure 4.6 below. In the last half of the experiments, the use of EBITDA and Free Cash Flow decreases, netting out at 25% and 22%, respectively, over the course of all of the experiments. We also notice that the balance between EBITDA and Free Cash Flow is even closer, now only 3% apart.

Yet, we seen a sharp increase in the prominence of EBIT, which grew from 11% to 17%. An important distinction between EBIT and EBITDA is the aspect of depreciation and amortization which allocate the costs of a depreciable asset, either tangible assets such as equipment or intangible assets such as patents, over the course of their useful lives. While EBITDA can be used as a proxy for cash flow, this does not hold true for companies with intensive tangible capital or intangible assets, as the "amortization and depreciation" reflect either portions or full amounts of past capital expenditures that the company has incurred. In these cases, a better proxy for cash flow is EBIT.

As such, there are several plausible reasons that EBITDA has decreased while EBIT has increased. Firstly, the change could be driven by the companies that were included in the random sample. If more capital and intangible companies were included in the latter sample after experiment 49, the increase in EBIT could be explained by it being a better proxy for cash flow. Secondly, the most common non-GAAP measure in any scenario is EBITDA. Given the scrutiny of the SEC on the use of non-GAAP measures, companies may be attempting to become less *noticeable* by using other non-GAAP measures. Finally, companies are required, under Regulations G and SK to provide transparent non-GAAP calculations that must be compared to the closest GAAP measure. Under this possibility, companies may be finding it easier to find a close comparable GAAP measure for EBIT rather than EBITDA, especially if they are trying to make additional modifications to the EBITDA measure that are not conventionally seen like "Further Adjusted EBITDA" or "Structuring Adjusted EBITDA" [74]. Normally, Net Income is the closest GAAP measure for EBITDA, but as can be seen from Figure 4.8 below, making additional (and very alternative, to

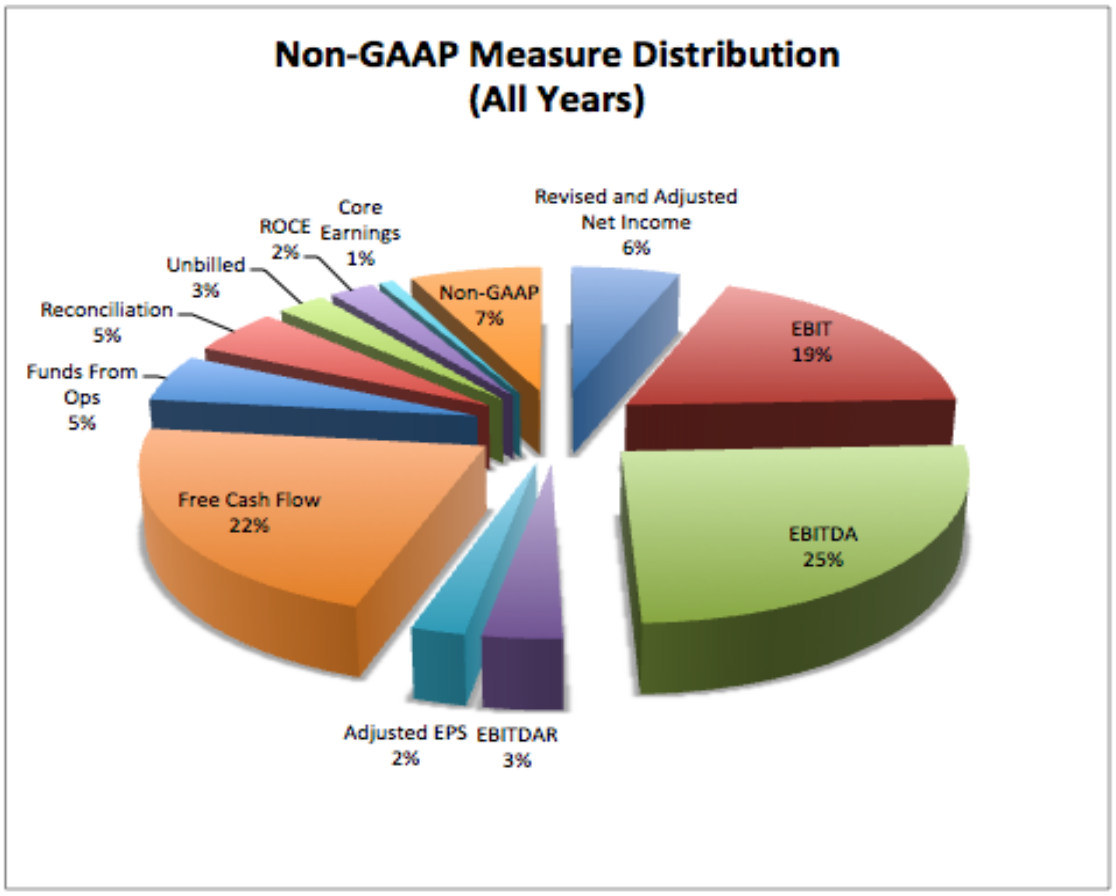

Figure 4.6: Non-GAAP Measure Distribution (All Years)

say the least) adjustments to EBITDA make it much harder to continue to compare such as EBITDA to Net Income.

In the example of Aleris International (Figure 4.7), EBITDA is adjusted, and then is further adjusted two more times. Although the company shows the reconciliation and the adjustments it has made, this non-GAAP measure in its three iterations of Adjusted EBITDA, Further Adjusted EBITDA, and Structuring Adjusted EBITDA, each is getting, in reality, further and further away from Net Income, which is supposed to be the closest comparator to EBITDA. Although the motivation for this company is to show the least amount of leverage, which is the amount of borrowing (i.e. debt) to fund assets, it highlights the issue with using EBITDA and then adjusting it further away from Net Income.

Hypothesis #2: Statistical Significance

Using the same four dictionaries, we tested the statistical significance using a

(Dollars in millions, metric tons in thousands)	For the Twelve Months Ended March 31, 2018
Other financial data:	
Net cash (used) provided by(b):	
Operating activities	$ (27.7)
Investing activities	(178.0)
Financing activities	207.1
Depreciation and amortization(b)	124.7)
Capital expenditures(b)	(175.2)
EBITDA(c)	116.0
Adjusted EBITDA(c)	202.4
Commercial margin(d)	1,230.3
Ratio of earnings to fixed charges(e)	—
Ratio of Adjusted EBITDA to cash interest expense(c)(f)	1.5x
Ratio of debt to Adjusted EBITDA(c)	9.1x
Ratio of net debt to Adjusted EBITDA(c)(g)	8.7x
Adjusted other financial data:	
Further Adjusted EBITDA(h)	$ 239
As adjusted ratio of Further Adjusted EBITDA to cash interest expense(f)(h)(i)	1.9x
As adjusted ratio of debt to Further Adjusted EBITDA(h)(j)	8.0x
As adjusted ratio of net debt to Further Adjusted EBITDA(g)(h)(j)	7.8x
Structuring Adjusted EBITDA(h)	$ 334
As adjusted ratio of Structuring Adjusted EBITDA to cash interest expense(f)(h)(i)	2.6x
As adjusted ratio of debt to Structuring Adjusted EBITDA(h)(j)	5.7x
As adjusted ratio of net debt to Structuring Adjusted EBITDA(g)(h)(j)	5.6x

Figure 4.7: Aleris International, Year Ended March 31, 2018

paired t-test, given that the distribution of the data for each dictionary was normal. We had hypothesized that the change in the mean of each dictionary, when we considered [X′ - X], that the change would be negative for each dictionary. As seen below in Table 4.2, the results for each dictionary were determined to be statistically significant at the 0.05 level, meaning that there is a 5% risk that we could erroneously conclude that there is a difference where none exists.

4.7 Paired T-Test Results

**significant at the 0.01 probability level*

We see that over 10,000 samples, that all of the dictionaries are statistically significant. This really underscores the importance of language. As we have seen, companies will disclose strategically, and have used non-GAAP measures alongside

Dictionary	Number of Samples	Mean	Std Deviation	T-Value	P-value
GI	10,000	−0.001757	0.006865	−25.6	<0.001*
QDAP	10,000	−0.00272	0.012801	−21.25	<0.001*
HE	10,000	−0.000081	0.002217	−3.66	<0.001*
LM	10,000	−0.000236	0.003083	−7.66	<0.001*

Table 4.2: Paired T-Test Results

GAAP measures in their communications to stakeholders. As we have extracted both the non-GAAP measures as well as the supporting words in the sentence, we see that the NGMs are having a pronounced effect for both the non-financial *and* the financial dictionaries, which act as proxies for the two different types of investors we identified. This is an important finding given that regardless of whether companies are using these non-GAAP measures altruistically or opportunistically, there **is** a quantifiable effect. We know from previous research that non-GAAP measures affect investors without financial training more than those who do. But, given these results, we also need to consider that non-GAAP measures are also having an effect, perhaps less pronounced, on the financially savvy as well — whether direct or indirect.

Chapter 5

Case Study: Sentiment of Securities Class Action Lawsuits

As has already been addressed, non-GAAP measures are a contentious tool that are often used opportunistically rather than altruistically. We also know that investors are affected by tone, and in the case of securities lawsuits, that investors are focused on the language of disclosures, particularly when a company's optimism is not followed by action [71]. Knowing this, we wanted to determine if the change in tone for firms under Securities litigation would prove statistically significant, and if we could use the change in tone in the MD&A and the Market Risks to predict the outcome of securities class action lawsuits under Rule 10b(5) of the U.S. Securities and Exchange Act of 1934.

5.1 Background and Related Work

Interest in applying machine learning and text analysis in the domain of law has been growing in recent years. Although the body of existing literature is still fairly small, there is related works that apply to our question of predicting the outcome of the securities litigation.

Citing that patent litigation is resource-heavy in both time and money, Wongchaisuwat, Klabjan, and McGinnis examined two main questions in their research: predicting the likelihood of patent lawsuits, and then if the lawsuit was likely, how much time the company had before the expected litigation would begin [99]. They used k-means clustering, random forest, and support vector machines to create both clustering and classification models on which to test their data. Using patent claims, they extracted features, and used a re-sampling method to prepare the dataset before fitting to the clustering and classification models to answer the first question — is litigation likely. [99]. To address the second question regarding lead time to litigation, Wongchaisuwat *et al.* used SEC financial data to estimate predict the timeline. They

measured model performance in all cases using a combination of precision, recall, and the F1 measure [99].

From a corporate point of view, these questions are critical for several reasons. A company will always be weighing the cost-benefit of proceeding with actions, even if there is a chance of being sued, it will largely depend on what the estimated likelihood of litigation it is. The higher the likelihood, the more likely it is that a company will change its course of action, or use more mitigating factors to guard against, or to positively affect (i.e. lower) the likelihood. However, a company can only amend its approach if they are aware of the likelihood in the first place. As well, if the company needs to make changes to either avoid litigation or lower the likelihood, it will most likely need to reallocate resources — either in time, money, or (most probably) a combination of both. This is something that a company cannot do instantaneously, and must prepare for. In addition, if the company is unable to reduce the likelihood or avoid the lawsuit altogether, it will need to allocate funding for its defence — something that companies prefer to budget for in advance. There is, of course, the other point of view, which is when the company itself is going to launch the lawsuit for infringement of intellectual property which is, again, something that companies may need time to allocate funding for. As such, knowing the probability and lead time to litigation is valuable information for companies.

Gruginskie and Vaccaro also researched lawsuit lead time to determine the quality of the court system in Região in Southern Brazil [38]. Data was provided by the Tribunal Regional Federal da 4ª Regiã from 2016, and used a range of categorical variables including electronic lawsuits, lead time, subject, and class [38]. Gruginskie and Vaccaro used four machine learning algorithms with k-fold cross validation: Support Vector Machines, Naive Bayes, Random Forest and Neural Networks. Accuracy, Precision, Recall, and the F1 measure were used for evaluation purposes [38].

The models were broken down into four time frames: Up to One Year; From 1 to 3 Years; From 3 to 5 Years; and More than 5 Years. Results up to One Year proved to be the most robust, with Support Vector Machines and Random Forest returning the best performance with F1 measures of 83.58 and 83.33, respectively [38]. Naive Bayes returned, consistently, the worst results over the different time periods, with the exception of the 3 to 5 year category where Random Forest performed the worst.

The variances between the models for each machine learning approach was also tested for statistical significance using Tukey's multiple comparisons test [38]. This test compares the means of paired samples while adjusting for multiple testing, resulting in either a confidence interval or a P-value [23]. Gruginskie and Vaccaro's results use P-value and indicate that each of the variances is statistically significant, with P-values ≤ 0.01 [38].

Part of of Alexander, al Jadda, Feizollahi, and Tucker's research examined features in lawsuits that could be used to predict the outcome of a lawsuit [2]. Using features that they had extracted from source documents such as the lawsuit itself and the docket sheet from trial, as well as ancillary documents such as summary judgements and the magistrate's report, if they were available, they used those to feed into a random forest model to predict the outcomes of a series of lawsuits [99]. The possible outcomes that were considered in their research was were "...dismissal, motion for summary judgment, pre-discovery settlement, and post-discovery settlement" [2]. Alexander *et al.* constructed four models based on the amount of information known throughout the lawsuit. The model that proved to be the most useful and insightful used the most amount of features thus providing the full range of information known in the lawsuit. From a machine learning perspective, this makes sense given that the model's performance will increase with the inclusion of more data. For each model, researchers identified the top four feature predictors. For the most informative model, the top predictors were dismissed that did not terminate litigation, as well as performance rates for both the plaintiffs and the defendants' attorneys, resulting in 94% accuracy [2].

Şulea, Zampieri, Malmasi, Vela, Dinu, and van Genabith conducted research to predict the accuracy of the Supreme Court Rulings in France, as well as the time period of the ruling, and the area of the law that the ruling pertained to [94]. It should also be pointed out that France's Supreme Court is also called the Court of Cassation. It is not meant to act as a final level of the judiciary, but rather a legal apparatus which determines if the lower courts applied judicial rules correctly [92]

5.2 Overall methodology for the Case Study

Rogers, Van Buskirk, and Zechman used plaintiff complaints to determine which corporate disclosures were most likely to put a firm at risk of litigation. They then selected 20 random lawsuits and examined a variety of disclosure types including earnings announcements, press releases and SEC filings [71]. The lawsuit information was provided by the Insurance Firm Woodruff-Sawyer, the last earnings announcement prior to the beginning of the alleged damage period was drawn from the Institutional Brokers Estimate System (commonly referred to as I/B/E/S), and the legal filings from the Stanford Class Action Clearinghouse (SCAC), focusing on complaints made under Rule 10b(5) [71].

The authors focused on establishing a benchmark around optimistic tone, defining it as "the extent to which managers frame their firms' results or outlook in a favorable manner". [71]. It is important to note that sentiment dictionaries in use today do not consider optimism in its true form (which is forward looking) but rather only consider positivity (which is rooted in present). Yet, parts of the MD&A are focused on the future, as management (usually) discusses the future, in relation to either past performance, future performance to improve on past performance, or what it sees as its expected performance. The discussion in the MD&A is usually a combination of all of these time periods.

While the research done by Rogers *et al.* used earnings announcements as well as the specific optimistic language from those disclosures that plaintiffs cited in their lawsuits [71], the underlying theme of potentially inflated optimism in disclosure was consistent with our research. The authors also used three word dictionaries to conduct their analysis - Diction, Loughran-McDonald, and Henry, with the former intended to represent a more generalized audience that would also include other non-financial professional domains such as law, and the latter two as indicative of financial professionals [71].

Rogers *et al.* did not disclose which companies were included in the dataset. Without that information, we could not fully follow their approach. We determined, however, that we would follow the main idea of the methodology, and use the Stanford Securities Class Action Clearinghouse (SCAC) to select random lawsuits enforced

under Rule 10b(5). Figure 5.1 shows the heat map of the SCAC by ranked from the most sued to least sued sector [95]. Following the method laid out in Rogers *et al.* [71], we randomly selected 96 lawsuits from the heat map — 16 from each of the top three sectors (Technology, Services, and Financial) and 16 from each of the bottom three (Utilities, Transportation, and Conglomerates) over the period from 1990 to 2017 that met certain criteria.

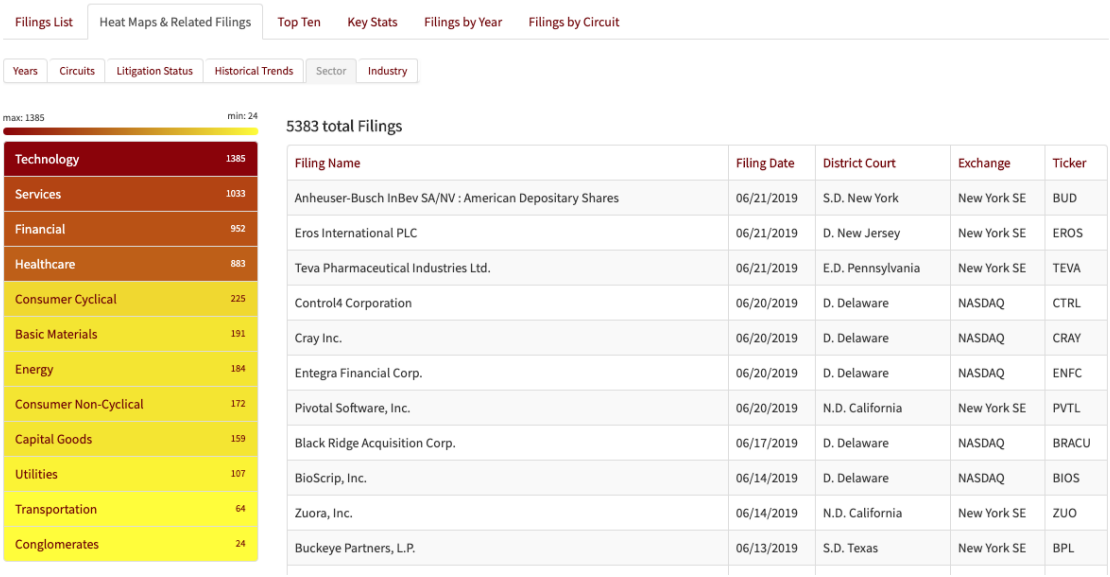

Figure 5.1: Heat Map by Sector

The requirements for inclusion in our dataset were that:

- the company had to be public so we would be able to access the company's 10-K and 10-Q reports from the dataset maintained by Bill McDonald
- the lawsuits had to be drawn equally from the top three and the bottom three sectors (as indicated on the Stanford heat map in figure 5.1);
- the class action lawsuit had to be promulgated under Rule 10b; and
- the lawsuit's status had to either be "settled" or "dismissed".

Rule 10b, which is most often addressed under section 5, addresses deception and making false statements, among other things.Rule 10b(5) states [21]

"It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails or of any facility of any national securities exchange,

(a) To employ any device, scheme, or artifice to defraud,

(b) To make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made, in the light of the circumstances under which they were made, not misleading, or

(c) To engage in any act, practice, or course of business which operates or would operate as a fraud or deceit upon any person, in connection with the purchase or sale of any security" [1]

The lawsuit status was also important as it had to be either "settled" or "dismissed". This means that a verdict had been rendered and the lawsuit had come to a close. Lawsuits with the status of either "remanded" (i.e. the lawsuit was sent back to a lower court for various reasons) or "ongoing" as the trial was still in progress were not included in the dataset as an outcome had not yet been reached. Including these lawsuits would have skewed the statistical significance, when examined in conjunction with lawsuits that had been decided, and it also would have impacted the machine learning results as well, as there was nothing concrete to yet predict.

Our next step was to review the class action court filings to identify the Class period Class Period, which is the alleged damage period over which the plaintiffs indicate that they incurred harm to their investment due to the company's statements and actions. It should not be confused with the length of the lawsuit. The damage period can be very short (such as one month) with the litigation lasting years, or vice versa. The Class Period directed us to which financial filings were going to be relevant to the sentiment of the lawsuit.

The launching of the lawsuit is (usually) the first notification to the company that shareholders believe that they have incurred damage. Companies carefully choose their words in communicating with stakeholders and will make critical changes in response to what is happening in the market. Rogers and Van Buskirk researched the change behaviour post-litigation and found that, in an effort to avoid future accountability, decrease the amount and quality of disclosure [70]. Therefore, the

first financial filing at the beginning of the damage period is critical as it provides the most informative view of how management views the company's performance, and its approach to disclosure at the time of the beginning of the alleged damage period.

If the beginning of the alleged damage period (class period) aligned with the date of the company's SEC filing, we used that report as the first filing. If it did not line up, we then considered the start of the class period in relation to when the last SEC filing had been provided, and used that filing for the beginning. For example, if the class period began on April 15th, which is in the second quarter, assuming a December 31st year end, then we would use the first quarter SEC filing as the start. We believe that using the second quarter filing as the critical "first" filing would not be appropriate because it is likely that the filing from the first quarter has influenced the alleged damage. Also, if the company had been legally served (or the threat of litigation had been brought up with the company), they would still have the remainder of the second quarter to adjust their disclosure strategy.

While important, subsequent filings are not as critical as the company will have changed its disclosure strategy and altered the language used in the filings that follow the commencement of litigation, as per Rogers and Van Buskirk [70]. While it was, in some cases, possible to link the beginning of the class period with a specific SEC filing date, it was much more challenging aligning it with the end of the alleged damage period. Therefore, we again considered the quarter in relation to the end of the class period and selected the most appropriate SEC filing as the final report. All filings between the beginning and the end of the alleged damage period were included as long as they included the MD&A and market risks, were 250 words long, and were either a 10-K or 10-Q report.

Using all of the information discussed above from the SCAC and the SEC filings, we created a dataset. From the sentiment analyses, we captured the date, the Central Index Key (CIK) (this is a unique company identifier assigned to each company by the SEC), and the change in tone (based on the movement in the sentiment prior to extraction of the sentences with the non-GAAP words and after) for each dictionary. From the SCAC, we retrieved the class action filing documents from the companies previously randomly selected from each sector to determine the class period (alleged damaged period) and the outcome of the lawsuit. Ligitation that, at the time of the

experiments, was still ongoing or remanded (i.e. sent back to a lower court for various reasons) were kept for the statistical analyses, but dropped from machine learning as there was no outcome yet to predict.

The attributes for the each dataset are as follows:

- sector
- date
- cik (central index key)
- cgi (change in the sentiment for the General Inquirer dictionary)
- cqdap (change in the sentiment for the QDAP dictionaries)
- che (change in the sentiment for Henry dictionary)
- clm (change in the sentiment for the Loughran-McDonald dictionary)
- period (class period - the alleged damage period)

The class used for prediction was the outcome of the lawsuits as either settled or dismissed.

5.3 Statistical Hypotheses for the Case Study

The use of non-GAAP measures have been shown to improve the tone of documents. Due to this, our hypothesis was that once the non-GAAP measures were extracted from the text, the tone would decrease. Therefore, our hypothesis is as follows:

Hypothesis 1. *Wilcoxon Signed Rank Test*

As our case study data did not follow normal distribution, which we demonstrate below, we decided to use the Wilcoxon Signed Rank test. It is considered the equivalent of the paired t-test, but for non-parametric data [93].

Null Hypothesis: The change in the sample mean $\overline{X} = 0$ for the dictionary under evaluation

Alternative Hypothesis: The change in the sample mean $\overline{X} < 0$ for the dictionary under evaluation

5.4 Statistical Experiments for the Case Study

We performed several statistical tests to evalaute the data:

E1: We evaluated the dataset by selecting each dictionary separately while considering the aggregate data from all sectors. By doing this, we investigated if the dictionary being tested was statistically significant across all of the sectors.

E2. The dataset was broken into its constituent parts of top 3 sectors and bottom three sectors. In this experiment, we evaluated the *top three* sectors to determine if the dictionary under evaluation was statistically significant.

E3. We performed the same tests as experiment 2, but this time evaluating the *bottom 3 sectors*.

We considered using the average tone change for each company over the class period rather than all of its financial filings during the alleged damage period. However, this approach created a data sparsity issue, which in turn provided extremely poor results which could not be relied on and which did not provided extremely limited information.

5.4.1 Wilcoxon Signed Rank Test

Before the statistical significance tests could be done, we had to determine the distribution of the data. We initially used histograms to plot the distribution, and if there was any doubt as to its normally, we then confirmed with the use of boxplots.

We first evaluated the aggregate of all the sectors, as can be seen below in Figure 5.2. Although the data looks normally distributed, we confirmed with boxplots that it is, in fact, non-parametric. See Figure 5.3.

We therefore conclude that for the aggregate of all sectors that we will be unable to use the paired t-test, as the data is not normally distributed.

The distribution was also checked for each constituent set (Top 3 and Bottom 3).

It is clear from Figure 5.3 that the General Inquirer and QDAP dictionaries do not follow a normal distribution. The Henry and Loughran-McDonald dictionaries first appeared to follow a normal distribution.

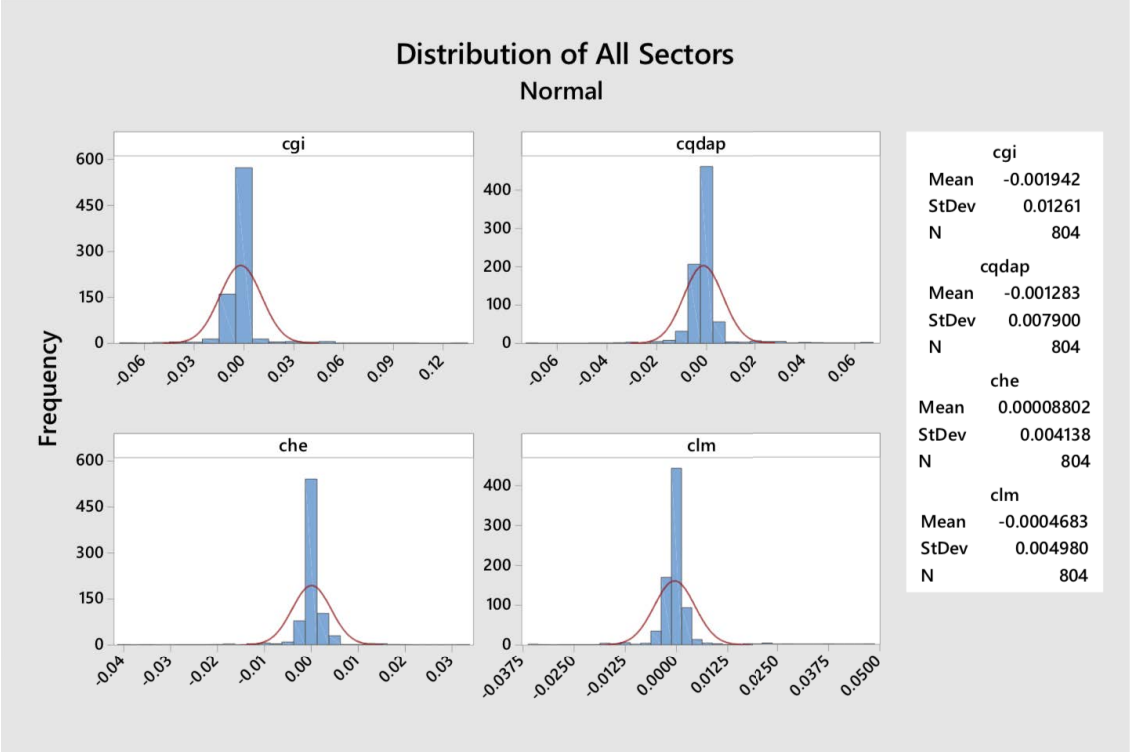

Figure 5.2: Distribution of All Sectors

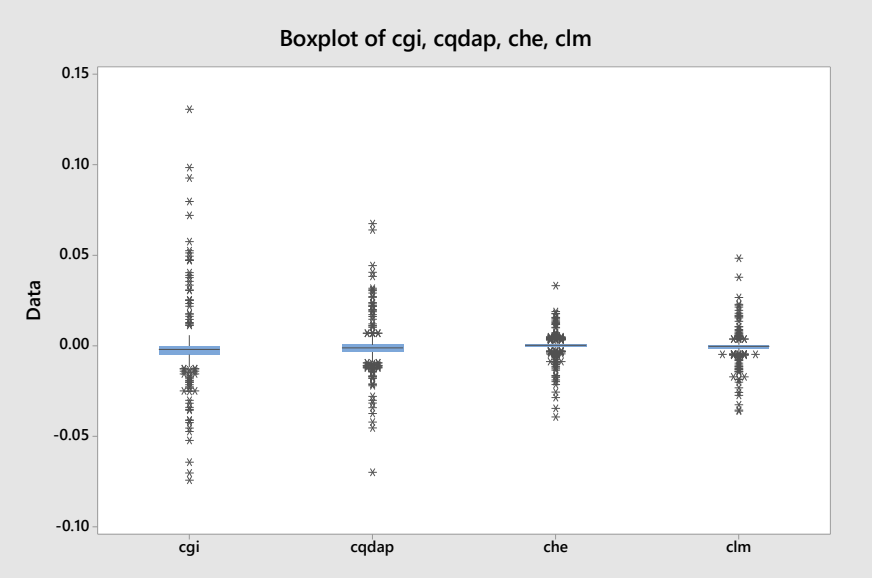

Figure 5.3: Boxplots of All Sectors

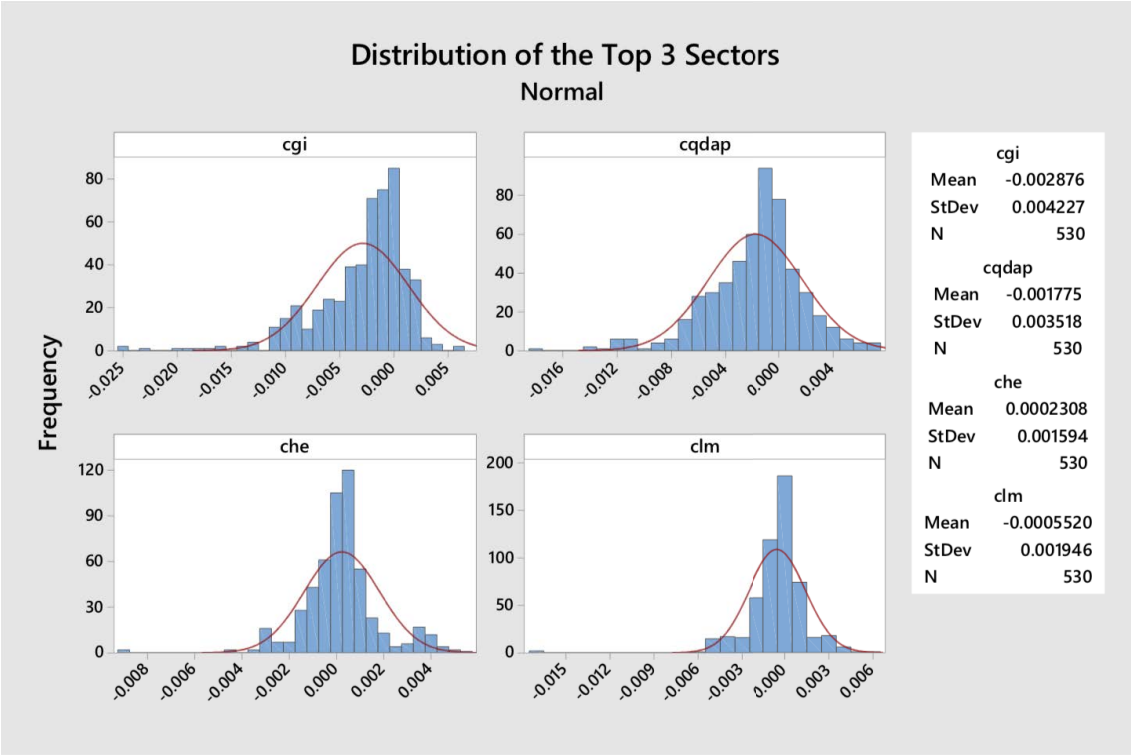

Figure 5.4: Distribution of Top 3 Sectors

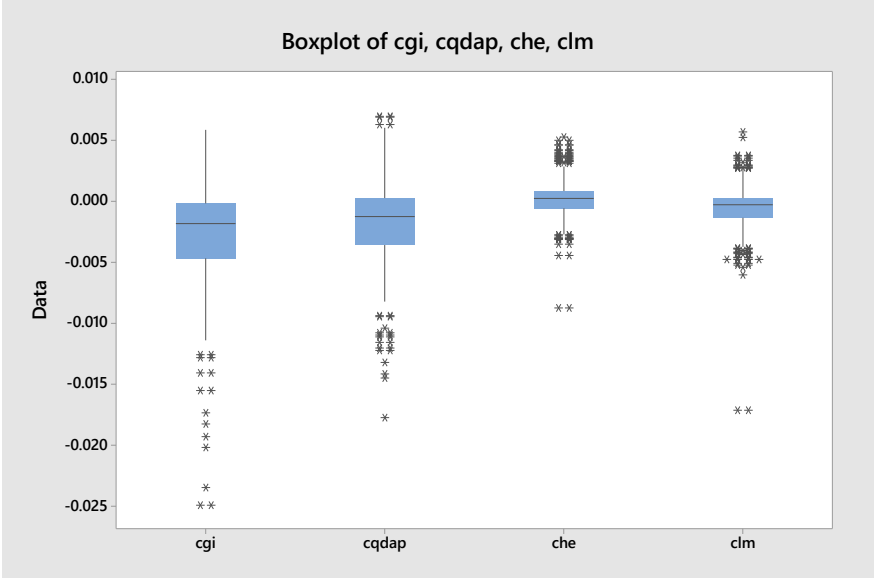

Figure 5.5: Boxplots of Top 3 Sectors

After reviewing the boxplots (Figure 5.5), we conclude that Henry and Loughran-McDonald are also non-parametric. Again, we will not be able to use the paired t-test in this case.

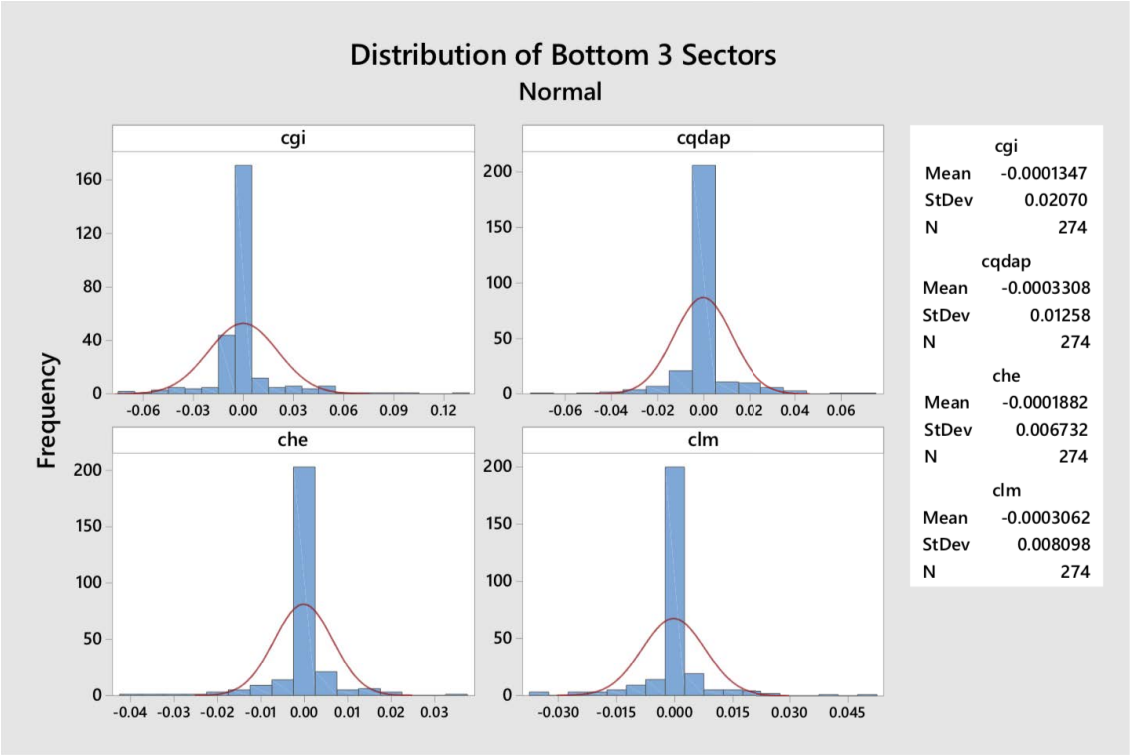

Figure 5.6: Distribution of Bottom 3 Sectors

We also reviewed the histograms for the bottom three sectors and confirmed with boxplots that the distribution is non-parametric. See Figures 5.6 and 5.7.

As can be seen from each boxplot, there are a number of outliers in the Aggregate of All Sectors, the Top 3 and the Bottom 3. Before conducting the statistical tests, we investigated the outliers to determine if these records were to be kept in the dataset or removed. The plots show that the outliers are most prevalent with the General Inquirer and QDAP dictionaries, which are used as proxies for investors with little to no financial training. In reviewing the associated filings for the outlier companies, we determined that these are extreme cases which had used a large number of non-GAAP measures in their reports. While the overall sentiment (in aggregate) dropped after the removal of the non-GAAP measures, that was not the case for every individual filing. If a company is performing very badly, non-GAAP measures will only have a marginal effect on increasing the tone, if at all. Therefore, some

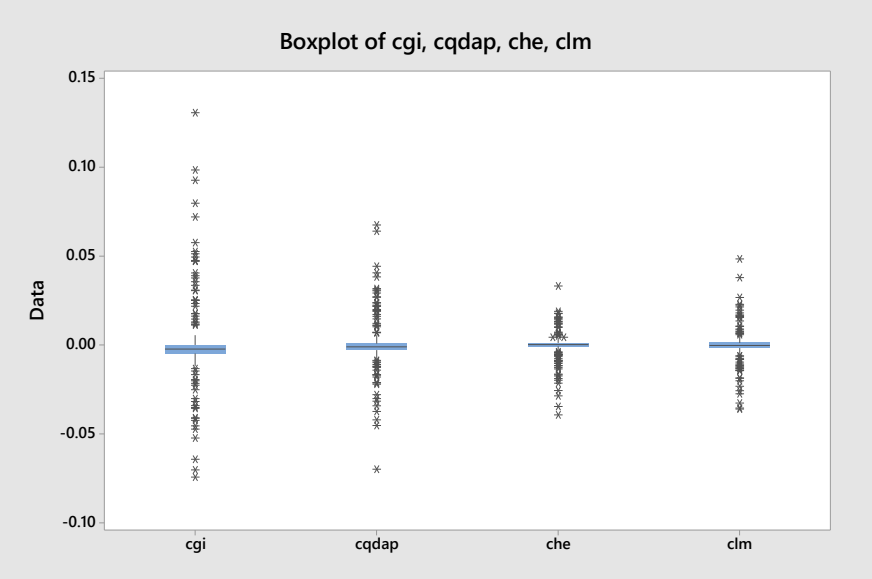

Figure 5.7: Boxplots of Bottom 3 Sectors

companies experienced an increase in the sentiment score after extraction. The larger majority of the extreme cases, however, were not performing badly necessarily, and were consistent with the drop in sentiment after extraction, and experienced a large decline due to the number of non-GAAP measures discussed.

As we concluded that our data was non-parametric, we were not able to used the paired t-test, as it is only used for normally distributed data. We then evaluated the assumptions for the Wilcoxon Signed Rank test, which is considered the non-parametric equivalent of the paired t-test [93].

In order to use Wilcoxon, we must first satisfy four assumptions [90]:

1. *Dependent samples* — The basis for the samples must be the same (i.e. dependent) as Wilcoxon will be comparing the samples under "before" and "after" scenarios. As the experiments for our data followed this convention, this assumption was satisfied.

2. *Independence* — The samples must be independently chosen, in that each report must have equal chance of being included in the sample. As we chose our sample randomly, equal chance was afforded to each report, satisfying this assumption.

3. *Continuous Dependent Variable* — the assumption here is that measurements can take on an infinite number of values. The measurement for the sentiment analysis ranges from -1 to 1, where results can be any number within that range, including -1 and 1. As such, this assumption is satisfied.

4. *Ordinal Level of Measurement* — We are able to determine if one value is greater, less, or equal to that of another. The resulting value of the change is a float, and therefore can be compared against another float to determine which value is greater, less, or equal to the other. As such, this criterion is met.

As we were able to satisfy all of the assumptions, we concluded that the use of the Wilcoxon Signed Rank test was appropriate to evaluate statistical significance.

5.5 Machine Learning for the Case Study

We chose three statistical algorithms, Naive Bayes, Support Vector Machines, and Random Forest, to determine how well our data could predict the outcome of the class action lawsuits.

5.5.1 Background on Algorithms Chosen

Naive Bayes

The Naïve Bayes (NB) classifier is based on Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

From this, we can determine the probability of A given B. What makes the classifer *naïve* is that it assumes that each variable is independent of one another in determining the dependent variable [98]. If there are strong dependencies between the variables, then NB is not appropriate, given the assumption of independence [60]. Given that, however, this classifier has shown to be quite robust in its ability to predict outcomes using this naïve approach because ideal decisions can still be made even if those decisions are not as accurate as they could be due to variable interdependency [60].

In the context of our dataset, any of the input variables such as change in tone (from any and all dictionaries), the (alleged damage) period, and the sector could be independently used to determine the dependent variable, which in our case, is the class or outcome of the securities lawsuit as either *settled* or *dismissed*. Practically speaking none of these variables are purely independent of each other, and we know from prior research and literature that tone effects investors. To our knowledge, however, the degree of effect has not been calculated; it would be very difficult to quantify the effect given that each person has a different level of financial sophistication, training, experience, and risk aversion — all of which would affect the degree to which financial tone would affect a particular investor. Given that we do not know to what degree the variables are interdependent, we determined that it was, therefore, appropriate to use the Naïve Bayes classifier.

Support Vector Machines

In the context of classification, Support Vector Machines are algorithms that work to find the best separation between the various classes. This separation is known as the maximum margin hyperplane. Figure 5.8 is an example of an SVM hyperplane [98]. In this particular case, the algorithm has determined that the diagonal (in this example) line is the best hyperplane to separate the two classes - light circles and dark circles. The figure has also identified the support vectors (three in this case); these are the points closest to the hyperplane. There is always a minimum of two support vectors - one on each side of the hyperplane, but there can be, and often there is, more, as we see in this example. The margin is the distance between the support vectors and the hyperplane. Therefore, the algorithm is searching for the hyperplane with the greatest margin between it and the support vectors [98]. Given that, we believed that SVM would be an appropriate algorithm to use for our research.

Random Forest

We are all familiar with how decision trees work — when you come to a decision point, you take the best decision among the inputs available, then follow the path chosen until you come to the next decision point, and repeat until you have gone as far as you can to ultimately make your decision. The Random Forest algorithm works largely the same way, but rather than taking the best decisions available at a particular node, the decision is made based on choosing the best among randomly

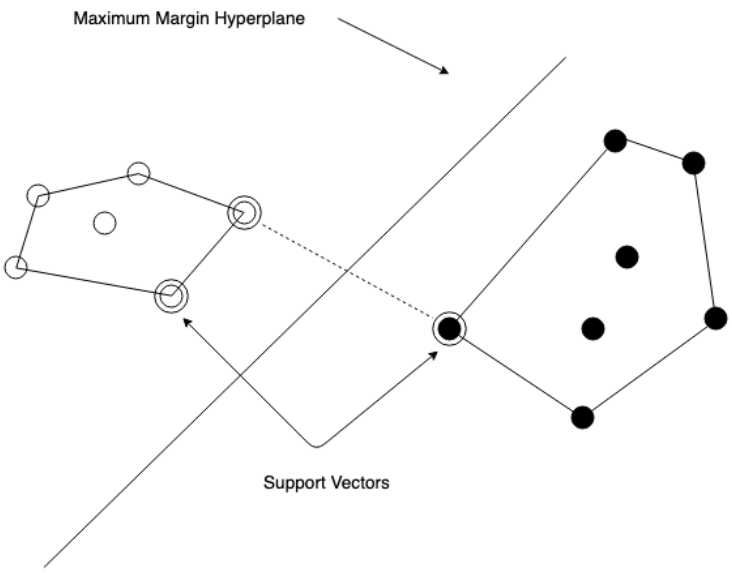

Figure 5.8: SVM Hyperplane

chosen predictors at each decision node [51]. This approach, Breiman indicates helps to prevent overfitting [16].

5.6 Statistical Results for the Case Study

All Sectors

Dictionary	Number of Samples	Median	Wilcoxon Statitstic	P-value
GI	804	−0.0022814	60282.00	<0.001*
QDAP	804	−0.0013595	87728.00	<0.001*
HE	804	0.0002021	188729.00	1.00
LM	804	−0.0003928	116941.00	<0.001*

Table 5.1: Wilcoxon Signed Rank Results - All Sectors

**significant at the 0.01 probability level*

As can be seen in Table 5.1, the General Inquirer (GI) and QDAP dictionaries are statistically significant. The GI and the QDAP dictionaries are general, and not specific to the domain of finance. Therefore, we believe that the changes in the sentiment of these two dictionaries reflect those of the lay person who are, based on the documentation that we reviewed, the lead plaintiffs in the cases. The Loughran-McDonald dictionary is also statistically significant. This suggests that companies

are using a sufficient amount of non-GAAP measures that, when extracted, cause a significant change in the tone. The Henry is not statistically significant, which is an important contrast to the other three dictionaries. We believe that this is due the small number of words in this particular dictionary, which makes it difficult to capture the change in the tone when the sentences are extracted.

Top 3 Sectors

Dictionary	Number of Samples	Median	Wilcoxon Statitstic	P-value
GI	804	−0.0022814	60282.00	<0.001*
QDAP	804	−0.0013595	87728.00	<0.001*
HE	804	0.0002021	188729.00	1.00
LM	804	−0.0003928	116941.00	<0.001*

Table 5.2: Wilcoxon Signed Rank Results - Top 3 Sectors

**significant at the 0.01 probability level*

Bottom 3 Sectors

Dictionary	Number of Samples	Median	Wilcoxon Statitstic	P-value
GI	804	−0.0022814	60282.00	<0.001*
QDAP	804	−0.0013595	87728.00	<0.001*
HE	804	0.0002021	188729.00	1.00
LM	804	−0.0003928	116941.00	<0.001*

Table 5.3: Wilcoxon Signed Rank Results - Bottom 3 Sectors

**significant at the 0.01 probability level*

Similarly, the results for the Top 3 Sectors and the Bottom 3 Sectors also show that the Henry dictionary is not statistically significant, when the other three are. Again, we postulate that the reasons for these results are the same as above for All Sectors.

5.6.1 Machine Learning Experiments

For the experiments, we took various approaches such as changing the number of independent variables (attributes) and tuning models to see what would produce the

best results. Each algorithm has been addressed below, detailing the experiments performed using each.

We performed a number of experiments that varying the amount of independent variables that the model had to work with, ranging from three attributes (change in sentiment for one dictionary, class period, and outcome) to all eight attributes listed below.

The common list of attributes available in our dataset were:

1. date (this is the date that the company filed the report with the SEC)
2. central index key (cik, which acts as the company number)
3. cgi (the change in tone measurement for the General Inquirer)*
4. che (the change in tone measurement for the Henry dictionary)*
5. clm (the change in tone measurement for the Loughran-McDonald dictionary)*
6. cqdap (the change in tone measurement for the QDAP dictionary)*
7. period (the alleged damage period by the plaintiffs)
8. outcome (the outcome of the class action lawsuit as either *settled* or *dismissed*)

** Note that the change in the tone is measured by subtracting the tone after extracting the non-GAAP sentences from the tone before extraction.*

At the time of the experiments, there were four lawsuits in the financial sector, one in the services sector, and one in the technology sector that had to be removed from consideration as the litigation was still ongoing. Therefore, the outcome (for prediction) was not yet known.

E1: This experiment was performed for the aggregate of sectors (Technology, Services, and Financial) and dictionaries (GI, HE, LM, and QDAP).

E2: This experiment was performed for the aggregate of sectors, as well as top 3 and bottom 3, but was done with each dictionary individually.

5.7 Machine Learnings Results

Here, the problem is predicting the outcome of securities class action lawsuits, which are inherently expensive (regardless of outcome) in the financial domain. So we offer contextualization for precision and recall in financial consequential terms. If the cost **of acting** is high — such as investing in a stock — then precision is the most important. If the cost **of not acting** is high - such as taking steps to avoid being overly optimistic in disclosure tone in order to mitigate or avert a lawsuit — then recall is the most important. With recall, there will be false negatives, but in protecting a company from the increased potential for litigation, those false negatives should not be viewed as problematic; it is far less costly, from a litigation standpoint, to amend the disclosure ahead of release, than to address a lawsuit (or the threat of one) after.

Securities legal action is a very costly endeavour, and often plaintiffs are not fully remunerated for the damage incurred [71]. We reviewed the court documents for the companies in our dataset to determine the extent of the settlements, fees, and costs. In the majority of settled cases in our dataset, counsel for the plaintiffs were awarded a percentage of the settlement amount, and an amount in addition up to a certain threshold, as well as ancillary costs. The following table outlines the lowest and highest settlement amounts, as well as the lowest and highest attorneys' fees and costs. Note that there is no correlation between the settlement and the attorneys' fees in table 5.1 below.

	Settlements	Attorneys' Fees and Costs
Lowest	$\$712,500$	$\$292,500$
Highest	$\$219,000,000$	$\$18,183,161$

Table 5.4: Settlements and Attorneys' Fees and Costs

This also highlights the fact that there is a tremendous amount of risk in disclosure; if a company is overly optimistic, they face litigation, but if disclosure is overly pessimistic, they run the risk of that pessimism affecting the market, and ultimately the share price. It is important to note that all of these lawsuits were promulgated under Rule 10-b, and investors pointed to overly optimistic disclosure. The cost

of dampening the optimism in the reports by eliminating or minimizing the use of non-GAAP measures, which have been proven to increase the tone of disclosure, is comparatively less — significantly so.

Keeping in mind that *recall* is the measure that we are focusing on, we see that Random Forest (RF) is predominantly the best algorithm for our dataset. When all of the features are considered for the aggregate of all of the sectors (both top and bottom), we see a result of 0.9142 (see Table 5.2).

Naive Bayes

Naive Bayes (NB) performed the best of all of the algorithms when classifying the Aggregate using the Sentiment Score, the Period, and the Outcome, resulting in a recall of 0.9794. NB also outperformed Random Forest again when classifying both the Top 3 and the Bottom 3 sectors using just the Sentiment and the Outcome. We believe that this is due the tenet of Naive Bayes, which is that all of the variables are assumed to be conditionally independent. It also works well with small datasets, which we have. See Table 5.2.

Random Forest

As can be seen in Table 5.2, the results returned using Random Forest are quite robust, returning a recall of 0.9142 for the Aggregate using all features, and 0.9938 and 0.9407 for the Top 3 and Bottom 3, respectively. At each node, this algorithm is designed to choose the best among randomly chosen predictors to make its decision, and then move on, to prevent overfitting. Random Forest also works well with both numerical and categorical data, which we have. As well, because it employs a bootstrapping method (i.e. that samples are selected and then replaced to be selected again the future), and therefore makes the random tree more robust.

When the RF has all available data, and thus performs the best, it is very informative to see what it splits on. As seen in Figure 5.9, the Central Index Key (CIK), which is the company number, proves to be an important feature, even though we did not anticipate that. The other feature that RF also zeroed in on was the period. As the sentiment was not a splitting feature, we were initially concerned that Random Forest may not be able to provide an acceptable recall result as features were removed. Yet, as seen in Table 5.2, Random Forest still performs very well, although

now not the best, with recalls of 0.9446 and 0.8207, as features are removed.

As can be seen in Table 5.2, the best result between the three algorithms is produced using Random Forest for the Top 3 sectors, achieving a recall of 0.9938. We varied the number of attributes used between tests, starting with all the attributes, and then removing one (company number) and then two (company number and period). Random Forest performed very robustly, never dropping below 0.7986, even though we had removed all attributes except for the sentiment and the outcome of the lawsuit. The ensemble nature of this algorithm is particularly well suited to this classification task as it uses prediction by committee to overcome the shortcomings of the individual trees.

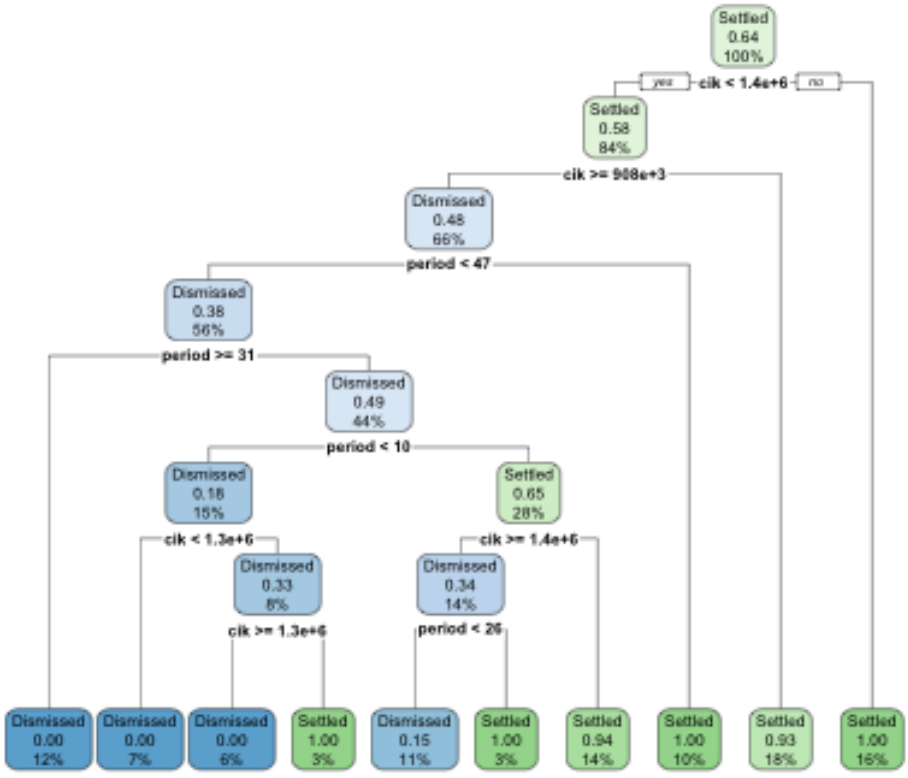

Figure 5.9: Random Forest - Case Study

Support Vector Machines

Support Vector Machines (SVM) performed the worst out of the statistical algorithms. The highest recall was 0.6600, was for the Bottom 3 Sectors using the

Sentiment, Period, and Outcome, but was still far off the best performing classifier — Random Forest with a recall of 0.9111, see Table 5.2. In our dataset, there are a number of filings where the change between the "before" and "after" was zero. This means that the company did not use any of the non-GAAP measures in our extraction list. We believe that due to the fact that this type of paired data cannot be easily separated, that Support Vector Machines is not well suited to our type of data.

Algorithm and Dataset	Precision	Recall	F1	Accuracy
Aggregate (All Features)				
Naive Bayes	0.6049	0.8822	0.7424	0.7973
Random Forest	0.9123	**0.9142**	0.9133	0.9210
Support Vector Machine	0.6610	0.6390	0.6100	0.6439
Top 3 (All features used)				
Naive Bayes	0.6409	0.8822	0.7424	0.7973
Random Forest	0.9493	**0.9938**	0.9710	0.9754
Support Vector Machines	0.6220	0.6240	0.6021	0.6137
Bottom 3 (All features used)				
Naive Bayes	0.6478	0.8067	0.7185	0.7798
Random Forest	0.8819	**0.9407**	0.9104	0.9104
Support Vector Machines	0.6884	0.6568	0.6453	0.6426
Aggregate (Sentiment, Period, Outcome)				
Naive Bayes	0.6597	**0.9794**	0.7884	0.8172
Random Forest	0.8100	0.8879	0.8472	0.8668
Support Vector Machines	4.6235	0.6280	0.6057	0.6090
Top 3 (Sentiment, Period, Outcome)				
Naive Bayes	0.6054	0.8761	0.7160	0.7812
Random Forest	0.8149	**0.9446**	0.8750	0.8990
Support Vector Machines	0.6235	0.6280	0.6057	0.6090
Bottom 3 (Sentiment, Period, Outcome)				
Naive Bayes	0.6149	0.8306	0.7067	0.7785
Random Forest	0.8542	**0.9111**	0.8817	0.8817
Support Vector Machines	0.6905	0.6600	0.6491	0.6555
Aggregate (Sentiment, Outcome)				
Naive Bayes	0.5581	0.7152	0.6270	0.6803
Random Forest	0.6977	**0.8207**	0.7542	0.7481
Support Vector Machines	0.2693	0.5194	0.3543	0.2734
Top 3 (Sentiment, Outcome)				
Naive Bayes	0.5248	**0.9610**	0.6789	0.7445
Random Forest	0.8089	0.8468	0.8274	0.8434
Support Vector Machines	0.5492	0.5492	0.4793	0.4636
Bottom 3 (Sentiment, Outcome)				
Naive Bayes	0.5248	**0.9610**	0.6789	0.7445
Random Forest	0.8156	0.7986	0.8070	0.7993
Support Vector Machines	0.4965	0.5119	0.4040	0.5002

Table 5.5: Case Study Machine Learning Results

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Our research presented a novel approach to quantitatively measure the effect on sentiment on 10-K and 10-Q reports filed with the U.S. Securities and Exchange Commission. Our approach used a specific list of non-GAAP measures that were extracted, along with the supporting words in the sentence where the non-GAAP measure appeared, from one copy of the report, which was then compared to the report as filed with the SEC. This provided quantitative measure of the effect that the selected non-GAAP measures and the supporting words had on the tone of the reports. The results of our experiments showed that after extraction, the sentiment decreased for each dictionary used. We also found that the decrease in the sentiment was statistically significant, as the 0.01 level.

We then built upon our findings and performed two case studies to determine if we could use the change in the extracted report versus the report as filed to predict the outcome of Securities lawsuits promulgated under Rule 10-b. We looked at a group of randomly selected companies in the top 3 sectors and the bottom three sectors listed on the Stanford Securities Class Action Clearinghouse. We found that we can use the change in sentiment to predict the aggregate outcome with a recall of 0.9142.

While the work on both of these questions add to the academic literature, we also believe that the datasets themselves that we have created for both research questions are significant contributions to the research community.

6.2 Future Work

The extractive sentiment approach that we introduced in this paper has opened up a number of new areas of research. We only applied this approach on 10-K and 10-Q

reports that use U.S. GAAP as the reporting framework. We would like to expand our experiments to also apply this approach to reports under the International Financial Reporting Standards (IFRS). It would also be of great benefit to determine if there are any differences presented between the two frameworks when extracting the non-GAAP measures, given that there is currently an effort to converge the accounting frameworks around the world to IFRS.

Another interesting area of research would be to perform topic modelling on the reports with the non-GAAP measures, and those without the non-GAAP measures to see if there is any changes in the topic focus, using Latent Dirichlet Allocation. This would provide comparative evidence as to what the major topics are under both report approaches.

Finally, summarization of financial reports has become a major focus of Natural Language Processing tasks in recent years. An interesting research path would be to look at how non-GAAP measures influence and effect those summarization tasks.

References

[1] 82nd Congress of the United States. § 240.10b-5 employment of manipulative and deceptive devices. https://www.law.cornell.edu/cfr/text/17/240.10b-5, 1951. Last Accessed: 2019-06-21.

[2] Charlotte S. Alexander, Khalifeh al Jaada, Mohammad Javad Feizollahi, and Anne Tucker. Using text analytics to predict litigation outcomes: A preliminary assessment, 2018.

[3] Kristian D. Allee, Nilabhra Bhattacharya, Ervin L. Black, and Theodore E. Christensen. Pro forma disclosure and investor sophistication: External validation of experimental evidence using archival data. *Accounting, Organizations and Society*, 32(3):201–222, 2007.

[4] David R. Anderson, Dennis J Sweeney, and Thomas A. Williams. *Statistics for Business and Economics*. South-Western Cengage Learning, eleventh edition, 1981.

[5] H. Scott Asay, Robert Libby, and Kristina Rennekamp. Firm performance, reporting goals, and language choices in narrative disclosures. *Journal of Accounting and Economics*, 65(2-3):380–398, 2018.

[6] M. Barth, I. Gow, and D. Taylor. Why do pro forma and street earnings not reflect changes in gaap? evidence from sfas 123. *Review of Accounting Studies*, 17:526–562, 2012.

[7] Jeremiah W. Bentley, Theodore E. Christensen, Kurt H. Gee, and Benjamin C. Whipple. Disentangling managers' and analysts' non-gaap reporting. *Journal of Accounting Research*, 56(4):1039–1081, 2018.

[8] Nilabhra Bhattacharya, Ervin L Black, Theodore E Christensen, and Chad R Larson. Assessing the relative informativeness and permanence of pro forma earnings and gaap operating earnings. *Journal of Accounting and Economics*, 36(1):285–319, 2003. Conference Issue on.

[9] Dirk E. Black, Theodore E. Christensen, Jack T. Ciesielski, and Benjamin C. Whipple. Non-gaap reporting: Evidence from academia and current practice. *Journal of Business Finance & Accounting*, 45(3-4):259–294, 2018.

[10] Robert J. Bloomfield. The "incomplete revelation hypothesis" and financial reporting. *Accounting Horizons*, 16(3):233–243, 2002.

[11] Financial Accounting Standards Board. Accounting standards codification, topic 280 segment reporting. https://www.iasplus.com/en-us/standards/fasb/presentation/asc280, n.d.

[12] Financial Accounting Standards Board. About the FASB. https://www.fasb.org/facts/, No Date. Last Accessed: 2019-06-03.

[13] Public Company Accounting Oversight Board. AS 3105: Departures from Unqualified Opinions and Other Reporting Circumstances. https://pcaobus.org/Standards/Auditing/Pages/AS3105.aspx, No Date. Last Accessed: 2019-06-03.

[14] Benoit Boyer, Ralph Lim, and Lyons Bridget. A case study in the use and potential misuse of non-gaap financial measures. *Journal of Applied Business and Economics*, 18(3):117–126, 2016.

[15] Mark T. Bradshaw and Richard G. Sloan. Gaap versus the street: An empirical assessment of two alternative definitions of earnings. *Journal of Accounting Research*, 40(1):41–66, 2002.

[16] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[17] Lawrence D. Brown and Marcus L. Caylor. A temporal analysis of quarterly earnings thresholds: Propensities and valuation consequences. *The Accounting Review*, 80(2):423–440, 2005.

[18] David Burgstahler and Ilia Dichev. Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics*, 24(1):99–126, 1997. Properties of Accounting Earnings.

[19] Samuel W.K. Chan and Mickey W.C. Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:128–147, 2017.

[20] Samuel W.K. Chan and Mickey W.C. Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64, 2017.

[21] United States Congress. § 240.10b-5 Employment of manipulative and deceptive devices. https://www.law.cornell.edu/cfr/text/17/240.10b-5, n.d. Last Accessed: 2019-06-04.

[22] Comprehensive R Archive Network (CRAN). DictionaryHE. https://rdrr.io/cran/SentimentAnalysis/man/DictionaryHE.html, 2019. Last Accessed: 2019-05-17.

[23] Nicola Crichton. Tukey multiple comparison test. http://www.blackwellpublishing.com, n.d.

[24] A. Curtis, S. McVay, and B. Whipple. The disclosure of non-gaap earnings information in the presence of transitory gains. *The Accounting Review*, 89:933–958, 2014.

[25] Angela K Davis and Isho Tama-Sweet. Managers' use of language across alternative disclosure outlets: Earnings press releases versus md&a. *Contemporary Accounting Research*, 29(3):804–837, 2012.

[26] Cambridge Dictionary. "jargon" in American English . https://dictionary.cambridge.org/dictionary/english/jargon, n.d. Last Accessed: 2019-06-21.

[27] Edward William Dolch. A basic sight vocabulary. *Elementary School Journal*, 36:456–460, 1936.

[28] Jeffrey T Doyle, Jared N. Jennings, and Mark T. Soliman. Do managers define non-gaap earnings to meet or beat analyst forecasts? *Journal of Accounting and Economics*, 56(1):40–56, 2013.

[29] W. Brooke Elliott. Are investors influenced by pro forma emphasis and reconciliations in earnings announcements? *The Accounting Review*, 81(1):113–133, 2006.

[30] Facebook. Facebook, Inc. Form 10-K. https://www.sec.gov/Archives/edgar/data/1326801, 2018. Last Accessed: 2019-09-13.

[31] Stefan Feuerriegel and Nicolas Proellochs. Package 'SentimentAnalysis'. https://cran.r-project.org/web/packages/SentimentAnalysis/ SentimentAnalysis.pdf, 2019. Last Accessed: 2019-06-19.

[32] Alex Fisher. Non-GAAP Measures - A 20-Year Echo. https://www.cpacanada.ca/en/business-and-accounting-resources/financial-and-non-financial-reporting/international-financial-reporting-standards-ifrs/publications/non-gaap-measures-academic-literature-review-1996-2016, 2016. Last Accessed: 2019-06-21.

[33] IFRS Foundation. Standards by Jurisdiction. https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/united-states/, n.d. Last Accessed: 2019-06-19.

[34] The IFRS Foundation. Who uses IFRS standards. https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/united-states, 2017. Last Accessed: 2019-05-19.

[35] Richard M. Frankel, Sarah McVay, and Mark T. Soliman. Street earnings and board independence. 2004.

[36] Raj Gnanarajah. Accounting and auditing regulatory structure: U.s. and international. *Congressional Research Service*, 2017.

[37] John R. Graham, Campbell R. Harvey, and Shiva Rajgopal. The economic implications of corporate financial reporting. *Journal of Accounting and Economics*, 40(1):3–73, 2005.

[38] Lúcia Adriana dos Santos Gruginskie and Guilherme Luís Roehe Vaccaro. Lawsuit lead time prediction: Comparison of data mining techniques based on categorical response variable. *PLOS ONE*, 13(6):1–26, 06 2018.

[39] Elaine Henry. Are investors influenced by how earnings press releases are written? *International Journal of Business Communication*, 45(4):363–407, 2008.

[40] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. *American Association for Artificial Intelligence*, 4(4):755–760, 2004.

[41] CFA Institute. Investor Uses, Expectations , and Concerns on Non-GAAP Financial Measures. https://www.cfainstitute.org/-/media/documents/support/advocacy/investor-uses-expectations-concerns-on-non-gaap.ashx, 2016. Last Accessed: 2019-06-21.

[42] Reviewed by Will Kenton Investopedia. SEC Form 10-K405. https://www.investopedia.com/terms/s/sec-form-10-k405.asp, 2018. Last Accessed: 2019-06-19.

[43] Helena Isidro and Ana Marques. The role of institutional and economic factors in the strategic use of non-gaap disclosures to beat earnings benchmarks. *European Accounting Review*, 24(1):95–128, 2014.

[44] Nirasimhan Jegadeesh and Di Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729, 2013.

[45] David Juckett. A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics*, 45:460–470, 2012.

[46] Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5):1449–1475, 2003.

[47] Taeyoung Kang, Do-Hyung Park, and Ingoo Han. Beyond the numbers: The effect of 10-k tone on firms' performance predictions using text analytics. *Telematics and Informatics*, 35(2):370–381, 2018.

[48] Kalin Kolev, Carol A. Marquardt, and Sarah E. McVay. Sec scrutiny and the evolution of non-gaap reporting. *The Accounting Review*, 83:157–184, 2018.

[49] Feng Li. Do stock market investors understand the risk sentiment of corporate annual reports? 2006. Last Accessed: 2019-06-21.

[50] Feng Li. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *The Journal of Accounting Research*, 48(5):1049–1102, 2010.

[51] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2/3:18–22, 2002.

[52] PricewaterhouseCoopers LLP. IFRS and US GAAP: similarities and differences. https://www.pwc.com/us/en/cfodirect/assets/pdf/accounting-guides/pwc-ifrs-us-gaap-similarities-and-differences.pdf, 2018.

[53] PricewaterhouseCoopers LLP. Non-gaap financial measures - issues in-depth. https://frv.kpmg.us/content/dam/frv/en/pdfs/2018/issues-in-depth-ngfm.pdf, 2018. Last Accessed: 2019-06-03.

[54] PriceWaterhouseCoopers LLP. Non-gaap measures and the ongoing dialogue: What you should know. https://www.pwc.com/us/en/cfodirect/publications/in-the-loop/non-gaap-measures.html, 2019.

[55] B. Lougee and C. Marquardt. Earnings informativeness and strategic disclosure: An empirical examination of "pro forma" earnings. *The Accounting Review*, 79:769–795, 2004.

[56] Tim Loughran and Bill McDonald. Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, 12(2):90–97, 2011.

[57] Tim Loughran and Bill Mcdonald. Documentation for the Loughran-McDonald Master Dictionary. https://drive.google.com/file/d/0B4niqV00F3msQ3lVeGpKSEg4QUU/view, 2011. Last Accessed: 2019-05-17.

[58] Tim Loughran and Bill Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[59] Tim Loughran and Bill Mcdonald. Measing readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671, 2014.

[60] Manning, Christopher D. and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[61] Ana Marques. Sec interventions and the frequency and usefulness of non-gaap financial measures. *Review of Accounting Studies*, 11(4):549–574, December 2006.

[62] Bill McDonald. Stage One 10-X Parse Data. https://sraf.nd.edu/data/stage-one-10-x-parse-data/, 01/01/2019. Last Accessed: 2019-06-19.

[63] Inc. Medco Health Solutions. Medco health solutions, inc. 10-k. https://www.sec.gov/Archives/edgar/data/1170650/000095012312002878/c25663e10vk.htm, 2011.

[64] Merriam-Webster. Something Wicked: The Story of an Adverb. https://www.merriam-webster.com/words-at-play/wicked-adverb-intensifier-usage, n.d. Last Accessed: 2019-09-13.

[65] Gautam Mitra and Leela Mitra, editors. *The Handbook of Sentiment Analysis in Finance*. Albury Books, 2016.

[66] Business Development Bank of Canada. How to find the optimal financing mix for your business. https://www.bdc.ca/en/articles-tools/start-buy-business/start-business/pages/sources-financing-business-project.aspx, n.d. Last Accessed: 2019-06-19.

[67] N. Pröllochs, S. Feuerriegel, and D. Neumann. Enhancing sentiment analysis of financial news by detecting negation scopes. In *2015 48th Hawaii International Conference on System Sciences*, pages 959–968, January 2015.

[68] Reviewed by James Chen. Philadelphia stock exchange (phlx). https://www.investopedia.com/terms/p/phlx.asp, 2019.

[69] Tyler Rinker. Package 'qdapDictionaries'. https://cran.r-project.org/web/packages/qdapDictionaries/qdapDictionaries.pdf, 2018. Last Accessed: 2019-06-21.

[70] Jonathan L. Rogers and Andrew Van Buskirk. Shareholder litigation and changes in disclosure behavior. *Journal of Accounting and Economics*, 47(1):136–156, 2009. Accounting Research on Issues of Contemporary Interest.

[71] Van Buskirk Andrew Rogers, Jonathan L. and Sarah L. C. Zechman. Disclosure tone and shareholder litigation. *The Accounting Review*, 86(6):2155–5183, 2011.

[72] Sugata Roychowdhury. Earnings management through real activities manipulation. *Journal of Accounting and Economics*, 42(3):335–370, 2006.

[73] M. Sarderlich and D. Kazakov. Extending the loughran and mcdonald financial sentiment words list from 10-k corporate filings using social media texts. *Proceedings of the 11th LREC Conference on Language Resources and Evaluation in Workshop on Financial Narratives*, 2018.

[74] Alexandra Scraggs and Jamie Powell. 'structuring adjusted ebitda' now exists. https://ftalphaville.ft.com/2018/06/11/1528741565000/-Structuring-adjusted-ebitda--now-exists/, 2018.

[75] SEC. What We Do. https://www.sec.gov/Article/whatwedo.html, 2013. Last Accessed: 2019-06-19.

[76] Securities and Exchange Commission. A Plain English Handbook - How to create clear SEC disclosure documents . https://www.sec.gov/pdf/handbook.pdf, 1998. Last Accessed: 2019-06-21.

[77] Securities and Exchange Commission. Beginners' Guide to Financial Statement. https://www.sec.gov/reportspubs/investor-publications/investorpubsbegfinstmtguidehtm.html, 2007. Last Accessed: 2019-06-19.

[78] Securities and Exchange Commission. Financial Reporting Manual - TOPIC 9 - Management's Discussion and Analysis of Financial Position and Results of Operations (MD&A). https://www.sec.gov/corpfin/cf-manual/topic-9, 2008. Last Accessed: 2019-06-03.

[79] Securities and Exchange Commission. TOPIC 8 - Non-GAAP Measures of Financial Performance, Liquidity, and Net Worth. https://www.sec.gov/corpfin/cf-manual/topic-8, 2008. Last Accessed: 2019-06-21.

[80] Securities and Exchange Commission. Accessing the U.S. Capital Markets — A Brief Overview for Foreign Private Issuers. https://www.sec.gov/divisions/corpfin/internatl/foreign-private-issuers-overview.html, 2013. Last Accessed: 2019-06-03.

[81] Securities and Exchange Commission. What We Do. https://www.sec.gov/Article/whatwedo.html, 2013. Last Accessed: 2019-06-03.

[82] Securities and Exchange Commission. Exchange Act Reporting and Registration. https://www.sec.gov/smallbusiness/goingpublic/exchangeactreporting, 2018. Last Accessed: 2019-06-03.

[83] Securities and Exchange Commission. Non-gaap financial measures. https://www.sec.gov/divisions/corpfin/guidance/nongaapinterp.htm, 2018. Last Accessed: 2019-06-03.

[84] Securities and Exchange Commission. A roadmap to non-GAAP financial measures. https://www2.deloitte.com/us/en/pages/audit/articles/a-roadmap-to-non-gaap-financial-measures.html, 2019. Last Accessed: 2019-06-21.

[85] Securities and Exchange Commission. Accessing EDGAR data. https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm, 2019. Last Accessed: 2019-06-21.

[86] Securities and Exchange Commission. Financial Reporting Manual - TOPIC 6 - Foreign Private Issuers & Foreign Businesses. https://www.sec.gov/corpfin/cf-manual/topic-6, 3/31/2009. Last Accessed: 2019-06-19.

[87] Securities and Exchange Commission. Forms 3, 4, 5. https://www.sec.gov/fast-answers/answersform345htm.html, n.d. Last Accessed: 2019-06-19.

[88] U.S. Securities and Exchange Commission. Topic 9 - management's discussion and analysis of financial position and results of operations (md&a). https://www.sec.gov/corpfin/cf-manual/topic-96, 2008.

[89] U.S. Securities and Exchange Commission. Topic 6 - foreign private issuers & foreign businesses. https://www.sec.gov/corpfin/cf-manual/topic-6, 2018.

[90] Statistics Solutions. Assumptions of the Wilcoxon Sign Test. https://www.statisticssolutions.com/assumptions-of-the-wilcox-sign-test/, 2019. Last Accessed: 2019-06-21.

[91] SourceForge. Release 4 of the 12dicts word lists. http://wordlist.aspell.net/12dicts-readme-r4/, 2003. Last Accessed: 2019-05-17.

[92] Staff. About the court. https://www.courdecassation.fr, n.d.

[93] Laerd Statistics. Wilcoxon Signed-Rank Test using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php, n.d. Last Accessed: 2019-06-21.

[94] Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. Exploring the use of text classification in the legal domain. *CoRR*, abs/1710.09306, 2017.

[95] Securities Class Action Clearinghouse Team. Securities Class Action Clearinghouse. http://securities.stanford.edu/sector.html?filter=Technology, n.d. Last Accessed: 2019-06-04.

[96] Tetlock, Paul C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 2007.

[97] M. Walker and E. Louvari. The determinants of voluntary disclosure of adjusted earnings per share measures by uk quoted companies. *Accounting and Business Research*, 33:295–309, 2003.

[98] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, fourth edition, 2017.

[99] P. Wongchaisuwat, D. Klabjan, and J. O. McGinnis. Predicting litigation likelihood and time to litigation for patents. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 257–260, New York, NY, USA, 2017. ACM.

[100] Steven Young. The drivers, consequences and policy implications of non-gaap earnings reporting. *Accounting and Business Research*, 44(4):444–465, 2014.

[101] D. Patrick Zimmerman. Effects of computer conferencing on the language use of emotionally disturbed adolescents. *Behavior Research Methods, Instruments, & Computers*, 19(2):224–230, March 1987. Last Accessed: 2019-05-19.