2015

Effects of Investor Sentiment Using Social Media on Corporate Financial Distress

Tarek Hoteit
Walden University

Walden University

College of Management and Technology

This is to certify that the doctoral dissertation by

Tarek Hoteit

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. David Bouvin, Committee Chairperson,
Applied Management and Decision Sciences Faculty

Dr. David Gould, Committee Member,
Applied Management and Decision Sciences Faculty

Dr. Walter McCollum, University Reviewer
Applied Management and Decision Sciences Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2015

Abstract

Effects of Investor Sentiment Using Social Media on Corporate Financial Distress

by

Tarek Adnan Hoteit

MBA, University of Dallas, 2006

BS, American University of Beirut, 1998

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Management and Technology

Walden University

April 2015

Abstract

The mainstream quantitative models in the finance literature have been ineffective in detecting possible bankruptcies during the 2007 to 2009 financial crisis. Coinciding with the same period, various researchers suggested that sentiments in social media can predict future events. The purpose of the study was to examine the relationship between investor sentiment within the social media and the financial distress of firms Grounded on the social amplification of risk framework that shows the media as an amplified channel for risk events, the central hypothesis of the study was that investor sentiments in the social media could predict t he level of financial distress of firms. Third quarter 2014 financial data and 66,038 public postings in the social media website Twitter were collected for 5,787 publicly held firms in the United States for this study. The Spearman rank correlation was applied using Altman Z-Score for measuring financial distress levels in corporate firms and Stanford natural language processing algorithm for detecting sentiment levels in the social media. The findings from the study suggested a non-significant relationship between investor sentiments in the social media and corporate financial distress, and, hence, did not support the research hypothesis. However, the model developed in this study for analyzing investor sentiments and corporate distress in firms is both original and extensible for future research and is also accessible as a low-cost solution for financial market sentiment analysis.

Effects of Investor Sentiment Using Social Media on Corporate Financial Distress

by

Tarek Adnan Hoteit

MBA, University of Dallas, 2006

BS, American University of Beirut, 1998

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Management and Technology

Walden University

April 2015

Dedication

I dedicate this work to my wife, Mayssaloun Tay, and my kids, Adam and Leen. Your support and patience were invaluable in the completion of my doctoral study. Thank you, and I love you.

Acknowledgments

I would like first to thank my wife and love of my life, Mayssaloun Tay, for her support and endless patience, my children, Adam and Leen, for their infinite curiosities about science and nature, and my entire family for all their support. I thank my late dad, Adnan, and my mother, Foze El-Amine, for their endless love and support. This work means to them as much as it means to me. I would like to thank my dissertation chair, Dr. David Bouvin, ex-dissertation chair and mentor, Dr. Reza Hamzaee, committee member and great inspiration, Dr. David Gould, and URR member of the committee, Dr Walten McCollum. Last but not least, I would like to thank everyone at Walden University who assisted me throughout the dissertation journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

The focus of this study was to determine the role of investor sentiment using social media on corporate financial distress levels. The need for such a study and its corresponding model is a result of limitations identified in the existing statistical models that were ineffective in predicting many corporate bankruptcies during the 2007 to 2009 financial crisis. Furthermore, existing literature has demonstrated the effects of investor sentiment and social media in the financial markets, but there is limited research on the relationship between investor sentiment in social media and financial distress levels in firms. In this study, I attempted to address such a gap by analyzing the relationship between sentiments extracted from textual messages using the social media website Twitter and the levels of financial distress of publicly held firms in the United States.

The positive implication of the study is that it provides a consolidated framework that investors and corporate institutions could use to evaluate both public opinion and corporate distress level that can be beneficial in financial decision-making. Chapter 1 includes the background of the research, the nature of the study, and key assumptions as well as limitations in the study.

Background of the Study

The major models in the finance literature for measuring corporate financial distress include only financial data about firms and markets to predict the likelihood of corporate bankruptcies and do not incorporate psychological factors that can influence investors' decisions in the financial markets. Since the 2007 to 2009 Great Recession, many of the quantitative financial distress prediction models have been under scrutiny for failing to predict the failures of many firms. For instance, Merton's (1974) distance-to-default model extended Black and Scholes's (1973) options pricing model that Harford (2012) in a BBC documentary considered as a primary cause of the 2007 to 2009 Great Recession. After the seminal work by Kahneman and Tversky (1979) on decision-making under risk, the behavioral finance theorists argued that the financial market participants exhibit psychological behaviors that are not explicit in the traditional corporate risk valuation models. Such actions include irrational decisions, bias, risk averseness, and overconfidence.

The effects of investors' behaviors can have contagious effects within financial markets. It can result in disruptive conditions that are similar to the downfall of Internet stocks in the 1990s and the global financial crisis between 2007 and 2009. Evidence of psychological attributes that govern the actions of market participants have challenged the core theoretical foundations of traditional finance theories and supported the call to extend the predominantly quantitative finance studies with qualitative research (Ardalan, 2008; Rambocas & Gama, 2013). A key psychological attribute is sentiment expressions through the media that can have a significant influence on investor decisions within the financial markets.

The news about corporate institutions and financial markets broadcast to large audiences through media outlets, including social media, could influence the decision-making of financial market participants. There is a growing body of research on the effect of media communication on investors' sentiments, which in turn affect asset valuations, market volatility, and investment risks. For example, Hafez and Xie (2012) demonstrated how changes in investor sentiment could influence the expected returns in financial markets. Akhtar, Faff, Oliver, and Subrahmanyam (2012) also showed the effects of corporate announcements about U.S. stocks and in the future market on consumer sentiment. One platform to express sentiment is social media that is accessible by over 67% of the United States population, according to a study by Duggan and Brenner (2013). In the last decade, social media platforms have become the major Internet-based destinations for information sharing and public communications.

Information about businesses and markets have expanded into informal and inexpert communications using Internet forums and social media platforms. In the recent period of technological innovations, corporate valuations are no longer limited to corporate quarterly updates that can drive financial risk analysis or expert opinions that can lead to investment decisions. The wealth of information available on the Internet, regardless of its accuracy, is easily accessible by anyone who may make uninformed financial decisions based on inaccurate information. Businesses and government organizations are also embracing Internet technologies to raise brand awareness and establish closer relationships with customers.

According to Montalvo (2011) and Culnan, McHugh, and Zubillaga (2010), companies are increasingly hiring social media specialists to raise brand awareness, monitor brand reputation, and collaborate with customers. Hanna, Rohm, and Crittenden

(2011) argued that even though firms understand the benefits of being active in social media, not all firms understand how to measure the performance indicators. The popularity of social media within businesses raises the question of the positive and negative implications of social media for the operations of corporate organizations.

Social media sites have become rich data sources for market analysis, consumer behavior, and sentiment information. Some studies showed that sentiments expressed in social media could predict future events. For instance, Asur and Huberman (2010) demonstrated how interactions between friends on social media platforms could predict future movies sales. Furthermore, Bollen et al. (2011) performed textual analysis of sentiment in the social media website Twitter to predict market mood. Moreover, Howard et al. (2011) argued that conversations about revolutions in social media preceded the Arab Spring mass uprising in 2010. However, in times of financial crises and market turmoil, limited literature is available on the possible implications between sentiments in the social media and the financial health of corporate institutions.

The lack of an integrated empirical framework that incorporates the classification of investor sentiment in the social media and the measurement of financial distress in corporate firms is also noticeable in the literature. The need for such a model that can bridge the gap between measuring the financial distress of firms and the investor sentiment toward such firms in social media can help improve the bankruptcy prediction models available in the literature. The next segment elaborates on such a problem, which formed the basis of the current study.

Problem Statement

The problem with the existing corporate financial distress prediction models is that they failed to provide timely signals toward the companies that came under financial distress or declared bankruptcy during the 2007 to 2009 Great Recession. It is evident by the lack of a consensus on a standard corporate financial distress prediction model in existing literature. The current predominant statistical approaches for predicting corporate distress would vary in statistical power among research in the academic literature. Such models within the field of corporate finance support Fama's (1970) efficient market hypothesis that assumes financial markets as information efficient markets and the value of firms is measurable using firms' disclosed financial data. Investors' panic and the mass sell-offs in the markets that occurred during the 2007 to 2009 Great Recession were evidence to counter support the efficient market hypothesis.

The financial crisis between 2007 and 2009 was evidence to support the behavioral finance theories that considered markets as inefficient and investors as irrational. Furthermore, the amplifications of financial crises regardless of magnitude, societal effect, or information accuracy traverse instantly through society using online social media, such as Facebook and Twitter. As a result, researchers such as Bollen et al. (2011), Chung (2011) as well as Asur and Huberman (2010) were able to demonstrate that certain future events could be predicted from the sentiments expressed in social media. However, such type of analysis has not been extensively applied or integrated into the statistical algorithms for predicting corporate financial distress.

Purpose of the Study

The purpose of this quantitative design study was to determine if investor sentiment through social media can significantly amplify the risk of bankruptcy for financially distressed firms. The overall theoretical framework of the study extends Kasperson's (2012) social amplification of risk framework. By controlling for the financial factors that are relevant to the measurement of financial distress in publicly held firms, I attempted to determine if investor sentiment, through textual analysis of user-generated content in social media, could significantly affect the level of financial distress of firms. The empirical findings of the study may help determine if there is an association between investor sentiment in social media and corporate financial distress.

Research Question(s) and Hypotheses

I examined the relationship between investor sentiment, as affected by social media, and the financial distress of firms by investigating the following hypotheses.

Research Question 1: What is the relationship between the financial distress of firms and the investor sentiment towards the firms in social media?

$H1_0$: There is no relationship between the level of financial distress of firms and the investor sentiment towards such firms in social media.

$H1_a$: There is a positive relationship between the level of financial distress of firms and the investor sentiment towards such firms in social media. Firms with higher degrees of financial distress positively correlate with higher negative investor sentiment in social media. Firms with lower degrees of financial distress positively correlate with higher positive investor sentiment in social media.

Research Question 2: How does sentiment in social media affect the risk of bankruptcy for financially distressed firms?

$H2_0$: There is no statistically significant relationship between sentiments in social media and the likelihood of financially distressed firms to declare bankruptcy.

$H2_a$: There is a statistically significant relationship between negative sentiments in social media and the likelihood of financially distressed firms to declare bankruptcy. Financially distressed firms with greater negative sentiments than positive sentiments are more likely to file for bankruptcy.

Research Question 3: What is the relationship between a firm's stock movement and the level of sentiment towards the firm in social media?

$H3_0$: There is no relationship between the firms' stock movement and the level of sentiment in social media.

$H3_a$: There is a positive relationship between the firms' stock movement and the level of sentiment in social media. Negative sentiment correlates with a decline in the stock value, and positive sentiment correlates with an increase in stock value.

Theoretical Framework

The theoretical frameworks that underpin this study are Altman's (1968) *Z*-Score approach for measuring the levels of financial distress of firms using a multivariate discriminant analysis approach, Manning et al.'s (2014) sentiment analysis method for textual data, and Kasperson's (2012) social amplification of risk framework. Altman developed the Altman *Z*-Score as an index for measuring the likelihood of firms to file for bankruptcy in 2 years. Equation 8 includes the *Z*-Score index that Altman developed using five financial ratios from the sampled firms' balance sheets and income statements. Various literature has demonstrated the effectiveness of Altman's *Z*-Score for predicting financial distress of firms, but alternative models also exist that counter support of the model. However, recent literature has not disregarded the model. To the contrary, the *Z*-Score model has become one of the standard corporate distress indicators at the professional level. In this study, I applied the model using the financial data of publicly held firms in the United States. A detailed explanation of the model is present in Chapter 2 of the study. The second framework that supports the study is Manning et al.'s (2014) sentiment classification model for textual data.

Manning et al.'s (2014) model is part of the literature on sentiment analysis and opinion mining in which a system or a human maps an opinion into one of the predefined labels, such as *negative* or *positive*, or on a continuum from one end to another, such as from 1 to 10 (Pang & Lee, 2008). The algorithm, made available in Manning et al.'s (2014) Stanford Core NLP, is a supervised machine-learning algorithm that, given a predefined training set of sentences classified as *very negative*, *negative*, *neutral*, *positive*, and *very positive*, attempts to detect the sentiment in new data sets. Chapter 2 of the study includes a detailed explanation of the framework. In this study, I applied the framework over the captured tweets from the social media website Twitter in order to determine the level of investor sentiment and its relationship to the level of financial distress of firms. I hypothesized that investor sentiment is associated with the financial distress of firms, where negative sentiment positively correlates with no financial distress of firms. I used Kasperson's (2012) framework that shows how social communication can amplify negative events, which could have ripple effects in society. As shown in Figure 1, Kasperson's social amplification of risk framework demonstrates that the communication stations, such as word-of-mouth and the media, can amplify risk-related events that can have ripple effects across society. As a result, social amplification of risk can lead to adverse effects, such as corporate bankruptcies or loss of sales. Chapter 2 of the study includes a more detailed description of the framework. In my research, I hypothesized that the investor sentiment in social media amplifies the firms' financial distress conditions and increases the likelihood of declaring bankruptcy after 2 years.

Nature of the Study

Chapter 3 focuses on identifying the relationship between corporate financial distress and investor sentiment on social media. The nature of the study is predominantly quantitative. I used a quantitative approach to extract and analyze corporate distress and investor sentiment for the firms sampled in the study. However, a qualitative segment of the study was necessary at the early stage of the research since the machine learning algorithm that was used in the sentiment analysis of public messages in the social media required supervised training on financial jargon and on the classification of sentences that represent either positive, negative, or neutral sentiments. The focus of the study was on two constructs, the level of investor sentiment using social media and the level of financial distress in the sampled firms.

The research variables in the quantitative study are corporate distress indicators for each of the sampled firms and investor sentiment in social media toward the sampled firms as the independent variables. I used NASDAQ (2015) to extract the list of companies trading on New York Stock Exchange (NYSE), American Stock Exchange (Amex), and the National Association of Securities Dealers Automated Quotations (NASDAQ) stock exchange that would form the population sample of the study. I then used the Yahoo (2015) finance website to extract third quarter 2014 corporate financial data and applied Altman's (2013) *Z*-Score index to determine the level of financial distress for each of the sampled firms. Subsequently, I extracted at different time intervals multiple subsets of public content in the social media website Twitter (2015) that references the stock symbol of each of the sampled firms. I later applied Manning et al.'s (2014) Stanford Core NLP natural language processing toolkit to determine the sentiments for each of the public responses. After I had completed the data collection process, I applied a statistical correlation analysis for nonparametric data using Spearman rank correlation coefficient analysis in order to validate the research hypotheses.

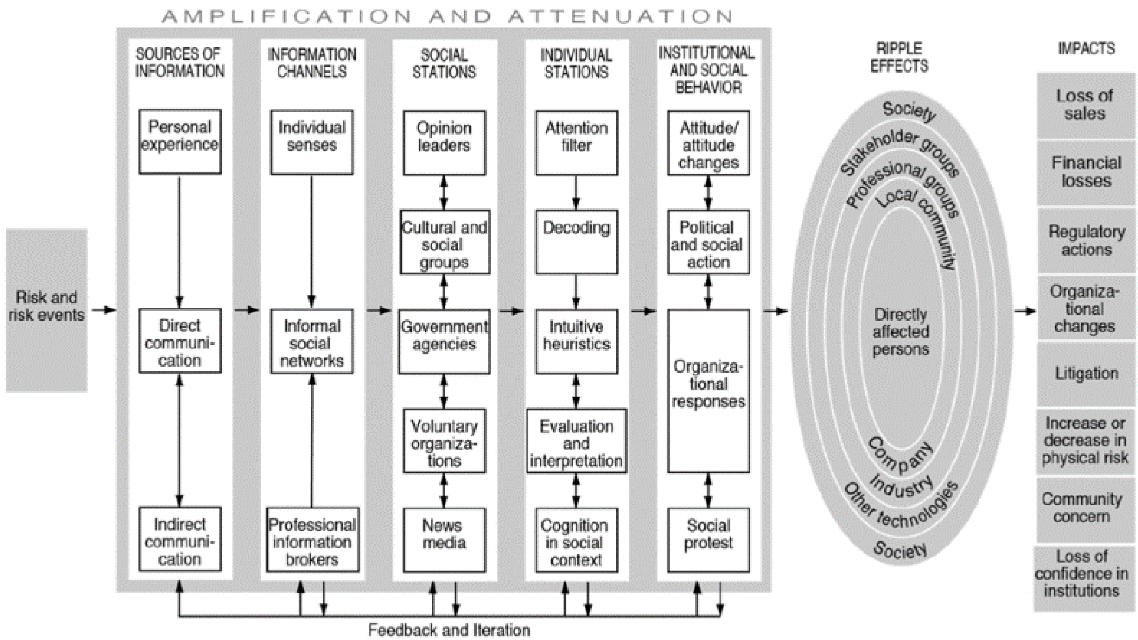


Figure 1. Social amplification of risk framework. From “A Perspective on the Social Amplifications of Risk.” by R. E. Kasperson (2012), *Bridge on Social Sciences and Engineering Practice, 42*(3), p. 25. Reprinted with permission.

Definitions

Corporate financial distress: A state of a firm that usually precedes its declaration of bankruptcy by some period (Platt & Platt, 2009).

Financial distress: The inability of a firm to pay its financial obligations (Beaver, 1966, p. 71).

Scipy: A Python-based library for mathematics, science, and engineering (SciPy Developers, 2014).

Sentiment analysis: A computational study of opinions, sentiments, and emotions usually expressed in any text (Liu, 2010).

Social media: "Web-based services that allow individuals to (a) construct a public or semipublic profile within a bounded system, (b) articulate a list of other users with whom they share a connection, and (c) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site." (Boyd & Ellison, 2007, p. 211)

Tweet: A message that is 140 characters or less in the social media website Twitter (O'Connor, Balasubramanyan, Routledge, & Smith, 2010).

Assumptions

There are two key assumptions in the study. The first assumption is the presence of a company's stock symbols accompanied by the symbol ($) in the tweets posted by investors. For example, the tweet "don't want to lose like you did with $OXY $APC $WFM $AMZN" assumes that it is an investor referring to the companies Occidental Petroleum (stock symbol $OXY), Anadarko Petroleum Corp ($APC), Whole Foods Market ($WFM), or Amazon.com Inc. ($AMZN). The second assumption is that public postings by social media users on the website Twitter represent their opinions regardless of the exact origin of their messages. In social media sites, many individuals can share the view of others without necessarily citing the source of the content. For example, the concept of *retweeting* or reposting someone else's messages on Twitter is a common practice on such platform (Boyd, Golder, & Lotan, 2010). Predicting the likelihood of retweeted messages is possible using conditional random field algorithms, as evident in Peng et al. (2011), but it is not relevant to the current study. In this study, I focused on the effect of public responses towards the financially distressed firms and not among one another. I assumed that if a person publicly posts a message that can be interpreted as a sentiment, then such a sentiment can be associated with that person regardless of whether it is his or her own. Furthermore, I assumed that each of the sentiments detected using the social media website Twitter is a representation of the overall sentiment for the social media user regardless of the social media platform.

Scope and Delimitations

The scope of the study is the role of investor sentiment in the social media website Twitter over the financial distress conditions of corporate firms. Even though individuals can express opinions across virtually any platform, such as Internet forums and opinion columns in newspapers, social media is a relatively new phenomenon in the last decade. Numerous studies on the relationship between media platforms and financial markets are available in the finance literature, but given that social media is a relatively new phenomenon, limited studies have addressed the role of social media in the analysis of financial distress of firms. As a result, I decided to examine the relationship between social media and the financial distress of firms.

For external validity of the research, I have limited the classification of investor sentiment in social media to the social media website Twitter. Individuals using the Twitter send over 50 million messages each day that demonstrates the relevance of such medium in society (Murthy, 2011). The website allows researchers to extract textual postings by its users, which makes the platform suitable for sentiment analysis. Social media platforms are also numerous but not homogenous. They include Facebook, Twitter, LinkedIn, Google Plus, Pinterest, Instagram, Tumblr, and many more that are accessible by different age groups and for varied reasons (Duggan & Brenner, 2013; Hanna et al., 2011). However, determining the difference in investor sentiment across all social media platforms is exhaustive and ineffective for the study. Even though the demographics of social media users vary across the platforms, as determined by Lenhart, Purcell, Smith, and Zickuhr (2010), I have focused on the effect of the sentiments using Twitter as the social media platform due to the popularity of the site and assumed that future research can extend the study to include other social media platforms.

Limitations

The design and methodology limitations of existing studies that measure the financial distress of firms using financial data apply to the current study as well. The most common statistical and machine learning models for measuring corporate distress use different financial ratios, such as debt ratio and asset liquidation formulas, to determine if firms are likely to default or not. I incorporated Altman's (1968|2012) *Z*-Score that consists of a weighted set of financial ratios calculated from firms' financial statements. The selection of the particular financial formulas in similar studies is usually associated with past literature that analyzed previously defaulted firms that already defaulted (Campbell, Hilscher, & Szilagyi, 2008). Since the study followed a similar quantitative approach as other studies within the field of finance, its design weakness can be in its confounding variables.

The circumstances that can lead some firms to fail may be unknown to the researchers at the time of their analysis. Firms' financial ratios may also fail to capture all the conditions that lead to corporate default. Furthermore, the various methodologies in the literature that measure the corporate distress and expected bankruptcies of firms vary in their predictive power (Zurada, Foster, Ward, & Barker, 2011). However, much of the recent research still references Altman's *Z*-score. In this study, I used Altman's *Z*-Score as the measurement of choice for corporate distress, as detailed in Chapter 3 of the study, but future research can incorporate other corporate distress prediction models.

Significance of the Study

The significance of the study is that it provides a novel approach to the study of corporate bankruptcies by considering investor sentiment from social media as a factor that can affect financial distress. Since the 2007 to 2009 financial crisis, the traditional finance literature methods have been under scrutiny for their ineffectiveness in predicting financial crises (Blackledge, 2010). Furthermore, behavioral finance theorists argue that markets are not as efficient as claimed by the traditional finance theories (Baker & Wurgler, 2011). As a result, many academics have called for a deeper integration of behavioral and psychological factors into the field of finance (Byrne & Brooks, 2008). The model in this study supports the behavioral finance school of thought by incorporating investor sentiment using social media as one of the exploratory variables that can predict financial distress.

With this study, the integration of investor sentiment into the study of financial distress in companies will help bridge the gap between behavioral finance and traditional finance literature. The tool developed for the study will also aid corporate firms with an early alert system that can provide early warning signals about the perceived status of their firms within the social media.

Significance to Theory

The significance of this study is that it directly incorporated the voice of the investor, whether it is a trader or a speculator, into the study of corporate financial distress that was generally associated with the analysis of corporate financial statements in order to predict the level of bankruptcy of firms. Before the advancements in technology and the media, the opinions of investors toward companies were not readily accessible unless possibly inferred indirectly through corporate sales or the demands of its stock. The major corporate distress models including Altman (1968), Ohlson (1980), Shumway (2001), and Campbell et al. (2011) and the various literature that expands on such models were solely based on financial variables that include companies' assets, liquidity, and equity but not investor sentiment. This study may help extend the body of literature on corporate distress analysis by incorporating the study of investor sentiment into the research.

Significance to Practice

The ability for virtually anyone in the world to share his or her opinions on just anything including companies using social media, such as Facebook, LinkedIn, and

Twitter, drove the need for companies to manage their online presence and monitor the public's opinion more assiduously (Culnan et al., 2010). The practicality of this study is that it incorporates the analysis of corporate distress and sentiment analysis into one framework that is both extensible and modifiable for companies' management to use in their decision-making processes. It makes the study beneficial not only for academic research but for application purposes within corporate institutions.

Significance to Social Change

The aftermath of the 2007 to 2009 financial crisis that saw many firms declaring bankruptcy drove the need for more corporate distress prediction algorithms. Moreover, the global rise of online social media during eh same period of the financial crisis called for a proactive monitor of public opinions towards corporate firms that are under financial distress. The study may help support both cases by providing a practical and inexpensive tool that firms can use as a proactive measurement tool on corporate financial distress that can potentially save them from the possibility of bankruptcies.

Summary and Transition

After the ineffectiveness of many traditional finance models in finance that use financial ratios to measure the expected bankruptcy of firms, the study incorporates investor sentiment within social media as an exploratory factor to predict corporate distress and potential bankruptcies. In this study, I used Kasperson's (2012) conceptual framework to hypothesize that media can amplify the risk of corporate distress and can ultimately lead to negative conditions including bankruptcy. The study is based on a quantitative approach using Altman's (2013) multivariate discriminant analysis model, *Z*-Score, for measuring financial distress of firms and Manning et al.'s (2014) machine learning algorithm to classify public messages in the social media website Twitter as either positive, negative, or neutral sentiments. The outcome of the research may help determine if investor sentiment within social media can improve the accuracy of predicting corporate bankruptcies. Chapters 2 and 3 of the study provide an in-depth review of the relevant literature and the design of the research.

Chapter 2: Literature Review

Two major phenomena occurred between 2007 and 2009: the global financial crisis that led many companies to fail and the rise in social media that has become an influential communication platform for investor sentiment. Academic research about each of the events remained largely separate until very recently when corporate firms began to embrace the social media for marketing and public communication purposes. However, academic studies that applied empirical models to predict corporate bankruptcies did not include investor sentiment within the social media as a possible contributor to corporate financial distress. The ineffectiveness of the traditional empirical models to predict the 2007 to 2009 financial crisis drove the call for the incorporation of psychological factors into the studies on financial markets (Baker & Wurgler, 2011). However, there is lack of information that leverages investor sentiment within the social media into the measurement of corporate distress or corporate bankruptcies. Chapter 2 of the study is an in-depth review of the current literature surrounding financial distress prediction and sentiment analysis using social media.

Literature Search Strategy

The literature review strategy in the study includes the selection of the databases and search engines to extract key research related to the corporate financial distress and investor sentiment analysis through social media. Table 1 lists the databases, search engines, and relevant search items in the study. Furthermore, a daily alert was set up using Google Scholar in order to email me a list of new articles published on either the subject of corporate financial distress or the topic of sentiment analysis. The volume of research captured was then stored and categorized using Roy Rosenzweig Center for History and New Media's (2014) research tool Zotero. The scope of research covered the period between 2008 and 2014 when many corporate firms filed for bankruptcy and, concurrently, social media became increasingly popular with the public and in academic studies. The next sections provide a detailed review of the theoretical foundations of the study followed by an exhaustive literature review on investor sentiment analysis and corporate bankruptcy predictions.

Table 1

Research Sources in the Study

Source	Type	Key search items
Google Scholar	Search engine	*Financial distress, sentiment analysis, investor sentiment, social media, bankruptcy prediction.*
Ebsco	Academic database	*Financial distress, sentiment analysis, investor sentiment, social media, bankruptcy prediction.*
Walden Library	Search engine	*Financial distress, sentiment analysis, investor sentiment, social media, bankruptcy prediction.*

Theoretical Foundation

Research Paradigms in Finance

The theoretical framework for the study of financial distress of firms and the effect of investor sentiment requires a concise understanding of the academic paradigms in finance. Academic finance is a social sciences' field that deals with the monetary aspects of governments, businesses, and society. Different paradigms in academic finance fall into Burrel and Morgan's (1985) matrix of sociological paradigms with the various levels of subjectiveness, objectiveness, regulative, and revolutionariness within the research. As shown in Figure 2, different research in academic finance falls within the four quadrants of paradigms, *functionalist*, *interpretist*, *radical structuralist,* and *radical humanist* paradigms (Ardalan, 2008; Burrell & Morgan, 1985). A regulatory paradigm consists of accepted principles and procedures where each study extends the work of previous research. A revolutionary paradigm includes new theories that defy existing laws and refute existing paradigms (Kuhn, 1996). An objective paradigm consists of research that maintains objectivity throughout the research by limiting researchers as only facilitators and observers with no influence on the outcome of their research. A subjective paradigm consists of research where scholars influence their research through personal experiences and subjective thoughts. The next section summarizes each of the four paradigms within the context of academic finance research.

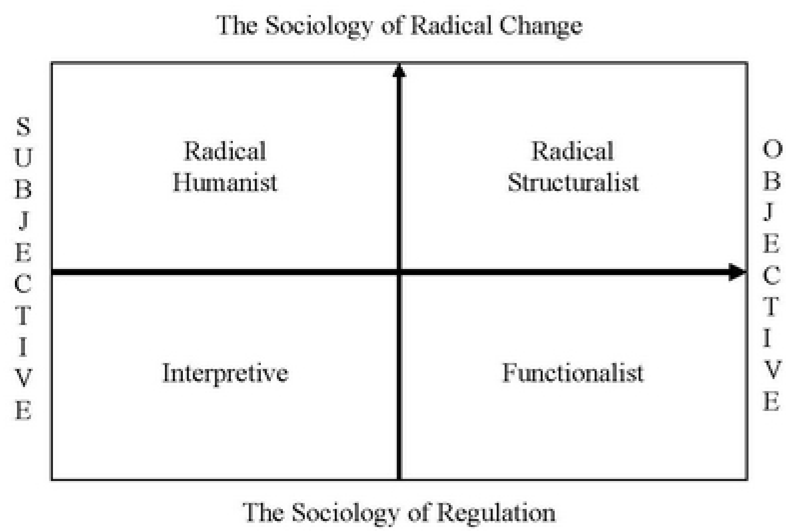


*Figure 2.*The four sociological paradigms as they apply to academic finance. From *Sociological paradigms and organizational analysis: Elements of the sociology of corporate life* by G. Burrell & G. Morgan (1985). Burlington: Ashgate Publishing, Ltd. Reproduced with permission.

Functionalist paradigm. The functionalist paradigm is a regulatory paradigm that branches from the positivism philosophy within the social sciences. It assumes that society has a concrete existence that follows a predictable order and produces explanatory knowledge (Ardalan, 2008). The functionalists conduct empirical research to support their hypotheses and assume objectivity when explaining societal issues (Ardalan, 2008). Within the field of finance, the functionalists consider uniformity and regularity within financial markets when conducting cause and effect research. The functionalist paradigm is the most common model found in finance research, publications, and academic teachings (Ardalan, 2008). Both the traditional finance and the behavioral finance belong to the functionalist paradigm.

The classical finance literature that supports the functionalist paradigm includes Markowitz's (1952) portfolio selection theory, Sharpe's (1964) capital asset pricing model, and Fama's (1970) efficient market hypothesis. Such theories assume uniformity and rationality within financial markets. The behavioral finance literature, on the other hand, also supports such paradigms including the Kahneman and Tversky's (1979) prospect theory that assumes normal decision-making rather than rational behavior in markets. Such studies are typically quantitative in their research methods. Alternatively, qualitative research in the academic finance literature that incorporates subjective elements in the research fall under the interpretist paradigm.

Interpretist paradigm. Unlike the functionalist paradigm that assumes objective research in a concrete society, the interpretist paradigm is a regulatory paradigm that assumes social reality as nothing but a subjective assessment of individuals. Interpretists argue that no firm structure exists in society that researchers can objectivity identify using hard, concrete, and tangible evidence; what exists in society is nothing but subjective interpretations that continuously change (Ardalan, 2008). Hence, researchers should take active roles instead of passive roles within the research and should include their personal experiences and frames of reference. Qualitative research follows the interpretist paradigm, and it includes narrative research, phenomenological research, grounded theory, ethnographic research, and case studies (Janesick, 2010). The interpretist paradigm is common in psychology, education, and sociology research. In finance, it is common in nonpeer-reviewed research, such as finance narratives or business-related case studies, and it is less favorable than the functionalist paradigm in peer-reviewed finance literature. Both the functionalist paradigm and the interpretist paradigm are regulatory paradigms that accept the existing laws and regulations in society. The radical humanist and the radical structuralist paradigms are alternative paradigms to the functionalist and the interpretist paradigms respectively that reject the status quo in society.

Radical humanist paradigm. The radical humanist is a revolutionary paradigm with a more extremist view than the interpretist paradigm. It views reality as being socially constructed and antihuman; it focuses its sources on what it considers alienations found in society that dominate human consciousness in the form of objective forces that individuals have no direct control over (Ardalan, 2008). Major concerns of radical humanists include the dominance of purposive rationality in corporations and financial markets, rules and control systems that monitor rational actions, the use of technology as a liberating force, and the presence of behaviors that govern relationships between individuals in the workplace (Ardalan, 2008). The radical humanist paradigm is nonexistent in the field of academic finance (Ardalan, 2008). Similarly, the radical structuralist paradigm is also a revolutionary paradigm that takes an extremist view of the functionalist paradigm.

Radical structuralist paradigm. The radical structuralist takes an objective but a more radical view on society than the functionalist paradigm. Such a revolutionary paradigm sees reality as concrete and objective, and it sees the societal world similar to the natural world as independently constructed from outside the minds of human beings (Ardalan, 2008). Scientists under the radical structuralist paradigm take an objective point of view as functionalists, but they see society as a dominating force and, as a result, are committed to radical changes in society. They view society as a whole and emphasize the need to study society in its totality and not through disparate data-centric and concrete problem-based research as seen by the functionalists (Ardalan, 2008). This paradigm looks at all main dimensions of society including totality, structure, contradiction, and crises (Ardalan, 2008). The radical structuralist paradigm, according to Ardalan (2008), is also nonexistent within the field of academic finance.

Theoretical Frameworks on Financial Risks

The majority of research in academic finance supports the functionalist paradigm. As noted previously, the functionalist paradigm is both a regulatory and a quantitative framework that considers society as concrete, predictable, and analyzable through objective research (Ardalan, 2008). Risk is a key attribute that affects nearly all segments of society and has been a major field of research in academic finance. Risk is “the probability of success or failure” (Lee, Lee, & Lee, 2010, p. 53). As the definition implies, risk is the possibility of a loss or a hazard that firms may encounter throughout its lifetime. Table 2 includes a comprehensive list of risks that can affect firms. The theoretical framework that forms the foundation of research in the study of risk falls between two distinctive approaches in understanding and analyzing financial risks–the classical finance approach and the behavioral finance approach.

Table 2

Common Types of Financial Risk

Type of financial risk	Description
Call risk	The potential effect on return when repurchasing an equity before its maturity date.
Convertibility risk	The potential effect on return from converting one type of financial instrument into another.
Credit risk	The risk of an institution losing its ability to take loans to fund its growth.
Default risk	The probability of a zero return if the issuer of the financial organization is unable to make payments.
Interest-rate risk	The potential effect on return due to changes in interest rates.
Management risk	The potential effect on return due to poor managerial decisions.

(continued)

Type of financial risk	Description
Operation risk	The risk associated with the day-to-day operations of a firm, such as sales, marketing, and customer support.
Political risk	The potential effect on return due to law changes by government entities.
Purchasing-power risk	The potential effect on return due to inflation that influences the return value of the assets.
Systematic risk	The potential effect on return due to the rise or fall of the financial markets.
Unsystematic risk	The potential effect on return due to specific factors related to a particular financial instrument.

Note. Certain types of risk such as call risk might not apply for firms that to do not possess financial instruments such as equities and bonds. Adapted from *Handbook of Quantitative Finance and Risk Management* by C.F. Lee, A.C. Lee, and J. Lee (Eds), 2010, New York, NY: Springer. Reproduced with permission.

Classical finance theories on financial risk. The mainstream theories in financial asset valuations, asset pricing, and investment portfolio selections assume that market participants make rational decisions to seek maximized profits at minimum risk of losses. The founding principles of the classical finance theories toward risk date back to Bernoulli's (1738|1954) expected utility maximization theory. Bernoulli defined *utility* as the intrinsic value of a good or service as it relates to every person. He later established the *utility maximization equation* as the valuation mechanism for selecting utilities that offer the most gain. Selecting a utility is an individual preference.

The key assumption in Bernoulli's theory is that it is irrational for individuals to select choices that do not maximize the expected utility. Morgenstern and Neumann's (1953) theory of games and economic behavior expands on Bernoulli's argument to show that the maximized utility among different individuals is not contradictory. Every person is rational and would always seek maximum utility (Morgenstern & Neumann, 1953). Hence, it is irrational for investors to seek either riskier or fewer profitable investments that do not maximize the utility of their investments.

The major theories in classical finance that assume rational behavior of investors include the discounted cash flow method, Modigliani and Miller's (1958) valuation method, Markowitz's (1952) portfolio theory, and Sharpe's (1964) capital asset pricing model. The empirical models behind such theories assist corporate managers and investors in asset valuations and the risks associated with each investment. The methods provide systematic approaches to calculating the value of assets or investments as a function of future cash flow, maximized returns, or minimized risk variance.

The discounted cash flow method in Equation 1 considers the current value of a financial asset as the present discounted value of the future dividends from owning the asset at a market interest rate. In Equation 1 D_t is the future dividend in period *t*, and *k* is the interest rate associated with the investment (Cheng-Few Lee et al., 2010). The limitation with the discounted valuation method is the overemphasis on future dividends and market interest rates when evaluating financial assets. Modigliani and Miller (1958) addressed such limitation by modifying the discounted valuation method to support a more generalized form of cash inlay and cash outlay.

$$P_0 = \frac{\sum_{t=1}^{\infty} D_t}{(1+k)^t} \tag{1}$$

Modigliani and Miller (1958) proposed a generalized form of the discounted cash flow method in Equation 2, which they labeled as the fundamental principle of valuation. In Equation 2 P_0 is the present value, D_1 and P_1 are the dividend and price for the subsequent period P_1, and *k* is the interest rate (Lee et al., 2010). Equation 3 is a more generalized form of the Modigliani and Miller valuation method, where V_0 is the current market value of the firm, X_t is the net operation earnings or the cash inlays at period *t*, I_t is the investment costs or the cash outlays at period *t*, and *k* is the interest rate (Lee et al., 2010). Merton (1974) later expanded on Modigliani and Miller valuation method and incorporated debt and taxes in valuation calculations.

$$P_0 = \frac{1}{(1+k)} \quad (D_1 + P_1) \tag{2}$$

$$V_0 = \sum_{t=0}^{\infty} \frac{1}{(1+k)^{t+1}} (X_t \quad I_t) \tag{3}$$

The asset valuation methods, including the discounted method and Modigliani and Miller's more generalized valuation method, help investors determine the assets with the highest net present value of the future cash flows. However, such models do not consider the investment risks associated with the assets. Furthermore, investing in an asset could yield to greater losses if investments did not implement any diversification strategies, such as investing in a diversified portfolio of financial securities. Markowitz (1952) portfolio theory and Sharpe (1964) capital asset pricing model are two key theories in the classical finance literature that incorporate risk-related variables in the financial decision-making process.

The portfolio theory and the capital asset pricing model are classical finance frameworks that incorporate different elements of risk in investment valuation strategies. The portfolio theory uses either the minimum variance or the maximum expected returns in order to determine the optimal portfolio weight for every instrument in a portfolio (Lee et al., 2010). Equation 4 displays the Lagrangian function *Min L* that corresponds to the minimum variance in a portfolio, where E^* is the target expected return for the portfolio, $E(R_i)$ is the expected return for each asset in a given portfolio, and W_i is the weight of each asset *i* in a given portfolio (Lee et al., 2010). *Min L* is subject to two constraints: $\sum_{i=1}^{n} W_i E(R_i) = E^*$ and $\sum_{i=1}^{n} W_i = 1.0$. "The first constraint simply says that the expected return on the portfolio should equal the target return determined by the portfolio manager. The second constraint says that the weights of the securities invested in the portfolio must sum to one." (Lee et al., 2010, p. 10) An alternative to the Min L is the maximum Lagrangian function *Max L*.

$$MinL = \sum_{i=1}^{n}\sum_{j=1}^{n} W_i W_j + \lambda_1 \sum_{i=1}^{n}\left[W_i E(R_i) \quad E^*\right] + \lambda_2 \left(\sum_{i=1}^{n} W_i \quad 1\right) \tag{4}$$

Equation 5 displays the Lagrangian function *Max L* that corresponds to the maximum expected returns in a portfolio, where $\overline{R}_i$ is the average rate of return of the portfolio with targeted standard deviations σ for the portfolio, W_i, is the weight of each asset *i*, and R_i is the expected return of asset *i* in the portfolio (Lee et al., 2010). *Max L* is subject to the constraints:

$$\left[\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(R_i, R_j\right)\right]^{\left(\frac{1}{2}\right)} - \sigma_p \text{ and } \sum_{i=1}^{n} W_i = 1$$

"The first constraint is to minimize the risk or variance of the portfolio, subject to the portfolio's attaining some target expected rate of return, and also subject to the portfolio weight summing to one" (Lee et al., 2010, p. 10). The Lagrangian functions *Min L* and *Max L* provide an empirical method to either minimize the variance or maximize the expected returns in a financial portfolio. Another method for incorporating risk within the valuation of assets is the capital asset pricing model.

$$MaxL = \sum_{i=1}^{n} W_i \overline{R}_i + \lambda_1 \left\{ \left[\sum_{j=1}^{n} \sum_{i=1}^{n} Cov\left(R_i, R_j\right) \right]^{1/2} - \sigma_p \right\} + \lambda_2 \left(\sum_{i=1}^{n} W_i - 1 \right) \qquad (5)$$

Advancements in Markowitz portfolio theory lead to Sharpe's (1964) capital asset pricing model, also known as the CAPM that became a major risk assessment model within the field of financial management. CAPM captures the relationship between market risk and expected returns (Finch, Fraser, & Scheff, 2011). Equation 6 displays the CAPM model, where R_j is the rate of return for security, β is the risk measure for security j in respect to the market, R_m is the overall market returns, and R_f is the rate of return for a risk-free financial instrument.

$$E\left(R_j\right) = R_f + \beta_j \left\lfloor E\left(R_m\right) \quad R_f \right\rfloor \qquad (6)$$

Unlike Modigliani and Miller's valuation method that depends on firms' cash flows, the CAPM model incorporates any investment securities, market returns, and market risks. In a portfolio of financial securities, such as common stocks, the total risk is the combination of *systemic* risk of returns and *unsystematic* risk of returns. Figure 3 represents the relationship between systemic and unsystematic risks. Systemic returns are the expected returns in relation to the expected returns of other firms in financial markets with a risk constant *beta*; unsystematic returns are the residuals between the expected market returns and the expected firm returns (Lee et al., 2010). As the portfolio of financial securities expands with additional securities, unsystematic returns diminish and the correlation between the overall expected return of the portfolio and the overall market returns increases. Only systemic risk is relevant in the CAPM model using the beta coefficient, β, shown in Equation 6.

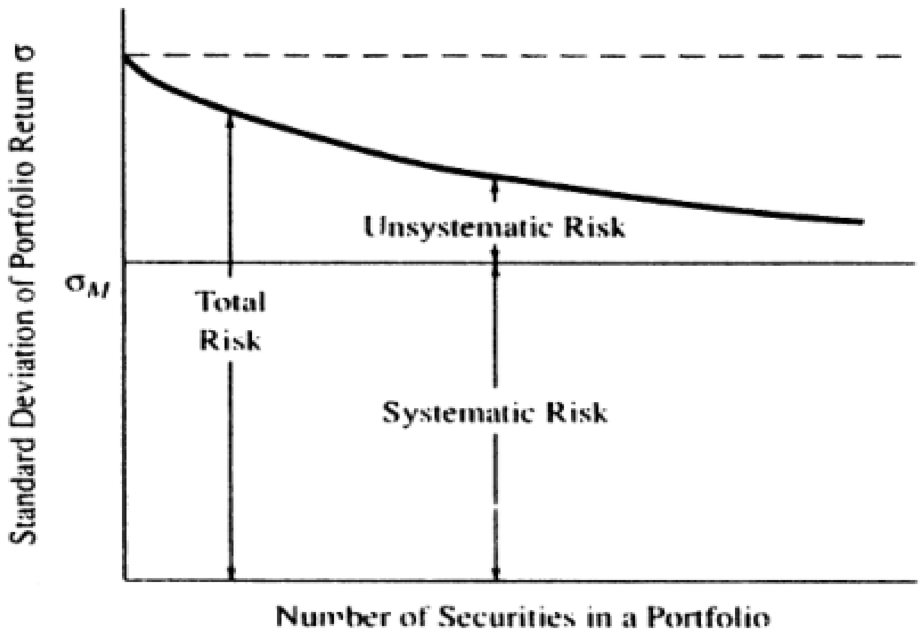


Figure 3. Diversification process in portfolio theory. Adapted from *Handbook of Quantitative Finance and Risk Management* by C.F. Lee, A.C. Lee, and J. Lee (Eds), 2010, New York, NY: Springer. Reprinted with permission

The primary assumptions of the classical finance theories are that markets are efficient and investors are rational since they invest in portfolios with the least possible risks and the greatest expected returns. Investors and financial managers would then calculate expected returns using mean-variance calculations. Hence, the level of expected returns is a function of risk only. Opponents of such assumptions, the behavioral finance scholars, take evidence from cognitive psychology to claim that investors and market participants make biased decisions rather than rational decisions (Byrne & Brooks, 2008). The next section presents the major theories of financial risk in the behavioral finance literature.

Behavioral finance theories on financial risk. The behavioral finance theorists argue that the behaviors of financial market participants are not rational as claimed by the classical finance theories. Instead, market participants have cognitive bias that includes *overconfidence, overoptimism, representativeness, conservatism, availability bias, frame dependence, mental accounting,* and *regret aversion* (Byrne & Brooks, 2008). Overconfidence and overoptimism occur when investors overestimate the information they have about a particular asset. Representativeness occurs when investors analyze asset information superficially rather than deeply. Conservatism applies when investors do not modify their investment strategies immediately after receiving new information about the investments. Availability bias occurs when investors overstate the probabilities of newly observed events to reoccur more often than older events. Frame dependence is the cognitive bias where any information is presented to investors in a way that influences their decision–making process. Mental accounting occurs when individuals mentally acquire and process different information without correlating the information. Regret aversion is also a cognitive bias where individuals make specific investment decisions in order to avoid future regret.

Behavioral finance theorists argue that prospective thinking rather than utility maximization drives the decision-making of individuals. The behavioral finance theories date back to Allais (1953), who argued that, during certain and uncertain times, psychological attributes are as important as monetary values when making decisions. Kahneman and Tversky (1979) extended such argument by proposing the prospect theory, a principal framework in behavioral finance literature, to demonstrate that the cognitive biases by investors refute the classical finance theories of rational behavior in financial markets. Its central premises are that individuals' subjective frames of reference influence the decision-making processes in the financial markets and investors are both risk seekers in the face of expected losses and risk averters in the face of expected gain (Byrne & Brooks, 2008). In contrast to the classical finance principles, such as the Sharpe's CAPM model which claim that higher (or lower) expected return is positively associated with higher (or lower) risk, Kahneman and Tversky demonstrated that individuals systematically violate the principles of the utility maximization theory.

According to the prospect theory, individuals make irregular decisions toward different levels of risky outcomes. Kahneman and Tversky (1979) argued that individuals overweigh outcomes they consider as certain and take more risks during events that have greater adverse results. They also observed that individuals tend to purchase insurance policies even during positive prospects. Since buying insurance is associated with the probability outcome of events, individuals are, therefore, not risk averters (Kahneman & Tversky, 1979). Moreover, since individuals are purchasing insurance against risky outcomes rather than attempting to avoid the risk-related events, they are not eliminating the risk. Moreover, individuals, according to Kahneman and Tversky, do not compare the different prospects before making decisions. According to Kahneman and Tversky, the empirical evidence shows that individuals possess a concave function for gain and a convex function for losses. The utility function in Figure 4 displays the prospect of losses having a steeper value than the prospect of gain. As a result, people are reluctant to gamble when the stake of losses is high, and they will avoid taking chances if both prospects for gains and losses are the same.

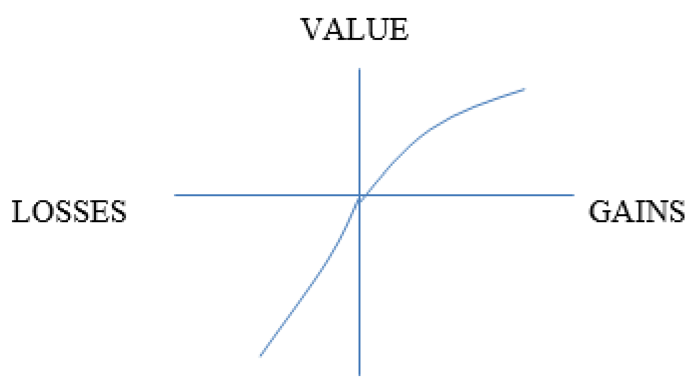


Figure 4. Utility function. Adapted from "Prospect Theory An Analysis for Decision Under Risk." by Kahneman, D., & Tversky, A. (1979). *Econometica*, *47*(2), p. 279. Reprinted with permission

According to the prospect theory, individuals make decisions according to what prospects they believe would have the highest value. Kahneman and Tversky (1979) showed that selecting the best prospect is a two-step decision-making process. First, individuals would identify and analyze the expected prospects using their subjective frame of reference by selecting the prospects with the highest value they perceive. Second, the editing step of the choice process consists of *coding*, *combining*, *segregating*, and *canceling* steps. During the coding and combining steps, the prospects are ordered and are associated with a frame of reference followed by combining the different set of references together. In parallel, individuals eliminate the riskless prospects. After the editing process, the decision maker selects the prospect with the highest value *V*, as shown in Equation 6.

For a decision *x* with a probability *p* and a decision *y* with a probability *q*, the prospect theory formula in Equation 6, according to Kahneman and Tversky, is the general form of the expected utility theory:

$$V(x, p; y, q) = \pi(p)v(x) + \pi(q)v(y) \tag{6}$$

where $\pi(p)$ is the probability *P* for a decision on scale π, $v(x)$ is the subjective value of the prospect, and $(x, p; y, q)$ is a regular prospect, such that [$p+q<1$] and [$x \geq 0 \geq y$] or [$x \leq 0 \leq y$] (Kahneman & Tversky, 1979). For strictly positive or negative prospects, the valuation equals the value of a riskless component and the difference in value between the outcomes multiplied by the weight associated with the extreme outcome. According to Kahneman and Tversky, Equation 6 can be rewritten as Equation 7 such that [$p+q = 1$] and [$x>y>0$] or [$x<y<0$]. The crux of the prospect theory is the incorporation of subjective perceptions into the valuation theory as opposed to the valuation theories in tradition finance that only assume objective outlooks and risk averseness by investors.

$$V(x, p; y, q) = v(y) + \pi(p)[v(x) - v(y)] \tag{7}$$

To support the prospect theory hypothesis, Kahneman and Tversky asked a group of research participants to select either Program *A* that would save 200 people from disease or program *B* with a one-third probability that 600 people could survive. Seventy two percent of the participants selected the certainty gain over the probability gain. Conversely, when they asked a separate group of research participants to choose either Program *C* that would kill 400 people or Program *D* with two-thirds probability that 600 people will die, 78% of the respondents preferred the probability risk to the certainty risk. All four programs, *A*, *B*, *C*, and *D*, have the same expected outcome of saving 200 lives but are framed differently–a sure gain of 200 lives, a one third probability gain of 600 lives, a certain loss of 400 lives, and a two thirds probability loss of 600 lives.

Kahneman and Tversky argued that if the traditional utility maximization is valid then the research participants should have selected the same program, *A*, *B*, *C*, or *D*, all of the time regardless of how they framed the questions. Furthermore, the authors conducted additional experiments to demonstrate the violations of Bernoulli's utility maximization theory as well as the rationality of human behavior that form the basic assumptions of the classical finance theories in risk management. Framing situations that can lead to irregular decisions, as the prospect theory demonstrated, could occur because of the communication about particular events as the case in the Program *A*, *B*, *C*, and *D* examples. Moreover, framing messages such as how a firm is not performing well in the market or how bad its product and services regularly occur through the media. The next section demonstrates the effectiveness of the media in society.

Theoretical Frameworks on Social Media

The media is an influential platform that can affect the decision-making of people because of the ways that it can broadcast information about events. From the earliest forms of gesture and language to the modern era of Internet and mobile communications, the creation, transportation, interpretation, and influence of communication have been critical aspects of societies (Thompson, 2013). The term *media* encompasses the multiple mediums, including television, radio, newspapers, and the Internet that different members of society use to communicate to the masses. Between the late fifteenth century and today, the industrialization of media has passed through major transformational periods.

The media in its various forms have existed as early as the 1500s. Such platforms include print media, such as books, newspapers, and journals, as early as the 1500s, films in the late 1800s, broadcasting media, including radio and television, in the 1900s, and the Internet media, including email, newsgroups, and websites, in the 1990s (Curtis, 2013; Wilke, 2010). As a social phenomenon, the media communication process reflects the social life through symbolic representations that are only meaningful for the individuals who produce or receive them (Thompson, 2013). Its universal form of power is equivalent to other cultural institutions, such as churches, schools, and universities. The messages broadcasted through the media by influential persons could potentially affect the decision-making of many individuals.

Media interaction frameworks. Media communication is an influential form of communication. It involves the production, transmission, and reception of symbolic forms through various technological mediums (Thompson, 2013). The power and influence of the media communication vary according to multiple societal factors. Some of these factors include the influence and social status of the speaker, the efficiency attributes of the media platform, and the message decoding abilities of the recipients. Government officials, corporate management, and individuals produce media information in the form of symbolic forms that include text, images, and videos through various media platforms, including television, newspapers, and the Internet. The receiving ends of the communications are the individuals who would analyze, interpret, and respond when necessary to the media messages. The interpretations, influences, and the actions of the media participants are the key focus in this study.

Interpreting the meaning of the messages received through the media is a subjective process that can vary in its output between the persons that are interpreting the messages. The theoretical implications of media reception, the receiving end of the media communication process, are that they are situated, routine, skilled accomplishment activity, as well as a *hermeneutic* process (Thompson, 2013). It is not a passive process as individuals receive the media information and analyze its symbolic materials. It is a situated activity since processing of media information is dependent on the particular social context. It is also a routine activity since media communication is part of everyday life. The reception process is also a skilled accomplishment since it requires higher levels of cognitive skills to interpret the information. Furthermore, media reception is a hermeneutic process where individuals interpret the symbolic signals according to their frame of mind.

The platforms used in the media communication process, whether it is physical or virtual, can influence the accuracy of interpreting the relevant information in the media. Three frameworks establish the different methods of social relationships formed by the media interactions in society–*face-to-face interactions*, *mediated interactions*, and *mediated quasi-interactions* (Thompson, 2013). Face-to-face interactions occur in the form of a dialogue between participants who are physically present next to one another. Nonverbal cues, such as winks, frowns, or smiles, supplement the interpretation of words between physically present participants. Mediated interactions are forms of dialogue between participants that cannot occur in a physical face-to-face environment. An intermediary platform is also necessary within the communication process. Examples of technology mediums as the mediated platforms include post letters, emails, and phones. Since the physical presence is absent in mediated interactions, nonverbal cues are not possible. On the other hand, the mediated quasi-interactions, including television, radio, and the printing press, are forms of mediated interactions that only provide monologues instead of dialogues during the communication process. Within the traditional media, corporate firms use mediated quasi-interactions to broadcast quarterly releases, and financial analysts use similar platforms to disclose their opinion about the financial health of companies or provide recommendations for stocks investments.

Crisis communication using the media. Corporate communication using the media has been an essential mechanism to broadcast corporate messages, such as new product offerings and financial information, to corporate shareholders and the public. Corporate management would also leverage the media to share information about major events, such as a crisis, a possible merger, or a corporate restructuring plan (Schultz, Utz, & Göritz, 2011). During a crisis, the information that stakeholders receive from interactions through the media and corporate management effects the reputation of the corporate under crisis (Schultz et al., 2011). Furthermore, the response strategies by corporate management could lead to negative reactions from the public. Negative words of mouth are unfavorable comments and opinions that spread from one person to another and ultimately hurt a corporate image. The literature shows that negative information, in comparison to positive information, attracts more attention, raises more questions, and triggers more behavioral responses (Akhtar et al., 2012). Hence, in time of crisis, negative sentiment toward corporate actions could have an adverse effect on the reputation of corporate institutions.

Social amplification of risk. Negative information through the media can have a detrimental impact on societal events including corporate distress. Kasperson et al.'s (1988) social amplification of risk framework, SARF for short, is a theoretical framework that demonstrates the effect of the media on risk-related events. "The social structures and processes of risk experience, the resulting repercussions on individual and group perceptions, and the impact of these responses on community, society, and economy compose a general phenomenon that we term the social amplification of risk" (Kasperson et al., 1988, p. 179). Its authors formulated the framework in 1998 to reconcile what they believed were fragmented frameworks on risk perception within the fields of the social sciences (Duckett & Busby, 2013). The central premise of their framework is that the public perception towards a risky event can generate behavioral responses that can amplify or attenuate the physical risk itself.

The SARF framework includes the signal amplification process, borrowed from communications theory, as a metaphor to describe the process of either intensified or attenuated signals that would occur during the decoding of messages as information transverses its source to its destination while passing through intermediary receivers (Kasperson et al., 1988). The different factual, inferential, valuable, and other key symbols within each message could affect the interpretation of the message and the reaction towards its content from the receiving end (Kasperson et al., 1988). Hence, the media can amplify or impede the interpretation of the messages by the receiving end of the communication process.

According to the SARF framework, the amplifying stations include experts or risk assessors, the media, and opinion leaders amongst societal groups. The key steps to amplified risk, according to Kasperson et al. (1988), are as follows: (a) Filter signals to process the relevant fraction of the information. (b) Decode each signal. (c) Infer the risk information within each of the signal. (d) Apply social values to each of the risk-related signals. (e) Interact and communicate amongst societal groups in order to interpret and validate the signals. (f) Determine if one can tolerate the risk or if an action is necessary to address the risk. (g) Take action, either as an individual or as a group, to accept, reject, tolerate, or change the risk. After the last step, according to the framework, the outcome of the process, if not controlled, could spread across society and ultimately lead to adverse events.

Under the SARF framework, the ripple effects of socially amplification risks start with those directly affected by the risks and then expand to the institutions, such as companies or even to larger dimensions in society. In the case of corporate institutions, the effect could trigger managerial interventions against the amplified risks or hinder any action towards the attenuated risks (Kasperson, 2012). In addition, the secondary effects of the amplified risk could have market impacts, such as loss of sales in the case of business risks or social disorder in the case of political hazards. It can also have contagious effects across organizations.

The spread of information and the presence of media as a risk amplification station are critical factors for the impact of the social media on the expected bankruptcy of firms in the current study. The SARF framework provides an overarching framework where societal perceptions towards risk and the roles of the media regardless of the technical assessment by experts could have amplified responses that may lead to adverse consequences (Kasperson, 2012). Social media is a relatively new form of the media, where the interexchange of expert and inexpert opinions in public forums can occur at larger and faster scale than the traditional and lesser interactive media (Curtis, 2013).

Under the SARF framework, the interpretation of the messages shared among individuals, including social media users; depends on how the receiving end of the communication process perceives the information. The media, such as the social media websites, would then serve as amplified stations.

Unlike traditional mediums, such as newspapers and television, where internal staff filters the information before broadcasting messages to the public, social media provides the communication base for social media participants to broadcast unfiltered information between each other. Social media users share any information among themselves regardless of whether the messages are factual or inaccurate. The volume and speed of information dissemination within the social networks depend on the weak and strong ties among the different network groups (Bakshy, Rosenn, Marlow, & Adamic, 2012). Using the SARF framework, messages related to risk-related events in Internet forums, such as social media, could lead to heightened public responses even if such responses contradict with technical assessments by experts (Duckett & Busby, 2013). Several studies have leveraged SARF in measuring amplified responses from the public on social media and various Internet-based forum. The approach can follow a qualitative or a quantitative method. An example of a qualitative study of investor sentiment is Larson, Cooper, Eskola, Katz and Ratzan (2011) case study of investor sentiment towards vaccination; an example of quantitative study on investor sentiment is Chung (2011) correlation analysis of between the volume of articles and the number of comments on online forums in South Korea. However, there is no evidence of recent literature that applied SARF within the context of corporate bankruptcies and social media.

The literature surrounding the social amplification of corporate financial distress as affected by sentiment in social media is limited, but the literature on measuring investor sentiments and expected bankruptcies independently is extensive. The next segment of the Chapter 2 in the dissertation includes the key studies and relevant methodologies for measuring financial distress of firms and investor sentiment in social media.

Literature Review

Corporate Financial Distress Analysis

Corporate bankruptcy risk. Measuring the risk of corporate bankruptcy is a critical task that corporate managers and finance speculators can do using statistical methods. The possibility of corporate bankruptcy is inherent in any business. Managing the exposure of different risk including operational risks, market risks, and credit risks could lead some firms to prosper and other firms to encounter financial distress and possible bankruptcy (Graham, Hazarika, & Narasimhan, 2011). The risks of financial distress and bankruptcies are the types of risks that are of interest in this study. Internal factors that can directly lead to financial distress of firms include reduced sales, excessive debt, and little analysts' coverage, while external factors that can have an indirect effect include macroeconomic conditions and financial market turmoil (Campbell et al., 2008; Cole & Wu, 2009; Graham et al., 2011). Predicting the likelihood of financial distress of firms and measuring their expected bankruptcy are important measurements for corporate management, investors, debtors, and creditors (Chava & Purnanandam, 2010). Various statistical methods exist that could help predict the likelihood of bankruptcy. The next section summarizes the common approaches found in the literature that can predict the likelihood of corporate financial distress.

Measuring corporate financial distress. The most common approach to predict the likelihood of corporate bankruptcy is to apply a statistical analysis method using the financial data of corporate firms. Both quantitative and qualitative methods can help determine the financial distress of firms, but, similar to other research in the field, the most common research model in academic finance is the quantitative method (Ardalan, 2008). The case studies on the operational activities of specific firms or the phenomenological studies on financial market contagions are some examples of a qualitative approach (Janesick, 2010). However, unless qualitative studies induce newly grounded theories, generalizing the research outcomes to support the overall population is not possible with qualitative research (Creswell, 2009). Quantitative research about the financial distress of firms is more common than qualitative research in academic finance.

The most common approach to measure corporate financial distress is to follow a quantitative research approach when a hypothesis is first proposed and is then followed by empirical evidence in order to support or reject the null hypothesis. The approach follows the general quantitative methodology, discussed in Creswell (2009) and summarized as follows:

1. Identify the problem statement, theoretical questions, and the research hypotheses.
2. Identify the research method and the variables (dependent, independent, and control variables) the study.
3. Identify the research sample and extract financial data about the sampled firms from financial databases.
4. Apply the research method using the identified research variables and the sample data.
5. Determine if the research results support the hypothesis after ensuring the reliability and validity of the data.

Selecting the exploratory variables that will determine the likelihood of financial distress and followed by identifying the type of data analysis are primary functions in the study of corporate financial distress. Table 3 includes the common exploratory variables applied in corporate financial distress studies. Researchers would extract the financial data about corporate firms from various databases, such as Standard and Poor's (2012) CRISP/Compustat database. They would then apply different exploratory variables, such as the financial ratios in Table 3, as the independent variables into their model in order to predict the likelihood of corporate financial distress.

Table 3

Key Exploratory Variables for Corporate Distress Predictions

Category	Covariate	Definition
Profitability	1.EBIT margin	EBIT/operating revenue
	2. Return on equity	NPAT before abnormals / (shareholders equity minus outside equity interests)
	3. Return on assets	Earnings before interest/(total assets minus outside equity interests)
Liquidity	4. Current ratio	Current assets/current liabilities
	5. Quick ratio	Current assets minus current inventory)/current liabilities
	6. Working capital/total asset	Working capital/total asset
Leverage	7. Debt ratio	Total debt/total asset

(Continued)

Category	Covariate	Definition
Activity	8. Capital turnover	Operation revenue/operating invested capital before goodwill
	9. Total asset turnover	Operation revenues/total asset
Company-specific	10. Size of company	Log of total asset
	11. Squared size	Square of log of total asset
	12. Age of company	Number of years since registration
Market-based	13. Excess returns (year t)	Company's stock return the previous year minus ASX200 index return the previous year

Note. Adapted from "Firms in Financial Distress, a Survival Model Analysis" by N. Chancharat, P. Davy, M. S. McCrae, and G. G. Tian, 2007, *SSRN eLibrary*. Reproduced with permission.

The most common method for predicting the financial distress of companies is by quantitative methods. Beaver's (1966) financial ratios model is one of the early quantitative models in measuring any financial distress of firms. The model relies on univariate financial ratios as the criteria for measuring the financial distress of firms.

Altman (1968) later argued that univariate ratios are ineffective in the financial distress predictions since some financial ratios are not applicable across all market sectors. Altman proposed the multivariate discriminant model, also known as Z-Score, which became one of the early benchmarks in measuring the risk of financial distress across firms in the United States and elsewhere (Altman, Iwanicz-Drozdowska, Laitinen, & Suvas, 2014). The limitation of the models from Beaver and Altman is that they assumed a linear path to bankruptcy using static accounting data that represent the condition of firms at a particular point in time (Bharath & Shumway, 2008). Firms, however, can experience rapid changes such as corporate restructuring that could alter their expected bankruptcy conditions.

Future publications later addressed such shortcomings, including Altman's (2013) revisit of the Z-Score model, Shumway's (2001) hazard model that includes time-dependent variables, and Campbell et al.'s (2011) logistic regression model. Alternatively, machine-learning models, such as artificial neural networks and the support vector machine that use complex computation algorithms also provided more powerful predictive modeling for measuring the financial distress of firms (Lee & To, 2010; Salehi 2013). The accuracy and the predictive power of all such models vary according to the data and the sampling approach used, but they all demonstrate more accurate results than the univariate financial ratios (Campbell et al., 2011; Lin, 2009). The next section includes three key statistical methods, the multivariate discriminant analysis, the logit/probit regression model, and the artificial neural network model within the field of measuring

Multivariate discriminant analysis. The multivariate discriminant analysis is one of the early but still popular statistical method for measuring corporate financial distress bankruptcies. Altman (1968) first used the multivariate discriminant analysis model to predict financial distress of firms instead of the univariate financial ratio approach previously used by Beaver (1966). Altman argued that estimating corporate failure using profitability, liquidity, and solvency ratios have been effective in previous studies, but such studies did not all apply the same financial ratios (Altman, 1968). The methodology used in earlier research solely focused on univariate variables and particular problem signals that may not necessarily apply to all companies. As a result, such type of research would cause ambiguities when comparing the relative performance of various firms (Altman, 1968). Instead, Altman implemented a multivariate discriminant analysis model, named as Altman *Z*-Score, which remains popular to this day in the finance literature.

Altman *Z*-Score is a simple yet useful statistical tool to measure the expected bankruptcy of firms using a weighted multivariate set of financial ratios. It includes five different financial ratios and their respective weights. The outcome of the model is a value that would determine the likelihood of a firm to go bankrupt in two years. The equation for Altman *Z*-Score is as follows:

$$\text{Z-Score} = 0.012*\text{WC} + 0.014*\text{RER} + 0.033*\text{EBIT} + 0.006*\text{MKLI} + 0.999*\text{SALES} \quad (8)$$

where WC = working capital / total assets, RER = retained earnings / total assets, EBIT = earnings before interest and taxes / total assets, MKLI = market value equity / book value of total liabilities, SALES = sales / total assets, and Z-Score = overall financial distress index. A Z-Score that is greater than 2.99 indicates a *safe zone* or the firm is at a lower risk of bankruptcy in the next two years. A Z-score that is between 1.8 and 2.99 indicates a *gray zone* or the firm is in an undermined state of possible bankruptcy. A Z-score that is less than 1.81 indicates a *distress zone* or the firm is at a higher likelihood of bankruptcy in the next 2 years. Altman constructed the coefficients of the model in Equation 8 after analyzing a number of companies that either defaulted or survived during the period between 1946 and 1965. Several research, including Altman's revisit of the model in Altman (2013) and in a study of international firms in Altman et al. (2014), demonstrated the effectiveness of the model in identifying the financial distress of firms. However, various literature has countered the effectiveness of Altman's model.

Extensive research supported Altman *Z*-Score as a reliable predictor of financial distress of firms, but numerous studies in recent literature questioned its efficiency and proposed alternative financial models. Hayes et al. (2010) applied Altman *Z*-Score over a sample of retail firms that defaulted between 2007 and 2008 and concluded that it predicted the bankruptcy of such firms 94% of the time. Apergis et al. (2011) observed a positive correlation between Altman *Z*-score and firms' stock prices. On the other hand, Mansi et al. (2010) argued that Altman's (2001) Z-Score has a lower predictive power than the logit models of Ohlson (1980) and Campbell et al. (2008). Similarly, Lin (2009) observed that the probit, logit, and neural network models have more predictive power than Altman *Z*-Score does. According to Hayes et al. (2010) and Lifschutz and Jacobi (2010), the multivariate discriminant analysis models, including Altman *Z*-Score, remain efficacious, nevertheless, in the academic literature and among financial organizations.

Even though Altman *Z*-Score and similar multivariate discriminant models were arguably effective in identifying financial distress of firms, they had two problems. The filtering of distressed firms from nondistressed firms was necessary during the development of the model, and the financial ratios used to develop the formulas needed to be independent of each other (Ahmadi, Soleimani, & Vaghfi, 2012; Zmijewski, 1984). These issues were associated with the nonrandom sampling procedure and the selection bias in the data when using multivariate discriminant analysis models (Zmijewski, 1984). Since the frequency of the defaulted firms in the overall population is low when compared to financially healthier firms, scholars tend to overstate the sample frequency rates for financially distressed firms in order to ensure that the variance and covariance matrices of the financial distress predictors are the same for both failed and nonfailed groups (Ohlson, 1980). In addition, a recent study of financial failure of publically traded firms in Israel by Lifschutz and Jacobi (2010) showed that the predictive ability of Altman *Z*-Score with respect to predicting bankrupt companies is very high but is less efficient when predicting stable companies. Alternatively, the regression analysis models including logit and probit regression models performed better in some studies than

Altman *Z*-Score and other multiple discriminant analysis models (Christidis & Gregory, 2010). The next section describes such regression models.

Logit and probit regression models. Applying regression analysis is an alternative approach to using multivariate discriminant analysis in the measuring of corporate bankruptcies. Notable regression models for predicting corporate bankruptcies extend Ohlson's (1980) logit model and Zmijewski's (1984) probit model. These models took an alternative approach to the multivariate discriminant models by ignoring any presumptions regarding prior probabilities of failure and the distributions of the prediction variables. Ohlson first constructed the logistic regression model using 105 failed and 2058 nonfailed firms for the sample period between 1970 and 1976 (Lin, 2009). Ohlson's logit model is as follows:

Given a set of observations as data points,

Let X_i be a vector of predictors for the *i*th observation.

Let β be a vector of unknown parameters.

Let $P(X_i, \beta)$ represent the probability of bankruptcy for any given X_i and β with $0 \leq P \leq 1$

The likelihood of any particular outcome is a function $l(\beta)$ where S_1 is the index set of bankrupt firms and S_2 is the set of nonbankrupt firms.

$$l(\beta) = \sum_{i \in S_1} \log P(X_i, \beta) + \sum_{i \in S_2} \log(1 \quad P(X_i, \beta)) \tag{9}$$

For any specified function P, the maximum likelihood estimates for $\beta_1, \beta_2, ...$, is the result of solving for $\max_{\beta} l(\beta)$. Since no predefined function P exists for defining a bankruptcy, Ohlson used the following function:

$$P = \left(1 + \frac{1}{e^{y_i}}\right) \text{, where } y_i = \sum_j \beta_j X_{ij} = \beta' X_i \tag{10}$$

where P increases in y and y is equal to log[P/(1-P)] (Lin, 2009, p. 3510; Ohlson, 1980, p. 118). In Zmijewski's probit regression model, according to Lin (2009), the logistic function $l(\beta)$ is the standard normal cumulative distribution function:

$$l(\beta) = \Phi(z) = \int_{\infty}^{\infty} \phi(v) dv \tag{11}$$

where $\phi(z)$ is the standard normal density function $\phi(z) = \frac{1}{\sqrt[2]{2\pi}} e^{\frac{z^2}{2}}$.

Both the logit and the probit regression model can help predict the expected bankruptcy of firms. The main difference between the logit and the probit model is in the cumulative distribution function. The logit model uses the standard logistic distribution function, whereas the probit model uses the standard normal distribution function (Van der Ploeg, 2010). Both models are regression models with continuous/categorical covariates, such as the variables in Table 3, and the dependent variable is discrete for the probit model and continuous for the logit model. Both also use the maximum likelihood estimator. However, the predictive power of both models is dependent on the assumptions used in their respective transformation functions (Van der Ploeg, 2010). Some researchers argue that both models have better prediction capability than Altman *Z*-Score (Lin, 2009; Van der Ploeg, 2010). However, similar to the multivariate discriminant analysis models, these models also assume that a bankruptcy occurs at a discrete point time and under a specified sample period. Hence, these models do not consider the dynamic changes in the financial conditions of distressed firms and the overall financial markets.

Initial signals of economic distress of companies are beneficial in the prediction algorithms, but, according to Van der Ploeg (2010), it is not observable when using logit, probit, and the multivariate discriminant analysis models. The dynamic logit model, as in Campbell et al. (2008, 2011), extended the models of Chava and Jarrow (2004) and Shumway (2001) hazard model to address the limitations of the static regression models by correcting for the periods at risk of financial distress and incorporating time-varying co-variables within the model. However, such type of regression models assumed a predefined set of financial ratios to help identify financial distress signals or predict the likelihood of bankruptcies. Some researchers saw the need to leverage sophisticated computational and artificial intelligence systems that would use larger sets of data in order to identify new insights on how to predict financial distress and the likelihood of corporate bankruptcies. Such systems include machine-learning algorithms, such as the artificial neural network model.

Artificial neural network models. Machine learning models are technically advanced statistical models that use more computational power than the traditional regression models, such as the logit and probit models discussed earlier. The models stem from the field of artificial intelligence, and they include decision trees, fuzzy set theory, case-based reasoning, genetic algorithms, Support vector machine, data envelopment analysis, rough sets theory, and various kinds of artificial neural networks (ANN). ANN models include the back propagation neural network, probabilistic neural network, self-organizing map, and cascade correlation neural network (Yu, 2013). In finance, notable machine learning models include ANN, support vector machine, radial basis function neural network (RBF), multiplayer perception (MLP), and self-organized competition or SOC (Lee & To, 2010). The most common of such models in the prediction of corporate financial distress is the artificial neural network model.

Artificial neural network models try to mimic the interactions between human neurons in order to generate more intelligent outcomes than other less complex statistical models. Artificial neural network models are nonrestrictive and nonparametric alternatives to statistical models, such as the linear regression models. First proposed by McCulloch and Pitts (1943), these models replicate the biological characteristics of human neural networks by learning from the data in order to make better predictions. Unlike the traditional statistical models that are entirely dependent on the human selection of data and methods, the neural network models require little training, also known as *supervised* learning, and sometimes no training at all, also known as *unsupervised* models, in order to develop the *intelligence* behind its results. Researchers have successfully applied artificial neural network models across various domains, including web page ranking for Internet search engines, face recognition, automated translations of multilingual documents, robot controls, and in the field of bankruptcy predictions (Yu, 2013). One effective mechanism in predicting the likelihood of corporate bankruptcy using an artificial neural network model is by applying a supervised learning algorithm using the back propagation neural model. Supervised learning for neural network models is the approach in which the system identifies relevant information in the training data set and then attempts to identify the missing information in the actual data set. The likelihood of corporate bankruptcy using neural network model is a supervised binary classification problem (Yu, 2013). Yu wrote:

> As to the bankruptcy prediction problems, it is always treated as a binary classification one. Each sample of the data belongs to a group of predefined classes, *Bankrupt* or *Nonbankrupt*, and the objective is to try to separate one class from the other with the minimum amount of error. (p. 33)

A back propagation neural network model (BPN) is a multilayered feedforward neural network in which the possible factors that can classify a set of firms as either likely or less likely to become bankrupt are first learned from a training data set, and the findings are then applied to a separate sample dataset. The model would find the set of weight of values that generate the output that best fits the existing data.

The general structure of the BPN model in Figure 5 shows the model as an *l-m-n* feed forward architecture, where the input layer constitutes of *l* input variables, the hidden layer constitutes of *m* hidden variables, and the output layer constitutes of *n* output variables. The model is applicable to multiple areas of the research including corporate bankruptcy predictions, where the input layer could include input variables, such as the financial ratios in Table 3 as exploratory variables, and the output layer could be a binary variable that would classify the firms as either likely or less likely to default. The hidden layer would do all the information processing by approximating the nonlinear relationship between the input and the output variables using weight adjustments of the variables in the model.

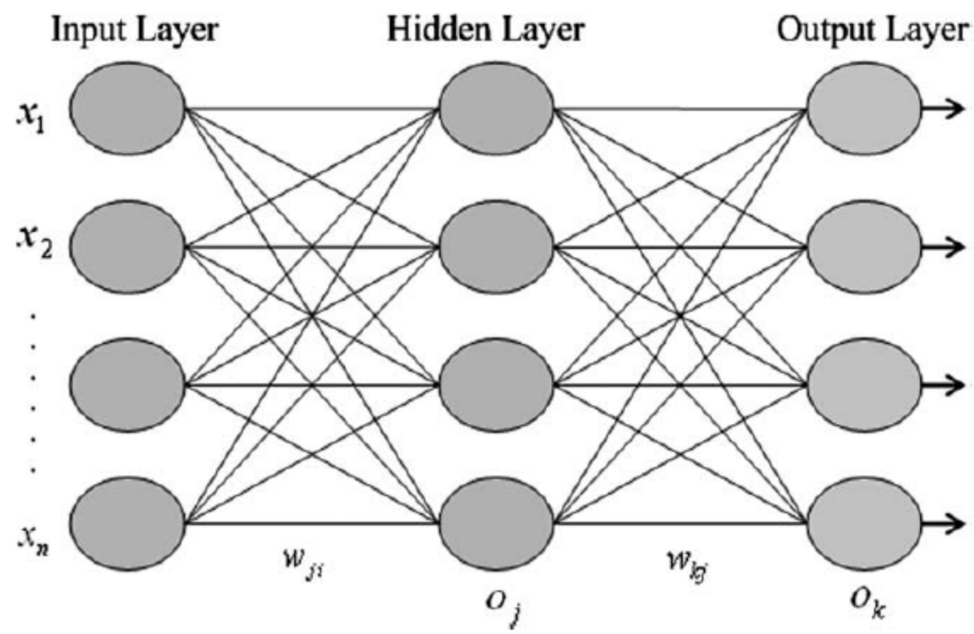


Figure 5. Back propagation neural network. Retrieved from Chen, W.-S., & Du, Y.-K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, *36*(2), 4075–4086. Reproduced with permission.

The key processing steps of the BPN model occur in its hidden layers. An interconnection weight, known as a weight factor w_{ji} that represents the interconnection between the *i*th node of the first layer and the *j*th node of the second layer adjusts the input signals from the input layer. Similarly, the output signals of the hidden layer are modified by the interconnection weight, w_{jk} , between the *k*th node in the output layer and the *j*th node in the hidden layer. A sigmoidal transfer function transfers the sum of modified signals across the layers and until output layer. In mathematical terms, Panda et al. (2008) defines the BPN model as follows:

Let $I_p = (I_{p1}, I_{p2,\ldots,} I_{pl})$, p=1,2,…N be the input vector with p^{th} among N input patterns.v

Let W_{ji} and W_{kj} be the interconnection weights between the i^{th} input node to j^{th} hidden neuron and between the j^{th} hidden neuron to the k^{th} output node respectively.

The output from an input layer node is O_{pi} , where $O_{pi} = I_{pi}$, i=1,2,….l.

The output from a node in the hidden layer is O_{pi} , where

$$O_{pi} = f(NET_{pj}) = f\left(\sum_{i=0}^{l} W_{ji} O_{pi}\right), \text{ j=1,2,…m.}$$

The output from a node in the output layer is O_{pk} , where

$$O_{pk} = f(NET_{pk}) = f\left(\sum_{j=0}^{m} W_{kj} O_{pj}\right), \text{ k=1,2,…n.}$$

The BPN model can support multiple hidden layers, but, according to Lee and To (2010), one hidden layer has shown to be sufficient. However, the number of nodes selected in the hidden layer can affect the performance of the model. Furthermore, the model also requires two parameters by the user, a learning rate of α (0 < α < 1) and a momentum rate η (0 < η < 1).

Applying the BPN model is a two-staging process. A supervised training process is first run followed by the actual testing process. During the training process, the interconnection weights, W_{ji} and W_{kj}, are first adjusted using the delta rule algorithm, and the predicted output is compared with the expected output. If the computed mean square is greater than the expected value, a readjustment process of the interconnection weights occurs using a back propagation from the output layer to the input approach until the error is minimized or until the expected iterations are within the expected limit (Panda et al., 2008). The mean square error, E_p, for a pattern *p* is defined as

$$E_p = \sum_{i=1}^{n} \frac{1}{2}\left(D_{pi} - O_{pi}\right)^2 \tag{12}$$

where D_{pi} is the targeted output and O_{pi} is the computed output for the i^{th} pattern. The weight change at any time *t* is $\Delta W(t)$, where $\Delta W(t) = -\eta E_p(t) + \alpha \Delta W(t-1)$, η is the learning rate, and α is the momentum rate (Panda, Chakraborty, & Pal, 2008). Once the training process is complete, the researcher feeds the testing data into the trained network in order to determine the percent variation in the predicted output and the actual output (Panda et al., 2008). The model would then solve for the minimum variation as an optimization problem in order determine the predictable variables that can assist in determining the likelihood of corporate bankruptcy.

The back-propagation neural network model is an effective machine-learning algorithm for predicting the probability of corporate bankruptcies. Researchers, including Zurada et al. (2011) and Lee and To (2010) have applied such model into corporate financial distress analysis. Zurada et al. (2011), Lee and To (2010), and Chancharat et al. (2007) argue that the machine learning models provide more accurate predictions than the multivariate discriminant analysis and other statistical models. However, not all research support the argument that the neural network models are more superior to the traditional statistical techniques. For instance, Altman et al. (1994) previously compared the logit model to the neural network model in the study of 1,000 Italian firms and found that both models have similar prediction accuracy. They also observed that the illogical weightings and the overfitting of data during the training stages of the neural network models negatively affect its statistical accuracy. Lin (2009) also showed that the probit model has a higher prediction accuracy than the neural network model, but the latter can be superior whenever data does not satisfy the assumptions of the statistical approach. Nevertheless, such statistical models demonstrate better accuracy than the univariate ratios to predict the likelihood of corporate bankruptcy.

In summary, the most common quantitative models in predicting the financial distress of corporate firms include the multivariate discriminant analysis, the logit and probit regression models, and the artificial neural network models. Such models typically rely on firms and market variables, such as the financial ratios in Table 3, as the exploratory variables in order to determine if companies are likely to default or not.

Furthermore, perceived market information about companies, following Kasperson (2012) social amplification of risk, can have a negative effect on the valuation and business transactions of financially distressed firms which can further lead to more distress (Hill, Perry, & Andes, 2011). Expressing the sentiment using textual methods such as posting messages on social media or participating in Internet-based forums can reflect the opinion of market participants towards firms. The next segment of the chapter includes the theoretical concepts related to sentiment analysis of public opinion as textual-based emotions within the media.

Investor Sentiment Analysis

The sentiments expressed by the public that had an effect on societal events have prompted the analysis of such phenomenon in order to help predict possible outcomes in the future. Seeking public opinion or the opinion of others is an information gathering behavior that exists in human civilizations as part of the cognitive decision-making process (Pang & Lee, 2008). Whether it is governments or political institutions seeking the opinion of the public or individuals looking for advice from others, expressing and interpreting ideas require subjective analysis. The process requires a subjective understanding of what individuals communicate using various forms of verbal and nonverbal expressions. Using the research paradigms matrix in Figure 2, the decoding of public opinions falls under the interpretist paradigm. The most common approach to express and interpret opinions is through language since the latter is the cognitive mechanism that humans use in the form of verbal conversions between one another using different communication mediums, such as face-to-face, correspondence, and social media.

The human language. A unique ability of individuals to express their sentiments is by using human language. Archaeological evidence suggests that human language evolved within the past 100,000 years and, from a biological perspective, no other specifies possess a communication capability that is equivalent to the human language (Berwick, Friederici, Chomsky, & Bolhuis, 2013). The basic design of the human language in Figure 6 consists of three components, *syntactic rules and representations*, *external sensory-motor interface*, and *internal conceptual-intentional interface*. The syntactic rules and representations together with the lexical items constitute the basis of the language system, and the two interfaces are the platforms where which our mental expressions connect with the outside and the internal world (Berwick et al., 2013). Individuals would create and produce an unlimited number of expressions that others with similar sets of knowledge can interpret. Using the external sensory-motor interface, such as vision and hearing, individuals produce and perceive sequences of words as they interact with the outer world. Through their conceptual-intentional interface that is internal to their cognitive brain functions, individuals would then conceive, reason, and interpret the words and the associated syntactical rules.

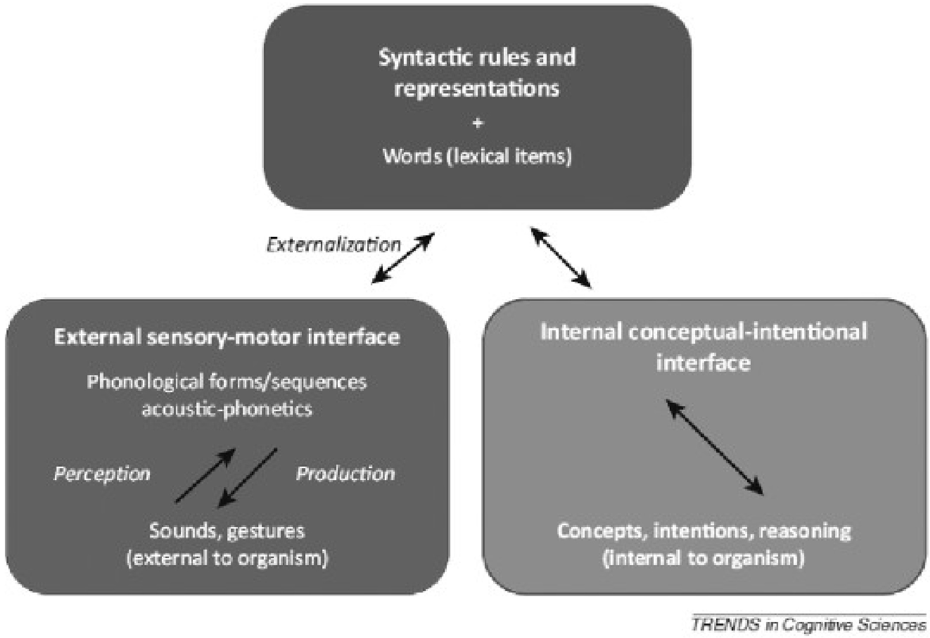


Figure 6. The basic design of the human language. Retrieved from Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17(*2), 89–98. Reproduced with permission

Advancements in the field of neuroscience, the study of the brain and its neurons, has made it possible to view the significant triggers in the human brain that relate to the construction and comprehension of language. However, the complete anatomy of the neurons in the human brain that relate to language and speech remains difficult in current literature (Friederici, 2011). The primary language—related areas in the human brain are the inferior frontal cortex (IFG), the superior temporal gyrus (STG), and the middle temporal gyrus (MTG), shown in Figure 7. After a person hears a speech sound, the process of sentence processing follows three linguistic processing phases in the following order: a) build the phrase structure based on word category information, b) compute the syntax and the semantics of the sentence, and c) comprehend the sentence (Friederici, 2011). The biological construction of human language aided the development of the natural language processing (NLP) field that began in the 1950s as an intersection between the study of linguistics and artificial intelligence within the field of computer science (Nadkarni, Ohno-Machado, & Chapman, 2011). The global transformation of societies into the era of computers, mobile phones, and the Internet where people would interact more frequently through written correspondence drove an increased interest in leveraging technologies to classify the textual messages using natural language processing mechanisms.

Natural language processing. Natural language processing, also commonly known as NLP, is the field of research that deals with modeling, constructing, interpreting, and predicting natural language using computer algorithms. This area of study was originally separate from the field of information retrieval (IR), which employs statistical-based techniques to extract large volumes of data. However, recently, both NLP and IP have somewhat converged since, recently, scientists and market researchers have been obtaining and analyzing large amount of data from different Internet-based mediums including social media (Nadkarni et al., 2011). The most common NLP algorithms are syntax-based that focus on single or adjacent groups of words using pattern-matching techniques without necessarily understanding the semantics of the formed sentences (Cambria & White, 2014). Advanced models are semantic-based and incorporate facts, such as inferring that a chair is furniture. They also add common sense knowledge, such as concluding that people smile when they are happy. These models are effective in deconstructing natural language into sentiments, such as an adverse feedback about a product or a positive feedback about a movie review (Cambria & White, 2014). Empirical frameworks that support both types of NLP models, the syntax-based and the semantic-based, are available in the academic literature.

The major frameworks in NLP since its inception in the early 1950s include production rules, semantic pattern matching, a first order logic, Bayesian networks, semantic networks, and ontology web language. The production rules, first proposed by Chomsky (1956), are independent sets of condition and action statements that combine words into phrases and phrases into group of sentences until the process ties all the words together. The first order logic, introduced in Barwise (1977), is a deductive system that specifies how symbols should be properly formed, the meaning of the formed expressions, and the method by what textual information can correlate effectively with one another (Cambria & White, 2014). The Bayesian network method, as suggested by Pearl (1985), uses probability distributions to predict word formations using prior knowledge. The ontology web language, as proposed by McGuinness and Harmelen (2004), is a comprehensive list of structured knowledge representing words that the machine-based systems could use to interpret and process content on the Internet. The semantic network, as suggested by Sowa (1987|2006), is a graphical representation of knowledge in the form of interconnected nodes and arcs. The six common kinds of semantic networks are *definitional networks*, *assertional networks*, *implicational* *networks*, *executable networks*, *executable networks*, *learning networks*, and *hybrid networks* (Sowa, 2013). Each of the networks represents knowledge in a different matter, but machines are capable of processing the semantics behind the information presented through logical means using such systems.

Like any machine-based systems, none of the current NLP models can perfectly interpret the subjectivity in human communications. For instance, Chomsky (1956) production rules can grow exponentially due to the vast size of parts of speech (nouns, verbs, and adjectives) that makes such system difficult to manage (Nadkarni et al., 2011). Some assertions in Barwise (1977) first order logic may not hold true. For example, if the statements *all birds fly* and *penguin is a bird* are facts it is not possible to say that penguins can fly (Camrbia & White, 2014). In fact, penguins are birds that do not fly. In the case of Pearl (1985) Bayesian network model, the system requires large tables of data that make it also difficult to manage for large-scale information processing. McGuinness and Harmelen (2004) ontology of web language also cannot handle the subjectivity within the content on the Internet, and it is only suitable for representing declarative knowledge (Camrbia & White, 2014). Early semantic network models also lacked the performance and capability to handle the complexity of human language when placed in a cognitive context (Borge-Holthoefer & Arenas, 2010). However, the omnipresence of modern day technologies including computers, mobile phones, and social media provide researchers with extensive databases of human interaction data that are interpretable and classifiable (Rambocas & Gama, 2013). A prevailing use of such data is in public sentiment analysis.

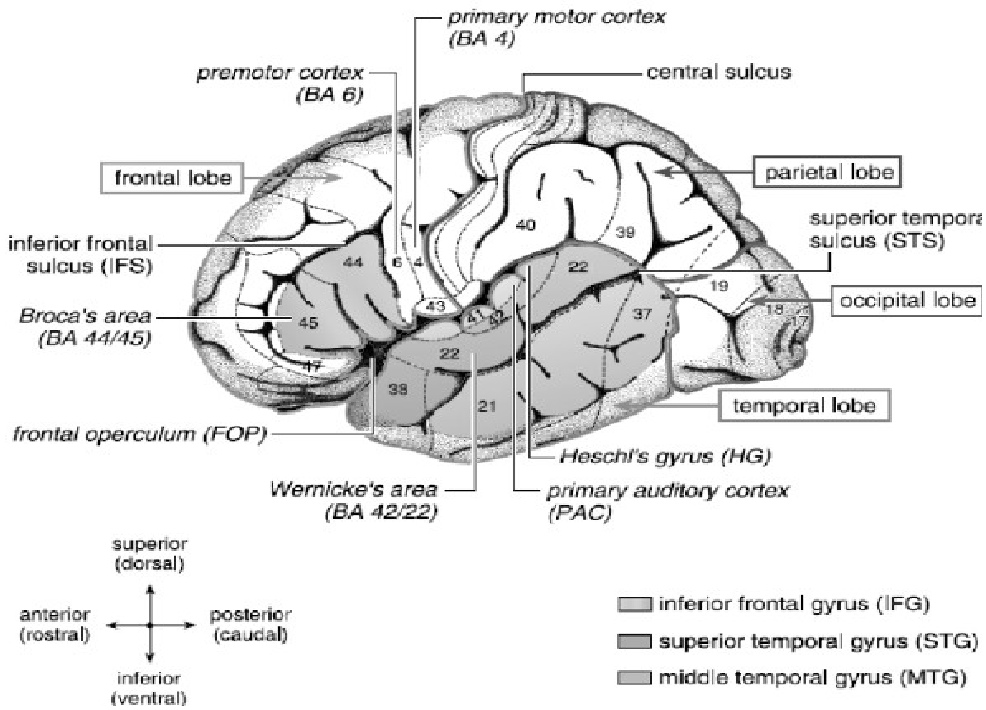


Figure 7. Left hemisphere of the human brain. Retrieved from "The brain basis of language processing: from structure to function" by Friederici, A. D. (2011). *Physiological Reviews*, 91(4), 1357–1392. Reproduced with permission

Public opinion & sentiment analysis. Sentiment analysis includes understanding public opinion on matters that concern governments or corporate institutions. From a political perspective, public opinion is a communication from citizens to their governments and between citizens themselves (Speier, 1950). Past scholars cited public opinion as one of the most valuable rights of men. Shakespeare called public opinion *the* *mistress of success*; Pascal called it *the queen of the world*; and John Locke considered the law of opinion as one of the three laws that men rectitude their actions (Speier, 1950). To this day, assessing public opinion and the opinion of others through different media platforms remain a key information-gathering behavior for institutions and decision makers (Pang & Lee, 2008). Since different facets of public opinion can exist in the form of positive or negative remarks or various facial expressions, the study of sentiments expands the field of research on the subjectivity of human behavior and human language. Classifying individuals' textual manifestations in order to recognize their sentiments borrows key concepts from the field of social psychology that associates human emotions with one of the six hierarchy classes of primary emotions, love, joy, anger, sadness, fear, and surprise (Shivhare & Khethawat, 2012). It requires a systematic interpretation of the textual expressions by individuals.

Identifying and interpreting individual's sentiments across any forums is a laborious effort if done manually rather than using automated systems and structured algorithms. Data sources for sentiment include product review websites, customer feedback forums, social media, or media surveys, such as the Wall Street Journal corpus, the Document Understanding Conference, and the Multi-Perspective Question Answering corpus (Prabowo & Thelwall, 2009). The main approaches for sentiment classification include NLP and machine learning algorithms. NLP techniques in sentiment analysis include the use of *Unigrams*, *N-grams*, *Lemma*, *Negation*, and *Opinion Words* (Rambocas & Gama, 2013). Unigrams is associated with the frequency of each word. N-grams is associated with words in sequences, such as words in pairs or triplets. Lemma is associated synonyms to classify similar words into the same category, such as words *better*, *good*, and *best*, fall within the same positive sentiment category. Negation is associated with sentences that fall into two opposing categories, such as *I like this company* and *I don't like this company*. Opinion words look at the verbs, adverbs, adjectives, and nouns in the text in order to describe people feels and views. Researchers would then use the corpus of data using one of the NLP techniques into the sentiment classification systems.

Machine learning for sentiment classification. Two machine-learning approaches support the sentiment classification of content, the *supervised learning* and the *unsupervised learning* approach. In supervised learning, the model *learns* the classification criteria that previously resulted in an expected outcome using a training data set. After a supervised learning cycle is complete, the researchers execute a classification function, such as naïve bayes, support vector machine, or maximum-entropy over the sample data in order to categorize each sentiment into its respective classification group (Rambocas & Gama, 2013). Supervised machine learning models are popular in sentiment analysis, but they require training data that is not always possible because of the considerable amount of time and resources necessary. Alternatively, unsupervised learning algorithms, such as deep learning, *bag of word* method, the use lexicons, and the Web Search algorithms, do not rely on previously trained data sets and can detect sentiment within text but at lesser efficiency (Rambocas & Gama, 2013).

Recent literature shows that the three machine-learning techniques, naïve bayes, support vector machines, and maximum-entropy, perform very well in the sentiment classification process and can sometimes outperform human classifiers.

Naïve bayes method for sentiment classification. The naïve bayes approach is a very simple probabilistic model that works well in textual classifications and requires lesser time and data to train compared to alternative machine learning models. The model uses Bayes rule, whereby if given two positive and negative classes, the words that belong to either one of the two categories are conditionally independent to one another. According to Narayanan et al. (2013), if given a word *x* in document *d* and a classifier class *c* (positive or negative), the probability of word *x* belonging to class c is as follows:

$$P(x_i / c) = \frac{countofx_i \in documentsofclassc}{totalnumberofwordsx \in documentsofclassc} \qquad (13)$$

Furthermore, the probability of a document belonging to class c, where c consists of words conditionally independent from each other is:

$$P(c_i / d) = \frac{\left(\prod P(x_i / c_j)\right) * P(c_j)}{P(d)} \qquad (14)$$

The model then returns the classifier class, c_i, with the maximum posterior probability. An alternative to the naïve bayes method is the support vector machine.

Support vector machine for sentiment classification. The support vector machine is another machine technique that some researchers argue that it is more efficient than the naïve bayes method, but it requires extensive training time. Given a category set C = {+1, -1} that corresponds to positive and negative classification classes, and given two preclassified training sets, such that: $T_r^+ = \sum_{i=1}^{n}(d_i, +1)$ is the positive sample set, and $T_r^- = \sum_{i=1}^{n}(d_i, -1)$ is the negative sample set: the support vector machine finds a hyperplane that separates the two sets at a maximum distance (Prabowo & Thelwall, 2009). During the training phase, each training sample converts into a real vector x_i that includes the significant words representing the document d_i. For the positive sample sets, the positive training set T_r^+ becomes $T_r^+ = \sum_{i=1}^{n}(d_i, +1)$, and the negative training set T_r^- becomes $T_r^- = \sum_{i=1}^{n}(d_i, -1)$. Figure 8 illustrates the support vector machine for sentiment analysis.

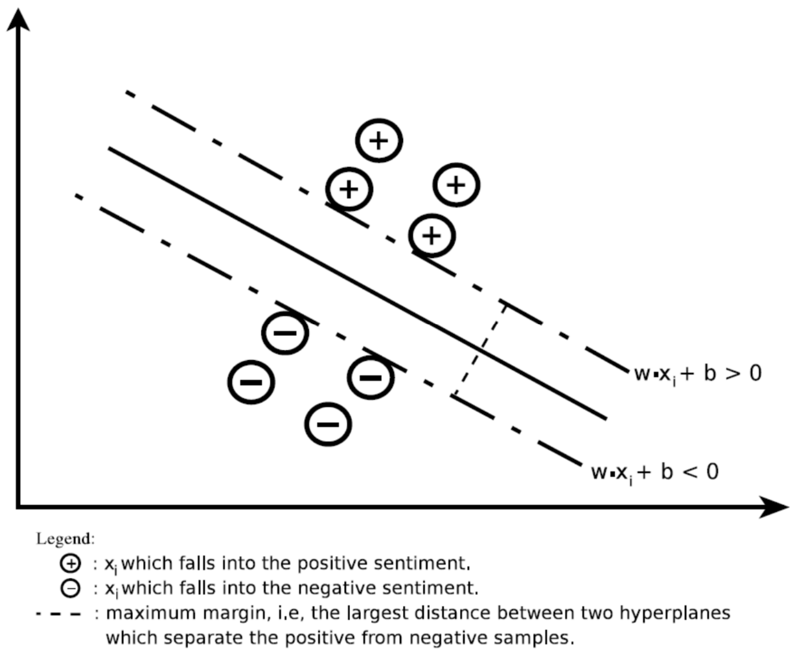


Figure 8. Support vector machine for sentiment classification. Retrieved from "Sentiment analysis: A combined approach" by Prabowo, R., & Thelwall, M. (2009). *Journal of Informetrics*, 3(2), 143–157. Reproduced with permission.

The sentiment classification approach using the support vector machine is to determine the side of the hyperplane a document associated with a sentiment falls into (Prabowo & Thelwall, 2009). Documents that include positive sentiments would fall on the positive side of the hyperplane; documents that include negative sentiments would fall on the negative side of the hyperplane. The process excludes documents that include both positive and negative features in order to avoid any skewness in the classification decisions. However, such approach would fail at extremely noisy data when the likelihood of a document belonging to either positive or negative sentiments are the same. An alternative to the support vector machine is the maximum entropy model.

Maximum entropy model for sentiment classification. The maximum entropy model is another machine learning classification technique that has proven to be effective in sentiment analysis. It follows the logistic regression approach, and its steps consist of the following. First, identify the set of features, such as words, that correspond to a particular category, then measure the expected value of the each feature over the training data and treat the results as the constraints for the model distribution (Lee & Renganathan, 2011). The goal of such classifier is to derive a model of maximum entropy in which all the constraints identified from the training data are satisfied. Paltoglou et al. (2010) wrote

> The idea behind this goal is that models with less entropy have added information beyond that in the training set, which are not justified by the empirical evidence. Thus, a maximum entropy model aims to preserve as much uncertainty as possible with the condition that the constraints of the problem (i.e. the training data set) are satisfied. (p. 15)

The key constraints to the model are that the feature used to identify the constraints should have uniformly distributed, and the events in the data are independent (Batista & Ribeiro, Lee & Renganathan, 2011). Once the maximum entropy model generates the rules that correspond to the process information derived from the inferred data set, the model would then predict the conditional probability of events given its corresponding features (Batista & Ribeiro, 2013). Along with the naïve bayes model and the support vector machine, shown previously, the maximum entropy model is an effective algorithm that can help classify investor sentiment.

Summary and Conclusions

The fields of research on corporate financial distress and investor sentiment have been relatively disparate in the literature. It reflects the distinctive differences between traditional and the behavioral finance theories in which the former assumes efficient markets while the latter assumes the irrationality in human behavior during the financial decision-making processes. The notable statistical models in the study of corporate bankruptcy including Altman (2013), Campbell (2011), and Zurada et al. (2011) solely focus on the financial data within corporate statements in order to determine the likelihood of financial distress. Such models do not explicitly incorporate investor sentiment as a factor that can influence the expected bankruptcy of firms, since it assumes that the market valuation of investment assets such as investments in corporate firms already encapsulate market information through sale demands and stock prices. However, behavioral finance theorists demonstrate that the irrationality of investors, such as emotional herding behavior and cognitive bias, would influence their actual valuation of financial assets (Byrne & Brooks, 2008). The disparity between the traditional finance and the behavioral finance theories is evident among the research in the field of finance.

Both, the traditional and the behavioral finance theories, have a strong opposing presence in the finance literature. The 2007 to2009 global crisis, where which many of the traditional risk management models have failed to either predict or prevent, drove the call for finance models that bridges the gap between the traditional finance and the behavioral finance models (Baker & Wurgler, 2011). Evidence of such consolidation is notable in the finance literature from the last five years that is also witnessing the growing popularity of Internet-based systems and social media websites as the primary communication platforms. The ability for researchers to measure and correlate individuals' behavior across social media platforms, such as Facebook and Twitter, lead to increased volume of research to predict financial markets using such platforms. Recent studies, such as Bollen et al. (2011), Chung (2011), as well as Asur and Huberman (2010) demonstrated the effectiveness of investor sentiment in predicting events within financial markets. However, the relationship between the investor sentiment using social media and the expected bankruptcy of firms is not explicit in the existing literature for measuring the financial distress of companies.

The current chapter included some of the leading techniques for predicting corporate bankruptcies and classifying sentiments in the form of textual expressions. However, a consolidated framework that includes the prediction of corporate bankruptcies with the aid of investor sentiment analysis is not evident in the current literature. By incorporating statistical methods from both fields in corporate bankruptcy and investor sentiment and with the help of common machine learning technologies that support both areas, the current study may help expand the field of research on corporate bankruptcy. The next chapter documents the design and the methodology proposed for the current study.

Chapter 3: Research Method

The purpose of the study was to analyze the relationship between investor sentiment in social media and the level of financial distress of firms. The current chapter provides a detailed description of the research and the applied methods in the study. The first segment of the chapter includes the research setting, the design, the rationale in selecting a particular design, and my role as a researcher. It includes an in-depth account of the methodology in the study that can be replicable by other researchers. The segment includes the sampling approach, the instrumentation used, the experiment, and the data analysis in the study. Threats to validity and issues of trustworthiness, including identifying all ethical procedures, will follow prior to concluding the chapter.

Research Design and Rationale

In order to support my hypothesis in the study that there is a relationship between investor sentiment in social media and the level of corporate distress in firms, I designed the research according to the following variables. The first independent variable is the level of sentiment (very positive, positive, neutral, negative, very negative) towards the publicly held companies in the United States and classified using public responses in the social media website Twitter. The second independent variable is the level of financial distress in corporate firms using Altman’s (2013) *Z*-Score scale. An Altman *Z*-Score less than 1.8 indicates that a company is under distress and has a high likelihood of bankruptcy in the next 2 years, a value greater than 2.99 indicates a safer zone, and a value between 1.8 and 2.99 indicates that the financial distress of the company cannot be determined. I analyzed the relationship between investor sentiment toward the firms sampled in the study and the level of financial distress in such companies.

I used a quantitative approach to support the research questions in the study. The research questions are as follows: (a) What is the relationship between the financial distress of firms and investor sentiment towards the firms in social media? (b) How does sentiment in social media affect the risk of bankruptcy for financially distressed firms? (c) What is the relationship between a stock's price movement on a given day and the level of investor sentiment in social media? I did not use a qualitative or mixed method study since my focus was to determine if the data collected in the study could support my hypothesis that there is statistically significant relationship between investor sentiment in social media and the financial distress of firms. In this study, I first identified all public companies that are trading on the three stock exchanges in the United States, NYSE, Amex, and NASDAQ, as of December 2014. For each of the companies, I performed a random selection of public postings in the social media website Twitter that references the companies' stock symbols in the content during the period between December 2014 and January 2015. Concurrently, I extracted the third for each of the companies sampled in the study. I then used the data to perform corporate distress and investor sentiment analysis for each of the sampled companies.

I applied Altman's (2013) *Z*-Score to classify the sampled companies as either under corporate distress or in a safe zone, and subsequently applied Manning et al.'s (2014) Stanford Core NLP natural language processing toolkit to determine the level of sentiment as very negative, negative, neutral, positive, or very positive for the messages collected from the social media website Twitter. I then conducted the data analysis to determine the relationship between the sentiments and the level of financial distress of firms. A detailed implementation of the study and the data analysis is included in the subsequent segments of the paper. It is important to note that I chose such a research design after evaluating the time and resource of constraints of implementing alternative techniques in establishing the relationship between investor sentiment in social media and the level of financial distress in corporate firms.

I could have taken different approaches to determine the relationship between investor sentiment in social media and the level of financial distress in corporate firms, but the approach used in the study is original and inexpensive when compared to other methods. It is extensible for commercial usage as well as future academic research. For example, instead of extracting real-time messages using from Twitter using my own code with a little to no cost, I could have accessed the Twitter archives using a third-party vendor at a cost of approximately $5,000 for 1 million tweets. Furthermore, instead of accessing the latest financial data for the sampled companies using my own code with a little to no cost, I could have purchased the financial archives from a third-party vendor at a cost that ranges between $17,325 and $39,750 annually. The approach used in the study is also extensible to support different measurements.

The model in the study is modifiable to use a logit or machine learning analysis of corporate distress instead of the Altman *Z*-Score. It can be modified to parse texts and classify the sentiments from other social media websites, such as Facebook and LinkedIn. Furthermore, I trained the sentiment classification model using Stanford Core NLP to classify financial lingo, such as a 48-week loss, as negative sentiment. None of the resources that I found on the Internet provided a similar solution or at least publicized it for academic access. Hence, this study is unique in its application and is extensible for future research. The next segment provides an in-depth description of the approach used in the study.

Methodology

Population

The targeted population of this study was the publicly held companies that are active in the three stock exchanges, Amex, NASDAQ, and NYSE in the United States. As of January 2014, 5,025 companies were trading publicly on the United States stock exchanges with a total market capitalization of 23 trillion dollars, which is 40% of the global stock market capitalization (World Federation of Exchanges, 2015). The website NASDAQ.com includes a company list that remains up to date whenever companies merge, new companies join, or some companies are withdrawn from active stock market trading.

In this study, I attempted to analyze the entire population of publicly held firms in the United States, excluding companies in the financial sector.

Sampling and Sampling Procedures

Since the research method was quantitative, the research design includes identifying the population sample and measuring the data in order to conclude if there is statistical evidence that supports the hypotheses. In this study, the estimated population size was 5,000 firms that were actively trading on any one of the three U.S.-based stock exchanges, NYSE, NASDAQ, and Amex, at the time of the study and were not part of the financial services sector. The sampling technique was a stratified nonprobability sampling, and the sample size included all 5,787 firms. I then performed a random sampling of tweets that mentioned the stock symbol of each of the firms in the population sample. It resulted in a data collection of 66,038 tweets associated with the 5,787 publicly held firms in the United States between December 7, 2014 and January 6, 2015. I excluded the companies that were associated with the finance sector because the Altman's (2013) *Z*-Score index for measuring corporate distress is not accurate for the financial sector. To assist in the data analysis process, I created the database table, shown in Table 4, to store the data samples that I would later use in the data collection process.

Table 4

Database Structure for Public Companies' Data

Database field name	Database field type (Variable type & size)	Database field description
Id	Unique integer	Unique identifier
Symbol	Varchar(10)	Stock symbol (eg., $VZ)
Name	Varchar(100)	Company name
Last Sale	Varchar(20)	Last stock price
Market Cap	Varchar(20)	Market capitalization
AdrTSO	Varchar(20)	American depository receipt
IPO Year	Varchar(20)	Year went public
Sector	Varchar(100)	Company sector
Industry	Varchar(100)	Company industry
Summary quote	Varchar(200)	URL for summary quote
Exchange	Varchar(10)	Stock market location (“NYSE,” “NASDAQ,” “Amex”)

Archival Data

For each of the firms sampled in the study, I collected tweets associated with the sampled companies from the social media website Twitter and subsequently extracted the last quarterly financial data for such firms using the site Yahoo.com. Access to the tweets is permissible through Twitter's application programming interface (API) under Twitter's (2012) terms of service. The access to the financial data for the publicly held firms sampled in the study is permissible through Yahoo! APIs under Yahoo's (2012) terms of service. Permission information is available in Appendix J. The procedure for extracting the tweets and the financial data is as follows.

Tweet extracts. The steps that I followed in obtaining the tweets for each of the public companies sampled in the study are as follows. I first established Twitter-based application permission at https://apps.twitter.com with read-only permission. The site supplied me with a consumer key and access token that granted my access to the Twitter stream of data using its API. I later created a database table using the database structure described in Table 4 that I would use to store the sampled companies' tweets. I then developed a computer script using the programming language Python monitored the Twitter data streams, identified tweets that included the trading symbols of the sampled companies in Table 4 and stored the results in Table 5. The general schema of the script is as follows, and the complete code is available in Appendix B.

1. Picked 200 random stock symbols associated with the sampled companies in Table 4.

2. Listened to Twitter stream for any tweets that included the stock symbols in (1).
3. Stored tweets into the database, up to 500 tweets in Table 5,
4. Repeated the process.

Note that the choice of limited 200 stock symbols for each iteration is due to the restrictions imposed by Twitter on the number of query parameters applicable for the data retrieval procedure.

Table 5

Database Structure for Public Companies' Tweets

Database field name	Database field type variable type (Size)	Database field description
Id	Unique identifier	Unique identifier
Twitter_User_Id	Varchar(255)	Twitter user identification
Twitter_User_Name	Varchar(100)	Twitter user name
Twitter_Text	Varchar(1024)	Twitter text
Twitter_Text_Id	Varchar(255)	Twitter text identification
Twitter_Text_Keyword	Varchar(1000)	Keywords found that match companies' stock symbols
Twitter_Text_Timestamp	Datetime	Time stamp of the tweet
Twitter_For_Training	Integer	Reserved for machine learning classification purposes
Twitter_Sentiment	Integer	Reserved for machine learning classification purposes
Training_User_Id	Integer	Reserved for machine learning classification purposes

Sampled companies' financials extract. The steps that I followed in extracting the financial data for each of the public companies sampled in the study are as follows. I created a database table with the schema included in Table 6. I then wrote a Python script that iterates through all the public companies listed in Table 4, accesses the financial data for each company using Yahoo (2015) APIs, Yahoo query language, and Yahoo (2013) Open Data Tables, and stores the results in the database under Table 6. The general schema of the script is as follows, and the full source code is included in Appendix E.

1. Picked a stock symbol associated with a company in Table 4.
2. Selected total assets, total liabilities, total current assets, total current liabilities, retained earnings from the balance sheet table.
3. Selected ebitda and sales from the income statement table.
4. Selected last trade price and market capitalization from quotes table.
5. Stored the results in Table 6.
6. Repeated the process.

After I completed the data collection process of the relevant firms' tweets and financial data, I administered the operationalization of the constructs in the study–the level of sentiments and the level of financial distress in the sampled firms.

Table 6

Database Structure for Public Companies' Financials

Database field name	Database field type	Database field description
Id	Unique identifier	Unique identifier
Company_id	Unique identifier	Unique identifier to company in Table 5
Quarter	Varchar(1)	Financial quarter: 1—first quarter, 2—second quarter, 3—third quarter, 4—fourth quarter
Year	Varchar(4)	Financial year, eg. 2014
Total_Assets	Varchar(20)	Total Assets
Total_Liability	Varchar(20)	Total Liability
Current_Assets	Varchar(20)	Current Assets
Current_Liability	Varchar(20)	Current Liability
Retained_Earnings	Varchar(20)	Retained Earnings
Market_Capital	Varchar(20)	Market Capital
Ebitda	Varchar(20)	Earnings before interest, tax, depreciation, amortization
Sales	Varchar(20)	Total sales
Stockprice	Varchar(20)	Current stock price
Date_extracted	Datetime	Timestamp of data extract

Instrumentation and Operationalization of Constructs

The different levels of firms' financial distress and the investor sentiment on social media toward such firms make the two major variables in this study. The constructs in this study for measuring corporate financial distress and investor sentiment are Altman's (2013) Z-Score and Manning et al.'s (2014) Stanford Core NLP natural language processing toolkit, respectively. A detailed description of each of the two instruments and its application in the study is below.

Altman *Z*-Score. Altman's (2013) *Z*-Score is a multivariate discriminant analysis equation for measuring corporate financial distress. First published in 1968, the Altman *Z*-Score is effectively used in recent literature to measure financial distress of firms. An example of such researchers included Hayes et al. (2010), Apergis et al. (2011), and Altman et al. (2014). The instrument, discussed in detail in Chapter 2 of this study, is as follows:

Z-Score = 0.012*WC + 0.014*RER+ 0.033*EBIT + 0.006*MKLI +0.999*SALES where WC = working capital / total assets, RER = retained earnings / total assets, EBIT = earnings before interest and taxes / total assets, MKLI = market value equity / book value of total liabilities, SALES = sales / total assets, and Z-Score = overall financial distress index. A Z-Score that is greater than 2.99 indicates a *safe zone* or the firm is at a lower risk of bankruptcy. A Z-score that is between 1.8 and 2.99 indicates a *gray zone* or the firm is in an undermined risk to bankruptcy. A Z-score that is less than 1.81 indicates a *distress zone* or the firm is at a higher likelihood of bankruptcy in the next 2 years. In this study, I leveraged the financial data collected for each of the firms to develop Altman *Z*-Score for each of the firms.

I used Altman's (2013) Z-Score model for corporate financial distress analysis. Altman granted me permission to use the instrument in this study. The approval letter is available in Appendix J. The tool was appropriate for this study because of the simplicity in determining the level of corporate distress in firms using a set of financial ratios and variables that are accessible for publicly held companies. The steps that I took to operationalize Altman *Z*-score for this study are as follows. I first created a database table, as shown in Table 7, to store Altman *Z*-Score values for each of the sampled firms in the study. Table 7 includes the database fields that are associated with such variables. I then wrote a Python script that iterates over each of the public firms and the financial data in Tables 4 and 6 respectively, then calculates and stores Altman *Z*-Score for each firm in Table 7. The general schema of the script is below, and the complete source code is available in Appendix G.

1. Get each company stored in Table 4.
2. Get the financial variables for each company in (2) from Table 6. The variables are current assets, current liability, total assets, total liability, retained earnings, ebitda, market capital, stock price, and sales.
3. Let Altman_Z-Score = 1.2*X1+1.4*X2+3.3*X3+0.6*X4+0.99*X5,

where X1=(current assets – current liability)/total assets, X2=retained earnings/total assets, X3=ebitda/total assets, X4=market_capital/total assets, and X5=sales/total assets.

4. Store Altman *Z*-Score in Table 7.

Table 7

Database Structure for Public Companies' Altman Z*-Score*

Database field name	Database field type	Database field description
Id	Unique identifier	Unique identifier
Company_Id	Unique identifier	Unique identifier to company in Table 5
Company_Financials_ID	Unique identifier	Unique identifier to company in Table 6
Z-Score	Varchar(10)	Altman *Z*-Score
date_updated	Timestamp	Last updated

Stanford core NLP natural language toolkit. I used Manning et al.'s (2014) Stanford Core NLP natural language processing toolkit to identify the sentiments from the collected tweets that are associated with the sampled public companies. Stanford Core NLP sentiment tool is based on *Stanford Sentiment Treebank* corpus of 215,154 unique phrases, each annotated by three human judges, and *Recursive Neural Tensor Network* machine-learning algorithm to classify sentences as either positive or negative (Socher et al., 2013). The model uses semantic vectors to transform sentences of any length into fully labeled parse trees in order to identify the different sentiments in word compositions of any complexity. Figure 9 demonstrates an example of the classifier in which each node in the sentence "this film does not care about cleverness, wit, or any other kind of intelligent humor" falls under one of the five sentiment classes, from very negative to very positive (--,-,0,+,++). According to Socher et al. (2013), the recursive neural tensor network provides better accuracy than alternative machine learning models, including standard recursive neural networks (RNN), matrix-vector RNNs, Naïve Bayes (NB), bi-gram NB, and support vector machine.

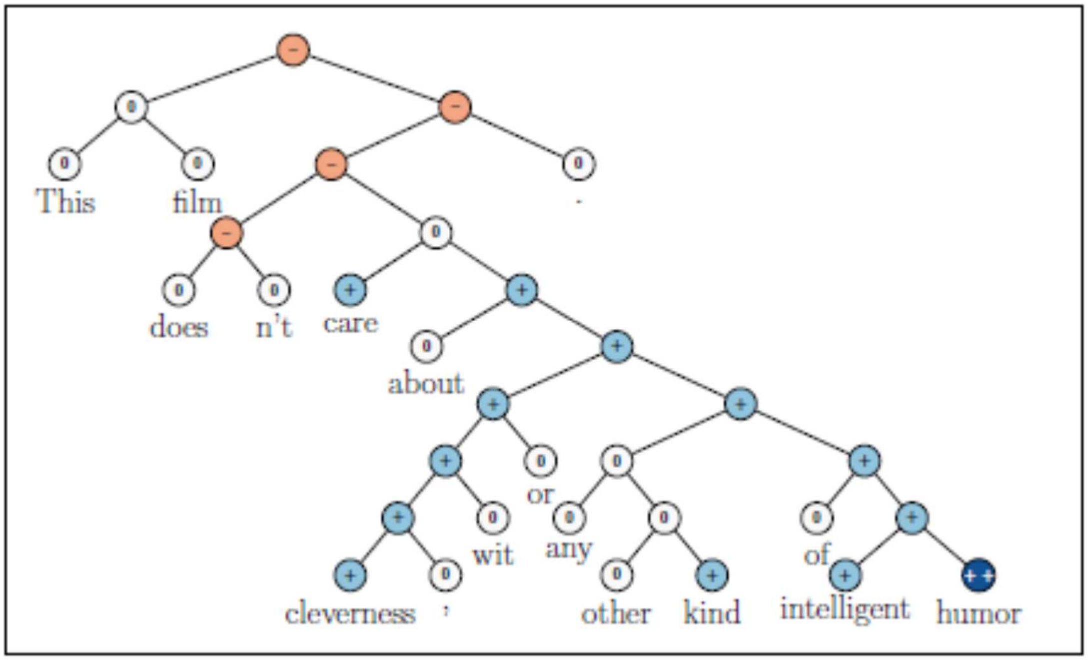


Figure 9. Neural tensor network example. Retrieved from "Recursive deep models for semantic compositionality over a sentiment treebank. " in In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) by Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). (pp. 1631–1642). Reproduced with permission

The Stanford Core NLP instrument has been frequently applied in multiple research studies in social media and sentiment analysis, such as Kucuktunc, Cambazoglu, Weber, and Ferhatosmanoglu (2012) and Go et al. (2009). The instrument is available under the GNU General Public License (2014) and does not require permission to use, since the instrument will be used for academic purposes and not within any proprietary software (Stanford NLP Group, 2014). The toolkit also includes the source code that is modifiable in order to become applicable for the study. The steps that I followed were as follows. I first created a database table, shown in Table 8, to store the sentiments for each of the sampled companies. Then I wrote a Python script that retrieves all the tweets collected for the sampled firms, execute Stanford Core NLP sentiment analysis for each tweet, and then stored the results in Table 8. The overall schema of the script is below, and the complete source code is available in the appendix. The script's execution steps are as follows:

1. Retrieve each company from Table 4.
2. Extract every tweet for each company in (1) except the tweets that were previously manually classified from Table 5.
3. Execute the retrained Stanford Core NLP sentiment classification algorithm for each tweet in (2).
4. Store the sentiment results into Table 8.

Table 8

Database Structure for Sentiment Analysis

Database field name	Database field type	Database field description
Id	Unique identifier	Unique identifier
Company_Id	Unique identifier	Unique identifier of each company in Table 4
Twitter_Text_Id	Unique identifier	Unique identifier for each tweet in Table 5
sentiment_prob_very_negative	varchar(10)	Probability of very negative sentiment
sentiment_prob_negative	varchar(10)	Probability of negative sentiment
sentiment_prob_neutral	varchar(10)	Probability of neutral sentiment
sentiment_prob_positive	varchar(10)	Probability of positive sentiment
sentiment_prob_very_positive	varchar(10)	Probability of very positive sentiment
sentiment_root_value	varchar(10)	Overall sentiment (0─very negative, 1─negative, 2─neutral, 3─positive, 4─very positive)
date_updated	timestamp	Date sentiment analysis

Initially, Stanford Core NLP sentiment analysis model did not accurately classify the sentiments in the tweets included financial jargon, such as *48-week low* or *I sold all my stocks*. Retraining of the model was necessary. The steps that I took to retrain Stanford Core NLP toolkit were as follows. I first picked 10,000 random tweets that are associated with the sampled companies and stored in Table 5. I then developed a web page for training the model. A screenshot of the web page is in Figure 10. With the help of family, I manually classified the 10,000 tweets as either very negative, negative, neutral, positive, or very positive sentiments. The results were first stored in the database under Table 6 before I imported them into the toolkit.

Retraining of Stanford Core NLP toolkit required that the input data be in Penn Treebank II annotation format, as documented in Bies et al. (1995). For example, the text "4 I made lots of profit with stock $vz" should be translated into Penn Treebank annotation format "(4 (4 I (4 made (4 lots (4 of (4 profit (4 with (4stock (4 $vz))))))))" before it is incorporated in the training process of the machine learning algorithm. I wrote a script using the Java programming language that would leverage the Stanford Core NLP class *edu.stanford.nlp.sentiment.BuildBinarizedDataset* to convert the manually trained sentiments and the associated tweets into Penn Treebank before executed the retraining process of Stanford Core NLP algorithm with the financial-related data.

After transforming all 10,000 manually trained tweets into Penn Treebank format in (e), I split the output into two files *train.txt* and *dev.txt* and executed the retraining module of Stanford Core NLP toolkit using the command *"java -mx8g edu.stanford.nlp.sentiment.SentimentTraining -numHid 25 -trainPath train.txt -devPath dev.txt -train -model model.ser.gz*". The output generated from the retraining of Stanford Core NLP process became the sentiment analysis model that can accurately classify tweets that include financial jargon. After completing the necessary steps to train Stanford Core NLP to classify finance-related terms as negative, positive, or neutral sentiments, the next step was to allow the algorithm to classify the sentiments automatically for all the tweets in the database.

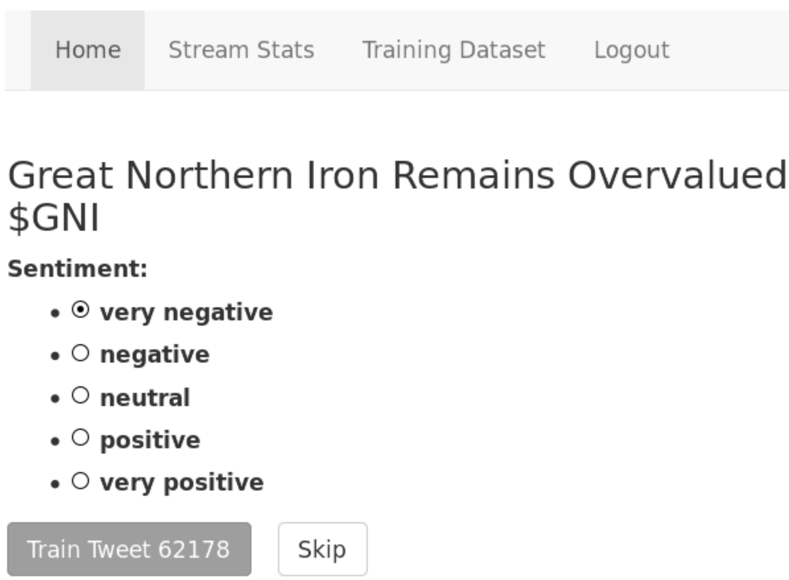


Figure 10. Screen shot of manually training tweets in the study.

After I had completed both the data collection process and the instrumentalization of the research constructs, I performed a data analysis over the data.

Data Analysis Plan

I used SciPy, the scientific computing tools for the programming language Python, for the data analysis in the study. SciPy consists of freely accessible scientific models for academic researchers in the field of mathematics, science, and engineering. It is an academic alternative to the commercial-based data analysis tools, such as Matlab (SciPy Developers, 2014). To prepare the research data for the appropriate data analysis, I first imported the research data into Oracle Corporation (2014) MySQL. I then screened the data and cleaned any irrelevant data in the database using PhpMyAdmin, an open source software that allows the administration of MySQL servers and databases using a web-based interface (PhpMyAdmin Contributors, 2014). After I extracted the sampling data and applied the sentiment analysis as well as financial distress analysis, I performed the data analysis to support each of the hypotheses proposed in the study.

To support the hypotheses in the study, I first determined the normality from the data set and applied the appropriate statistical tool using SciPy. Each of the hypotheses in the study and accompanied by the appropriate statistical tests are as follows.

Research Question 1: What is the relationship between the financial distress of firms and the investor sentiment towards the firms in social media?

$H1_0$: There is no relationship between the level of financial distress of firms and the investor sentiment towards such firms in social media.

$H1_a$: There is a positive relationship between the level of financial distress of firms and the investor sentiment towards such firms in social media. Firms with higher levels of financial distress positively correlate with higher negative investor sentiment in social media. Firms with lower levels of financial distress positively correlate with higher positive investor sentiment in social media.

Research Question 2: How does sentiment in social media affect the risk of bankruptcy for financially distressed firms?

$H2_0$: There is no statistical difference between the presence and absence of investor sentiment in social media in affecting the likelihood of bankruptcy by financially distressed firms.

$H2_a$: There is a statistical difference between the presence and absence of investor sentiment in social media in affecting the likelihood of bankruptcy by financially distressed firms. Financially distressed firms with investor sentiment in social media are more for bankruptcy at higher rates than if investor sentiment in social media is not present.

Research Question 3: What is the relationship between a firm's stock movement and the level of sentiment towards the firm in social media?

$H3_0$: There is no relationship between the firms' stock movement and the level of sentiment in social media.

$H3_a$: There is a positive relationship between the firms' stock movement and the level of sentiment in social media. Negative sentiment positively correlates with a decline in the stock value, and positive sentiment positively correlates with an increase in stock value.

The steps that I took to support the each of the hypotheses are as follows. I first determined that there is no normality in the data set. To support the research question 1, I applied Spearman (1904) rank correlation for nonparametric data to determine the relationship between investor sentiment and the level of financial distress of firms. The steps that I took were as follows:

- Set FINDISTESS = Altman *Z*-Score financial distress index, where FINDISTRESS<1.8 indicates financial distress, 1.8<FINDISTRESS<2.99 indicates that financial distress measurement is inconclusive, and FINDISTRSS>2.99 indicates no financial distress.
- Set SENTIMENT = (Positive, Negative, Neutral), where SENTIMENT = Positive to indicate an overall positive investor sentiment in social media, SENTIMENT = Negative to indicate an overall negative investor sentiment in social media, and SENTIMENT = Neutral to indicate that the presence of investor sentiment in social media cannot be determined.
- Applied Spearman correlation using Scipy (2009) where Y is FINDISTRESS and X is SENTIMENT at a 95% confidence level and determined if I can reject the null hypothesis when the P-value is <0.05.

The result of the correlation analyzed helped determine if I can reject the null hypothesis $H1_0$.

In order to determine if sentiment in social media can affect the risk of bankruptcy for financially distressed firms under the research question 2, I applied the similar steps for Research Question 1 using the subset of data with Altman *Z*-Score < 1.8. I also used Spearman rank correlation to determine if I can reject the null hypothesis $H2_0$. To support the research question 3, I first derived the average sentiments per day and the opening and close of the stock price for each firm sampled in the study and then applied Spearman rank correlation in order to determine if I can reject the null hypothesis $H3_0$. Chapters 4 and 5 include the results of the study. The threats to validity and the trustworthiness of the study are next.

Threats to Validity

External Validity

Threats to external validity are the conditions that could make it wrong to generalize the results of the research for the entire population. External conditions to the study, including the timing of firms' disclosing financial outcomes and the public reaction towards corporate firms between corporate financial releases could influence the level of sentiment from the public. The timing of extracting sentiment information for one period and then reporting the results in another time could lead to the problem of endogeneity (Akhtar et al., 2012). In order to control for such conditions, I followed the approach by Akhtar et al. by focusing on short intervals between firms disclosing financial information and the tweets relevant to such firms. I used the third quarter 2014 financial data, daily tweets, and stocks' daily price movement for each of the sampled firms that I believe have made the study externally valid.

Internal Validity

Issues of internal validity could exist in the form of the statistical measurements used in the research. The study is limited to the data collected and the timing of the data collection. The possibility of erroneous scripts that could extract incorrect data could be a threat to the internal validity of the study. To control for such issues, I continuously monitored the data collection process, confirmed that the output of the scripts that I developed and executed for the study are not returning incorrect data. On the other hand, other possible conditions of internal validity issues, including selection bias, timing of the study, researcher bias, and changes in subject conditions during the study maturation period are limited in the study. However, I maintained objectivity throughout the study in order to control for the threats to the internal validity of the study. I also followed the same approach to the operationalization of the research constructs. Hence, I consider the study as internally valid.

Construct Validity

Threats to the validity of the corporate distress and investor sentiment constructs exist in the study. Manning et al.'s (2014) first trained Stanford Core NLP model for measuring using a corpus of movie reviews. Like other supervised machine learning logarithms, the accuracy of the model is in the training procedure. Inaccurate or false retraining of the model may lead to incorrect classification of sentiments. To ensure validity of the investor sentiment construct, and similar to the original study by Manning et al. (2014), I enlisted other individuals besides myself to classify the sentiments of the training subset of the data in order to improve the accuracy of the supervised machine learning algorithm. On the other hand, the threat to the validity of the existing Altman's (2013) Z-Score construct to measure the level of corporate distress in firms is discussed in Chapter 2. Altman *Z*-Score for measuring the level of corporate distress was initially published in 1968 and is extensively referenced in academic research and applied within corporate markets to this day (Altman, 2013; Altman et al., 2014). Since it attempts to predict to possibility for a firm to declare bankruptcy in the next two years, its validity is dependent on both the accuracy of previous corporate financial data used to original develop the construct and on the current corporate financial information that is used to apply the construct. However, in this study, the threats to the validity of both constructs can be mitigated by qualitative analysis of the data and the results of the study.

Ethical Procedures

I conducted the study in full compliance with the ethical procedures as required by the Institutional Review Board (IRB). I complied with all the terms of service provided by the data sources that I used for the data collection using NASDAQ (2015), Yahoo (2015), and Twitter (2012). I also received permissions to reproduce the figures in the study, and I have enclosed the approval letters in the appendix. I have permission from the Free Foundation Software (2014) GNU General Public License to use Stanford NLP Group's (2014) Stanford Core NLP toolkit for the sentiment analysis. I also have permission by Altman to use Altman's (1968, 2013) Z-Score for the corporate distress analysis. All financial data collected on the firms sampled in the study are publicly available data and are not proprietary. Furthermore, all the tweets collected for the sampled firms are publicly accessible data and are not restricted under Twitter (2012) terms of service. In addition, all the data gathered for the purpose of this study are stored in an encrypted storage system. At the time of the dissertation proposal, the IRB reviewed and approved the research proposal before I proceeded with the study. The IRB approval number for this study is 08-29-14-0070929.

Summary

Chapter 3 of the study includes the research approach that I used to determine the relationship between investor sentiment in the social media website Twitter and the level of corporate distress in US-based publicly held firms. I had first set up a computer database to store the data for data analysis purposes. I then identified the sources and the interfacing procedures that I would use for the sample and the data collection. I also developed software scripts including a web application that would assist in the data collection and the operationalization of the research constructs. After I had completed the data collection process, I conducted a data analysis on the data. Chapter 4 includes the results of the study

Chapter 4: Results

Chapter 4 includes the results of the research. To review, the purpose of the study is to examine the relationship between investor sentiment in social media and the level of corporate distress in firms. The research questions were as follows: What is the relationship between the financial distress of firms and the investor sentiment towards the firms in social media? How does sentiment in social media affect the risk of bankruptcy for financially distressed firms? What is the relationship, if any, between firms' stock movement and the level of investor sentiment towards such firms in social media? I hypothesized that there is a statistically significant relationship between the level of investor sentiment in social media and the level of corporate distress in firms. Negative sentiments are positively associated with increased levels of corporate distress in firms, whereas positive sentiments are positively associated with decreased levels of corporate distress in firms. To support the study, public tweets for the period between December 7, 2014 and January 6, 2015 from the mainstream social media website Twitter that were associated with publicly held firms in the United States were extracted and compared with third quarter 2014 financial conditions of such firms. The data collection steps, intervention procedures, and the study results are available in this chapter, and the interpretation of the findings are available in Chapter 5.

Data Collection

The 66,038 tweets that mention the stock symbols of 5,787 publicly held firms in the United States were collected during the period between December 7, 2014 and

January 6, 2015. At the time of the data collection process, NASDAQ.com included 6,674 stock symbols that were associated with 5,787 public companies being traded on one of the three stock exchanges, NASDAQ, Nyse, and Amex. Table 9 contains the count of companies that were actively trading on the stock exchange market at the time of the study. As noted in Chapter 3, the population sampling was a nonprobability sampling that included the population of all actively traded companies in the US under one of the three stock exchanges. Table 10 includes the count of all publicly held firms by their corresponding sector.

Table 9

Population of Companies Actively Trading on Nyse, NASDAQ, and Amex Stock Exchange

Stock exchange	Active companies	Stock symbols
Amex	372	414
NASDAQ	2,815	2,954
Nyse	2,600	3,306

Note. Data are based on NASDAQ (2015, January 1). Company List. Retrieved December 7, 2014, from http://www.NASDAQ.com/screening/company-list.aspx. Reproduced with permission.

Table 10

Population Sample of US Publicly-traded Companies in the study

Sector	Public companies
Basic industries	350
Capital goods	390
Consumer durables	153
Consumer nondurables	226
Consumer services	800
Energy	362
Finance	839
Health care	676
Public utilities	250
Technology	678
Transportation	110
Various	953

Note. Data are based on NASDAQ (2015, January 1). Company List. Retrieved December 7, 2014, from http://www.NASDAQ.com/screening/company-list.aspx. Reproduced with permission.

For each of the firms sampled in the study, a random set of tweets that include the stock symbol of each firm were captured at various time intervals during the period between December 7, 2014 and January 6, 2015. The data were stored in the database table, shown in Table 4, and they included a total of 66,038 tweets associated all the firms sampled in the study. It was not possible to capture every tweet associated with the firms sampled in the study due to the limitations imposed by the Twitter website as noted in Chapter 3 of the study. Table 11 includes a descriptive statistics for the number of tweets collected for each of the firms sampled in the study. Forty four tweets were collected on an average for each of the sampled firms during the data collection period. However, the number of tweets per sampled company was not consistent across the entire population sample.

Table 11

Descriptive Statistics of Tweets Collected by Sampled Company

Statistic	Value
Mean	43.8
Standard error	1.99
Median	8
Standard deviation	126.5
Minimum	1
Maximum	4,356
Range	4,355
Mode	1
Kurtosis	398.24
Skewness	15.24
Confidence level (95.0%)	3.9

The frequency of tweets collected for each of the sampled firms did not fall under a uniform distribution as evident by the positive skewness in Figure 11 and the failure of Anderson-Darling (1952) normality test in Figure 12. The nonuniformity of the frequency of tweets collected for each of the sampled firms is not surprising. The volume of tweets collected for each of the firms is dependent on the probability that each of the sampled firm is picked up for the data collection, the timing of the data collection, the presence of the firms' stock symbols in the tweets, and the popularity of the firm within the social media. After I have completed the data collection of tweets for each of the sampled company, I collected the latest quarter financial data for each of the firms in order to populate the database table shown in Table 6. At this stage of the study, I collected all the necessary data prior to conducting the data analysis on the relationship between the level of investor sentiment in social media and the level of financial distress in corporate firms. However, the analysis of investor sentiment of the sampled firms in the study required a modification of the sentiment analysis model in order to support the study.

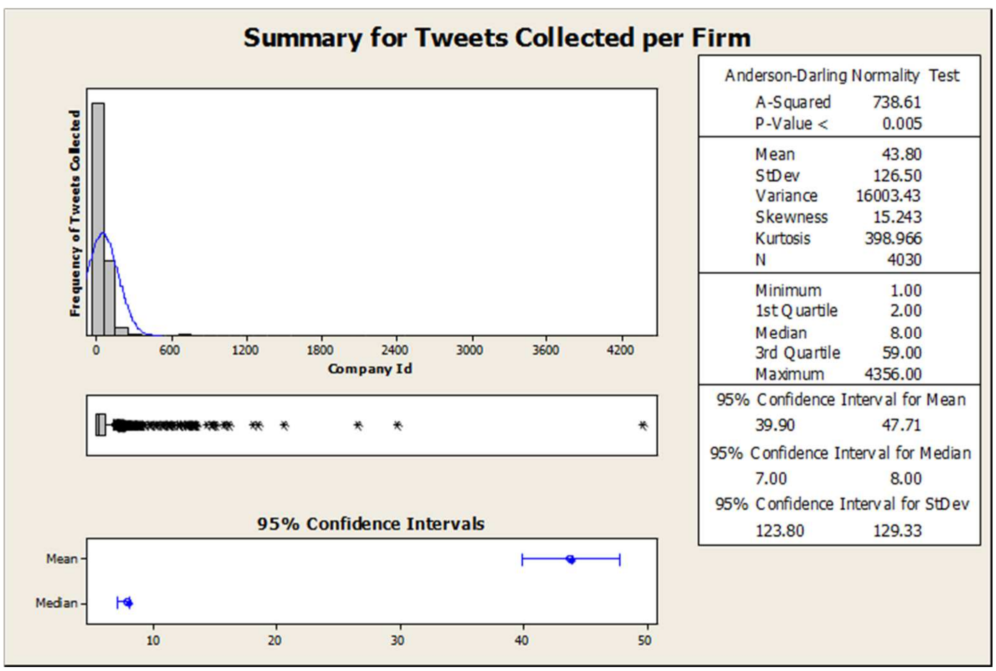


Figure 11. Frequency of tweets collected for the sampled firms.

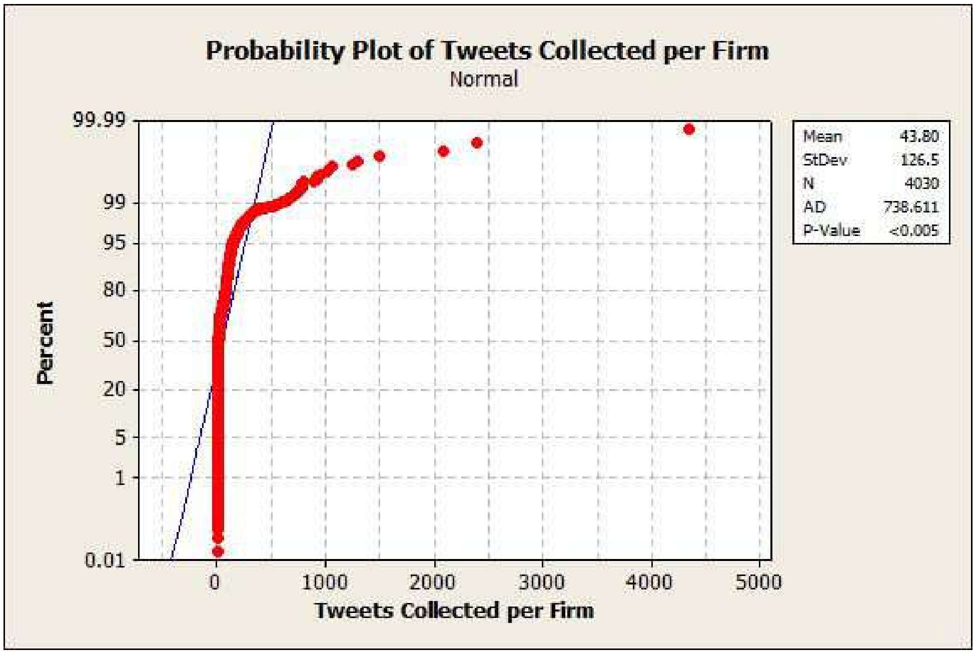


Figure 12. Normality test of tweets collected by sampled firms.

Study Results

The results of the study after completing the data collection include the sentiment analysis using the collection of tweets associated with the sampled firms, the corporate distress analysis of the sampled firms using the financial data collected, and the relationship between both constructs. The descriptive statistics associated with the research constructs is below and then followed by the statistical analysis.

Sentiment Analysis

Using the trained Stanford Core NLP sentiment classification algorithm, I applied the algorithm over all the tweets captured for each of the 5,787 firms sampled in the study. Table 12 includes the descriptive statistics for each sentiment associated with every tweet collected.

Table 12

Descriptive Statistics of Sentiment Analysis of Tweets

Sentiment	Mean (percentage)	Std Dev	Variance	Min	Max	Skewness	Kurtosis	Q1	Q3
Very negative	0.0144	0.059	0.0034	0.000	0.9576	9.61	113.17	0.000	0.003
Negative	0.1110	0.25	0.0620	0.000	0.9931	2.51	4.85	0.004	0.04
Neutral	0.6615	0.363	0.1319	0.000	0.99	-0.68	-1.22	0.283	0.971
Positive	0.1507	0.237	0.06	0.000	0.98	1.87	2.33	0.015	0.158
Very positive	0.0623	0.159	0.0255	0.000	0.96	3.57	12.69	0.001	0.03

I used the schema below to calculate the overall sentiment for each of the firms, and I have included the complete source code in Appendix I.

1. Looped within each firm sampled in the study.
2. Obtained all tweets stored in the database that include the stock symbol firm (1).
3. Obtained the sentiment probability for each tweet in (3) as very negative, negative, neutral, positive, or very positive.

4. Detected the sentiment with the maximum cumulative probability between the very negative, negative, neutral, positive, or very positive in (4).

Table 13 includes the breakdown of the sentiments as an aggregate for all the 5,787 firms sampled in the study.

Table 13

Percentage of Sentiments for 5787 Sampled Firms

Sentiment level	Cumulative count	Aggregate percentage
Very negative	1,004	0.01
Negative	18,945	0.11
Neutral	123,543	0.7
Positive	24,001	0.14
Very positive	9,033	0.05

Corporate Distress Analysis

Using the financial data collected for the third quarter of 2014 for each of the firms sampled in the study, I applied Altman *Z*-Score to determine the level of financial distress in each firm. As a reminder, an Altman *Z*-Score of less 1.8 indicates that a firm is in financial distress and is likely to file for bankruptcy in 2 years; a *Z*-Score of greater than 3.0 indicates that a firm is in a safe zone, and a *Z*-Score between 1.8 and 3.0 indicates that the financial distress of a firm cannot be determined. As noted previously, the Altman *Z*-Score does not yield accurate results for firms in the financial sector. Hence, I excluded the 839 companies that are in the financial sector, as listed in Table 10. Table 14 includes the results of the Altman *Z*-Score.

Table 14

Altman Z-Score for Sampled Firms

	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Z-Score <1.8	1,032	-1.96	9.81	-157.2	-1.09	0.66	1.25	1.25
1.8<*Z*-Score<3.0	591	2.37	0.33	1.8	2.09	2.37	2.66	2.99
Z-Score>3.0	1,690	9.93	20.51	3	4	5.4	8.88	454

Research Constructs Data Analysis

The focus of the study is to determine the relationships between the levels of financial distress and the investor sentiment in public firms in order to support the hypotheses of the study. The primary statistical assumption is to ensure that both financial data and investor sentiment exist for the firms sampled in the study. Furthermore, to ensure the reliability of the Altman *Z*-Score for measuring corporate distress, I selected the firms that are not associated with the financial sector. After I had considered the following assumptions into the data set, I identified 2,618 nonfinancial companies that I managed to construct an Altman *Z*-Score and a sentiment scale. Table 15 includes the descriptive statistics for such firms, where the sentiment indicator is a discrete variable (0 = *very negative*, 1 = *negative*, 2 = *neutral*, 3 = *positive*, and 4 = *very positive*) and Altman *Z*-Score is a continuous variable (<1.8 is financial distress, > 3 is safe, and between 1.8 and 3.0 is undetermined). To support the correlation analysis between the level of investor sentiment and the level of sentiments, I applied a normality test on both constructs.

Table 15

Descriptive Statistics of Nonfinancial Firms

	Last stock price ($)	Market cap	Sentiment	Altman *Z*-Score
Count	2,618	2,618	2,618	2,618
Mean	36.62	9.04e+09	2.16	4.64
Std Deviation	51.21	3.00e+10	0.66	12.87
Min	0.10000	2.10e+06	0.00	-153.81
25%	6.96	2.73e+08	2.00	1.37
50%	22.61	1.33e+09	2.00	3.036
75%	49.59	5.09e+09	2.00	5.49
Max	1,135.97	6.74e+11	4.00	269.37

Note. For sentiment, 0 = *very negative*, 1 = *negative*, 2 = *neutral*, 3 = *positive*, and 4 = *very positive*. For the Altman *Z*-Score, <1.8 is financial distress, > 3 is safe, and between 1.8 and 3.0 is undetermined.

A correlation analysis between the levels of investor sentiment and the levels of financial distress first required tests for normality in order to determine the type of correlation analysis to be applied. Since Altman *Z*-Score is a continuous variable, I used SciPy *normaltest* function that is based on D'Agostino (1971) and D'Agostino and

Pearson (1973) test for normality. It tests the null hypothesis that a sample of continuous is from a normal distribution (The Scipy community, 2014). After applying SciPy *normaltest* on the data that constitutes Altman *Z*-Score, the probability that the data follow a normal distribution yielded a p-value=0, which means that such data are not formed from the normal distribution. On the other hand, the test for normality for the sentiment variable is not necessary since the sentiment variable is discrete. After determining that the data associated with the continuous Altman *Z*-Score variable and the discrete sentiment variables are not normally distributed, I applied Chi-Square to determine if there is a correlation between both variables. The results of the correlation analysis helped determine if I can reject the null hypothesis $H1_0$.

Statistical Analysis for Hypothesis #1

The first hypothesis in the study is that there is a positive relationship between the level of financial distress and the level of investor sentiment in social media for corporate firms. I argued that there is a positive correlation between negative sentiments and financial distress, and there is a positive correlation between positive sentiments and no financial distress in corporate firms. To recall, Altman *Z*-Score is the continuous none normally distributed variable in this study that formed the basis of whether a corporate firm is under financial distress or not, and Stanford Core NLP sentiment variable is the discrete variable that is used to determine the levels of investor sentiment using the social media website Twitter. To allow a comparability between both variables in order to support a correlation analysis, I transformed the constructs in the study to two variables with a scale of three values each (-1, 0, 1):

1. Let variable *SENTIMENT* be the sentiment variable, such that *SENTIMENT* = -1 if Stanford CoreNLP sentiment returned 0 (very negative) or 1 (negative) for a sampled firm; *SENTIMENT* = 0 if Stanford CoreNLP sentiment returned 2 (neutral); and *SENTIMENT* = 1 if Stanford CoreNLP returned 3 (positive) or 4 (very positive).
2. Let variable *DISTRESS* be the financial distress variable, such that *DISTRESS* = -1 if Altman *Z*-Score returned <1.8 (distress zone) for a sampled firm, *DISTRESS* = 0 if Altman *Z*-Score was between 1.8 and 3.0 (unknown zone), and *DISTRESS* = 1 if Altman *Z*-Score > 3.0 (safe zone)

Table 16 includes the results of applying the transformation of the sentiment and the financial distress variables over the final sample set of 2618 firms from all sectors except the financial sector. The data helped determine if there is a relationship between the SENTIMENT and the DISTRESS variables.

Table 16

Sampled None-Financial Firms Sentiment & Distress Indicator

	-1 (Negative)	0 (Neutral)	1 (Positive)
Sentiment	155	2,094	369
Distress	825	478	1,315

I used Scipy for Spearman (1904) rank correlation between nonparametric variables. A Spearman's correlation that is equal to zero means that that no correlation exists between the data samples, whereas a correlation that is close to +1 or -1 indicates the significant correlation and the direction of the correlation. The result of the statistic on the DISTRESS and SENTIMENT variables using the 2,618 firms in the data set yielded a Spearman's correlation (p) of -0.03024 and a p-value of 0.1219 that indicated no correlation between DISTRESS and SENTIMENT. Similar analysis were applied at a sector level, and the results were similar. Table 17 includes the application of Spearman coefficient over the firms grouped by sector.

Table 17

Spearman Correlation for Sampled Firms by Sector

Sector	Firms count	Spearman correlation	2 tailed *p*-value
Basic industries	233	-0.0576	0.3814
Capital goods	253	0.0669	0.2886
Consumer durables	100	0.0556	0.5822
Consumer nondurables	170	-0.0861	0.2637
Consumer services	397	-0.0202	0.6881
Energy	212	-0.2193	0.0013
Health care	427	-0.0472	0.3300
Miscellaneous	102	-0.0482	0.6301
Public utilities	138	-0.0262	0.7599
Technology	499	-0.0532	0.2353
Transportation	81	-0.0471	0.6759

For all the sectors except the energy sector in the sample set, the correlation between the *distress* and the *sentiment* variables were close to zero. I detected a weak inverse relationship with a Spearman correlation coefficient of -0.2103 and a p-value of 0.0013 for the firms in the energy sector. Such result infers that a -20% correlation exist between the investor sentiment and the financial distress in the sample sets. Even though the correlated is weak at 20%, the positive sentiments in the Energy sector correlated with the financial distress of firms, while negative sentiments correlated with safe firms. Further discussion of the finding is available in Chapter 5 of the study. However, the overall finding using Spearman correlation between SENTIMENT and DISTRESS is that there is not statistical evidence to support the hypothesis that a correlation exists between the level of financial distress of firms and the level of investor sentiment using social media. Hence, I could not reject the null hypothesis $H1_0$.

Statistical Analysis for Hypothesis #2

The second hypothesis in the study is that there is a statistical relationship between negative sentiments and the risk of bankruptcy for financially distressed firms. A Z-Score of less than 1.8 indicates that a firm is likely to report bankruptcy in two years (Altman, 2013). In order to determine if the data can support the hypothesis, I took the 825 firms with an Altman *Z*-Score of less than 1.8 as listed in Table 16. The Spearman correlation coefficient yielded a -0.058 with p-value=0.092 which infer that there is no statistical significant correlation between investor sentiment and the risk of bankruptcy of financial distress for firms that are under financial distress.

Hence, I could not reject the null hypothesis $H2_0$.

Statistical Analysis for Hypothesis #3

To support the third hypothesis if there is relationship between a firm's stock movement and the level of sentiment towards the firm in social media, I applied Spearman correlation analysis between the average sentiment per day and the gain/loss stock value for each of the sampled firms in the study. The steps that I took were as follows:

1. Let DAILYSENTIMENT be the aggregate average of sentiments per day for a sampled firms, where DAILYSENTIMENT = 0 is very negative sentiment, DAILYSENTIMENT=1 is negative sentiment, DAILYSENTIMENT = 2 is neutral sentiment, DAILYSENTIMENT=3 is positive sentiment, and DAILYSENTIMENT = 4 is very positive sentiment.
2. Let OPENSTOCK = opening stock value for a given day, and CLOSESTOCK = closing stock value for a given day
3. Let DAILYSTOCKDIRETION be the overall positive or negative direction between the start and close of the firm's stock per day, where DAILYSTOCKDIRECTION = +1 if CLOSESTOCK > OPENSTOCK, DAILYSTOCKDIRECTION = -1 if CLOSESTOCK < OPENSTOCK, and DAILYSTOCKDIRECTION = 0 if CLOSESTOCK = OPENSTOCK.

To determine if there is a relationship between stock movement and sentiments, I first extracted daily stock data for each of the sampled firms using Yahoo (2013) open tables for the period between 2014-12-06 and 2014-1-05 and calculated DAILYSENTIMENT and DAILYSTOCKDIRECTION for each of the sampled firms. Table 18 includes the data set per each day period in the final sample set after excluding firms that had no sentiments collected or had no stock trading during the sample period.

After I had identified the daily sentiments and daily stock movement for each of the sampled firms, I applied the Spearman rank correlation.

Table 18

Stock Sentiments & Trading Movement per Date Period

Date	Stock	Negative sentiment	Neutral sentiment	Positive sentiment	Negative stock movement	Neutral stock movement	Positive stock movement
2014-12-08	495	48	382	65	144	11	340
2014-12-09	555	35	469	51	470	4	81
2014-12-10	589	48	467	74	71	11	507
2014-12-11	510	33	398	79	229	5	276
2014-12-12	528	54	380	94	184	8	336
2014-12-15	451	30	352	69	61	6	384
2014-12-16	512	45	393	74	242	11	259
2014-12-17	544	28	461	55	495	7	42
2014-12-19	487	42	392	53	287	10	190
2014-12-22	290	22	224	44	160	7	123
2014-12-23	324	30	225	69	145	9	170
2014-12-24	353	42	239	72	175	10	168
2014-12-26	358	27	277	54	181	12	165
2014-12-29	316	12	219	85	177	12	127
2014-12-30	439	23	319	97	185	17	237
2014-12-31	364	38	247	79	106	8	250

The Spearman correlation analysis between the DAILYSENTIMENT and the DAILYSTOCKDIRECTION returned a -.005 and P=0.6724 that indicates a none statistically significant correlation between the stock movement and the sentiments per day. It indicates that I cannot reject the null hypothesis $H3_0$. Hence, there is a lack of a statistical relationship between investor sentiment and daily stock movements.

Summary

In the data analysis of the study, I attempted to determine if there is a statistical relationship between the financial distress of firms and investor sentiment towards the firms in social media. The statistical analysis between the levels of investor sentiments using the social media website Twitter and the level of financial distress using Altman *Z*-Score for the sampled firms yielded a nonstatistical significant relationship. Similarly, I did not find a significant relationship between negative sentiments and financially distressed firms. In addition, after I determined the stock price movements for each of the stocks associated with the sampled firms, I did not find a significant relationship between daily stock movements and daily sentiments for the sampled firms. The next chapter expands on the findings and includes the conclusions and recommendations from the study.

Chapter 5: Discussion, Conclusions, and Recommendations

The purpose of this study was to examine the relationship between investor sentiment using social media and the level of financial distress of firms. With the ubiquity of social media such as Facebook and Twitter within the marketplace, recent research, such as Bollen et al. (2011) and Asur and Huberman (2010) argued that sentiment in social media can be a predictor of market moods, investor sentiment, and future events. In this study, the relationship between investor sentiments and the prediction of financial distress using the Altman *Z*-Score and stock market movements was applied for the majority of publicly held firms in the United States. The findings of the study revealed a nonsignificant relationship between investor sentiments in the social media website Twitter and the levels of financial distress in the sampled firms as well as firms' stock movements. This chapter includes additional insights about the study findings.

Interpretation of Findings

The findings did not confirm with various literature that considered sentiments expressed in social media as a predictor of future events. For instance, Bollen et al. (2011) provided statistical evidence to demonstrate that changes to public moods matches 86.8% of the shifts in the Dow Jones Composite Index. They analyzed approximately 10 million tweets posted by 2.7 million users in 2008 and cross-validated the results with a questionnaire for detecting the public moods during the presidential election and Thanksgiving Day in 2008. Luo, Zhang, and Duan (2013) compared consumer ratings between online blogs and traditional websites about products associated with firms in the computer hardware and software sector and found that social media can be a leading indicator of firm equity value. However, in their study of how social media can be used to forecast future outcomes, Asur and Huberman (2010) found that sentiments provide improvements in predicting future sales of movies but not as much as the rate of tweets themselves. The findings as they relate to the study of corporate financial distress did not confirm with the literature that sentiments in social media could be a predictor of future events. The findings also do not add support to Kasperson's (2012) social amplification of risk framework.

The lack of a statistical relationship between negative sentiments in the social media website Twitter and the expected bankruptcy of financially distressed firms shows no empirical evidence in support of Kasperson's (2012) social amplification framework. According to Kasperson, amplified negative responses in the media can have detrimental effects on the society including the possibility of corporate bankruptcies. During the proposal of the study, I hypothesized that negative sentiments positively correlate with the Altman's (2013) *Z*-Score of less than 1.8 that indicate the possibility of bankruptcies in 2 years by financially distressed firms. Once I executed the research study, I found no statistical relationship between negative sentiments and the Altman *Z*-Score across the sampled firms from the different sectors except a weak correlation in the energy sector. A weak correlation of -21.9% using Spearman rank correlation analysis with $p < 0.001$ was detected between the level of sentiments and the level of financial distress, while no significant correlation was detected for firms in the basic industries, capital goods, consumer durables, consumer nondurables, consumer services, healthcare, public utilities, technology, transportation, and miscellaneous sectors.

By interpreting the tweets associated with negative, positive, and neutral sentiments, the public postings in social media platform might a misrepresentation of actual investor sentiment towards the value of the targeted firms or the stock movements or market fundamentals may not necessary reflect investor sentiment. Using the data collected in Chapter 4, Figure 13 includes a matrix that shows the relationship between the tweet sentiments and the same day stock movements for the tweets and stock trading associated with the financially distressed firms sampled in the study. The percentage of stocks gain or loss for firms associated with positive sentiments are approximately 9% each, and the percentage of stocks gain or loss for firms associated with negative sentiment are approximately 5% each. Such findings support the statistical analysis using Spearman correlation that shows no statistically significant correlation between both variables. To provide some insights on why such relationship does not exist, I extracted a random sample of the data collected.

Tweets Broken Down by Stock Movements & Stock Movements for Publicly Traded Firms between 12-6-2014 and 1-7-2015 under Financial Distress (Altman *Z*-Score<1.8)		Stock Movement (Same Day Stock Close vs Same Day Stock Open)			
		Stock Loss	No Change	Stock Gain	Total
Tweet Sentiments	Negative	1276, 5.13%	32, 0.13%	1293,5.20%	2601,10.45%
	Neutral	9248, 37.16%	207, 0.83%	8049,32.34%	17504,70.33%
	Positive	2429,9.76%	89,0.36%	2266,9.10%	4764,19.22%
	Total	12953,52.04%	328,1.32%	11608,46.64%	24889,100%

Figure 12. Matrix representing the collected tweet sentiments and same day trading gain or losses for sampled firms under financial distress (Altman *Z*-Score <1.8) In this figure, I considered tweets that mention more than one stock symbol as multiple tweets, where each tweet represents a single firm.

Table 19 includes a random sample set of the tweets collected and associated with the stock movements of the firms for interpretation. Each row in the table includes a sample tweet, the stock symbol associated with the firm that is included in the tweet, the date of the tweet collection, the opening and closing stock prices for the firm during the given day, the stock movement, and the sentiment captured by the Stanford Core NLP algorithm. The stock movement is positive if the closing price exceeds opening price, is negative if the opening price exceeds the closing price and is neutral if no change occurred between the opening and closing stock price. There is a discrepancy in the data. In some instances, the stock movement for a firm was positive even if the tweet implied a negative sentiment, such as the 12/19/2014 tweet *11:53 Drops: $NVGN -4.8%*. The tweet is associated with the biotechnology company Novagen that announced on the same day an issuance of securities plan for existing institutional investors in the United States.

Novagen's stock rebounded in value from $2.56 to 2.58 on at the close of trading on that same day. This is an example of a tweet with a negative sentiment that had a positive outcome on a given day. Similarly, some tweets include positive sentiments but trading stock experienced losses. For instance, the 12/15/2014 tweet "$UNP Seeking the next hot #pennystock $AIV $NVE $QLYS Visit" is a positive sentiment, but the stock value for $AIV declined that day. The decline in the stock value could have been a result of financial trades and market demand for the stock. After I have considered only the tweets associated with financially distressed firms, I observed the pattern in Figure 19 that signifies randomness in the distribution of sentiments and stock movements for the firms under financial distress.

The breakdown of sentiments by stock movements in Figure 19 shows that there are slightly more positive sentiments than negative sentiments in the collected tweets, but the breakdown of the stocks that either gained or lost on the days of each tweet is approximately 50% respectively. I performed the same analysis for the firms that are not under financial distress, and the results had a nonstatistically significant marginal different from the data. This further implies that there is no correlation between the sentiments in the collected tweets and the stock movements between the opening and closing stock market prices for the publicly held firms in the United States. The data analysis in this study would not support the hypothesis that the sentiments within the social media has a relationship with the financial distress conditions of the firms or the stock movements associated with such firms.

Table 19

Sampled Tweets With Stocks Movement

Stock open ($)	Stock close ($)	Stock symbol	Tweet	Date tweet & stock	stock movement	Sentiment
37.52	38.1	MS	Are you Investing in $SEIC $KORS $MS $CRAY Must have	12/9/2014	positive	positive
36.62	35.94	AIV	$UNP Seeking the next hot #pennystock $AIV $NVE $QLYS Visit	12/15/2014	negative	positive
6.83	7.19	GRPN	Are you considering buying $JEC $GRPN $FTI $AMZN #financialnews	12/17/2014	positive	positive
2.56	2.58	NVGN	11:53 Drops: $NVGN -4.8%	12/19/2014	positive	negative
83.36	82.11	VAL	#Stocks you might want to sell $INGR $AET $VAL $DO #investing	12/10/2014	negative	negative
51.82	53.02	TMK	$SBUX Are you hanging on to $FDS $TMK $PCLN Try this	12/17/2014	positive	negative
22.7	22.58	CNP	Avoid mistakes like $STR $HHC $CNP $BLOX #StockMarket	12/11/2014	negative	negative
144.94	140.25	COST	Significant activity on social media: $TITN $LAYN $KKD $CYRN $VRA $COST $TOL. Trade social buzz @	12/10/2014	negative	negative
2.1	2.07	VRML	$VRML Run continues and option activity today in July Calls	12/29/2014	negative	positive

Limitations of the Study

Several limitations in the research design and the data collection may have affected the outcome of the study. The primary limitation is that the data collected were only for the tweets that mention the stock symbol of the sampled firms. They do not include the wider audience of public users in social media who may mention the firm by its name rather than its symbol. Other limitations include the chosen instruments for the study. The two instruments used for the analysis of sentiments and financial distress of firms were the Altman's (2013) *Z*-Score and the Manning et al.'s (2014) CoreNLP. As discussed in depth in Chapter 2 of the study, the Altman *Z*-Score is a notable indicator for predicting the level of financial distress of firms. However, there are alternative instruments that may yield a different outcome than the Altman *Z*-Score, and it includes logit, probit, and machine-learning algorithms. None of such financial distress prediction models has consistently demonstrated more accurate prediction than the rest. On the other hand, the Manning et al.'s (2014) CoreNLP model for sentiment analysis is a supervised learning machine-learning instrument that required training of the model. The accuracy of the model is, hence, dependent on the quantity and quality of training data used in the system. In this study, 10,000 random tweets were used to manually classify the tweets as *very positive*, *positive*, *neutral*, *negative*, or *very negative* sentiments before applying executing the automated sentiment analysis over the entire population of tweets collected. At the time of the study, I could not find any existing trained sentiment analysis model for financial data, and, therefore, I resorted to training the system on my own.

A possibility of training the system with a much larger training dataset and by multiple participants could return more accurate sentiment predictions. Moreover, the limitations imposed by the social media website Twitter in accessing every tweet associated with the sampled firms required an innovative way to extract tweets even though the volume of tweets were not exhaustive. However, the study included the development of a software model that can support different instrumentation methods and data collection procedures, which can help future research in the fields of sentiment analysis using social media and corporate financial distress analysis.

Recommendations

The evidence in the recent literature, as previously discussed, is that sentiments captured from the social media can be a predictor of future events. Even though the current study did not find a statistical relationship between investor sentiments using the social media website Twitter and the level of corporate distress in firms, the software that I developed in this study can support alternative instruments that might yield different results. For instance, a future study can use a machine-learning model, such as Lee and To (2010) back propagation neural network model instead of the Altman's (2013) *Z*-Score for measuring corporate financial distress. Manning et al.'s (2014) supervised machine learning model could be replaced with an unsupervised deep learning model as proposed by Ribeiro and Lopes (2011). The data collection process can include tracking of all tweets associated with each firm instead of only the tweets that include the stock symbols of the relevant firms. Furthermore, the study can include collecting data from not only the social media website, Twitter, but from other social media platforms, such as Facebook, LinkedIn, and other social media websites as well. With the continuous commercialization of social media and financial data for business purposes, the strength of the framework developed in the study is that it is a cost—effective approach to analyzing sentiments and detecting corporate distress in firms.

All the software that I developed or leveraged in the study is open source technologies that do not require license purchases. For instance, I used the open source programming language Python for the data collection and analysis segments of the study. I also used the open source Oracle Corporation (2014) MySQL database for storing and querying the data collected in the study. Moreover, the approach that I used to collect the financial data using Yahoo (2013) community open tables and a live stream of tweets using Twitter (2014) developer modules was with permission but at relatively no cost, except the time spent to develop the algorithms. The possibilities of extending the framework in the study to further support sentiment analysis within the financial markets are numerous and can have positive implications within corporate institutions and among investors.

Implications

The implication of this study is on the individual and the organizational level. On the individual level, investors could benefit from the framework as a low-cost solution that can detect sentiments from other investors about specific firms. Corporate firms, on the other hand, could embed the framework within their social media departments in order to help detect market mood changes after special events, such as a quarterly release update, or during a marketing campaign event. Even though the study did not find a statistically significant relationship between the levels of investor sentiment in the social media and the level of financial distress of firms, the model developed in the study is innovative and extensible in order to support further research within the field. Practical applications of the study include public sentiment analysis by financially distressed and stock screening systems or financial tradition recommendation systems for investors or traders.

Conclusions

Contrary to the evidence in the literature that supports the relevance of sentiments in the social media to predict future events, in the study, I did not find a relationship between investor sentiments and the level of financial distress of firms. As Luo et al. (2013) argued, sentiment analysis could be very effective in assessing public opinion toward product and services that could affect the investor's decision to buy or sell a stock. However, the lack of sufficient information about the financial health of companies can make it difficult to investors to make short selling decisions to the distressed firms (Campbell et al., 2011). Even with the calls of further integrating the social media into the study of financial markets, the effects of sentiments in the social media, as determined in the study, remains to be challenged within the study of corporate financial distress and the prediction of corporate bankruptcies. However, the model developed for the study can be a key instrument in the continuous research on social media and its role in the financial markets.

References

Ahmadi, A. P. S., Soleimani, B., & Vaghfi, S. H. (2012). Corporate bankruptcy prediction using a logit model: Evidence from listed companies of Iran. *World Applied Sciences Journal*, *17*(9), 1143–1148. Retrieved from idosi.org/wasj/wasj17(9)12/11.pdf

Akhtar, S., Faff, R., Oliver, B., & Subrahmanyam, A. (2012). Stock salience and the asymmetric market effect of consumer sentiment news, *36* (12). *Journal of Banking & Finance*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378426612001999

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine. *Econometrica*, *21*(4), 503–546. doi:10.2307/1907921

Altman, E. (2013). Predicting financial distress of companies: Revisiting the Z-Score and ZETA models. In A. Bell, C. Brooks, & M. Prokopczurk (Eds.) *Handbook of Research Methods and Applications in Empirical Finance* (pp. 428–457). Northampton, MA: Edward Elgar Publishing.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*(4), 589–609. Retrieved from http://www.twd2u.com/wp-content/uploads/2008/11/z-score.pdf

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2014). Distressed firm and bankruptcy prediction in an international context: A review and empirical analysis of Altman's Z-Score model. *Available at SSRN 2536340*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2536340

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, *18*(3), 505–529. Retrieved from http://www.sciencedirect.com/science/article/pii/0378426694900078

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, *23*(2), 193–212. doi:10.1214/aoms/1177729437

Apergis, N., Sorros, J., Artikis, P., & Zisis, V. (2011). Bankruptcy probability and stock prices: The effect of Altman *Z*-score information on stock prices through panel data. *Journal of Modern Accounting and Auditing*, *7*(7), 689–696. Retrieved from http://www.davidpublishing.com/davidpublishing/Upfile/7/20/2011/2011072071549545.pdf

Ardalan, K. (2008). *On the role of paradigms in finance*. Hampshire, United Kingdom: Ashgate Publishing, Ltd. Retrieved from http://books.google.com/books?id=k0GqEHAKQRkC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Web Intelligence and Intelligent Agent Technology, 1*, 492-499. doi: http://10.1109/WI-IAT.2010.63

Baker, M., & Wurgler, J. (2011). *Behavioral corporate finance: An updated survey*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17333

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In A. Mille et al. (Eds.) *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). Retrieved from http://dl.acm.org/citation.cfm?id=2187907

Barwise, J. (1977). An introduction to first-order logic. *Studies in Logic and the Foundations of Mathematics*, *90*, 5-46. Retrieved from http://www.sciencedirect.com/science/article/pii/S0049237X08710978

Batista, F., & Ribeiro, R. (2013). Sentiment analysis and topic classification based on binary maximum entropy classifiers. Retrieved from http://rua.ua.es/dspace/handle/10045/27867

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, *4*, 71-111. doi:10.2307/2490171

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society*, *22*(1), 23–36. doi:10.2307/1909829

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89-98. Retrieved from http://people.psych.cornell.edu/~jec7/pcd%202012-13%20pubs/berwicketal.pdf

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, *21*(3), 1339-1369. doi:10.1093/rfs/hhn044

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., & Tredinnick, V. (1995). Bracketing guidelines for Treebank II style Penn Treebank project. *University of Pennsylvania*, *97*, 100. Retrieved from http://www.stanford.edu/~bresnan/128_2002/Docs/prsguid1.pdf

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81* (3), 637-654. Retrieved from http://www.jstor.org/stable/1831029

Blackledge, J. (2010). Systemic risk assessment using a nonstationary fractional dynamic stochastic model for the analysis of economic signals. *ISAST Transactions on Computers and Intelligent Systems*, *2*(1), 76–94. Retrieved from http://arrow.dit.ie/cgi/viewcontent.cgi?article=1021&context=engscheleart2

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint arXiv:0911.1583*. Retrieved from http://arxiv.org/pdf/0911.1583

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8. Retrieved from http://arxiv.org/pdf/1010.3003.pdf

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*(5), 1264-1302. doi:10.3390/e12051264

Boyd, D. M. & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer‐Mediated Communication*, *13*(1), 210-230. doi:10.1111/j.1083-6101.2007.00393.x

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. *System Sciences (HICSS), 2010 43rd Hawaii International Conference, 1,* (pp. 1-10). doi:10.1109/HICSS.2010.412

Burrell, G., & Morgan, G. (1985). *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Burlington, VT: Ashgate Publishing, Ltd.

Byrne, A., & Brooks, M. (2008). Behavioral finance: Theories and evidence. The Research Foundation of CFA Institute. Retrieved from http://dgase.googlecode.com/svn/trunk/latex/papers/rflr%252Ev3%252En1%252E1%5B1%5D.pdf

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48-57. Retrieved from http://www.sentic.net/jumping-nlp-curves.pdf

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, *63*(6), 2899-2939. doi:10.1111/j.1540-6261.2008.01416.x

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2011) Predicting financial distress and the performance of distressed stocks. *Journal of Investment Management*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1829622

Chancharat, N., Davy, P., McCrae, M. S., & Tian, G. G. (2007, August). Firms in financial distress, a survival model analysis. *20th Australasian Finance & Banking Conference*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1009385&

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, *8*(4), 537-569. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=287474

Chava, S., & Purnanandam, A. (2011). The effect of banking crisis on bank-dependent borrowers. *Journal of Financial Economics*, *99*(1), 116-135. doi:10.1093/rfs/hhp107

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, *2*(3), 113–124. Retrieved from http://noam-chomsky.org/articles/195609--.pdf

Christidis, A., & Gregory, A. (2010). Some new models for financial distress prediction in the UK. *Xfi-Centre for Finance and Investment Discussion Paper*, *1* (10). Retrieved from https://ore.exeter.ac.uk/repository/bitstream/handle/10036/114713/1004.pdf?sequence=1

Chung, I. J. (2011). Social amplification of risk in the Internet environment. *Risk Analysis*, *31*(12), 1883-1896. Retrieved from http://www.kapa21.or.kr/data/data_download.php?did=5076&origin=publication_detail

Cole, R. A., & Wu, Q. (2009). Predicting bank failures using a simple dynamic hazard model. Retrieved from http://hdl.handle.net/1959.13/919847

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. (3rd ed.) Thousand Oaks, CA: Sage.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive*, *9*(4), 243-259. Retrieved from http://misqe.org/ojs2/index.php/misqe/article/view/342</ee>

Curtis, A. (2013). The brief history of social media. Retrieved March 24, 2014, from http://www2.uncp.edu/home/acurtis/NewMedia/SocialMedia/SocialMediaHistory.html

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*(2), 341–348. Retrieved from http://webspace.ship.edu/pgmarr/geo441/readings/d%27agostino%201971%20-

%20an%20omnibus%20test%20of%20normality%20for%20moderate%20and%20large%20size%20samples.pdf

D’Agostino, R., & Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of b2 and√ b1. *Biometrika*, *60*(3), 613–622. Retrieved from http://biomet.oxfordjournals.org/content/60/3/613.short

Duckett, D., & Busby, J. (2013). Risk amplification as social attribution. *Risk Management*, *15*(2), 132–153. doi:10.1057/rm.2013.2

Duggan, M., & Brenner, J. (2013, February). The demographics of social media users — 2012. Pew Research Center’s Internet & American Life Project. Retrieved from http://pewinternet.org/Reports/2013/Social-media-users/Social-Networking-Site-Users/Overview.aspx

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383–417. Retrieved from http://down.cenet.org.cn/upfile/36/20058187426139.pdf

Finch, J. H., Fraser, S. P., & Scheff, S. R. (2011). Teaching the CAPM in the introductory finance course. *Journal of Economics and Finance Education*, *10*(1), 77. Retrieved from http://www.economics-finance.org/jefe/fin/10%20-%20Teaching%20the%20CAPM.pdf

Free Software Foundation. (2014). GNU general public license, version 2. Retrieved April 22, 2014, from http://www.gnu.org/licenses/gpl-2.0.html

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392. Retrieved from http://physrev.physiology.org/content/91/4/1357.short

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*, 1–12. Retrieved from http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf

Graham, J. R., Hazarika, S., & Narasimhan, K. (2011). Financial distress in the Great Depression. *National Bureau of Economic Research Working Paper Series*, (No. 17388). Retrieved from http://www.nber.org/papers/w17388

Hafez, P. A., & Xie, J. (2012). Factoring sentiment risk into quant models. *Available at SSRN 2071142*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2071142

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, *54*(3), 265–273. doi:10.1016/j.bushor.2011.01.007

Harford, T. (2012, April 27). One maths formula and the financial crash. *BBC*. Retrieved from http://www.bbc.co.uk/news/magazine-17866646

Hayes, S. K., Hodge, K. A., & Hughes, L. W. (2010). A study of the efficacy of Altman's Z to predict bankruptcy of specialty retail firms doing business in contemporary times. *Economics & Business Journal: Inquiries & Perspectives*, *3*(1). Retrieved from http://ecedweb.unomaha.edu/EBJI%202010HayesHodgeHughes.pdf

Hill, N. T., Perry, S. E., & Andes, S. (2011). Evaluating firms in financial distress: An event history analysis. *Journal of Applied Business Research (JABR)*, *12*(3), 60–71. Retrieved from http://cluteinstitute.com/ojs/index.php/JABR/article/viewFile/5804/5886

Howard, P. N., Duffy, A., Freelon, D., Hussain, M., Mari, W., & Mazaid, M. (2011). Opening closed regimes: What was the role of social media during the Arab Spring? Retrieved from http://ictlogy.net/bibliography/reports/projects.php?idp=2170

Janesick, V. J. (2010). *"Stretching" Exercises for Qualitative Researchers*. Thousand Oaks, CA: Sage.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometica*, *47*(2), 263–291. Retrieved from http://pages.uoregon.edu/harbaugh/Readings/GBE/Risk/Kahneman%201979%20E,%20Prospect%20Theory.pdf

Kasperson, R. E. (2012). A perspective on the social amplifications of risk. *The Bridge on Social Sciences and Engineering Practice*, *42*(3). Retrieved from http://www.clarku.edu/departments/marsh/faculty/kasperson.cfm

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R. & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk analysis*, *8*(2), 177-187. Retrieved from http://elib.uni-stuttgart.de/opus/volltexte/2010/5307/pdf/ren27.pdf

Kucuktunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! answers. *Proceedings of the fifth ACM international conference on Web search and data mining*, *1*, 633–642. Retrieved from http://dl.acm.org/citation.cfm?id=2124371

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions.* (3rd ed.). Chicago, IL: University of Chicago Press.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., & Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet*, *378*(9790), 526–535. Retrieved from http://lib.kums.ac.ir/documents/10129/57858/526.pdf

Lee, Cheng-Few, Lee Alice,& Lee, John. (2010). Foundation of portfolio theory. In *Handbook of Quantitative Finance and Risk Management*. New York, NY: Springer.

Lee, H. Y., & Renganathan, H. (2011). Chinese sentiment analysis using maximum entropy. Retrieved from http://acl.eldoc.ub.rug.nl/mirror/W/W11/W11-37.pdf#page=105

Lee, M.-C., & To, C. (2010). Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress. *International Journal of Artificial Intelligence & Applications*, *1*(3), 31–43. doi:10.5121/ijaia.2010.1303

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social media & mobile internet use among teens and young adults. millennials. *Pew Internet & American Life Project*. Retrieved from http://files.eric.ed.gov/fulltext/ED525056.pdf

Lifschutz, S., & Jacobi, A. (2010). Predicting bankruptcy: Evidence from Israel. *International Journal of Business and Management*, *5*(4), 133–141. Retrieved from http://www.ccsenet.org/journal/index.php/ijbm/article/download/4933/4618?origin=pub

Lin, T.-H. (2009). A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. *Neurocomput.*, *72*(16-18), 3507–3516. doi:10.1016/j.neucom.2009.02.018

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing* (Vol. 2, p. 568). Retrieved from http://people.sabanciuniv.edu/berrin/proj102/1-BLiu-Sentiment%20Analysis%20and%20Subjectivity-NLPHandbook-2010.pdf

Luo, X., Raithel, S., & Wiles, M. A. (2013). The impact of brand rating dispersion on firm value. *Journal of Marketing Research*, *50*(3), 399–415. Retrieved from http://www.researchgate.net/profile/Sascha_Raithel/publication/255977947_The_Impact_of_Brand_Dispersion_on_Firm_Value/links/0deec521c5fc30f34a000000.pdf

Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research*, *24*(1), 146–163. doi:10.1287/isre.1120.0462

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. Retrieved from http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

Mansi, S., Maxwell, W. F., & Zhang, A. (Jianzhong). (2010). Bankruptcy prediction models and the cost of debt, *Forthcoming*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1622407

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, *7*(1), 77–91. Retrieved from http://media.blubrry.com/institutcoppet/p/www.math.ust.hk/~maykwok/courses/ma362/07F/markowitz_JF.pdf

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. Retrieved from http://cns-classes.bu.edu/cn550/Readings/mcculloch-pitts-43.pdf

McGuinness, D. L., & Harmelen, F. van. (2004). OWL web ontology language overview. *W3C Recommendation*, *10*(2004-03), 10. Retrieved from http://202.119.112.136/pweb/~zhuoming/teachings/MOD/N4/Readings/5.3-B1.pdf

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, *29*, 449–470. Retrieved from https://stat.fsu.edu/~jfrade/PAPERS/Credit%20risk%20modeling%20papers/SWP-0684-14514372.pdf

Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review*, *48*(3), 261–206. Retrieved from http://www.academia.edu/download/30867396/modiglianiandmiller1958.pdf

Montalvo, R. E. (2011). Social media management. *International Journal of Management & Information Systems*, *15*(3). Retrieved from http://cluteinstitute.com/ojs/index.php/IJMIS/article/download/4645/4734

Morgenstern, O., & Neumann, J. V. (1953). *Theory of games and economic behavior* (3rd ed.). Princeton, NJ: Princeton University Press.

Murthy, D. (2011). Twitter: Microphone for the masses? *Media Culture and Society*, *33*(5), 779. Retrieved from http://www.dhirajmurthy.com/wp-content/uploads/2012/04/Twitter-Microphone-for-the-masses.pdf

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. Retrieved from http://171.67.114.118/content/18/5/544.full

Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. Retrieved from http://arxiv.org/abs/1305.6143

NASDAQ. (2015, January 1). Company List. Retrieved December 7, 2014, from http://www.NASDAQ.com/screening/company-list.aspx

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1536/1842

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. Retrieved from http://www.math.ryerson.ca/ramlab/projects/crd/ohlson1980.pdf

Oracle Corporate. (2014). MySQL. Retrieved April 24, 2014, from http://www.mysql.com/about/

Paltoglou, G., Thelwall, M., & Buckley, K. (2010). Online textual communications annotated with grades of emotion strength. Retrieved from http://lrec-conf.org/proceedings/lrec2010/workshops/W24.pdf#page=33

Panda, S. S., Chakraborty, D., & Pal, S. K. (2008). Flank wear prediction in drilling using back propagation neural network and radial basis function network. *Applied Soft Computing*, *8*(2), 858–871. Retrieved from http://dspace.nitrkl.ac.in/dspace/bitstream/2080/546/1/sspanda-JOSC-1.pdf

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135. Retrieved from http://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf

Pearl, J. (1985). Bayesian networks: A model of ‘self-activated memory for evidential reasoning. Retrieved from ftp://ftp.cs.ucla.edu/pub/stat_ser/r43-1985.pdf

Peng, H.-K., Zhu, J., Piao, D., Yan, R., & Zhang, Y. (2011). Retweet modeling using conditional random fields. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6137399

PhpMyAdmin Contributors. (2014). PhpMyAdmin. Retrieved April 24, 2014, from http://www.phpmyadmin.net/home_page/index.php

Platt, H. D., & Platt, M. B. (2009). Financial distress comparison across three global regions. *Journal of Risk and Financial Management*, *1*(1), 129–162. Retrieved from http://www.mdpi.com/1911-8074/1/1/129/pdf

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.3029&rep=rep1&type=pdf

Rambocas, M., & Gama, J. (2013). *Marketing research: The role of sentiment analysis* (No. 489). Universidade do Porto, Faculdade de Economia do Porto. Retrieved from http://ideas.repec.org/p/por/fepwps/489.html

Ribeiro, B., & Lopes, N. (2011). Deep belief networks for financial prediction. *Neural Information Processing, 7064,* 766-773. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24965-5_86

Roy Rosenzweig Center for History and New Media. (2014). Zotero. Retrieved from http://zotero.org/support/quick_start_guide

Salehi, M. (2013). Bankruptcy prediction by using support vector machines and genetic algorithms. *Studies in Business and Economics*, *8*(1), 104-114. Retrieved from http://profdoc.um.ac.ir/paper-abstract-1034039.html

Schultz, F., Utz, S., & Göritz, A. (2011). Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media. *Public Relations Review*, *37*(1), 20–27. Retrieved from http://dtserv2.compsy.uni-jena.de/ws2014/medikfme_ue/36211177/content.nsf/Pages/64894BDCFE428C00C1257D80005C84CA/$FILE/perceptions%20and%20reactions%20to%20crisis%20commuinication%20via%20twitter.pdf

SciPy Developers. (2014). SciPy. Retrieved April 24, 2014, from http://www.scipy.org/

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*(3), 425–442. Retrieved from http://e-m-h.org/Shar64.pdf

Shivhare, S. N., & Khethawat, S. (2012). Emotion detection from text. Retrieved from http://arxiv.org/abs/1205.4944

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, *74*(1), 101–124. Retrieved from http://www.rcg.ch/resources/Forecasting_Bankruptcy__Simple_Hazard_Model1.pdf

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment Treebank. Retrieved from http://oldsite.aclweb.org/anthology-new/D/D13/D13-1170.pdf

Sowa, J. (2013, October 19). Semantic networks, Encyclopedia of Artificial Intelligence, edited by SC Shapiro. Retrieved April 15, 2014, from http://www.jfsowa.com/pubs/semnet.htm

Sowa, J. F. (2006). Semantic networks. *Encyclopedia of Cognitive Science*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0470018860.s00065/full

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101. doi:10.2307/1412159

Speier, H. (1950). Historical development of public opinion. *American Journal of Sociology*, *55*(4), 376–388. Retrieved from http://www.jstor.org/stable/2772299

Standard & Poor's. (2012). CRSP/Compustat merged database. Retrieved July 8, 2012, from http://www.compustat.com/CRSP_Compustat/

Stanford NLP Group. (2014). Stanford CoreNLP. Retrieved April 22, 2014, from http://nlp.stanford.edu/software/classifier.shtml

The Scipy community. (2009, March 17). scipy.stats.spearmanr — SciPy v0.7 Reference Guide (DRAFT). Retrieved January 14, 2015, from https://docs.scipy.org/doc/scipy-0.7.x/reference/generated/scipy.stats.spearmanr.html

The Scipy community. (2014, May 11). scipy.stats.mstats.normaltest — SciPy v0.14.0 Reference Guide. Retrieved January 14, 2015, from http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mstats.normaltest.html

Thompson, J. B. (2013). *Media and Modernity: A Social Theory of the Media*. California: John Wiley & Sons.

Twitter. (2012, June 25). Twitter Terms of Service. Retrieved June 22, 2014, from https://twitter.com/tos

Twitter. (2014). Twitter developers. Retrieved April 22, 2014, from https://dev.twitter.com/

Twitter. (2015, January 1). Twitter. Retrieved January 1, 2015, from https://twitter.com/

Van Der Ploeg, S. (2010). Bank default prediction models: A comparison and an application to credit rating transitions. Retrieved from http://oaithesis.eur.nl/ir/repub/asset/6470/294726ploegma0110.pdf

Wilke, J. (2010). Media genres. Retrieved March 24, 2014, from http://ieg-ego.eu/en/threads/backgrounds/media-genres/juergen-wilke-media-genres

World Federation of Exchanges. (2015). Monthly reports. Retrieved January 10, 2015, from http://www.world-exchanges.org/statistics/monthly-reports

Yahoo. (2012, March 16). Terms of service. Retrieved January 19, 2015, from http://info.yahoo.com/legal/us/yahoo/utos/en-us/

Yahoo. (2013). YQL - Community open data tables. Retrieved January 19, 2015, from http://www.datatables.org/

Yahoo. (2015, January 1). Yahoo finance. Retrieved January 1, 2015, from http://finance.yahoo.com/

Yu, Q. (2013). Machine learning for corporate bankruptcy prediction. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-16943-4_6

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59. doi:10.2307/2490859

Zurada, J. M., Foster, B. P., Ward, T. J., & Barker, R. M. (2011). Neural networks versus logit regression model for predicting financial distress response variables. *Journal of Applied Business Research (JABR)*, *15*(1), 21–30. Retrieved from http://cluteinstitute.com/ojs/index.php/JABR/article/viewFile/5685/5762

Appendix A: Database Model Source Code

```
__author__ = 'Tarek Hoteit'

from django.db import models
from django.contrib.auth.models import User
class Company(models.Model):
    # Each company under analysis
    symbol = models.CharField('Stock Symbol', max_length=10)
    name = models.CharField('Company Name', max_length=100)
    lastSale = models.CharField('Last Sale', max_length=20, null=True, blank=True)
    marketCap = models.CharField('Market Cap', max_length=20, null=True, blank=True)
    adrTso =  models.CharField('Adr Tso', max_length=20, null=True, blank=True)
    ipoYear = models.CharField('IPO Year', max_length=4, null=True, blank=True)
    sector = models.CharField('Sector', max_length=100, null=True, blank=True)
    industry = models.CharField('Industry', max_length=200, null=True, blank=True)
    summaryQuote = models.CharField('Summary Quote', max_length=200, null=True, blank=True)
    exchange = models.CharField('Stock Exchange', max_length=10, null=True, blank=True)
    date_extracted = models.DateTimeField('Date extracted', null=True, blank=True)
    class Meta:
        verbose_name = "Company"
        verbose_name_plural = "Companies"
        ordering = ["name"]
    def __str__(self):
        return self.name+'('+self.symbol+')'

class CompanyFinancials(models.Model):
    company=models.ForeignKey(Company)
    quarter=models.CharField('Quarter', max_length=1)
    year = models.CharField('Year', max_length=4)
    total_assets = models.CharField('Total Assets', max_length=20, null=True, blank=True)
    total_liability = models.CharField('Total Liability', max_length=20, null=True, blank=True)
    current_assets = models.CharField('Current Assets', max_length=20, null=True, blank=True)
    current_liability = models.CharField('Current Liability', max_length=20, null=True, blank=True)
    retained_earnings = models.CharField('Retained Earnings', max_length=20, null=True, blank=True)
    market_capital = models.CharField('Market Capital', max_length=20, null=True, blank=True)
    ebitda = models.CharField('EBITDA', max_length=20, null=True, blank=True)
    sales = models.CharField('Sales', max_length=20, null=True, blank=True)
    stockprice = models.CharField('Last Stock Price', max_length=20, null=True, blank=True)
    date_extracted = models.DateTimeField('Date extracted', null=True, blank=True)
    class Meta:
        verbose_name = "Company Financial"
        verbose_name_plural = "Company Financials"
    def __str__(self):
        return self.company.name+" "+self.quarter+" - "+self.year

class TwitterText(models.Model):
    # Each sentiment related to the companies under analysis and captured from the social media platforms
    twitter_user_id = models.CharField('Social User ID', max_length=255)
    twitter_user_name = models.CharField('Social User Name', max_length=100)
    twitter_text = models.CharField('Social Updates', max_length=1024)
```

```
    twitter_text_id = models.CharField('Message Id', max_length=255)
    twitter_text_timestamp = models.DateTimeField('Message Timestamp')
    twitter_text_keyword = models.CharField('Keyword', max_length=1000)
    twitter_for_training = models.IntegerField('training value', max_length=1, default=9) #9 untrained #0:
trained dataset,1: dev-test dataset, 2: test dataset #4 applied dataset
    twitter_sentiment = models.IntegerField('sentiment', max_length=1, default=2)
    training_user = models.ForeignKey(User, null=True, blank=True)
    class Meta:
        verbose_name = "Twitter Text"
        verbose_name_plural = "Twitter Texts"
    def __str__(self):
        return self.twitter_text+" keyword: "+self.twitter_text_keyword

class CompanyAltmanZscore(models.Model):
    company=models.ForeignKey(Company)
    company_financials = models.ForeignKey(CompanyFinancials)
    zscore = models.CharField("Altman Z-Score", max_length=10, null=True,blank=True)
    date_updated = models.DateTimeField('Date Calculated', auto_now=True)
    class Meta:
        verbose_name = "Altman Score"
        verbose_name_plural = "Altman Scores"
    def __str__(self):
        return self.company.name+" - zscore:"+self.zscore

class CompanySentiment(models.Model):
    company = models.ForeignKey(Company)
    twitter_text = models.ForeignKey(TwitterText)
    sentiment_prob_very_negative = models.CharField('Very Negative Probability', max_length="10",
null=True, blank=True)
    sentiment_prob_negative = models.CharField('Negative Probability', max_length="10", null=True,
blank=True)
    sentiment_prob_neutral = models.CharField('Neutral Probability', max_length="10", null=True,
blank=True)
    sentiment_prob_positive = models.CharField('Positive Probability', max_length="10", null=True,
blank=True)
    sentiment_prob_very_positive = models.CharField('Very Positive Probability', max_length="10",
null=True, blank=True)
    sentiment_root_value = models.CharField('Sentiment', max_length="20", null=True, blank=True)
    date_updated = models.DateTimeField('Date Calculated', auto_now=True)
    class Meta:
        verbose_name = "Company Sentiment"
        verbose_name_plural = "Company Sentiments"
    def __str__(self):
        return self.company+" - tweet: "+self.twitter_text.id+" - sentiment: "+self.sentiment_root_value

class CompanyQuoteHistory(models.Model):
    company=models.ForeignKey(Company)
    date = models.DateField('Quote Date',null=True, blank=True)
    open=models.FloatField('Open', null=True, blank=True)
    high = models.FloatField('High',null=True, blank=True)
    low = models.FloatField('Low', null=True, blank=True)
```

```
    close = models.FloatField('Close', null=True, blank=True)
    volume = models.FloatField('Volume', null=True, blank=True)
    adjs_close = models.FloatField('Adj Close', null=True, blank=True)
    symbol = models.CharField('Symbol', null=True, blank=True, max_length="10")
    class Meta:
        verbose_name = "Company Quote History"
        verbose_name_plural = "Company Quote History"
    def __str__(self):
        return self.company.name+self.date+" trend: "+(self.close - self.open)

class CompanyStocksSentimentHistory(models.Model):
    company=models.ForeignKey(Company)
    symbol = models.CharField('Stock Symbol',max_length="10", null=True, blank=True )
    date = models.DateField('Quote Date',null=True, blank=True)
    tweet_count =   models.IntegerField('Tweets', null=True, blank=True)
    sentiment = models.IntegerField('Sentiment', null=True, blank=True)
    stockopen = models.FloatField('Stock Open', null=True, blank=True)
    stockclose = models.FloatField('Stock Close', null=True, blank=True)
    stockdirection = models.IntegerField('Stock Direction', null=True, blank=True)
    class Meta:
        verbose_name = "Company Tweet Quote "
        verbose_name_plural = "Company Tweet Quotes"
    def __str__(self):
        return self.company.name+" - tweet id: "+self.twitter_text.id+" "+self.date+" trend: "+(self.close -
self.open)
```

Appendix B: Tweets Collector Source Code

```
__author__ = 'Tarek Hoteit'
#the purpose of this script is to first capture all tweets on publicly held firms and store them in the database
import pandas as pd
import tweepy
from tweepy import StreamListener
from twitterSentiment import models
from datetime import datetime
from pytz import timezone
import json, time, sys
from numpy import random
import re
from django.contrib.auth.models import User
def timeupdate(twitterdate):
    # method to return a django-supported time from twitter-based time entry
    # input comes in the following fashion: Tue Jul 02 14:33:59 +0000 2013
    # return 2013-06-18 18:23:22-04:00
    central = timezone('US/Central')
    return central.localize(datetime.strptime(twitterdate, '%a %b %d %H:%M:%S +0000 %Y'))
class TwitterListener(StreamListener):

  def __init__(self, api = None, fprefix = 'streamer'):
    self.api = api or API()
    self.counter = 0

  def on_data(self, data):
    global tweetsmax
    global tweetscount
    global alltweetscount
    global iterationcount
    print ("tweets count:",tweetscount, "/",tweetsmax,". Iteration: ",iterationcount," Total tweets: ",
alltweetscount)
    if (tweetsmax == tweetscount):
      tweetscount=0
      iterationcount = iterationcount+1
      return False
    else:
      tweetscount = tweetscount+1
      alltweetscount = alltweetscount +1
    try:
      tweet = json.loads(data) #convert twitter stream in json into Python dictionary
      if isinstance(tweet, dict):
        if tweet['user']['lang'] != 'en':
          return
        else:
          print ("tweet: ", tweet['text'])
          TwitterDatabase(tweet)
    except:
      print ("Error in Twitter listener. Error message:", sys.exc_info())
    return
```

```
    def on_limit(self, track):
        print(">> limit")
        return
    def on_error(self, status_code):
        print(">>> error: ", str(status_code) + "\n")
        return
    def on_timeout(self):
        print(">>> timeout Sleeping for 60 seconds...\n")
        time.sleep(60)
        return

def CompanyNamesbyStocks(tweet):
    #extract the company behind each stock as the latter is mentioned in a tweet
    try:
        match = re.findall("\$\w+",tweet) #captures the stock symbols
        global sp500_companies
        return (",".join(sp500_companies.Name.get(symbol,symbol)+"("+symbol+")" for symbol in
match))
    except:
        return ("na")
def TwitterDatabase(tweet):
    ## take Twitter data in jsonformat and insert it into the database
    try:
        global systemid
        aTweet = models.TwitterText(twitter_user_id=tweet['user']['id'],
                                twitter_user_name=tweet['user']['screen_name'],
                                twitter_text=tweet['text'],
                                twitter_text_id=tweet['id'],
                                twitter_text_timestamp=timeupdate(tweet['created_at']),
                                twitter_text_keyword=CompanyNamesbyStocks(tweet['text']),
                                training_user_id=systemid)
        aTweet.save()
    except:
        print ("Error in Django insert tweet. error message:", sys.exc_info())
    return

def PickRandomCompanies(allstocks, stocks_count):
    #since Twitter does not allow an exhaustive keyword searcj
    stocks_random = random.choice(allstocks, stocks_count)
    return (",".join("{0}".format(stocks.strip()) for stocks in stocks_random))
def TwitterStreaming(stocks):
    ## twitter authentication keys

    consumer_key       = "yoWOau00G19Q81WKeVZ6g60zU"
    consumer_secret    = "A0rJ4XlMndHv2xTeQlA2t7N9thBr3FDRu6vkrCy5ab7KAiKmNB"
    access_token       = "16859687-bt1jbTlHUXO39n114gWEpg24VlKZQVbaF4AgXs4ha"
    access_token_secret = "kLMge9f3GypNwHv6N9uMCuUdLS7kr5gfR5lzTXEmwMyfi"
    global keywords
    keywords = [stocks]
    try:
```

```
        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        api = tweepy.API(auth)
        listener = TwitterListener(api, "test")
        print ("Begin Twitter streaming for ", stocks)
        stream = tweepy.Stream(auth, listener)
        stream.filter(track=[stocks])
    except:
        print ("Error in Twitter streaming",sys.exc_info())
    return True
tweetsmax = 500
tweetscount = 1
alltweetscount = 1
iterationcount = 1
def run():
    try:
        systemid = User.objects.get(username="system").id
        all_companies = pd.DataFrame.from_csv("data/allpubliccomp.csv", index_col=['Symbol']) ##used
a script from the internet that scrapes  spindices
        stocks = all_companies.index
        infinite = True
        while (infinite == True):
            TwitterStreaming(PickRandomCompanies(stocks,200))
    except:
        print ("Error occured: "+sys.exc_info())
```

Appendix C: Tweets Sentiment Classifier

```
__author__ = 'Tarek Hoteit'
from twitterSentiment.models import Company, TwitterText, CompanySentiment
import subprocess
import re
import sys
import mysql.connector
from django.contrib.auth.models import User
from django.core import exceptions
corenlpcp = "/home/tarek/phd/nlp/stanford-corenlp-full-2014-10-31/*"
model = "/home/tarek/phd/nlp/stanford-corenlp-full-2014-10-31/model.ser.gz"
sentimentlevel = {'Very negative': 0,'Negative' :1, 'Neutral' : 2, 'Positive' : 3, 'Very positive' :4 }
sentimentregstr = re.compile(r"\s+(Positive|Negative|Neutral|Very negative|Very
positive)\n\(.*?\n\s+0\:\s+(\w.*?)\s+(\w.*?)\s+(\w.*?)\s+(\w.*?)\s+(\w.*?)\n",re.DOTALL)
companykeywordreg = re.compile(r"\((.*?)\)", re.DOTALL)
def sentiments(text):
    # function that takes text, runs java corenlp and return output in string format
    text = subprocess.Popen(['echo', text], stdout=subprocess.PIPE,)
    sentiment = subprocess.Popen(['java','-cp',corenlpcp,'-mx5g',
                                  'edu.stanford.nlp.sentiment.SentimentPipeline',
                                  '-sentimentModel', model, '-stdin', '-output',
                                  'ROOT,PROBABILITIES'], stdin=text.stdout,
                                  stdout=subprocess.PIPE,)
    text.stdout.close()
    output = sentiment.communicate()[0]
    #sentiment.stdout.close()
    text.kill()
    return (output.decode("utf-8"))  # decode is necessary for the pattern match
def update_tweet(id, sentiment, user):
  # update tweet with overall sentiment
    try:
        atweet = TwitterText.objects.get(id=id)
        atweet.twitter_sentiment = sentiment
        atweet.training_user_id = user
        atweet.save()
        print("tweet", id, "updated.")
        return True
    except:
        print("errors occured", sys.exc_info())
        return False
def run():
    try:
      systemid = User.objects.get(username="system").id  # this account is used to update Twitter table
      tweets = TwitterText.objects.filter(twitter_for_training=2)  # get all tweets that are ready to be
trained.
      tweets_count = tweets.count()
      if tweets_count > 0:
        for twt in range (0, tweets_count):
          #do sentiment analysis
```

```
            tweet = tweets[twt]
            tweet_text = str(tweet.twitter_text).encode()
            tweet_id = tweet.id
            print("tweet id: ", tweet_id, "tweet text:", tweet_text)
            sentiment = sentiments(tweet_text)
            sentimentreg = sentimentregstr.match(str(sentiment))
            if sentimentreg is not None:
              print ("sentiment: ", sentimentreg.group(1))
              print ("very negative: ", sentimentreg.group(2))
              print ("negative: ", sentimentreg.group(3))
              print ("neutral: ", sentimentreg.group(4))
              print ("positive: ", sentimentreg.group(5))
              print ("very positive: ", sentimentreg.group(6))
            else:
              print ("no pattern match")
            #pick the companies to store the sentiment
            tweet_companies_symbols = companykeywordreg.findall(tweet.twitter_text_keyword)
            if tweet_companies_symbols is not None:
              for company_symbol in tweet_companies_symbols:
                #get company
                print(company_symbol)
                try:
                  company = Company.objects.get(symbol=company_symbol)
                except Company.DoesNotExist:
                  company = None
                if company is not None:
                  companysentiment =
CompanySentiment(company_id=company.id,twitter_text_id=tweet_id,
                                  sentiment_root_value=sentimentreg.group(1),
                                  sentiment_prob_very_negative=sentimentreg.group(2),
                                  sentiment_prob_negative=sentimentreg.group(3),
                                  sentiment_prob_neutral=sentimentreg.group(4),
                                  sentiment_prob_positive=sentimentreg.group(5),
                                  sentiment_prob_very_positive=sentimentreg.group(6))
                  companysentiment.save()
                  print("added tweet sentiment for company:",company_symbol)
                else:
                  print("symbol ",company_symbol," does not exist in the database")
            update_tweet(tweet_id,sentimentlevel[sentimentreg.group(1)],systemid)
          TwitterText.objects.filter(twitter_for_training="2").update(twitter_for_training="4") #flag tweets
as classified
        else:
          print ("no tweets remmaining to be classified for sentiments")
    except mysql.connector.Error as err:
      print("MySQL error: ", err.msg)
    except:
      print("error occured", sys.exc_info())
```

Appendix D: Company Stock & Sentiment History

```
__author__ = 'Tarek Hoteit'
import pandas as pd
from django.db.models import Q
##load the django model
from twitterSentiment.models import Company, CompanyFinancials, TwitterText,
CompanySentiment, CompanyQuoteHistory,CompanyStocksSentimentHistory
from datetime import date, timedelta
from time import strftime
def company_quote_calculations(companyid, adate):
    try:
        company_stock = CompanyQuoteHistory.objects.filter(company_id=companyid, date=adate)
        if company_stock.count() >0:
            company_stock_values = company_stock.values()
            stockopen = company_stock_values[0]['open']
            stockclose = company_stock_values[0]['close']
            if stockopen > stockclose:
                direction = -1
            elif stockopen < stockclose:
                direction = 1
            else:
                direction = 0
            return (stockopen, stockclose,direction)
        else:
            return (None, None, None)
    except:
        print ("quote error occured for company:", companyid, "error:",sys.exc_info())
        return (None, None, None)
def company_sentiment_calculations(companyid, date):
    try:
        company_sentiments =
CompanySentiment.objects.filter(company_id=companyid,twitter_text__twitter_text_timestam
p__startswith=date).values()
        if company_sentiments.count() >0:
            company_sentiments_list = pd.DataFrame(list(company_sentiments))
            averages =
pd.Series([company_sentiments_list['sentiment_prob_very_negative'].astype(float).sum(),
                company_sentiments_list['sentiment_prob_negative'].astype(float).sum(),
                company_sentiments_list['sentiment_prob_neutral'].astype(float).sum(),
                company_sentiments_list['sentiment_prob_positive'].astype(float).sum(),
                company_sentiments_list['sentiment_prob_very_positive'].astype(float).sum()],
                index=[0,1,2,3,4])
```

```
        return (averages.idxmax(axis=1), company_sentiments.count()) #pick the index with the
maximum probability
      else:
        return (None,None)
  except:
      print ("sentiment error occured for company:", companyid, "error:",sys.exc_info())
      return (None,None)
def daterange(start_date, end_date):
  for n in range(int ((end_date - start_date).days)):
      yield start_date + timedelta(n)
def run():
  start_date = date(2014, 12, 6)
  end_date = date(2015, 1, 5)
  companies = Company.objects.filter(~Q(marketCap=0.0))
  row = 0
 # df = pd.DataFrame(columns=('Date', 'Company', 'Symbol', 'Sentiment', 'Stock Open', 'Stock
Close', 'Direction'))
  for company in companies:
      for single_date in daterange(start_date, end_date):
        querydate = strftime("%Y-%m-%d", single_date.timetuple())
        sentiment,tweet_count = company_sentiment_calculations(company.id,querydate)
        stock_open,stock_close,stock_direction  =
company_quote_calculations(company.id,querydate)
        aCompanyStocksSentimentHistory =
CompanyStocksSentimentHistory(company_id=company.id,
                                                    symbol=company.symbol,
                                                    date=querydate,
                                                    tweet_count=tweet_count,
                                                    sentiment = sentiment,
                                                    stockopen = stock_open,
                                                    stockclose = stock_close,
                                                    stockdirection = stock_direction
                                                    )
        aCompanyStocksSentimentHistory.save()
        print (row,":company",company.symbol," date,", querydate," complete.")
        row = row +1
```

Appendix E: Company Financials Yahoo Extract

```
__author__ = 'Tarek Hoteit'
import pandas as pd
import urllib
import time
import sys
#script that loops over all the public companies in allpubliccomp.csv, extracts the financials using YQL
then store the
#the results in a new csv file
#note the financials are annual
count = 1
total = 1
def financial_variables(stock):
    #function to query Yahoo YQL and return the financial data for a stock based on stock symbol.
    #the variables are selected according to Altman Z-Score
    stock = stock.strip('$')
    total_assets=""
    total_liability=""
    current_assets=""
    current_liability=""
    retained_earnings=""
    market_capital = ""
    ebitda=""
    sales=""
    stockprice=""
    try:
        baseurl = "https://query.yahooapis.com/v1/public/yql?"
        #extract data from balance sheet
        yql_bs_query = "select * from yahoo.finance.balancesheet where symbol in ('"+stock+"') and
timeframe='annual'"
        yql_bs_url = baseurl + urllib.parse.urlencode({'q':yql_bs_query}) +
"&format=json&diagnostics=true&env=store%3A%2F%2Fdatatables.org%2Falltableswithkeys&c
allback="
        bs_json = pd.io.json.read_json(yql_bs_url)
        if bs_json["query"]["results"]["balancesheet"]["statement"] is not None:
            try:

total_assets=bs_json["query"]["results"]["balancesheet"]["statement"][0]['TotalAssets']['content']
            except:
                total_assets="0"
            try:

total_liability=bs_json["query"]["results"]["balancesheet"]["statement"][0]['TotalLiabilities']['conte
nt']
            except:
                total_liability="0"
            try:

current_assets=bs_json["query"]["results"]["balancesheet"]["statement"][0]['TotalCurrentAssets']['c
ontent']
```

```
        except:
           current_assets="0"
        try:

current_liability=bs_json["query"]["results"]["balancesheet"]["statement"][0]['TotalCurrentLiabiliti
es']['content']
        except:
           current_liability = "0"
        try:

retained_earnings=bs_json["query"]["results"]["balancesheet"]["statement"][0]['RetainedEarnings']
['content']
        except:
           retained_earnings="0"
         #extract data from income statement
        yql_is_query = "select * from yahoo.finance.incomestatement where symbol in ('"+stock+"')
and timeframe='annual'"
        yql_is_url = baseurl + urllib.parse.urlencode({'q':yql_is_query}) +
"&format=json&diagnostics=true&env=store%3A%2F%2Fdatatables.org%2Falltableswithkeys&c
allback="
        is_json = pd.io.json.read_json(yql_is_url)
        try:

ebitda=is_json["query"]["results"]["incomestatement"]["statement"][0]['EarningsBeforeInterestAn
dTaxes']['content']
        except:
           ebitda="0"
        try:

sales=is_json["query"]["results"]["incomestatement"]["statement"][0]['TotalRevenue']['content']
        except:
           sales="0"
        #extract data from finance quotes
        yql_qt_query = "select * from yahoo.finance.quotes where symbol in ('"+stock+"')"
        yql_qt_url = baseurl + urllib.parse.urlencode({'q':yql_qt_query}) +
"&format=json&diagnostics=true&env=store%3A%2F%2Fdatatables.org%2Falltableswithkeys&c
allback="
        qt_json = pd.io.json.read_json(yql_qt_url)
        try:
           stockprice = qt_json["query"]["results"]["quote"]["LastTradePriceOnly"]
        except:
           stockprice = "0"
        try:
           market_capital = qt_json["query"]["results"]["quote"]['MarketCapitalization']
        except:
           market_capital = "0"
  except:
     print ("error on ", stock, ". Error message",sys.exc_info())
  global count
  count = count +1
  global total
```

```
    print ("\n",count,"/",total,":",stock, "|", total_assets, "|", total_liability, "|",
current_assets,"|",current_liability,"|",retained_earnings,"|",market_capital,"|",ebitda,"|",sales,"|",
stockprice)
    return ([total_assets, total_liability,
current_assets,current_liability,retained_earnings,market_capital,ebitda, sales, stockprice])
def run():
    firms_fin_location = "data/allpubliccomp.csv"
    firms = pd.read_csv(firms_fin_location)
    total = len(firms.index)
    firms['altman_variables']=firms['Symbol'].apply(financial_variables)
    firms.to_csv('data/altman_results4.csv')
```

Appendix F: Company Historical Quotes Yahoo Extract

```
__author__ = 'Tarek Hoteit'
import pandas as pd
import urllib
import time
import sys
from twitterSentiment.models import Company, CompanyQuoteHistory
#script that loops over every company  and then retrieve the start quote and end quote for the time of the
tweets
def company_quote_history_data(company_id, stock, startDate, endDate):
    #function to query Yahoo YQL and return the financial data for a stock based on stock symbol.
    #the variables are selected according to Altman Z-Score
    stock = stock.strip('$')
    baseurl = "https://query.yahooapis.com/v1/public/yql?"
    #extract data from balance sheet
    yql_bs_query = "select * from yahoo.finance.historicaldata where symbol = '"+stock+"' and
startDate= '"+startDate+"' and endDate = '"+endDate+"'"
    yql_bs_url = baseurl + urllib.parse.urlencode({'q':yql_bs_query}) +
"&format=json&diagnostics=true&env=store%3A%2F%2Fdatatables.org%2Falltableswithkeys&c
allback="
    as_json = pd.io.json.read_json(yql_bs_url)
    if as_json["query"]["results"] is not None:
      if as_json["query"]["results"]["quote"] is not None:
        for quote in as_json["query"]["results"]["quote"]:
          print(quote)
          if quote is not None:
            try:
              aCompanyQuote = CompanyQuoteHistory(company_id = company_id,
                                date = quote["Date"],
                                low = quote["Low"],
                                high = quote["High"],
                                close = quote["Close"],
                                open = quote["Open"],
                                volume = quote["Volume"],
                                adjs_close = quote["Adj_Close"],
                                symbol = quote["Symbol"])
              aCompanyQuote.save()
            except:
              print ("error with stock:", stock, "error:",sys.exc_info())
          else:
            print('done')
    return True
def run():
    get_companies = Company.objects.all()
    for company in get_companies:
        company_quote_history_data(company.id, company.symbol, '2014-12-06', '2015-1-05')
```

Appendix G: Altman Z-Score Calculations

```
__author__ = 'Tarek Hoteit'
### the script will retreive the financials from the database for each of the firms, calculate altman zscore
### and upload it into the database
import sys
import re
from twitterSentiment.models import Company, CompanyFinancials, CompanyAltmanZscore
# add some more to powers as necessary
import sys
def largenumbers(numstring):
  powers = {'B': 10 ** 9, 'M': 10 ** 6, 'T': 10 ** 12}
  try:
    numberregex = re.match("(\w.*?)([B|M|T])", numstring)
    decimal = numberregex.group(1)
    large = numberregex.group(2)
    return float(decimal) * powers[large]
  except:
    print (sys.exc_info())
    return numstring
def run():
  companies = Company.objects.filter()
  if companies is not None:
    for company in companies:
      try:
        companyFinancials = CompanyFinancials.objects.get(company__id = company.id, year="2014",
quarter="3")
        current_assets = int(companyFinancials.current_assets)
        current_liability = int(companyFinancials.current_liability)
        working_capital = current_assets - current_liability
        total_assets = int(companyFinancials.total_assets)
        total_liability = int(companyFinancials.total_liability)
        retained_earnings = int(companyFinancials.retained_earnings)
        ebitda = int(companyFinancials.ebitda)
        market_capital = largenumbers(companyFinancials.market_capital)
        stockprice = float(companyFinancials.stockprice)
        sales = int(companyFinancials.sales)
        altmanX1 = working_capital / total_assets
        altmanX2 = retained_earnings / total_assets
        altmanX3 = ebitda / total_assets
        altmanX4 = market_capital /total_liability
        altmanX5 = sales / total_assets
        z=(1.2*altmanX1) + (1.4*altmanX2) + (3.3*altmanX3) + (0.6*altmanX4) + (0.999*altmanX5)
        print("current assets:", current_assets)
        print("current liability:", current_liability)
        print("total assets:", total_assets)
        print("total liability:", total_liability)
        print("retained earnings:", retained_earnings)
        print("ebitda:", ebitda)
        print("market capital:", market_capital)
        print("stock price:", stockprice)
```

```
        print("sales:", sales)
        print("x1:", altmanX1, "x2:",altmanX2, "x3:", altmanX3, "x4:",altmanX4, "x5:",altmanX5)
        print("Z Score: ", z)
        companyAltmanZscore = CompanyAltmanZscore(company_id=company.id,
company_financials_id=companyFinancials.id, zscore=round(z,3))
        companyAltmanZscore.save()
        #sudo wait = input("press ENTER to continue")
    except:
        z=0
        print("error:", sys.exc_info())
```

Appendix H: Sentiment Training Source Code

Views.py

```
def pick_trainingDatum(request):
  """
  the purpose of this function is pick a Twitter training message in order to be classified
  between 0 and 4 for sentiment
  :param request:
  :return:
  """
  if request.method=='POST':
    form = TrainingDatumForm(request.POST)
    if form.is_valid():
      tweet = TwitterText.objects.get(id = request.POST['tweet_id'])
      tweet.twitter_for_training = "0"
      tweet.training_user = request.user
      tweet.twitter_sentiment = form.cleaned_data['twitter_sentiment']
      tweet.save()
      return HttpResponseRedirect(reverse_lazy('trainingDS'))
    return render (request,  'twitterSentiment/trainingds.html',
      {'form' : form})
  else:
    get_untrained_sentiments = TwitterText.objects.filter(twitter_for_training="9")
    num_untrained_sentiments = get_untrained_sentiments.count()
    #pick a tweet randomly
    if num_untrained_sentiments > 0:
      atweet_rnd_loc = random.randint(0, num_untrained_sentiments-1)
      pick_object = get_untrained_sentiments[atweet_rnd_loc]
      tweet_id = pick_object.id
      form = TrainingDatumForm(instance=pick_object,
                  initial={'twitter_for_training':"0", 'training_user': request.user})
      #note twitter_for_training = 0 means the tweet is being trained
      return render(request, 'twitterSentiment/trainingds.html',
            {'form': form, 'tweet_id': tweet_id, 'message': pick_object.twitter_text})
    else:
      return render(request, 'twitterSentiment/done.html')
```

forms.py

```
__author__ = 'tarek'
from django import forms
from django.forms import ModelForm
from twitterSentiment.models import TwitterText
class CompaniesImport(forms.Form):
  companiesList = forms.CharField( widget=forms.Textarea(attrs={'rows': 20, 'cols': 30}))
CHOICES = (('0', 'very negative',), ('1', 'negative'), ('2', 'neutral'), ('3', 'positive'), ('4', 'very positive'))
class TrainingDatumForm(ModelForm):
  class Meta:
    model = TwitterText
    fields = ('twitter_sentiment',)
```

```
    widgets = {
      'twitter_sentiment': forms.RadioSelect(choices=CHOICES),
    }
```

Trainingds.html

```
{% block content %}
<h3>{{ message }}</h3>
<form action="{% url 'trainingDS'%}" method="post">
   {% csrf_token %}
   {% if form.errors %}
   {% for field in form %}
      {% for error in field.errors %}
         <div class="alert alert-error">
            <strong>{{ error|escape }}</strong>
         </div>
      {% endfor %}
   {% endfor %}
   {% for error in form.non_field_errors %}
      <div class="alert alert-error">
         <strong>{{ error|escape }}</strong>
      </div>
   {% endfor %}
{% endif %}
   {{  form.as_table }}
   <button type="submit" class="btn btn-success">Train Tweet {{ tweet_id }}</button>
      <button type="button" class="btn btn-default"
onclick="window.location.reload()">Skip</button>
   <input type="hidden" name="tweet_id" value="{{ tweet_id }}"/>
</form>
{% endblock %}
```

Appendix I: Data Analysis Source Code & Results

In [1]:

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
import pylab as pl
import numpy as np
import scipy as sc
from twitterSentiment.models import (CompanyStocksSentimentHistory,Company, CompanyFinancials,
TwitterText,
CompanyAltmanZscore, CompanySentiment, CompanyKeyStats)
from django.db.models import Q
```

In [2]:

```
def company_sentiment_calculations(companyid):
    try:
        company_sentiments = CompanySentiment.objects.filter(company_id=companyid).values()
        if company_sentiments.count() >0:
            company_sentiments_list = pd.DataFrame(list(company_sentiments))
            averages =
pd.Series([company_sentiments_list['sentiment_prob_very_negative'].astype(float).sum(),
              company_sentiments_list['sentiment_prob_negative'].astype(float).sum(),
              company_sentiments_list['sentiment_prob_neutral'].astype(float).sum(),
              company_sentiments_list['sentiment_prob_positive'].astype(float).sum(),
              company_sentiments_list['sentiment_prob_very_positive'].astype(float).sum()],
              index=[0,1,2,3,4])
            return (averages.idxmax(axis=1)) #pick the index with the maximum probability
        else:
            return
    except:
        print ("sentiment error occured for company:", companyid, "error:",sys.exc_info())
        return (-1)


def company_zscore_calculations(companyid):
    try:
        company_zscore =
CompanyAltmanZscore.objects.all().filter(company_id=companyid).values('zscore')
        r = list(company_zscore[:1])
        if r:
            return r[0]['zscore']
        return

    except:
        print ("zscore error occured for company:", companyid, "error:",sys.exc_info())
        return (-1)
```

```
def build_data_file():
    ##use this method to create the companysentimentzscore for the analysis
    companylist = pd.DataFrame(list(Company.objects.filter(~Q(marketCap=None).values())))
    print("total companies listed",companylist.count())
    companylist['sentiment']=companylist[['id']].apply(company_sentiment_calculations, axis=1)
    companylist['altman']=companylist[['id']].apply(company_zscore_calculations, axis=1)
    companylistix = companylist.set_index(['id'])
    companylistix.to_csv("companysentimentszscore.csv")
```

Altman *Z*-Score for Sampled Firms (Table 14)

In [3]:

```
companyinfo= pd.read_csv("companysentimentszscore.csv")
print ("Total Firms: ", companyinfo.count())
altmanzscore = companyinfo.altman
distress = companyinfo.altman[companyinfo.altman<=1.8]
safe = companyinfo.altman[companyinfo.altman>=3.0]
neutral = companyinfo.altman[companyinfo.altman.between(1.8,3)]
print ("****Distress:", distress.count(), ",Safe:",safe.count(), ",Neutral:",neutral.count())
print ("\n\n****Description for Financially-Distressed Firms Altman < 1.8")
print (distress.describe())
print ("\n\n****Description for Neutral Firms Altman > 1.8 & < 3.0")
print (neutral.describe())
print ("\n\n****Description for Safe Firms >= 3.0")
print (safe.describe())
```

```
Total Firms:  id              5787
adrTso          5787
date_extracted    5787
exchange         5787
industry         5787
ipoYear          5787
lastSale         5787
marketCap        5787
name             5787
sector           5787
summaryQuote     5787
symbol           5787
sentiment        3985
altman           3313
dtype: int64
****Distress: 1032 ,Safe: 1690 ,Neutral: 591


****Description for Financially-Distressed Firms Altman < 1.8
count    1032.000000
mean       -1.961134
std         9.818075
min      -157.210000
25%        -1.091750
```

50%   0.663500
75%   1.255500
max   1.798000
Name: altman, dtype: float64

****Description for Neutral Firms Altman > 1.8 & < 3.0**
count   591.000000
mean   2.374411
std   0.337121
min   1.801000
25%   2.092500
50%   2.375000
75%   2.663500
max   2.997000
Name: altman, dtype: float64

****Description for Safe Firms >= 3.0**
count   1690.000000
mean   9.925705
std   20.506765
min   3.001000
25%   4.009500
50%   5.404500
75%   8.883500
max   454.079000
Name: altman, dtype: float64

Descriptive Statistics of Nonfinancial Firms

In [4]:

```
companies = companyinfo[~(companyinfo.sentiment.isnull()) & ~(companyinfo.altman.isnull()) &
~(companyinfo.sector.isin(['Finance']))]
print (companies.describe())
```

	id	lastSale	marketCap	sentiment	altman
count	2618.000000	2618.000000	2.618000e+03	2618.000000	2618.000000
mean	3118.103132	36.625553	9.044559e+09	2.158136	4.643611
std	1921.975390	51.212159	3.000794e+10	0.667646	12.866519
min	6.000000	0.100000	2.102689e+06	0.000000	-153.814000
25%	1490.750000	6.960000	2.736628e+08	2.000000	1.372000
50%	2955.500000	22.610000	1.333308e+09	2.000000	3.035500
75%	4747.500000	49.590000	5.094740e+09	2.000000	5.489500
max	6674.000000	1135.970000	6.744566e+11	4.000000	269.368000

In [5]:

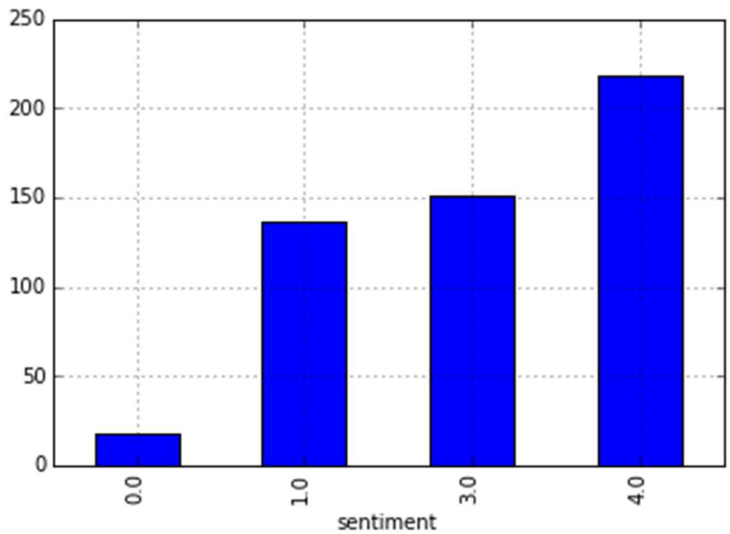
```
def altmanscale(value):
    if value <=1.8:
        return -1
```

```
    elif (value>1.8) & (value<3.0):
        return 0
    else:
        return 1
def sentimentscale(value):
    if value <2:
        return -1
    elif value ==2:
        return 0
    else:
        return 1
def sentimentsstr(value):
    if value <2:
        return "negative"
    elif value ==2:
        return "neutral"
    else:
        return "positive"

def stockmovementstr(value):
    if value ==1:
        return "negative"
    elif value ==0:
        return "neutral"
    else:
        return "positive"


##sentiments plots excluding neutral
sentiments = companies.groupby('sentiment')['sentiment']
print ("\n**** Sentiments Breakdown",sentiments.count())

companies.altmanscale = companies.altman.apply(altmanscale)
companies.sentimentscale = companies.sentiment.apply(sentimentscale)
print ("\n****Sentiments by Main
Categories",companies.sentimentscale.groupby(companies.sentimentscale).count())
print ("\n**** Sentiments Breakdown Excluding Neutral")
print(companies[~(companies.sentiment==2)].groupby('sentiment')['sentiment'].count().plot(kind="bar"))
```

****** Sentiments Breakdown sentiment**
0 18
1 137
2 2094
3 151
4 218
Name: sentiment, dtype: int64

******Sentiments by Main Categories sentiment**
-1 155
0 2094
1 369

Name: sentiment, dtype: int64



<u>Research Hypothesis 1</u>

In [6]:

```
from scipy.stats.mstats import normaltest
print ("\n****Test for Normality:")
print ("normal test for sentiment",normaltest(companies.sentiment))
print ("normal test for altman",normaltest(companies.altman))
print ("a P close to zero indidicated that the data are not normal")
```

******Test for Normality:**
normal test for sentiment (679.21183752063507, 3.2435713873994668e-148)
normal test for altman (3404.2039999564081, 0.0)
a P close to zero indidicated that the data are not normal

<u>Spearman Correlation Analysis</u>

In [7]:

```
print ("Spearman correlation between sentiment and altman
scales:",sc.stats.spearmanr(companies.sentimentscale, companies.altmanscale))
print ("a P>0.05 indicates no statistically significant correlation")
```

Spearman correlation between sentiment and altman scales: (-0.030236069863026162, 0.12193910694076453)
a P>0.05 indicates no statistically significant correlation

In [8]:

```
print ("correlation analysis by sector")
```

```
print ("For basic industries:","(",companies.sector[companies.sector=="Basic Industries"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Basic
Industries"].rank(method="average", ascending=True),
    companies.sentimentscale[companies.sector=="Basic Industries"].rank(method="average",
ascending=True)))
print ("For capital goods: ","(",companies.sector[companies.sector=="Capital Goods"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Capital Goods"],
            companies.sentimentscale[companies.sector=="Capital Goods"]))
print ("Consumer Durables: ","(",companies.sector[companies.sector=="Consumer Durables"].count(),"
firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Consumer Durables"],
            companies.sentimentscale[companies.sector=="Consumer Durables"]))
print ("Consumer NonDurables: ","(",companies.sector[companies.sector=="Consumer
Nondurables"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Consumer Nondurables"],
            companies.sentimentscale[companies.sector=="Consumer Nondurables"]))
print ("Consumer Services: ","(",companies.sector[companies.sector=="Consumer Services"].count(),"
firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Consumer Services"],
            companies.sentimentscale[companies.sector=="Consumer Services"]))
print ("Energy","(",companies.sector[companies.sector=="Energy"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Energy"],
companies.sentimentscale[companies.sector=="Energy"]))
print ("Healthcare","(",companies.sector[companies.sector=="Health Care"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Health Care"],
companies.sentimentscale[companies.sector=="Health Care"]))
print ("Miscellaneous","(",companies.sector[companies.sector=="Miscellaneous"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Miscellaneous"],
companies.sentimentscale[companies.sector=="Miscellaneous"]))
print ("Public Utilities","(",companies.sector[companies.sector=="Public Utilities"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Public Utilities"],
companies.sentimentscale[companies.sector=="Public Utilities"]))
print ("Technology","(",companies.sector[companies.sector=="Technology"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Technology"],
companies.sentimentscale[companies.sector=="Technology"]))
print ("Transportation","(",companies.sector[companies.sector=="Transportation"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="Transportation"],
companies.sentimentscale[companies.sector=="Transportation"]))
print ("N/a","(",companies.sector[companies.sector=="n/a"].count()," firms)",
    sc.stats.spearmanr(companies.altmanscale[companies.sector=="n/a"],
companies.sentimentscale[companies.sector=="n/a"]))
correlation analysis by sector
```

For basic industries: (233 firms) (-0.057604509123855778, 0.38141460299857033)
For capital goods: (253 firms) (0.066972348438911422, 0.28860940659082662)
Consumer Durables: (100 firms) (0.055661386982693427, 0.58229018517890074)
Consumer NonDurables: (170 firms) (-0.086183593701254654, 0.26378632495570026)
Consumer Services: (397 firms) (-0.020205811438471499, 0.68814813451527768)
Energy (212 firms) (-0.21930040287038272, 0.0013117767964922885)
Healthcare (427 firms) (-0.04725238360919258, 0.33000480612483729)
Miscellaneous (102 firms) (-0.048251155119557368, 0.63009898687019883)

Public Utilities (138 firms) (-0.026246287152440182, 0.75993037837830979)
Technology (499 firms) (-0.053218159188588389, 0.23536046951436518)
Transportation (81 firms) (-0.047156968264577326, 0.6759114012086983)
N/a (6 firms) (-0.33333333333333331, 0.51851851851851827)

<u>Research Hypothesis 2</u>

In [9]:

```
print ("****correlation analysis for financially distressed firms:")
print (sc.stats.spearmanr(companies.sentimentscale[companies.altmanscale==-1],
companies.altman[companies.altmanscale==-1]))
print ("p>0.05 indiciates no significance")
```

******correlation analysis for financially distressed firms:**
(-0.058611960064478115, 0.092492396411079425)
p>0.05 indiciates no significance

<u>Research Hypothesis 3</u>

In [10]:

```
dailydata = pd.DataFrame(list(CompanyStocksSentimentHistory.objects.filter(~Q(sentiment=None) &
~Q(stockdirection=None)).values()))
dailydata.sentimentsstr = dailydata.sentiment.apply(sentimentsstr)
dailydata.stockmovementstr = dailydata.stockdirection.apply(stockmovementstr)
dailydata.sentimentscale = dailydata.sentiment.apply(sentimentscale)
#print ("Dataset on Sentiment excluding days with no sentiments or no stock movements:\n",
#       dailydata.sentimentsstr.groupby([dailydata.date,dailydata.sentimentsstr]).count())
#print ("Dataset on Stock Movement excluding days with no sentiments or no stock movements:\n",
#       dailydata.stockmovementstr.groupby([dailydata.date,dailydata.stockmovementstr]).count())
print ("total tweets with sentiments and stock value: ", dailydata.tweet_count.sum())
print ("correlation between sentiment and stock direction:",sc.stats.spearmanr(dailydata.sentiment,
dailydata.stockdirection))
print ("correlation between sentiment scale and stock direction:",sc.stats.spearmanr(dailydata.sentiment,
dailydata.stockdirection))
print ("correlation between tweets count and stock direction:",sc.stats.spearmanr(dailydata.tweet_count,
dailydata.stockdirection))


dailydata =
pd.DataFrame(list(CompanyStocksSentimentHistory.objects.filter(~Q(sentiment=None)).values()))
print ("total tweets with sentiments: ", dailydata.tweet_count.sum())
print ("correlation between sentiment and tweet count:",sc.stats.spearmanr(dailydata.sentiment,
dailydata.tweet_count))
dailydata.sentimentscale = dailydata.sentiment.apply(sentimentscale)
print ("correlation between distress/safe stocks and stock direction",
    sc.stats.spearmanr(dailydata.sentimentscale[dailydata.sentimentscale != 2], dailydata.stockdirection))
```

total tweets with sentiments and stock value: 35726
correlation between sentiment and stock direction: (-0.0050123386944458297, 0.67249850944176182)

correlation between sentiment scale and stock direction: (-0.0050123386944458297, 0.67249850944176182)
correlation between tweets count and stock direction: (0.029289761020337871, 0.01348486080620942)
total tweets with sentiments: 66038
correlation between sentiment and tweet count: (-0.12071759517158794, 2.3157593866886628e-35)
correlation between distress/safe stocks and stock direction (0.014313482237576216, 0.14261916055802118)

Appendix J: Instruments & Citations Permissions

1) permission to use financial data using Yahoo

Yes, you may.

Regards,

Brian Coleman
Account Management, EDGAR® Online, a division of R.R. Donnelley & Sons Company

On Mon, Jan 5, 2015 at 2:55 PM, Case Notification <noreply@salesforce.com> wrote:

Case Number: 00150214

Web Name: Tarek Hoteit
Webemail:
Web Phone:
Company: Walden University

Account:

This case has been assigned to you by EOL Admin Admin.

Description:

Dear Sir/Madam,
I am made aware that Yahoo Finance posts financials from Edgar online. I am currently pursing a dissertation at Walden University where I would like to incorporate financial information about public companies for the sole purpose of the dissertation. Please let me know if I can have your permission to use Yahoo Finance financial information for my academic study.

2) Permission to use research paradigm figure
Keith Towndrow via waldenu.edu

Jan 5

to Tarek

Dear Tarek,

We are happy for you to use one figure taken from, Sociological Paradigms and Organisational Analysis by Gibson Burrell and Gareth Morgan, 1985, Ashgate, in your dissertation only. If the material is to be stored on an openly accessible public server, i.e. uploaded on to the internet, or if you decide to publish your dissertation you will need to request further permission from Ashgate.

Permission can only be granted for this use, figures cannot be adapted without further permission, and permission cannot be granted to allow others to reprint the material. Ashgate cannot grant permission for any material from third-party copyright holders, and it is the responsibility of the requestor to determine the copyright holder of the material.

Please can you ensure the original publication the material is taken from is fully acknowledged.
Please let me know if you have any questions.
Best,
Keith

Keith Towndrow

Ashgate, Gower & Lund Humphries Publishing

www.ashgate.com www.gowerpub.com www.lundhumphries.com

From: Tarek Hoteit [mailto:]
Sent: 02 January 2015 13:26
To: Keith Towndrow
Subject: Fwd: permission to use research paradigms in finance figure

---

This email has been scanned by the Symantec Email Security.cloud service.

Tarek Hoteit

Jan 6

to Keith
Thank you.
I will make sure the original reference is properly cited. As for the point about publishing my dissertation, I will only be publishing my dissertation within my academic institutions and not as a book publication. In other words, my dissertation will be published as a dissertation only.
Let me know if this is an issue or not.

--
Regards,
Tarek Hoteit

Jan 7

to Tarek

Dear Tarek,
Thank you for letting me know. We have no problem with you publishing it in this way. You would need to write back and seek further permission if you plan to publish commercially.
Best,
Keith

3)
Tarek Hoteit

Jan 6

to richard
Hello Prof Socher. I would like your permission to reproduce the figure 1 in page 1 of your paper Socher et al. (2013) where you show an example of the Recursive Neural Tensor Network model. I would like to include it in my dissertation paper. Please let me know if I have your permission

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1631–1642). Retrieved from http://oldsite.aclweb.org/anthology-new/D/D13/D13-1170.pdf
Richard Socher via waldenu.edu

Jan 6

to Tarek

Yes

4) Tarek Hoteit

Jan 5

to ealtman
Professor Altman,
I am currently completing my dissertation at Walden University on the impact of public sentiment in social media onto the corporate distress of firms. I would like your permission to use the Altman Zscore in an instrument that I am developing along with a sentiment analysis index for the sole purpose of the dissertation. Please let me know if I can have your approval.
edu

Jan 5

to Tarek
Yes, you may but could you show me the exact usage in your text? B est,

EA

Sent from my BlackBerry 10 smartphone.
From: Tarek Hoteit
Sent: Monday, January 5, 2015 9:44 AM
To:
Subject: permission to use Altman Zscore
Tarek Hoteit >

Jan 5

to ealtman
will do. Thank you.
5)permission to use table of variables
Tarek Hoteit

11/7/14

to gtian
Dr Tian,
I would like your permission to use the Table 1 list of covariates in Chancharat, N., Davy, P., McCrae, M. S., & Tian, G. G. (2007). Firms in Financial Distress, a Survival Model Analysis. This is needed as part of my dissertation paper at Walden University on the influence of public sentiments in social media over corporate financial distress. Please let me know if I have your approval

Regards,
Tarek Hoteit

Gary Tian

Hi Tarek, That is fine to me, Gary Sent from my iPhone
11/7/14

6) permission to use the SARF Framework figure
Walden University
x

Tarek Hoteit

11/7/14

to rkasperson
Professor Kasperson,
I would like your permission to use the SARF figure in my PhD dissertation paper at Walden University. My dissertation is about the influence of public sentiments in social media over corporate financial distress. The image that I am referencing is extracted from your article, "a perspective on the social amplification of risk" in The Bridge journal (Fall 2012). The link to the article is
http://www.nae.edu/Publications/Bridge/62556/62562.aspx

Please let me know if I have your permission.
Roger Kasperson

11/7/14

to Tarek
You have my permission--no problem.

Roger
From: Tarek Hoteit
Sent: Friday, November 07, 2014 7:18 AM
To: Roger Kasperson
Subject: permission to use the SARF Framework figure
Tarek Hoteit

11/7/14

to Roger
Thank you.

Click here to Reply, Reply to all, or Forward

7)

This is a License Agreement between Tarek Hoteit ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Tarek Hoteit
Customer address	
License number	3503861279052

License date	Nov 07, 2014
Licensed content publisher	Elsevier
Licensed content publication	Journal of Informetrics
Licensed content title	Sentiment analysis: A combined approach
Licensed content author	Rudy Prabowo,Mike Thelwall
Licensed content date	April 2009
Licensed content volume number	3
Licensed content issue number	2
Number of pages	15
Start Page	143
End Page	157
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	The Impact of Public Sentiment, Influenced by Social Media, on Expected Bankruptcy of Financially Distressed Firms
Expected completion date	Mar 2015
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD
Terms and Conditions	

8)

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Tarek Hoteit
Customer address	
License number	3503861010075
License date	Nov 07, 2014
Licensed content publisher	Elsevier
Licensed content publication	Trends in Cognitive Sciences
Licensed content title	Evolution, brain, and the nature of language

Licensed content author	Robert C. Berwick,Angela D. Friederici,Noam Chomsky,Johan J. Bolhuis
Licensed content date	February 2013
Licensed content volume number	17
Licensed content issue number	2
Number of pages	10
Start Page	89
End Page	98
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	The Impact of Public Sentiment, Influenced by Social Media, on Expected Bankruptcy of Financially Distressed Firms
Expected completion date	Mar 2015
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	

9)

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Tarek Hoteit
Customer address	
License number	3503860716629
License date	Nov 07, 2014
Licensed content publisher	Elsevier
Licensed content publication	Expert Systems with Applications

Licensed content title	Using neural networks and data mining techniques for the financial distress prediction model
Licensed content author	Wei-Sen Chen,Yin-Kuan Du
Licensed content date	March 2009
Licensed content volume number	36
Licensed content issue number	2
Number of pages	12
Start Page	4075
End Page	4086
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	The Impact of Public Sentiment, Influenced by Social Media, on Expected Bankruptcy of Financially Distressed Firms
Expected completion date	Mar 2015
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

10)

License Number	3503750881382
License date	Nov 07, 2014
Order Content Publisher	Springer
Order Content Publication	Springer eBook
Order Content Title	Research Methods in Quantitative Finance and Risk Management
Order Content Author	Wei-Peng Chen
Order Content Date	Jan 1, 2010
Type of Use	Thesis/Dissertation
Portion	Figures
Author of this Springer article	No
Order reference number	None

Original figure numbers	types of risk table in page 55, diversification process in portfolio theory figure in page 11
Title of your thesis / dissertation	The Impact of Public Sentiment, Influenced by Social Media, on Expected Bankruptcy of Financially Distressed Firms
Expected completion date	Mar 2015
Estimated size(pages)	200
Total	0.00 USD

Appendix K: Sampled Tweets With Negative Sentiments Detected

User Name	**Tweet**	**Date**
shelby_jody	Avoid mistakes like $WBC $ABT $CWH $CBM #Research	12/7/2014 1:08
Arm15Edgar	Regretting your investment in $GRA $EIX $WRB $S #wallstreet	12/9/2014 21:45
lawrenceimagin1	Regretting your investment in $GDI $HME $MMC $QLYS Visit	12/10/2014 1:20
DNev47	Is This Really Why Ford's Sales Are Down? -	12/30/2014 17:45
GrassFat	Are you Bearish on these #stocks $TMUS $WPO $ADI $ISRG #moneymanagement	12/6/2014 17:42
Coach23Malachy	Dont want to lose like you did with $UNP $RF $AGCO $QLYS #economy	12/7/2014 12:20
rootfour	Are you considering selling $KRFT $REGN $MDR $CI #StockMarket	12/9/2014 10:18
iHangout	chartguy89: Stock Charts: TBBK -0.70% Stock Charts $TBBK The Bancorp Inc.	12/13/2014 18:04
DailyContracts	$GCA News: Execution Version Guaranty This...	12/23/2014 21:42
RoxannBayer	Are you Bearish on these #stocks $FSL $CLH $AVP $DF #singedinvestor	12/7/2014 1:07
NatalieSlow	Sick and Tired of losing money on $SEIC $WM $GNTX $CSIQ Try this	12/7/2014 19:34
Stock_Tracker1	Ultrapar Participacoes S.A. (UGP) 19.30 $UGP Ultrapar Participacoes Has Returned 13.4% Since SmarTrend Recomme	12/25/2014 15:48
biofan1	â€œ@gilmoreport: $SPLK working on an outside reversal day on heavy volume...â€ /: don't deserve this price. Overvalued!! . Not surprised at all!	12/8/2014 23:37
HotKeyTrading	Price Declines: $NXTM $LFL $FNF $AUO $MCHX $RDN $CY $PMCS $SAPE $RF $SAND $ARNA $MTG $VIP $CYTK	12/27/2014 10:30
chartguy89	AMAT +2.62% Stock Charts $AMAT Applied Materials Inc. (NASDAQ) Last Updated: 12/07/2014 20:25:39 AMAT Stock Ch	12/8/2014 8:25
plasticlaughing	RT @OptionsHawk: Well, two deals that were announced months ago finally get offer terms $PETM $RVBD , get them off the watchlists	12/15/2014 19:47
DayTraderChoice	power_forward: $CYCC at low of day	12/17/2014 2:53
ChaytonFalke	Top Twenty Recent Exits #5: Sold $ASBCW long for a 47.40% gain in 24 days. #trading #ASBCW #forex #stocks	12/15/2014 11:45
LouisHarter	Dont want to lose like you did with $GRA $UHS $SWI $SLW #NSE	12/9/2014 17:17
JanyceLaflamme	Avoid mistakes like $ASBC $MSM $ECL $DF #equity	12/7/2014 15:31

Appendix L: Sampled Tweets With Positive Sentiments Detected

User Name	Tweet	Date
craigbuj	Avg Booking $ for Travel: iPhone, iPad, Android, Desktop $AAPL $GOOG $PCLN $TRIP $OWW $EXPE $LONG $QUNR $CTRP	12/7/2014 22:23
tixoqibymyge	RT @greatspoke: Stock Analyst Finds Best Options Strategies $AAPL $BABA $AMZN $FNMA	12/7/2014 16:26
BuckTwig	$DRI Seeking the next hot #pennystock $ZBRA $CMS $SLW #investing	12/16/2014 0:17
cakesward49	$ANAT Are you Investing in $DNR $G $IBM #NSE	12/17/2014 3:17
Eye13Beebe	Are you putting your money in $CVC $LNT $LM $PCLN Worth a look	12/7/2014 15:16
Benson53Bradley	Are you putting your money in $CB $MHFI $ORLY $QLYS View now	12/6/2014 15:08
micenter	Potential Acorda Therapeutics $ACOR Trade Has 7.03% Downside Protection	12/24/2014 22:56
BoB2Trader	$LINN (& $LNCO) and $BBEP are hedged for the vast majority of their production through and including 2016, at upwards of $91/barrel.	12/14/2014 21:36
tyswsugar	RT @fwpharma: EMA committee backs approval of Orexigen's #obesity therapy Mysimba	12/20/2014 0:11
1LuckyJimmy	CIR Stock up +3.22% percent Today $CIR High is at 59.04 and the Low 57.64 with current volume of 123,550. Circ	12/8/2014 22:04
Ball34Sammy	#Stocks you might want to hold onto $EW $APC $TFM $DO #wallstreet	12/13/2014 3:05
TrudiGary	#Stocks to hold onto $LSI $TMO $CE $IR View now	12/11/2014 3:03
LulaMock	Are you looking for winners like $VIAB $RF $WYNN $SLW View now	12/7/2014 21:09
i_Know_First	Warren Buffett Holdings: Up To 26.64% Return In 3 Months	12/31/2014 6:05
MetAlaArgLys	@grdollasign @zDonShimoda $MGNX has been around since 2000. They have margetux and a few phase 1 compounds, track record of complete failure	12/8/2014 23:57
AlysiaSarkisian	RT @SeekingAlpha: 3 Reasons Why Potash Corp's Dividend Is Safe	12/23/2014 19:02

Appendix M: Sampled Tweets With Neutral Sentiments Detected

User Name	Tweet	Date
PristineTrading	$HLF trying to pop above 5 min highs after dropping most of the day	1/6/2015 2:11
AkinLyman	Looking for more info on $ROK $HPT $ETFC $ABX #singedinvestor	12/6/2014 23:06
DoverPerry	Todays movers to research $EBAY $BOKF $KSU $AAPL Give it a try	12/7/2014 20:27
BelindaCardosi	RT @YahooFinance: .@ampressman Why it will be tough for Apple to hit the $1 trillion dollar mark in 2015 $AAPL	12/26/2014 19:23
jakedakota11	RT @MarketCurrents: Report: Google to invest in Elon Musk's SpaceX	1/20/2015 3:31
JarredTrax	RT @MarkMcCabe95: ICYMI: For those that follow #cybersecurity equities, $QLYS has broken out today.	1/21/2015 1:23
Brock0Melva	Looking for more info on $CTRX $AKAM $BBBY $CADX #money	12/20/2014 3:59
TonyaJenee	$XEL #Stocks to hold onto $IBKR $HAL $GLD #financialnews	12/6/2014 21:09
Mkt_educator	RT @StockSignaling: Pre-Market Price Advances: $SIMG $SYNA $CLDX $VTSS $FRAN $ULTA $GERN $AAPL $YHOO $FB $STRZA $RYAAY $CRUS $CIDM $RJET htâ€¦	12/7/2014 21:32
ordinary614	Looking for the next winners like $CBS $KEY $AMZN $YOD #investing	12/6/2014 20:10
thetafire1	$PT If you like bottom fishing, this may be for you.	12/21/2014 15:23
melmehdi	RT @adamfeuerstein: $AMGN Kyprolis ASPIRE study OS curves #ASH14	12/6/2014 23:42
eyed88	$PWR Is it breakout time on $ERIE $TDC $AAPL #economy	12/6/2014 18:11
TickerReport	Intel $INTC Releases Earnings Results, Beats Estimates By $0.08 EPS	1/16/2015 18:12
BlalockJaydin	Should you buy $TROW $SD $CTXS $BWP #NSE	12/17/2014 9:09
micenter	Beazer Homes USA $BZH Trading Near $16.51 Support Level	12/16/2014 19:45
RhonaThings	5 Stocks you should be watching $SWN $WEC $CTSH $DO #Research	12/7/2014 2:44