BETWEEN THE HEDGES: A COMPUTATIONAL ANALYSIS OF SENTIMENT AND LINGUISTIC HEDGING IN FINANCIAL DOCUMENTS

by

CAITLIN CASSIDY

(Under the Direction of Frederick W. Maier)

ABSTRACT

Each year, publicly incorporated companies are required to file a Form 10-K with the United States Securities and Exchange Commission. These documents contain an enormous amount of natural language data and may offer insight into financial performance prediction. This thesis attempts to analyze two dimensions of language held within this data: sentiment and linguistic hedging. An experiment was conducted with 325 human annotators to manually score a subset of the sentiment words contained in a corpus of 106 10-K filings, and an inference engine identified instances of hedges having governance over these words in a dependency tree. Finally, this work proposes an algorithm for the automatic classification of sentences in the financial domain as *speculative* or *non-speculative* using the previously defined hedge cues.

INDEX WORDS: Sentiment, Linguistic Hedging, Corpus Linguistics, 10-K

BETWEEN THE HEDGES: A COMPUTATIONAL ANALYSIS OF SENTIMENT AND LINGUISTIC HEDGING IN FINANCIAL DOCUMENTS

by

CAITLIN CASSIDY

B.B.A, The University of Georgia, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

BETWEEN THE HEDGES: A COMPUTATIONAL ANALYSIS OF SENTIMENT AND LINGUISTIC HEDGING IN FINANCIAL DOCUMENTS

by

CAITLIN CASSIDY

Major Professor:	Frederick W. Maier
Committee:	Janine E. Aronson
	William A. Hollingsworth
	Walter D. Potter
	Paula J. Schwanenflugel

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

DEDICATION

To my parents, with love.

I told you I could do it.

ACKNOWLEDGEMENTS

I would like to thank the following faculty members for serving on my committee and providing so much help and support throughout my time at IAI:

Dr. Aronson, my mentor and friend for almost five years now, who never stopped believing in my ability to accomplish great things,

Dr. Hollingsworth, for guiding me through the treacherous world of NLP and relentlessly encouraging me to finish this thesis,

Dr. Maier, who was indispensable in getting my experiment off the ground,

And Drs. Potter and Schwanenflugel, for agreeing to read and critique a 50-page paper they know nothing about.

I would also like to acknowledge the contributions of the student members of CompLing: Matthew Lisivick, Nicholas Moss, Brittany Norman, and Clarice Reid. I could not have done it without you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The use of historical data to make predictions about future performance is the predominant focus of research in the field of finance. Because these data are so widely available and simple to process, enormously complex forecasting models are being developed every day. Yet still no perfect mechanism exists. It is our belief that it is time to look beyond historical numbers and into the psychological aspects of financial research. One meaningful way to study these effects is through language.

In this thesis, the language we explore is that contained within annual financial reports filed with the United States Securities and Exchange Commission, 10-Ks. These documents each contain a summary and analysis of a publicly incorporated firm's financial situation for the previous fiscal year and are required annually. Because of this legal requirement, an extensive corpus of such documents is available to the public through the EDGAR website (https://www.sec.gov/edgar/searchedgar/companysearch.html), providing a rich body of unstructured text for research in this domain.

1.1 PURPOSE OF THE STUDY

This thesis seeks primarily to increase our knowledge of two linguistic phenomena within the field of finance. The first is sentiment, the emotion contained in text. Our study builds upon the work of other researchers in the field by evaluating a preexisting financial sentiment dictionary (Loughran & McDonald 2011) through a human annotation experiment. The ultimate goal of our work in sentiment is the creation of a separate dictionary specific to the field of finance and with

gradient sentiment values. It is our belief that this dictionary will prove useful for making predictions about financial performance.

The second phenomenon, hedging (also referred to as *epistemic modality* or *speculative language*), occurs when a speaker or writer seeks to mitigate the force of a proposition. Thus far, hedging has been studied almost exclusively within academic and scientific writing. This thesis attempts to give a detailed account of hedging within the financial domain, while analyzing its relation to sentiment and providing a method for classifying speculative sentences. We believe the classification of speculative language indicates the degree to which these documents are hedged and that hedging, like sentiment, has implications for financial forecasting.

Beyond its use in predicting corporate performance, this work contributes to the literature in sentiment and hedging by providing evidence for a gradient—rather than binary—characterization of both. In our experiment, sentiment is scored on a 5-point scale from -2 to +2, while our method for determining hedge scope characterizes sentences as totally, partially, or not-at-all speculative.

1.2 Expected Results

We expect the presence of both positive and negative words within the corpus, with a bias toward negative sentiment due to the high number of negative words in the Loughran-McDonald dictionary. In the annotation experiment, we predict mid to high inter-rater reliability, though perhaps not for gradient categorization. Finally, we believe that there will be some sentiment variance within each word, i.e. the same word may be considered very negative in some contexts and somewhat negative in others.

In the hedging dimension, we believe hedged sentences will have a significant presence within these documents. We expect our lexicon and various methods for categorizing it to shed

light on the nature of epistemic modality within financial text and its relation to sentiment. We anticipate that our computation for identifying the scope of a hedge cue will be a positive addition to the literature on the computational aspects of this phenomenon.

Altogether, it is our hope that this work will open doors to new works in the rapidly growing fields of financial text analysis and natural language processing.

CHAPTER 2

LITERATURE REVIEW

The scope of this work covers two methods for analyzing text, with the eventual purpose of predicting financial performance. It is therefore necessary to present a literature review of methods for making financial predictions, current work in sentiment analysis, and linguistic accounts of hedging. This review is a small subset of the work that has been done in these areas and is intended only as an overview to frame the subsequent chapters. Section 2.1 presents some popular methods for forecasting performance using only historical data. The most prevalent of these methods include linear models and artificial neural networks. Section 2.2 gives an account of prediction based on a textual analysis of financial documents, primarily using readability indices. Section 2.3 discusses the role of sentiment analysis, particularly with regard to Twitter, and Section 2.4 details linguistic hedging, with an emphasis on the biomedical domain.

2.1 FINANCIAL PREDICTION BY HISTORICAL DATA

Many of the most popular methods for predicting financial performance are based on historical data. The Capital Asset Pricing Model (CAPM) developed by Sharpe (1964), for example, is a simple calculation of expected return that is still popular today. It states that the rate of return on a capital asset can be estimated as a function of the risk-free rate (return on assets that are considered "risk-free", such as government bonds), the market rate of return (often approximated by indices such as the S&P 500), and the asset's β, a measure of sensitivity to market changes. More specifically,

$$R_{asset} = R_{risk-free} + \beta\left(R_{market} - R_{risk-free}\right).$$

The β coefficient is defined as the asset's covariance with the market over its variance and is thus a measure of how drastically fluctuations in the market affect the return of the asset. A β value of 1 indicates that the assets moves in tandem with market forces, a value of 0 specifies no correlation, and a value of 3 suggests that for every one point of increase in market return, the asset will increase by three points. Likewise, as the market goes down, the asset will decrease in value by triple the margin. Some assets even have a negative β, meaning they move in opposition to market forces.

Understanding each asset's relation to systematic changes is an excellent foundation for predicting future performance. However, the CAPM's reliance on rigorous assumptions about investors makes it an imperfect measure of theoretical returns. Still, it remains a common baseline for building new models, such as that specified by Fama & French (1993).

The Fama-French Three-Factor Model is an expansion of the CAPM to include two additional factors, SMB (small-minus-big) and HML (high-minus-low), each with its own β coefficient. These two parameters account for the researchers' observation that small-capitalization and high book-to-market stocks consistently outperform the market. Though not without faults of its own, the Fama-French model is still considered an excellent starting point for new computations and was shown in the 1993 paper to reliably beat the CAPM in U.S. markets.

More recently, Artificial Neural Networks (ANNs) have become prevalent tools in financial prediction (Desai & Bharati 1998, Jasic & Wood 2004, Ticknor 2013). ANNs are machine learning algorithms that can be used to develop non-linear functions to categorization and prediction problems. Though researchers employing these methods have met with great success, ANN parameters can be quite complex.

Desai & Bharati (1998) employed the classic feed forward-back propagation (FFBP) ANN model to predict excess returns on large stocks, over both long and short periods. With a network of 11 input parameters and 2 hidden layers of 15 nodes each, the researchers were able to out-predict linear models over a short time, but not over longer periods.

Jasic & Wood (2004) enhanced the previous algorithm by normalizing the training data to values between 0 and 1. This process helps prevent the model from giving too much weight to more variable input parameters, allowing it to make more accurate predictions when extrapolated to testing data. Again, the researchers achieved better performance over short time periods with ANNs than with linear models.

Finally, Ticknor (2013) focused not on predicting exact stock prices, but rather on classifying the price as increasing or decreasing in the next period. Assuming no trading costs, this model is useful in executing a trading strategy in which increasing stocks are bought and decreasing stocks are sold. Another altered feature is the type of ANN used. Ticknor implemented a Bayesian classification model, as opposed to FFBP, which was intended to reduce over-fitting, a phenomenon which occurs when the model becomes too accustomed to the training data and cannot be expanded to the test set.

2.2 Financial Prediction by Textual Analysis

Not all prediction methods utilize historical pricing data. Increasingly, researchers are using financial texts such as 10-Ks and analyst reports to glean information on how humans express their views of a company through natural language. The Gunning Fog Index, a measure of readability developed by Gunning (1952), is quite popular among researchers seeking to identify a correlation between financial performance and the language of various financial reports (Courtis 1986; Li 2008).

The Fog Index is designed to indicate reading ability by U.S. public school grade level. For instance, a passage with a Fog Index of 11 would be readable by most students in 11th grade. The calculation itself is very simple:

$$Fog = 0.4 * (avg.sentence\ length\ + \%\ of\ "hard"words * 100)$$

where "hard" means any word of more than two syllables which does not have the following properties:

- A proper name
- Compound words composed of short, "easy" words (e.g. *bookkeeper*)
- Verb forms ending in *–ed* or *–es* for which the root is a short, "easy" word (e.g. *created*)

As an example, let's look at the following passage, wherein "hard" words are displayed in bold:

> *We **operate** in a highly **competitive environment**. Our **competitors** include banks, thrifts, credit unions, **investment** banking firms, **investment advisory** firms, **brokerage** firms, **investment companies**, **insurance companies**, mortgage banking **companies**, credit card **issuers**, **mutual** fund **companies** and e-commerce and other **internet**-based **companies**. We compete with some of these **competitors globally** and with others on a **regional** or product basis.* (Bank of America 10-K, 2013)

The three sentences contain 58 words, for an average length of 19.33, of which 21 (36.21%) are "hard." We then add the two figures to get 55.54, which gives us a Fog index of 22.22. According to Gunning, this passage should be well beyond the reading level of a college graduate, but the average college freshman (and perhaps even high school students) would be able to read it with ease. The Fog Index is clearly an imperfect measure of readability, especially at higher levels, but what matters for the studies discussed below is how financial documents compare to each other.

Courtis (1986) compared financial risk across companies based on the Flesch Reading Ease score and the Fog Index. The researcher examined The Chairman's Address and footnotes from ninety-six 1983 Canadian annual reports and found an average Index of 19.48 and 20.32,

respectively, both of which are considered too advanced for a layperson to understand. However, no significant correlation between measures of financial risk and the readability scores of these reports was found, suggesting that the reading level of financial documents is not indicative of perceived risk.

Li (2008) found more promising results by looking into the future, instead of the present. In a corpus of over 50,000 10-Ks, Li found significant evidence that reports with a lower Fog Index were linked with persistent, positive earnings, while the opposite was true of more difficult documents. Because this study focused on firm performance *after* the release of the statements, we can infer that the documents themselves are actually affecting public perception. Li suggests that longer, more complicated documents require more processing power for the reader, deterring further investigation.

In addition to the evidence presented in these papers, Loughran & McDonald (2014) show that the Fog Index is an unreliable measure of analysts' prediction ability and that file size, instead, shows a significant relationship. The researchers believed that more readable documents would lead to better understanding of the content, so financial analysts would, in turn, make better predictions. Instead, they found the opposite: documents with a higher Fog Index lead to better predictability. Perhaps the higher reading level can be associated with management's higher confidence in their future earnings. This is one possibility that can be explored within the context of hedging.

2.3 Previous Work in Sentiment Analysis

The field of sentiment analysis, the study of the emotion contained in text, has gained mainstream popularity through its applications in movie reviews, online articles, and social media. Sometimes referred to as opinion mining, sentiment analysis has become a well-known buzzword

among companies seeking to identify public opinion of their products and services. For example, Opinion Finder (OF), a tool developed by Wilson et al. (2005), is an open source program for identifying various aspects of subjectivity. Written in Java, the latest version is capable of identifying subjective sentences, agents who are the source of opinion, and sentiment expressions.

Bollen et al. (2011) used this tool to correlate public mood with the Dow Jones Industrial Average (DJIA). They calculated this measure as a daily ratio of positive vs. negative sentiment in Twitter feeds. Though the researchers achieved significant results when they added the 1-day-lagged positive:negative measure to a Self-organizing Fuzzy Neural Network (SOFNN), the better input seems to be the *Calm* measure from their Google-Profile of Mood States (GPOMS). GPOMS is similar to OF, but measures six dimensions of mood—Calm, Alert, Sure, Vital, Kind, and Happy—rather than positive and negative. This result is compatible with the belief that stockholders buy and sell shares as a result of their confidence level.

Mittal & Goel (2012) built upon this work by creating their own sentiment measures for Calm, Alert, Happy, and Kind, of which they found Calm and Happy to have significant Granger Causality. Though this study only achieved a maximum accuracy of 75.56% in the predicted DJIA price (as opposed to Bollen et al.'s 87%), it makes a substantial contribution by including information on a concrete trading strategy, listed below:

> *We maintain a running average and standard deviation of actual adjusted stock values of previous k days.*
>
> *If the predicted stock value for the next day is n standard deviations less than the mean, we buy the stock else we wait.*
>
> *If the predicted stock value is m standard deviations more than the actual adjusted value at buy time, we sell the stock else we hold.*

in which an experiment showed the optimal parameters to be $n = m = 1$ and $k = 7$ or 15. This strategy earned a profit of 543.65 Dow Points, which very roughly translates to 5.44%.

Chen & Lazer (2013) used SentiWordNet (Esuli & Sebastiani 2006) to generate a list of over 400 thousand sentiment words. This list was used to calculate log probabilities of each Tweet being "happy" or "sad", and the average of this measure for all Tweets became the sentiment input for the algorithm. Two models were developed for classification and regression. The classification model predicted whether the share price would go up or down. If the price increased, the researchers purchased as many shares as possible given their resources; if it decreased, no shares were purchased, and the state of the market was reevaluated the next day. The regression model, however, attempted to predict an absolute value of future earnings, and the trading strategy is reproduced below:

$$invest = \begin{cases} 100\% & if\ .05\% < (\text{predicted \% change}) \\ 25\% & if -.1\% \leq (\text{predicted \% change}) \leq .05\% \\ 0\% & if\ (\text{predicted \% change}) < -.1\% \end{cases}$$

Compared to the benchmark strategy of simply buying as many shares as possible each day, this study performed quite well, with the classification model gaining 5.32% over a period of thirty-five days (regression: 4.91%), compared to the benchmark of 3.49%.

The previous studies used Twitter sentiment to estimate overall market performance, i.e. the DJIA. However, much like the studies using the Fog Index, some researchers have used firm-specific financial documents to gauge the future performance of individual companies. Loughran & McDonald (2011) (L&M), for example, used the Harvard Psychosociological Dictionary (Stone et al. 1966) to create six lists of sentiment words as they pertain to financial documents. This dictionary was created by members of Harvard's Department of Social Relations as part of the General Inquirer project, a computational approach to content analysis. It was meant to be a general purpose dictionary for researchers interested, as the name suggests, in psychology and sociology and consists of 3,564 terms and 83 tags, ranging from *affection* to *recreational*.

L&M hoped to expand upon contemporary research using negative word counts to gauge the tone of a text. Having observed that multiple dimensions of sentiment can exist within financial writing and that a majority of the Harvard negative words were not negative in a financial context, the researchers set about creating their own set of six word lists: positive, negative, strong modal, weak modal, uncertainty, and litigious. With the addition of a term-weighting scheme to reduce noise from high-frequency words, L&M found a significant correlation between their negativity measure and firm returns. Additionally, the authors found a weaker, though still significant, relationship between returns and the other five measures developed as well as links to firms accused of fraud and weak accounting controls. It is on these words lists that we base our own sentiment measure, discussed in Chapter 3.

2.4 Previous Work in Linguistic Hedging

Another method for analyzing text is linguistic hedging, for which quite a bit of the body of work is in the biomedical domain. Later, we explore how this work can be translated into an analysis of financial documents. A hedge is a linguistic device used to indicate uncertainty or objectivity, even politeness. Researchers often use hedging to distance themselves from their data,

*These data **suggest**…*

while politicians may use it to deceive or misdirect,

*That is not the case **to my knowledge**.*

and laypeople to qualify their exaggerated statements.

*She's **practically** ten feet tall!*

In the above examples, a hedge cue—the source of speculation—has been highlighted in bold, but not all instances of hedging are so obvious.

Well, I wouldn't say that.

Though there are many occurrences like the above which are difficult to identify even with a human eye, many are more akin to the first three examples. With an adequate lexicon, we may be able to use measures of hedging within a text to determine the level of confidence expressed by the author. In the case of annual reports, a high amount of hedging could indicate that the firm's managers, who know more than anyone about the state of the company, are very uncertain about the future.

Ken Hyland, a linguist at Hong Kong University, gives a very detailed account of hedging in the hard sciences in his 1998 book, *Hedging in Scientific Research Articles*. Though distinct from financial language, scientific writing does have some properties in common with the domain explored in this thesis. One is the goal of objectivity: both scientists and managers want to give the impression that they are removed from the material at hand and are simply stating the facts. In a similar vein, both types of writers do not wish to over-commit to any proposition. A scientist who makes a false statement faces backlash from his or her academic community, while a firm's manager could be subjected to legal and monetary consequences.

In Hyland's above-mentioned seminal work, he divides hedging devices into two broad categories. The first, making up about 85% of the hedges in his corpus is *lexical* hedges, which can be identified by specific *hedge cues*. The rest can be summarized as *strategic* hedges. Because this work focuses on instances of the former category, we only briefly describe the latter. Hyland explores three main uses for this type of hedge:

1. Reference to limited knowledge, e.g. *We do not know whether…*
2. Reference to limitations of model, theory, or method, e.g. *Viewed in this way…*
3. Reference to experimental limitations, e.g. *Under these conditions…*

Such instances are outside the scope of this project, but it is important to note that the lexicon used here is not exhaustive and cannot account for every type of epistemic modality.

Hyland separates the lexical hedges by grammatical classes, in turn discussing modals, lexical verbs, adjectives, adverbs, and nouns. In his analysis of modal verbs, Hyland acknowledges that their rampant polysemy makes it difficult to study these lexical items strictly in a hedging context. Referencing the findings of (Coates 1983), Hyland lists the percentage of each modal that appears in a hedging context, with *might* conveying epistemic modality in 78.5% of instances on the high end, and could as a hedge in 7.1% of occurrences on the low end (*can*, of which no instances were hedges, is excluded from the lexicon for the present study).

Lexical verbs, the most common hedge terms in Hyland's corpus, can be described, according to (Palmer 1986), as speculative, deductive, quotative, or sensory. Speculative verbs are verbs of prediction or subjectivity (e.g. *believe*). They convey the author's opinion on a matter without committing to stating it as truth. Deductive verbs suggest that the author arrived at a proposition through logical reasoning (e.g. *conclude*). These verbs, when used as hedges, are intended to outline a path from an observation to a conclusion. Quotative verbs express attribution to a source other than the author (e.g *suggest*). These verbs can reference another person or the data obtained by the author. Finally, sensory verbs describe the author's perception (e.g. *appear*). Such terms suggest that trust in the author's perceptive abilities is necessary to accept the conclusions drawn from these propositions.

Epistemic adjectives, adverbs, and nouns are used in similar contexts. In the speculative category, we have terms such as *likely*, *probably*, and *possibility*. These hedges express prediction of future events based on past experience. Deduction hedges outside of verbs are rarer. In this category, we see items like *consistent*, *conclusion*, and *essentially*. Under the quotative umbrella,

similar and *evidently* refer to past and current evidence to make claims about new data, and finally, sensory lexical items such as *apparent*, *feel*, and *quite* demonstrate the author's judgement of what he senses or perceives.

One additional use that Hyland separates into its own discussion is the hedging of numerical data. In our lexicon, such hedges are given the label *approximation* and are used in financial literature to indicate commitment to a general amount, but not an exact number, e.g. "Profits for the third quarter reached *almost* $10 million." Hyland denotes these terms as "degree of precision" adverbs, and they are often used as a means of presenting a useful figure without constraining the number to an exact measure.

The above discussion enumerates Hyland's categories for hedging attribution. However, in addition to a description of the source of information described by each type of hedge, Hyland categorizes lexical hedges by the author's attitude toward how the proposition in question will be interpreted. In the 1998 book, lexical hedges are also subdivided by the hierarchy displayed in Figure 1.

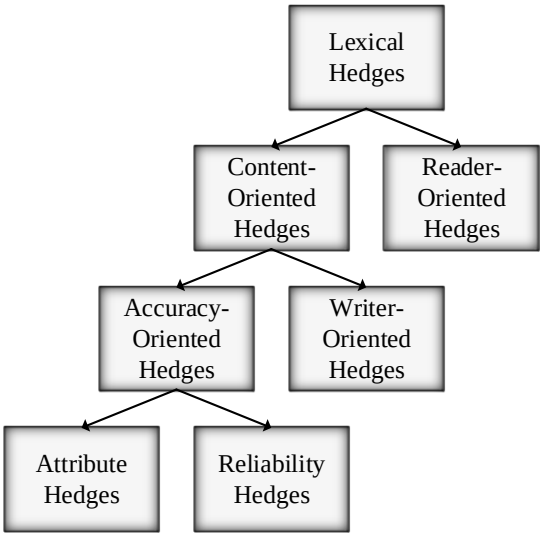


Figure 1: Hyland's Hierarchy of Lexical Hedges

Content-oriented hedges, according to Hyland, "serve to mitigate the relationship between propositional content and a non-linguistic mental representation of reality." They can be used, as with *accuracy-oriented* hedges, to indicate the level of precision with which a proposition is stated. Under the accuracy-oriented umbrella, we see both *attribute* and *reliability* hedges. Attribute hedges "allow deviations between idealized models of nature and instances of actual behavior to be accurately expressed." Take the following example:

*Although variable, the isoelectric point of kunitz seed inhibitor is **generally** lower…*

In this passage, *generally* is being used to describe the phenomenon we would expect to occur, without commitment to the proposition that this is the case in *every* instance. Also categorized as accuracy-oriented hedges are *reliability* hedges. Reliability hedges express the author's confidence in his or her claim and convey the extent to which a proposition is true. An example of a reliability hedge is given below:

> *…it appears **possible** that the mechanism causing the light-activated fluorescence quenching **may** be triggered by either photosystem.*

Here, *possible* conveys the author's judgement that there is some set of circumstances that could produce results consistent with the proposition that either system could trigger fluorescence quenching. Reaching back to the content-oriented node in the above tree, we see that a second subcategory of this type of hedge exists, which Hyland calls *writer-oriented*. These hedges have the primary goal of protecting the author from backlash, should the claim in question later be proven false. The following passage is an example of this type of hedge:

*It **seems** that the stomata do not use the Calvin cycle…*

The use of *seem* demonstrates the author's desire to assert a claim based on his or her intuition without being culpable for its falsehood if new evidence comes to light. Finally, we discuss the second main category of hedges under this hierarchy: *reader-oriented* hedges. Reader-oriented

hedges "address the various dimensions of the social relationship between writer and reader." Their purposes include politeness, deference, and clarification. In the example below,

Our interpretation *of these results is that the total level UV-B…*

our interpretation indicates to the reader that, though the author sees the evidence one way, other analyses may be equally plausible. It is a phrase of deference, allowing the reader to draw other conclusions as he or she sees fit. Now that we have seen a linguistic overview of hedging in scientific writing, we can analyze Hyland's detailed lexicon in the context of financial text.

Humpherys (2009) attempted to automatically classify financial statements as *fraudulent* or *non-fraudulent* based on the number of hedge cues employed in these documents. Specifically, the researcher was interested in three primary hypotheses:

1. Deceivers use hedging devices in fraudulent 10-Ks more frequently than in the control group.
2. Deceivers will use more hedging modal verbs and fewer certainty modal verbs than the control group in 10-Ks.
3. Deceivers use hedging conjunctions, hedging adjectives, hedging adverbs, and hedging nouns at a greater frequency in 10-Ks than the control group.

In order to test these theories, Humpherys developed a financial hedging lexicon—on which this study is based—containing modals, verbs, adjectives, adverbs, and nouns. Though this lexicon was inspired by the 1998 Hyland book on scientific research, Humpherys intended to create a hedging dictionary that was specific to finance. Therefore, items such as *should*, which is listed in Hyland's work as a modal hedge, were excluded from Humpherys' lexicon based on their infrequent hedging use within the financial domain.

The results of this study were mixed. Humpherys found significant evidence that fraudulent statements contain more hedges overall, but no difference in the ratio of hedge cues to total words. This suggests that fraudulent managers simply write more "fluff," possibly with the intent of distracting their audience from inconsistent figures. With regard to the second hypothesis, only the more frequent use of *could* and less frequent use of *will* returned statistical significance. Finally, the researcher found no significant evidence for the third hypothesis. This evidence indicates that hedging alone is not a proper indicator of fraud in 10-Ks.

Humpherys' work provides a domain-specific, yet inclusive hedging lexicon from which to conduct further studies. While Humpherys was not entirely successful in using hedging cues to identify fraud, this study attempts to use these cues to identify future outlook. To take this work a step further, we categorize each item in Humpherys' lexicon, based on the most frequent intended use of the hedge term. This parameter, along with part-of-speech and root information, is described in Chapter 4 and analyzed for differences in usage within this corpus.

A great deal of the computational work in hedging is in the context of biomedical texts. Because claims without significant statistical evidence require careful footing, it is unsurprising that these texts are so rife with qualifying and mitigating statements. Financial texts are similar, in that strong statements are also subject to backlash, namely legal actions. However, these fields differ in two significant ways.

The first is that financial texts do not seem to contain nearly as much jargon. After all, these documents are written with the goal of keeping shareholders informed on company issues, and most investors are laypeople with little industry knowledge. On the contrary, biomedical texts are written for an audience with extensive training in chemistry, biology, and anatomy, making specialized language the norm.

Secondly, corporate managers are very reluctant to disclose negative news, as their job security is directly dependent on their ability to appease shareholders. Much of the hedging in these documents is used either deceptively or with the purpose of redirecting the reader's focus. In contrast, hedges in the biomedical domain are used most often to convey humility or to allow the facts to speak for themselves. They take the focus away from the author, allowing this individual to claim misinterpretation in the face of contradictory evidence.

Light et al. (2004) explored the biomedical domain through a corpus of Medline abstracts. In order to determine how well humans can identify instances of speculation, a preliminary annotation experiment was conducted, in which the researchers obtained an average Kappa value of 0.655 for distinguishing between speculative and definite language. According to the paper, "Kappa scores between 0.6 and 0.8 are generally considered encouraging but not outstanding," so this experiment was successful at creating a baseline while also being indicative of the difficulty of assessing speculative language. Furthermore, this same annotation experiment sought to determine the readers' ability to distinguish between high and low speculation, for which the Kappa value was 0.03, indicating that different levels of speculation are not reliably distinguishable.

The researchers then ran a support vector machine (SVM) classifier on the corpus to sort sentences into *speculative* and *non-speculative* classes, along with a lexicon-based method, which placed sentences in the *spec* class if they contained one of fourteen strings identified during the annotation experiment. They found no significant difference in quality between the two methods.

Medlock & Briscoe (2007) expanded upon this work by developing a weakly supervised, probabilistic classifier. Obtaining a Kappa of 0.985, Medlock & Briscoe's annotation guidelines were much more specific. A sentence is speculative if it is a(n):

1. *Assertion relating to a result that does not necessarily follow from work presented, but could be extrapolated from it (Light et al.).*
2. *Relay of hedge made in previous work.*
3. *Statement of knowledge paucity.*
4. *Speculative question.*
5. *Statement of speculative hypothesis*
6. *Anaphoric hedge reference*

These guidelines were the basis for the seed data annotated by hand. The corpus was then run through a Bayesian classifier, which produced the label *spec* if P(*spec*|*x*) was greater than an arbitrary threshold, σ. Medlock & Briscoe's method achieved a BEP (break-even-point) precision/recall of 0.76. For comparison, Light et al. achieved a 0.60 on the same measure.

Szarvas (2008) went a step further by eliminating 90% of Medlock & Briscoe's seed data in order to weed out potentially speculative terms that often appear *near* hedge cues, but do not themselves carry speculation. Running the data through a maximum entropy classifier produced a BEP of .7868, which increased to .8202 after manual feature selection.

Ganter & Strube (2009) wanted to create a classifier that was more generalizable, rather than those mentioned in the previous three examples, which were strictly applicable to the biomedical domain. These researchers used Wikipedia *weasel* tags to create their seed data. *Weasel* tags appear in Wikipedia text when the editors believe a passage needs improvement, most often applied when specification or author attribution is needed. Their method reaches a BEP of 0.70, which is lower than the previous methods, but this can be attributed to the domain non-specificity. However, because the *weasel* tag is commonly applied to propositions that need clarification, this method often misses instances of modal hedging, prediction, and probability judgments.

CHAPTER 3

SENTIMENT

Sentiment analysis is a method for exploring the emotion contained in text. It can exist on many dimensions (positive, negative, nervous, deceptive, etc.) and may help identify the author's attitude toward the material at hand, even in supposedly unbiased circumstances. Financial documents, for example, are commonly supposed to be objective accounts of relevant quantitative data. In this chapter, we discuss sentiment as it pertains to annual financial reports, as well as an annotation experiment designed to aid in the creation of a domain-specific sentiment dictionary.

3.1 SENTIMENT LEXICON

As noted in Chapter 2, Loughran & McDonald (2011) created their own financial sentiment lexicon from the Harvard Psychosociological dictionary. We inspected these lists—positive, negative, weak modal, strong modal, litigious, and uncertainty—and found them lacking in several ways. Firstly, some words contained in these lists do not actually hold sentiment in financial contexts, e.g. *benefit*, which is used in this corpus to mean *pension*:

> *Sustained increases in costs of pension and healthcare* ***benefits*** *may reduce our profitability.* (GE 10-K, 2013)

While *benefit* in everyday English usually means *advantage*, here it is simply part of an employee's pay package.

Secondly, these lists are missing some words that appear in financial texts and that we believe hold sentiment, e.g. *quality*, which, in the example on the next page, describes something valuable or well-made:

> *Rather than compete primarily on price, we emphasize the **quality** of our products and services, the reputation of our brands and our knowledge of customers' fire and security needs.* (Tyco 10-K, 2013)

Finally, these lists are based on binary membership: they contain no gradient information, though it is intuitive to believe that some sentiment words are stronger than others. Take, for example, the contrast between *enhance* and *improve*. Both of these words carry positive sentiment, but the former implies very mild changes, while the latter indicates tangible results.

These issues, taken together, imply the need for a more robust, gradient-based financial sentiment dictionary, which we attempt to create with the aid of human annotations.

3.2 10-K CORPUS

The corpus used is a collection of two sections from corporate financial reports (10-Ks) submitted annually to the Securities and Exchange Commission (SEC). These documents provide an overview of their respective firms' activities for the previous fiscal year and can range from twenty to hundreds of pages in length.

For this project, we chose to sample only Item1 and Item 7, because these sections contain the most text and are therefore believed to be the most sentiment-rich. Item 1, *Business*, is the opening section, which provides an overview of the firm and its products and discloses any risk factors such as new government regulation or intense industry competition. This is the place for the company to introduce a positive image to its shareholders, and it is rife with terms such as "innovative" and "commitment." Item 7, *Management's Discussion and Analysis of Financial Condition and Results of Operations*, is a detailed account of the firm's operations and financial health for the past fiscal year. Here, the manager reviews key events, emphasizing successes and offering explanations for any shortcomings.

The lengths of the passages chosen for this corpus are based on the number of sentiment words, as defined by Loughran & McDonald (2011) (L&M), contained therein. From each

document, segments containing 100 sentiment words from Item 1 and 250 from Item 7 were extracted to be presented to annotators. This ratio reflects both the lengths of these sections and the amount of sentiment information we expect to find in each. The entire corpus contains 60,521 L&M positive and 138,865 L&M negative sentiment words, of which 1,317 unique words were evaluated.

3.3 Annotation Experiment

In an experiment conducted with 325 business and pre-business students, sentiment in 106 statements from 10 Global Industry Classification Standard (GICS) sub-sectors was manually scored on a scale from -2 (very negative) to +2 (very positive). Because Items 1 and 7 contain the most relevant information, selections containing 100 and 250 sentiment words from these respective sections were extracted and presented to the subjects (a pilot experiment among the researchers showed that passages containing 350 sentiment words were manageable in the allotted one-hour time period).

For each document, the text of both items was manually extracted to a text file (excluding tables and figures), converted to XML with paragraph tags[1], tagged for sentiment and hedging terms in an XML parser[2], and displayed to annotators in our Data Gathering Software (DGS) created with Java Swing[3]. The items were extracted by hand because of the great inconsistency with which these documents are formatted, not only between firms, but within individual filings as well. The large number and unpredictability of these differences made it impractical to automatically extract these sections.

[1] Programming credit: William A. Hollingsworth, Skimcast
[2] Programming credit: Nicholas Moss
[3] Programming credit: Matthew Lisivick

Each document averaged 45 minutes of extraction time, with a range of 30 minutes to 1 hour and 15 minutes. As noted above, tables and figures were excluded, as well as footnotes and leading statements, such as “The following table shows…”

These text files were then converted to XML documents with item and paragraph tags. An XML parser took this XML document as input, and produced an XML file tagged with the following attributes for each paragraph:

1. # L&M positive words
2. # L&M negative words
3. # hedge words
4. % L&M positive words
5. % L&M negative words
6. % hedge words
7. list of L&M positive words
8. list of L&M negative words
9. list of hedge words

The output was then transferred to the DGS, which used the word count information to display a selection containing the above-mentioned sentiment word quotas.

The DGS includes features intended to make the process as simple as possible for subjects, such as different font sizes and scoring mechanisms. The scoring mechanisms include a drop-down menu, clickable buttons, or selection of the {1,2,3,4,5} keys corresponding to sentiment scores of {-2,-1,0,+1,+2}. In the default display, each sentiment term is highlighted in yellow, and the preceding and following five words are bolded. This allows the user to take context into

consideration when annotating, while not being overwhelmed by a wall of text. Screenshots are provided in Appendix A.

In addition to the annotations, subjects were asked to complete a short demographic survey, intended to gauge their comfort with the English language and knowledge of finance, accounting, and linguistics. This survey and the experiment script are reproduced in Appendices B and C, respectively. For the purposes of this thesis, two separate pilot sessions were held in a university-sponsored computer lab. This lab holds 46 desktop computers running Windows 7, with permissions to download and run an executable .jar file hosted on an Institute for Artificial Intelligence (IAI) server.

The subjects for these sessions were 22 and 9 upperclassmen, majoring primarily in Management Information Systems and with credit for finance and accounting prerequisites. In exchange for their participation, these students were given extra credit in their MIST 4600: Computer Programming in Business course. Subjects ranged in age from 20 to 27 and comprised 19.35% females and 80.65% males. All but four were United States citizens, and of those four, only two attended a high school at which the primary language was not English. 21 subjects reported English as their first language, and 87.2% stated that their fluency in English was "above average for a U.S. college student."

The first session, held for 22 subjects, resulted in 7 annotated 10-Ks, with 3 or 4 raters for each document. Subjects were each given a notecard containing an experiment code, a place to write his or her name (in order to receive credit), and space to write comments and suggestions for future experiments. The experiment code serves two purposes: to link the subject's demographic survey information to his or her annotation output, and to indicate to the DGS which 10-K to load. Each code contains two capital letters representing an S&P 500 corporation, the year of the filing,

and a unique identifier representing one subject. The time for subjects to complete the experiment ranged from 25 to 38 minutes, with non-native speakers of English taking the longest amount of time.

A second session was held for the 9 students who could not participate in the first experiment but still wished to receive the extra credit. This session had a similar timeline and comparable results. In both sessions, a small selection of subjects left feedback on their code cards, which is reproduced in Appendix D. Because of the short time frame between sessions, many of the requested changes could not be implemented in time for the second set of subjects. The only difference in conditions was a more experienced proctor for the second group.

After the success of these two pilot sessions, we decided to expand the experiment to a larger subject pool, namely 294 undergraduate students from two Terry College of Business (Terry) classes. The first, MIST 2090: Introduction to Information Systems in Business, is an introductory course required for all underclassmen wishing to enroll in Terry. These students may or may not have been exposed to accounting or finance courses and have likely not yet declared a major. They were given 1 percentage point of extra credit on their final grade, in exchange for their participation.

The second course, MGMT 3000: Principles of Management, is one of the first courses required for all new business students, i.e. students recently accepted into Terry. All of these students have completed or are in the process of completing two college-level accounting courses and have declared their Terry major. Students in this class are required to participate in a certain number of research hours, and those who participated in our experiment each received 1 hour toward that total.

In this second set of 25 experiment sessions, subjects ranged in age from 17 to 26 and comprised 52.66% females and 47.04% males (1 subject marked "prefer not to answer"). 315 were United States Citizens, and among those that were not, 15 did not attend an English-speaking high school. 294 subjects reported English as their first language, and 97.34% stated that their fluency in English was at least "average for a U.S. college student."

3.4 Results, Pilot Experiment

In the pilot sessions, a total of 9 documents were annotated by at least three subjects each. In calculating inter-rater reliability, the measure described by (Fleiss, 1971) was applied to 5-category (-2, -1, 0, +1, +2), 3-category (negative, neutral, positive), and 2-category (negative, positive) classifications. The 5-category kappa was computed on the raw data retrieved from the experiment, while the 3- and 2-category measures were calculated by assigning -2 and -1 to the category "negative", 0 to the category "neutral" and +1 and +2 to the category "positive." Under the 2-category system, all tokens that were assigned a 0 by one or more annotators were removed from the data set. These values for each document are reproduced in Table 1. Documents used in the second experiment are denoted by *, and the highest Kappa in each column is highlighted in bold.

Table 1: Inter-Annotator Agreement

Document	# Raters	5-Category Kappa	3-Category Kappa	2-Category Kappa
*AG2013	5	0.2774	0.6830	0.9388
CF2013	3	0.3316	0.6936	0.9587
CP2013	3	0.4164	0.6439	0.8610
GE2013	3	0.2231	**0.7749**	0.9758
IN2013	4	0.1789	0.3730	0.4184
MA2013	3	0.1634	0.5648	0.8219
RE2013	3	0.2910	0.6818	0.8981
*TY2013	4	**0.4264**	0.6474	**0.9852**
VI2013	3	0.3340	0.5831	0.9316

From this summary, we can see some trends that indicate important information about the raw data, the conditions of the experiment, and the nature of sentiment analysis. The first is that, as the number of categories decreases, the inter-annotator agreement increases. This shows that although raters agree quite well on whether words are positive or negative, assigning the same intensity score is less common, indicating the difficulty in assigning gradient values to a sentiment dictionary.

Secondly, the IN2013 document has Kappa values well below the average for all 9 documents. We first attempted to account for this discrepancy by comparing the data to demographic information. However, all four raters who annotated IN2013 are U.S. citizens with a first language of English. Upon further investigation into the raw data, it became apparent that one particular annotator (IN2013001) was consistently marking a -1 for words the other annotators gave positive scores. It is our belief that this annotator was assigning scores using the 1-5 keys, for which the 2 key corresponds to a -1, intending to mark these words with a +2. For all subsequent sessions, special emphasis during training was placed on the correct use of these keys. As evidenced by the average and above average Kappa values for the AG2013 and TY2013 documents, respectively, stressing the proper use of keys seems to have rectified this particular issue.

Finally, it should be noted that having a high Kappa in one categorization scheme does not necessarily imply a high Kappa in all schemes. Take, for example, TY2013, which has the highest 5-category and 2-category Kappas, but has a 3-category Kappa that is nearly identical to the average. This implies that the TY2013 raters could agree on gradient ratings more easily than raters for the other documents but had a more difficult time agreeing on whether each word carried any sentiment. There is no pattern in the demographic data that explains this phenomenon, and

annotations on documents from previous years would be required to link it to the writing style of Tyco managers.

There does not exist universal agreement on the interpretation of Fleiss' Kappa. A common approach is that which is presented by Landis & Koch (1977), detailed in Table 2. By this interpretation, our subjects achieved "substantial agreement" on two-thirds of the documents in the 3-category system and "almost perfect agreement" on all but one document in the 2-category system. For the 5-category system, the one intended for use in creating a gradient sentiment dictionary, Kappa values are much more modest, indicating that the average and standard deviation of scores for each word would be a more accurate indication of sentiment value.

Table 2: Fleiss's Kappa Interpretation

Kappa	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

The 10 sentiment words with the highest average score are displayed in Table 3, along with their standard deviations and frequency. Only two words (*excellence, rewards*) received a score of +2 with no disagreement. While both carry a positive connotation in everyday language, *reward* contains special meaning in the finance world, which is heavily influenced by the terminology of economics. In a domain dominated by using incentives to manipulate behavior, it is unsurprising that a word synonymous with *prize* should be ranked so highly.

Conversely, the 25 lowest values are displayed in Table 4. We can clearly see that extremely negative words are much less controversial, as the lowest 17 values have an average score of -2 with no deviation. These results suggest that our raters had a tendency to perceive negative words as generally more polarized than positive words. This is as expected, because

humans feel negative emotions more strongly than positive emotions of an equal intensity, which Kanouse & Hanson (1972) call the "negativity bias."

Table 3: 10 Highest Average Sentiment Scores, Pilot

Token	Average Score	Standard Deviation	Number of Ratings
excellence	2.0000	0.0000	12
rewards	2.0000	0.0000	3
leadership	1.8889	0.3333	9
strongest	1.8750	0.3378	24
advantageous	1.8333	0.4082	6
innovation	1.8140	0.3937	43
optimistic	1.8000	0.4472	5
strengthening	1.7778	0.4410	9
innovative	1.7353	0.5110	34
successful	1.6721	0.4733	61

Table 4: 20 Lowest Average Sentiment Scores, Pilot

Token	Average Score	Standard Deviation	Number of Ratings
hazards	-2.0000	0.0000	12
terminate	-2.0000	0.0000	9
unexpected	-2.0000	0.0000	6
laundering	-2.0000	0.0000	6
disasters	-2.0000	0.0000	6
unreimbursed	-2.0000	0.0000	5
illegal	-2.0000	0.0000	3
abandon	-2.0000	0.0000	3
danger	-2.0000	0.0000	3
destroy	-2.0000	0.0000	3
fails	-2.0000	0.0000	3
insufficiency	-2.0000	0.0000	3
forfeited	-2.0000	0.0000	3
underfunded	-2.0000	0.0000	3
fines	-2.0000	0.0000	3
penalty	-2.0000	0.0000	3
threats	-2.0000	0.0000	3
penalties	-1.9167	0.2887	12
bankruptcy	-1.8750	0.3416	16
suffered	-1.8333	0.4082	6

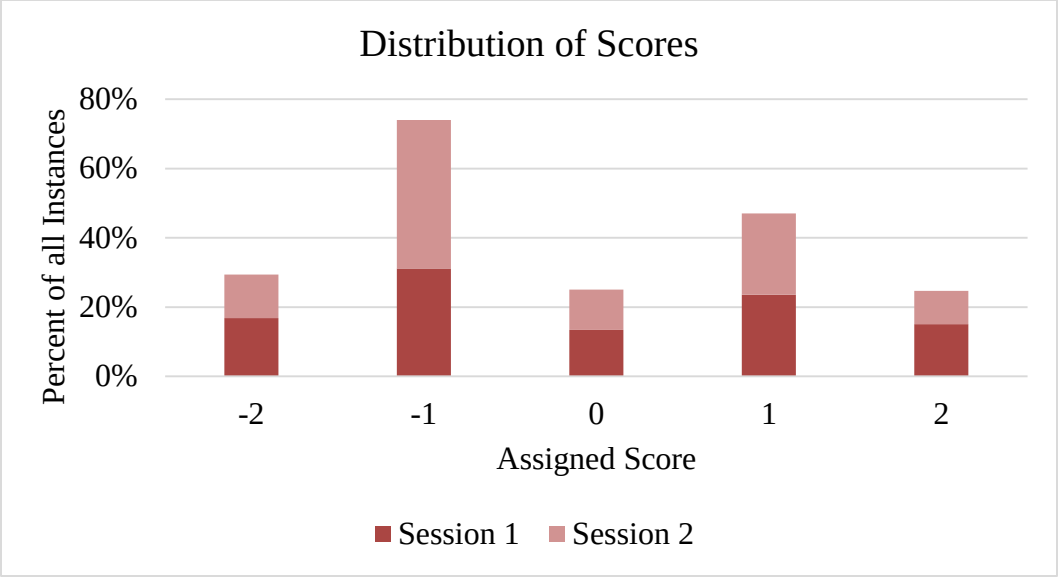


Figure 2: Distribution of Sentiment Scores, 5-Category, Pilot

In Figure 2, we investigate the frequency of each score. Because "0" is the score assigned to instances which do not carry sentiment, it is predictable that this category would be the least frequent. However, we expected a more even distribution of the other four types. It is possible that the raters were reserving their extreme scores of -2 and +2 for very strong words and therefore marked a disproportionate number of instances as only *somewhat* positive or negative. This is promising, because words that were assigned stronger scores will uncontroversially be given a sentiment value of +/- 2 in the final dictionary.

Finally, we visualize the scoring frequency under the 3-category system in Figure 3. We expected this disparity between positive and negative words, as the negative L&M list is more than six times as long as the positive list (2,329 words and 354 words, respectively). However, a surprising number of the total word count was scored as positive. This suggests that these companies write their 10-Ks filings not just to comply with SEC regulations, but as a form of propaganda for their shareholders.

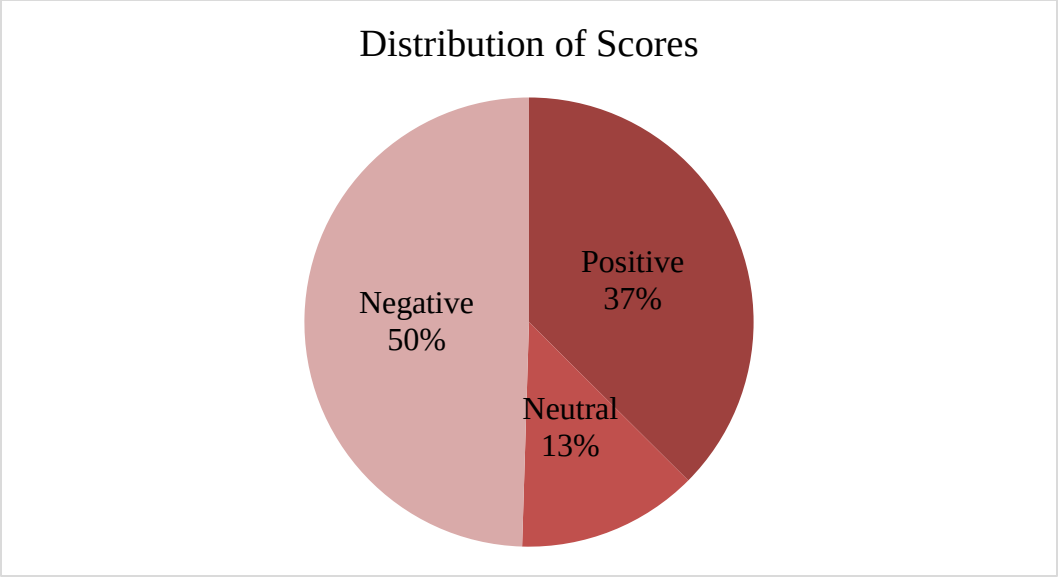


Figure 3: Distribution of Sentiment Scores, 3-Category, Pilot

Table 5 contains the distribution of scores by firm. If we apply the conclusions drawn by other researchers in Chapter 2, Intuit and GE should have above-average performance for the fiscal year 2013, while Tyco and Regions should not have performed as well. However, linking financial wellbeing measures to the results of the annotation experiment is beyond the scope of the current work and will be the subject of further study.

Table 5: Distribution of Sentiment Scores by Firm, 3-Category, Pilot

Row Labels	Negative	No Sentiment	Positive
IN2013	37.00%	9.92%	53.08%
GE2013	41.15%	7.07%	51.79%
AG2013	49.50%	8.38%	42.12%
CF2013	48.42%	11.68%	39.90%
CP2013	47.66%	16.71%	35.63%
VI2013	38.08%	28.26%	33.66%
MA2013	55.73%	14.14%	30.13%
TY2013	57.61%	12.70%	29.69%
RE2013	67.52%	8.12%	24.36%

As evidenced by the results of our experiment, sentiment can be difficult to quantify, even for human annotators. However, because the 2-category agreement measures were so promising, we feel confident we can combine the results of these experiment sessions with those of the

expanded experiment. The following section explores these new results and synthesizes both sets for a complete picture of sentiment within our corpus.

3.5 Results, Expanded Experiment

Overall, the results of the expanded study were quite similar to those of the pilot. Due to the large volume of annotated documents, it is impractical to calculate inter-annotator agreement for each. Instead, we note that standard deviations in scores for each word are comparable.

In Table 6, we see that 4 words were assigned a sentiment score of +2 with no disagreement. Relative to the size of the corpus, these results are similar to those of the pilot experiments. We note that 2 of the top 10 are what we consider "superlatives"—*perfect* and *favorite*—while many of the rest are related to the concept of standing out from a crowd—*exceptionally*, *outperformed*, *revolutionized*. This speaks to a common philosophy in the business world that perfect output is unnecessary, as long as you are outpacing the competition.

Table 6: 10 Highest Average Sentiment Scores, Expanded

Token	Average Score	Standard Deviation	Number of Ratings
pleasure	2.0000	0.0000	3
bolstered	2.0000	0.0000	3
perfect	2.0000	0.0000	6
accomplishing	2.0000	0.0000	3
exceptionally	1.7500	0.4330	4
benefitting	1.6667	0.4714	3
outperformed	1.6667	0.4714	3
revolutionized	1.6667	0.4714	6
versatility	1.6667	0.4714	3
favorite	1.6667	0.4714	3

Additionally, the lowest values are displayed in Table 7. Among these, we see that sentiment words relating to uncertainty are considered the most devastating in a financial context. This reinforces the belief that risk can be more detrimental than negative news or underwhelming performance.

Table 7: 15 Lowest Average Sentiment Scores, Expanded

Token	Average Score	Standard Deviation	Number of Ratings
manipulates	-2.0000	0.0000	3
unreimbursed	-2.0000	0.0000	5
repossessions	-2.0000	0.0000	2
deadlocks	-2.0000	0.0000	1
scrutinize	-2.0000	0.0000	2
fined	-2.0000	0.0000	1
severities	-2.0000	0.0000	4
devastating	-2.0000	0.0000	1
victims	-2.0000	0.0000	3
unfounded	-2.0000	0.0000	1
forbids	-2.0000	0.0000	1
unreliable	-2.0000	0.0000	5
worst	-1.8750	0.3307	8
worthless	-1.8571	0.3499	7
calamities	-1.8333	0.3727	6

The distribution of sentiment scores shown in Figure 4 is similar to that in the previous experiment sessions, with the exception of a disproportionately high number of -1's. We believe that, despite our best efforts, some annotators continued to incorrectly use the 1-5 keys to score words, leading to instances being scored -1, when the annotator intended a score of +2. However, as evidenced by the average scores for each word, we believe these mistakes were not common enough to significantly influence the construction of our dictionary.

Additionally, the distribution of 3-category scores for the larger experiments, displayed in Figure 5, is closer to what we expected than that of the smaller experiments, which indicates that the L&M dictionary may be more appropriate than the pilot experiments would lead us to believe.

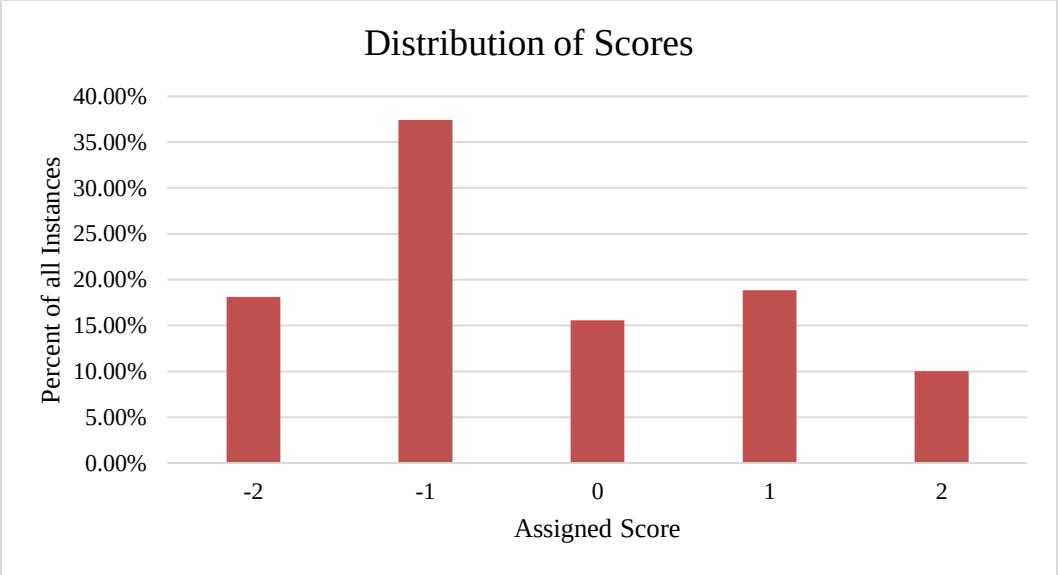


Figure 4: Distribution of Sentiment Scores, 5-Category, Expanded

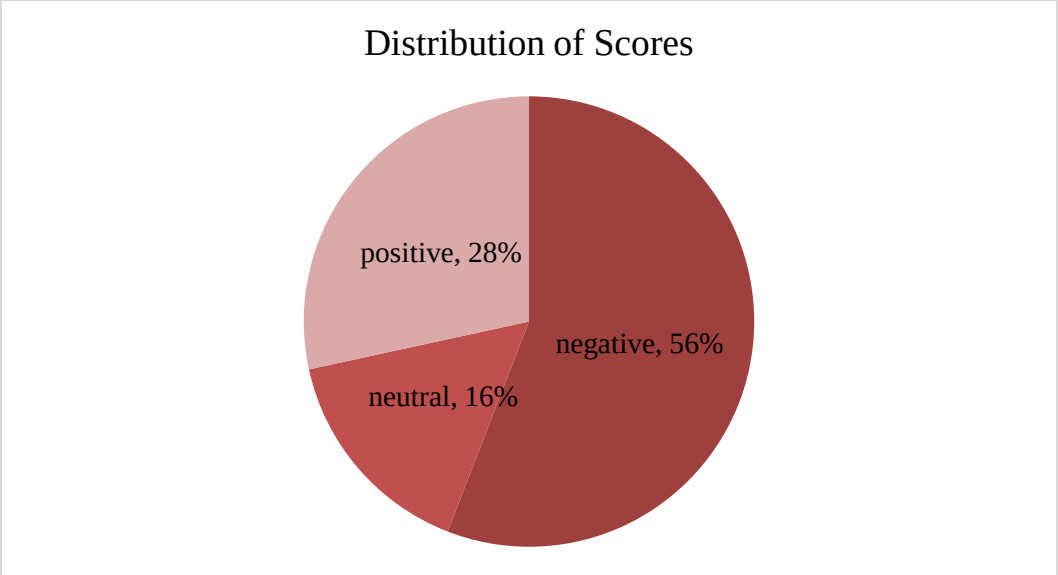


Figure 5: Distribution of Sentiment Scores, 3-Category, Expanded

3.6 Dictionary Evaluation and Limitations of the Study

Comparing the total body of human annotations to the L&M dictionary reveals a precision measure across all words of 77.53%, meaning that of the sentiment words tagged using their dictionary, over 75% of instances were marked as carrying sentiment. 205 words had a precision of 100%. The 10 words with the highest imperfect precision are reproduced in Table 8.

Table 8: 10 Highest Precision Values

Word	Sentiment	No Sentiment	Precision
corruption	61	1	98.39%
invalidated	57	1	98.28%
malicious	55	1	98.21%
sabotage	41	1	97.62%
misconduct	77	2	97.47%
unsaleable	35	1	97.22%
defects	135	4	97.12%
illegal	32	1	96.97%
unrecoverable	29	1	96.67%
rejected	26	1	96.30%

Recall cannot be computed without a separate experiment, allowing annotators to tag sentiment words that do not appear in the dictionary, e.g. *quality*. This measure would be calculated as the number of items in the L&M dictionary, over the number of unique items tagged by annotators.

Furthermore, we can evaluate the L&M dictionary based on the number of dictionary items that appeared at least once in our corpus. Of the 2,683 items listed in their positive and negative word lists, 1,327 occurred in the sample, just over 49%. These figures, taken together, indicate that the L&M dictionary is a good indication of the sentiment contained in financial text, but may be too large. A smaller dictionary, including only those words that actually appear in financial text, would be more domain-specific.

In this study, we examine sentiment on a negative-to-positive scale, in the hopes of creating a gradient-based dictionary that will be used for many textual analysis projects to come. However, this project is not without its limitations. As noted above, recall information is not available for assessing the L&M dictionary, because a separate tagging experiment would be necessary to identify sentiment words that are not included. Also, all of the human annotators who participated in this study are either currently enrolled in business school or are planning to enroll. This may not

be ideal, because the target demographic of 10-Ks are laypeople who might interpret the content of these documents differently. Finally, the L&M dictionary on which our study is based in un-stemmed and could be evaluated differently if we analyze the presence of linguistic roots, instead of identifying the instances of each unique variant. For example, the word *abnormally*, which shares a root with *abnormality*, appears in our corpus, though its sister does not. If we were to consolidate all L&M words sharing a root, these sentiment words may have different distributional properties.

CHAPTER 4

HEDGING

Hedging, also denoted to as *epistemic modality*, refers to a linguistic phenomenon in which a speaker or writer attempts to mitigate the force of a proposition in order to display humility, objectivity, or fallibility. Within financial texts, they are most often used to make assertions about the state of a firm, without committing to potentially false propositions. In this chapter, we first define a hedging lexicon based on that proposed by Humpherys (2009), described in Chapter 2. In section 4.2, we define a method for determining the scope of a hedge cue, which we apply to the corpus to classify speculative sentences. The final two sections contain distributional information about these hedge cues and an analysis of the classification algorithm, respectively.

4.1 HEDGING LEXICON

In contrast to the Medlock & Briscoe (2007) and Szarvas (2008) papers, which used a detailed algorithm to tag training sentences, this study, similar to Light et al., seeks to classify speculation based on a handful of terms commonly used in the financial domain. Listed in Appendix E, this lexicon is meant to be as inclusive as possible given the limitations of this corpus. It is an expansion on the work of Humpherys (2009), described in Chapter 2, and is investigated in conjunction with sentiment.

In addition to simply analyzing this lexicon, we seek to elaborate on each item by marking it as a member of one of several categories. These categories are meant to indicate the most likely purpose of the hedge term, though we recognize that this is not always the case. The categories and their descriptions are displayed in Table 9.

Table 9: Hedge Function Categories

Approximation:	indicates that the proposition is an estimate; e.g. *almost*	**Modality:**	modal verbs which decrease a proposition's certainty value; e.g. *might*
Degree:	indicates how well the proposition fits in to category membership; e.g. *primarily*	**Objectivity:**	indicates that the author is allowing the data set to speak for itself; e.g. *indicate*
Frequency:	indicates how often the proposition occurs; e.g. *occasionally*	**Prediction:**	indicates a judgment about the future; e.g. *anticipate*
Intention:	indicates future plans; e.g. *seek*	**Probability:**	indicates a proposition's likelihood; e.g. *possibility*
Logic:	indicates that a proposition follows logically; e.g. *conclude*	**Subjectivity:**	indicates that a proposition is based on assumptions or impressions; e.g. *presumptive*

However, many hedge terms are not used exclusively in these contexts. Take, for example, *almost* in the following excerpt:

> *[…] testing multiple parameters of the printed circuit boards used in* ***almost*** *every electronic device[.]* (Agilent 10-K, 2013)

Almost is listed in the lexicon as a term of approximation, but in this instance, it is expressing the *degree* to which the circuit boards are used in every electronic device. Whereas in this next example, the term is clearly expressing numerical rounding:

> *We hold a 16 percent working interest in the Waha concessions, which encompass* ***almost*** *13 million acres located in the Sirte Basin of eastern Libya.* (Marathon 10-K, 2013)

Almost only appears in the corpus only 4 times, so determining the most frequent use was manageable by investigating each instance and making a human judgment. For more common terms such as *may* or *approximately*, a random subset of 50 occurrences was analyzed, and the term was classified accordingly.

This lexicon is used to analyze the presence of hedging in our corpus, as well as to measure its relation to any sentiment terms that fall under the scope of a hedging cue. While it is not exhaustive, each item has a non-trivial presence in the corpus and may indicate important information about content mitigation in our domain.

4.2 Determining the Scope of a Hedge

In order to measure how the presence of hedging impacts sentiment, we first need to determine the *scope* of a hedge cue and whether a sentiment term falls within this boundary. For this thesis, we define scope as all items that fall under the parent node (hedge) of a dependency tree. To accomplish this, we made use of the Stanford Parser (de Marneffe et al. 2006), a statistical dependency parser which produces a set of grammatical relations, given a grammatical sentence. To demonstrate, the following example is taken from the corpus:

> *When there **appears** to be a range of **possible** costs with equal likelihood, liabilities are based on the low-end of such range.* (GE 10-K, 2013)

And its tree is illustrated in Figure 6[4]. The hedge cue *appears* is highlighted in the example sentence and has scope over the sentiment word *costs*. We would therefore expect *costs* to carry a more negative sentiment value in this sentence than in another context.

This computation was run for each sentence in the corpus, and the output was analyzed by an inference engine written in Prolog, which produces a list of all dependencies containing a hedge item. The following section contains distribution information and a discussion of these results.

[4] Image credit for dependency trees: William A. Hollingsworth

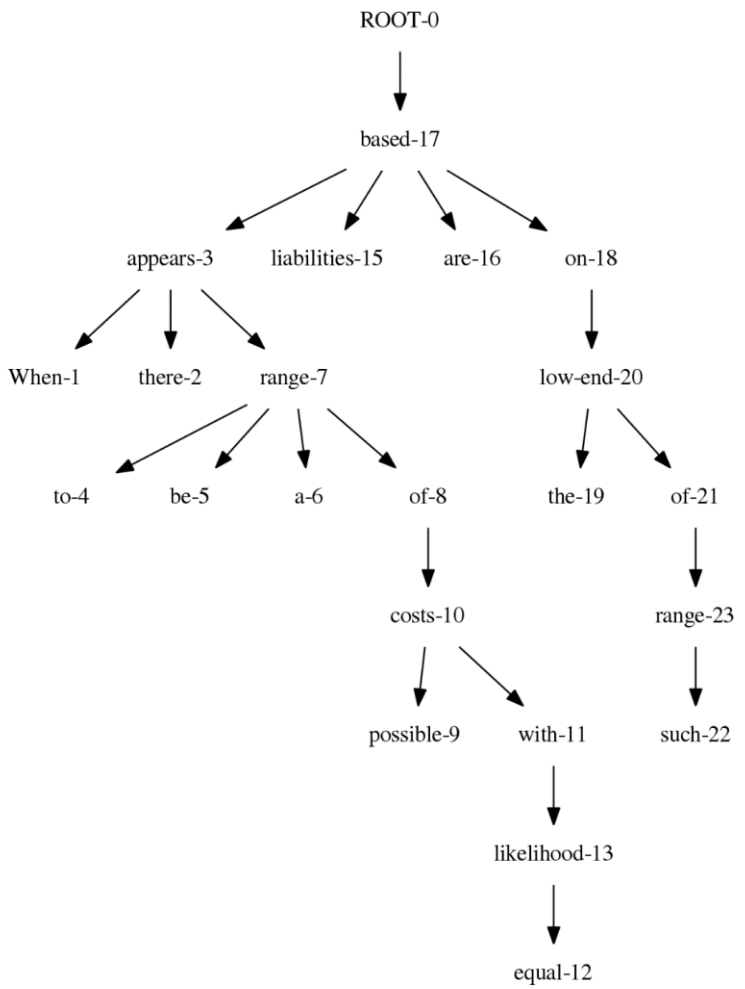


Figure 6: Example Dependency Tree

4.3 HEDGE FREQUENCY AND DISTRIBUTION

Within the lexicon developed by Humpherys (2009), only 58 items (out of 103) appear in this corpus. The 10 most frequent are listed in Table 10, along with their frequencies and average occurrences per document.

We can see a lot of variation in frequency within the lexicon. At the top of this list is *may*, which, in Hyland's study, is "the only modal which figures significantly more often in academic than in other genres." Because academic writing and financial writing share the goal of stating only true propositions, it is unsurprising that a word that means *I believe* or *perhaps* is used so

frequently in both. On the lower end, with 4 occurrences each, are *felt*, *seem*, and *suggested*. All three relate to perception and have a specialized meaning. Thus, it stands to reason that each only occurs in a select number of contexts.

Table 10: Hedge Frequency

Hedge	Number of Occurrences	Average Occurrences per Document
may	11503	123.76
could	8750	93.29
approximately	4633	51.86
would	3020	31.15
generally	2860	27.79
plan	2062	25.13
partially	2480	24.25
some	2076	21.45
potential	1910	20.13
most	1955	19.82

In addition to distributional information by hedge cue, the same information is available by grammatical category and hedge function and is displayed in Table 11 and Table 12, respectively.

Although adverbs and verbs make up most of the lexicon, modals occur most frequently within the corpus. This is because modals serve as auxiliary verbs and may sometimes occur together or with another hedge cue. Discounting modals, then, leaves adverbs the most frequent part-of-speech, possibly due to their usefulness in describing values. A calculation that is *approximately* $20 million may be factual while still leaving room for error.

Among the varying functions of hedging, modality is again a frequent purpose. The second highest measure belongs to hedges of approximation. Given the nature of this domain, hedges meant to allow for miscalculating a dollar amount are naturally common within the corpus.

Table 11: Grammatical Category Distribution

Part-of-Speech	Lexicon Items	Number of Occurrences	Occurrences per Item
adjective	16	8083	505.19
adverb	34	18883	555.38
modal	6	27546	4591.00
noun	4	1338	334.50
verb	43	8040	186.98

Table 12: Hedge Function Distribution

Hedging Function	Lexicon Items	Number of Occurrences	Occurrences per Item
approximation	11	12602	1145.64
degree	8	4125	515.63
frequency	6	3638	606.33
intention	4	3491	872.75
modality	5	27546	5509.20
objectivity	6	920	153.33
prediction	16	3501	218.81
probability	17	5061	297.71
subjectivity	12	3006	250.50

We can also investigate the hedge cues' distribution as governors in dependency trees, given in Table 13. The 10 items that occur most frequently as a governor do so in 100% of the occurrences in this corpus. This demonstrates the nature of dependency trees, in that all of these words are frequently used to introduce a dependent clause. What is more useful is the number of dependents per occurrence, which could indicate that these words, on average, have scope over a larger section of the dependency tree, making sentences containing them more speculative.

Table 13: Distribution of Governor Hedges

Hedge	Number of Sentences	Number of Sentences as Governor	Governor Sentence Frequency	Dependents per Occurrence
indicate	124	124	100.00%	3.42
suggest	32	32	100.00%	1.64
belief	32	32	100.00%	1.24
appear	25	25	100.00%	4.20
show	19	19	100.00%	3.79
imply	8	8	100.00%	1.77
imply	8	8	100.00%	1.77
seem	4	4	100.00%	2.00
suggestive	2	2	100.00%	2.00
felt	3	3	100.00%	2.00
seek	561	559	99.64%	3.29
predict	608	600	98.68%	3.93
expect	1129	1101	97.52%	3.26
assurance	913	869	95.18%	4.42
predicted	53	49	92.45%	4.34

Finally, we explore the co-occurrence of hedge cues and sentiment terms. Listed in Table 14 are the 10 hedge cues with the highest number of sentiment dependents. Dependents carrying sentiment make up a very small percentage of all words that occur under the scope of a hedge. Despite this infrequency, it is interesting to note the distribution between positive and negative among the sentiment terms which do fall under modality. Similar to the results of overall sentiment frequency, hedged sentiment words are overwhelmingly negative. Managers are naturally more reluctant to disclose negative news and therefore strive to mitigate its impact as much as possible. The hedge cues which most often govern positive sentiment words (*assurance*, *predict*, *plan*, *potential*) are mostly future-oriented, because even the most negative of documents express belief in improving circumstances.

Table 14: Hedges with Scope over Sentiment

Hedge	Total Dependents	Sentiment Dependents	Sentiment Frequency	Negative Percent	Positive Percent
assurance	4100	252	6.15%	12.30%	87.70%
plan	3506	147	4.19%	46.26%	53.74%
expect	3818	132	3.46%	64.39%	35.61%
seek	2181	132	6.05%	62.12%	37.88%
potential	764	72	9.42%	59.72%	40.28%
possible	1914	69	3.61%	91.30%	8.70%
predict	2497	62	2.48%	30.65%	69.35%
claim	431	50	11.60%	88.00%	12.00%
likely	2641	44	1.67%	93.18%	6.82%
most	830	38	4.58%	84.21%	15.79%

4.4 Classifying Speculative Sentences

In this thesis, we classify a sentence as *speculative* is it contains a hedge cue in a governing position. The following example has been taken from the corpus:

> *As a result, we have significantly modified our debit strategy and continue to renegotiate **some** portions of our contracts with our financial institution clients.* (Visa 10-K, 2013)

We can clearly see that this is not a hedged sentence, but if we tag sentences as *speculative* based only on the presence of a hedge cue, this example is included in the set. However, under the strategy proposed here, this is not the case. Because the highlighted hedge cue *some* does not appear as a governor within this sentence, the entire example is given the label *non-speculative*. For comparison, take this sentence from the corpus used by Medlock & Briscoe (2007):

> *Dl and Ser have been **proposed** to act redundantly in the sensory bristle lineage.*

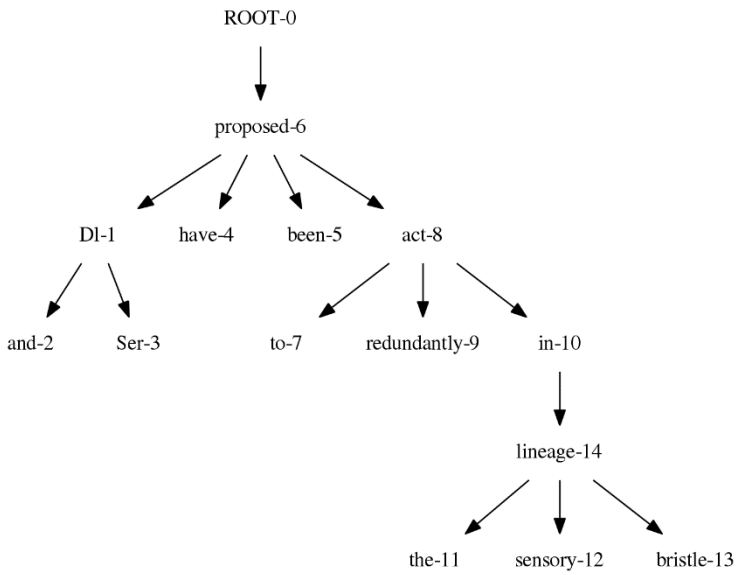


Figure 7: Medlock & Briscoe Dependency Example

In Medlock & Briscoe (2007)'s tagging algorithm, this sentence is classified as *speculative*, and the same is true under this algorithm. The hedge cue *proposed* is in a governing position, so this is a hedged sentence.

Within this corpus, 68,630 sentences contained a hedge term, but this term was in a governing position in only 11,315 of them. This means that only 16.49% of the sentences that would be classified as hedged using Light et al.'s system are considered speculative by this algorithm. The definition at hand clearly pares down the misclassifications quite a bit, but does not eliminate them entirely. Let us revisit the GE example from the beginning of the chapter:

> *When there **appears** to be a range of **possible** costs with equal likelihood, liabilities are based on the low-end of such range.*

The sentence as a whole is simply a statement of how a figure is calculated, and it is unhedged. However, the presence of *appears* as a governor would lead to a false positive classification as speculative. If we remove the second branch, the sentence becomes

> *There appears to be a range of possible costs.*

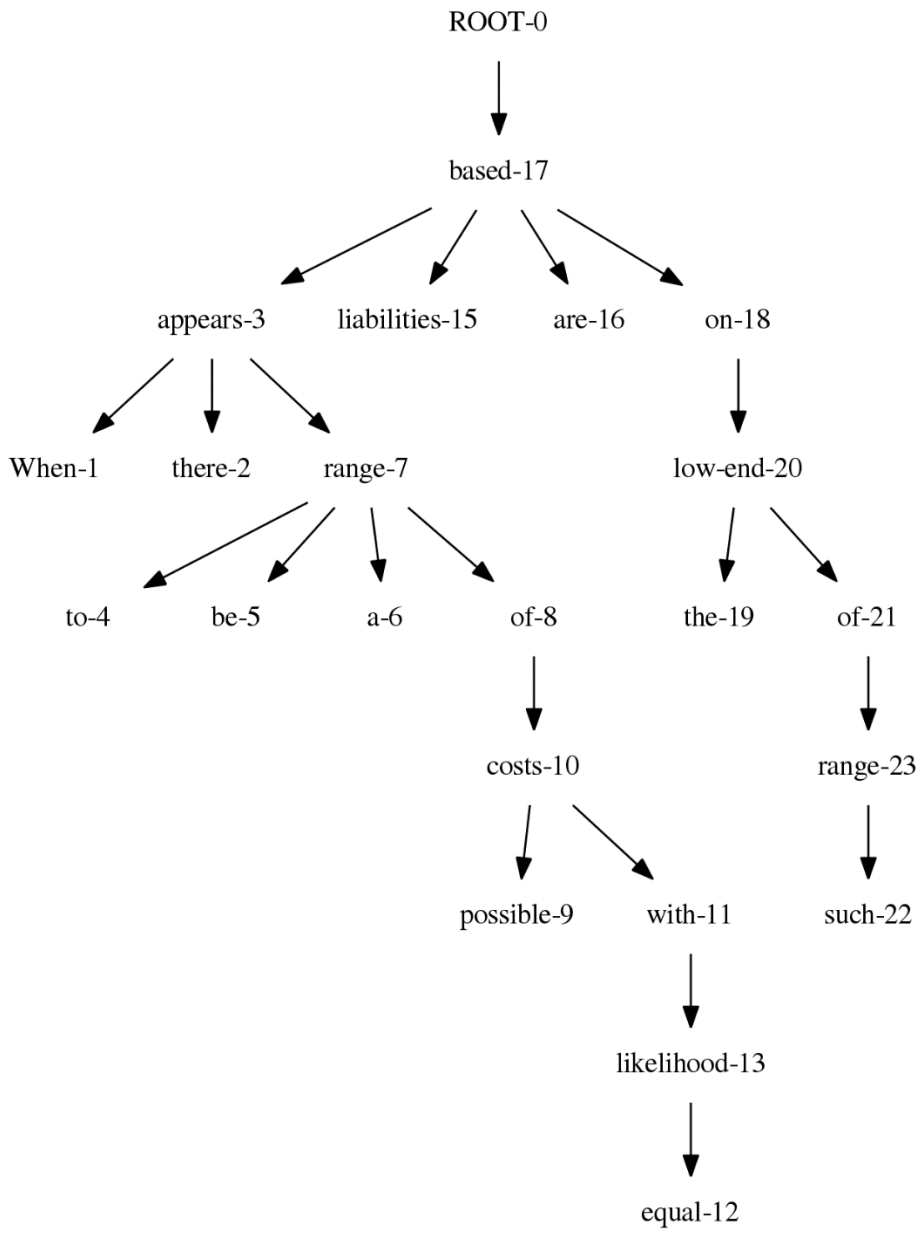


Figure 8: Example Dependency Tree, Revisited

which expresses much more epistemic modality. In this instance, we would say that everything under the scope of *appears* is hedged, but the entire sentence is not. However, in the following example,

> *Other clients and merchants are **likely** to take similar actions in the future.* (Visa 10-K, 2013)

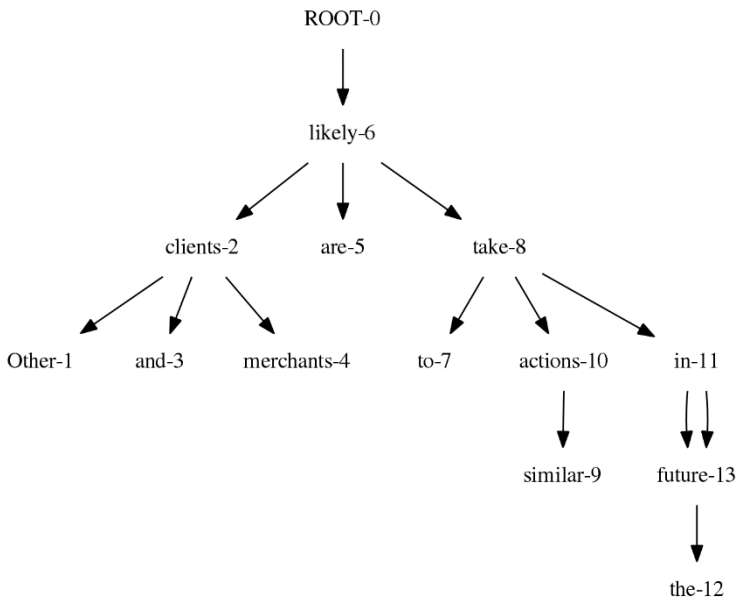


Figure 9: Hedge at Root of Dependency Tree

Likely has governance over the entire sentence, and we can clearly see that the sentence is hedged. The issue with the definition that a sentence is hedged if its root is a hedge cue is that, within this corpus, such a phenomenon only occurs 2,614 times. This could indicate that fully-hedged sentences are rare in the selection, but sentences such as

> *However, the risk of environmental liabilities cannot be completely eliminated and there can be no* ***assurance*** *that the application of environmental and health and safety laws to Agilent will not require us to incur significant expenditures.* (Agilent 10-K, 2013)

suggest otherwise. A process to circumvent these issues would be the result of future studies.

Hedged language is unsurprisingly prevalent in financial texts, which have simultaneous goals of mitigating negative news while not legally committing to positive propositions. In this chapter, we have seen how hedging cues are distributed within our corpus, how they interact with sentiment terms, and how the grammatical properties of hedge cues can be used to determine the speculative value of a sentence. In Chapter 5, we conclude our discussion of these linguistic properties with a summary of the work still to be done within this field.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

The language of financial text is multi-faceted and difficult to quantify. Because a great deal of text in this domain is dedicated to influencing public perception, it can be quite challenging to interpret the intentions of its authors. This study offers insight into two revealing aspects of financial reports, in the hope of opening doors to further research in the language of business.

5.1 DISCUSSION

Overall, the results of this study were unsurprising. We anticipated a significant presence of both sentiment and hedging, and these phenomena are both clearly represented across industries and years. In Chapter 3, we saw how financial texts can express emotion, despite their ostensible objectivity, and in Chapter 4, we explored the ways in which epistemic modality is represented in these texts and how it interacts with sentiment.

As anticipated, the binary dictionary created by Loughran & McDonald (2011) was insufficient to account for the various degrees to which certain words carry positive or negative sentiment. In order to develop a—more appropriate—gradient dictionary, we used the input of over 300 human annotators to produce an average score and standard deviation for each L&M lexical item. Because the annotators in our pilot studies achieved relatively high inter-rater reliabilities, we feel confident that the product of these experiments is a useful and much-needed tool in the field of sentiment analysis, though it is important to recognize the difficulty inherent in quantifying such a subjective aspect of language.

Additionally, our work in hedging, a field most often explored in the context of academia or the hard sciences, breaks ground by adding the domain of finance to a discussion of content mitigation. Like researchers, business managers stand to lose a great deal by committing to uncertain propositions. It is natural, then, that so much hedged language exists in financial documents. The lexicon developed for this project is one effective way to quantify modality and does not require the prosodic cues normally associated with non-lexical hedging.

Finally, our method for classifying speculative language incorporates knowledge of grammatical relations and subjective assessments of hedge cues to determine how much of a sentence can be considered hedged. Though imperfect, it is a new approach which does not require the use of complicated learning algorithms and which produces fewer misclassifications than methods which simply tally the presence or absence of a hedge cue.

5.2 Future Work

We have now seen how sentiment and hedging are represented within our corpus of corporate annual reports. Though each has a non-trivial presence and frequently interacts with the other, it remains to be seen how the presence of a hedging governor impacts the sentiment value of a word. One future study could combine the results of the annotation experiment with our scope algorithm to determine if there is a difference in the sentiment of a word that falls under the governance of a hedge and that same word without such a governor. If this is the case, we would expect a word that is assigned a sentiment score of -2 in most contexts to be given a -1 under the scope of a hedging word.

Another expansion of the sentiment work is to develop our own sentiment dictionary based on the work of Loughran & McDonald, excluding words that do not contain sentiment in a financial context or that do not appear within financial documents. This dictionary would also be

based on positive and negative sentiment, but would exist on a gradient from -2 to +2, possibly ranked based on the average annotations given by our experiment. We believe the standard deviation in sentiment scores is a good measure of the confidence we have in assigning each word to a gradient value.

Beyond positive and negative, Loughran & McDonald also created lists for modal, litigious, and uncertain dimensions of sentiment. Using the same experiment described in the thesis, a simple study would be used to create gradient dictionaries for these categories and others. Additionally, all of these dictionaries can be modified for other domains or enhanced to be domain-nonspecific.

Within hedging, our work could take the direction of perfecting the classification algorithm. As discussed in Chapter 4, our method has its merits but is certainly imperfect. We would hope to combine ours with other methods to improve the performance of all of these approaches. Similar to sentiment, this hedge research would benefit from an expansion to multiple domains. Furthermore, both hedging and sentiment measures could be compared to firm performance in order to make predictions about future outlook.

Finally, there remains much room for improvement in textual analysis in finance. Sentiment and hedging are two very illuminating aspects of natural language, but there are many more to explore—readability, deception, etc. Countless studies could be conducted on such a large and ever-evolving data set.

REFERENCES

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

Chen, Ray and Lazer, Marius, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement, " stanford. edu, 2013. [Online]. Available: http: //cs229. stanford. edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.

Coates, J. (1990). Modal Meaning: The Semantic–Pragmatic Interface. *Journal of Semantics*, *7*(1), 53-63.

Courtis, J. K. (1986). An investigation into annual report readability and corporate risk-return relationships. *Accounting and Business Research*, *16*(64), 285-294.

Desai, V. S., & Bharati, R. (1998). The Efficacy of Neural Networks in Predicting Returns on Stock and Bond Indices*. *Decision Sciences*, *29*(2), 405-423.

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417-422).

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Ganter, V., & Strube, M. (2009, August). Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In*Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 173-176). Association for Computational Linguistics.

Gunning R. (1968). *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Humpherys, S. L. (2009). Discriminating Fradulent Financial Statements by Identifying Linguistic Hedging. *AMCIS 2009 Proceedings,* 400.

Hyland, K. (1998). *Hedging in scientific research articles* (Vol. 54). John Benjamins Publishing.

Jasic, T., & Wood, D. (2004). The profitability of daily stock market indices trades based on neural network predictions: Case study for the S&P 500, the DAX, the TOPIX and the FTSE in the period 1965–1999. *Applied Financial Economics, 14*(4), 285-297.

Kanouse, D. E., & Hanson, L. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, S. Valins, H. H. Kelley, R. E. Nisbett, & B. Weiner (Eds.), Attribution: Perceiving the causes of behavior. Morristown, NJ: General Learning Press.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, *45*(2), 221-247.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004, May). The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users* (pp. 17-24).

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

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643-1671.

Medlock, B., & Briscoe, T. (2007, June). Weakly supervised learning for hedge classification in scientific literature. In *ACL* (Vol. 2007, pp. 992-999).

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis.*Standford University, CS229(2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*.

Palmer, F. R. (1986). *Mood and modality*. Cambridge University Press.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*. *The journal of finance*, *19*(3), 425-442.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

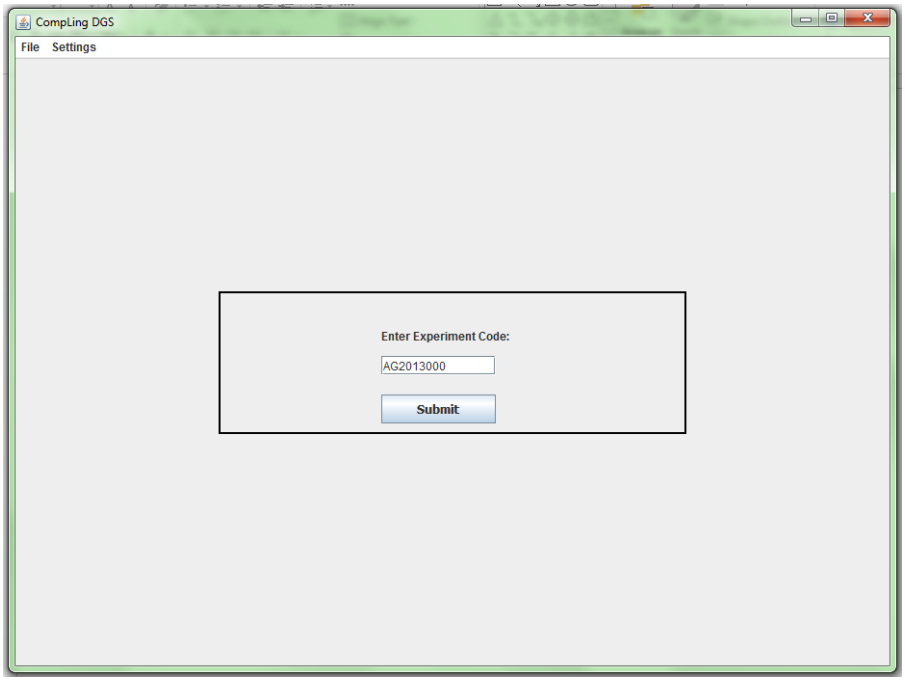
Szarvas, G. (2008, July). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of 46th Meeting of the Association for Computational Linguistics*.

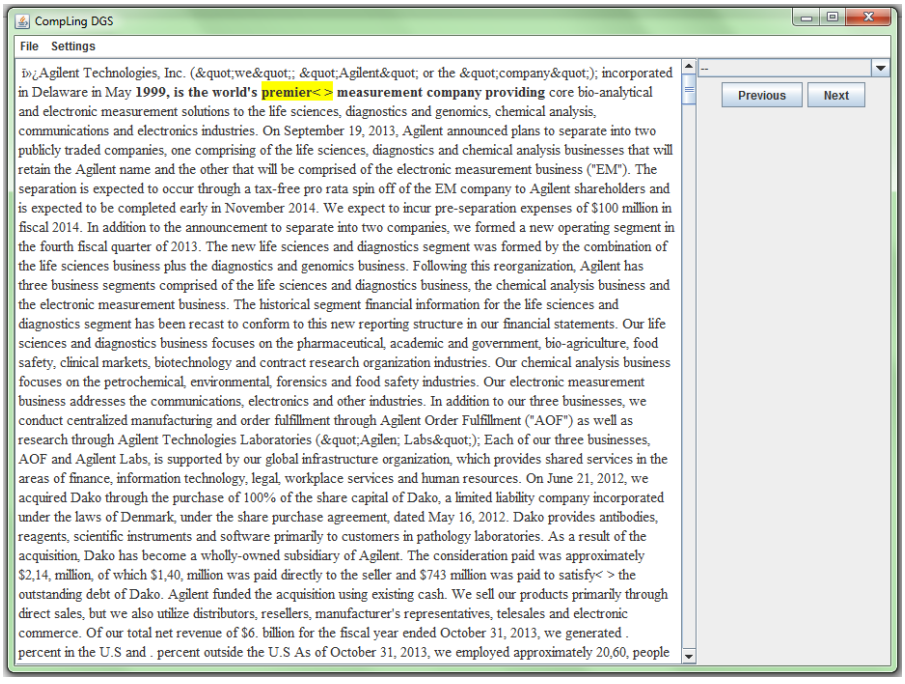
Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, *40*(14), 5501-5506.

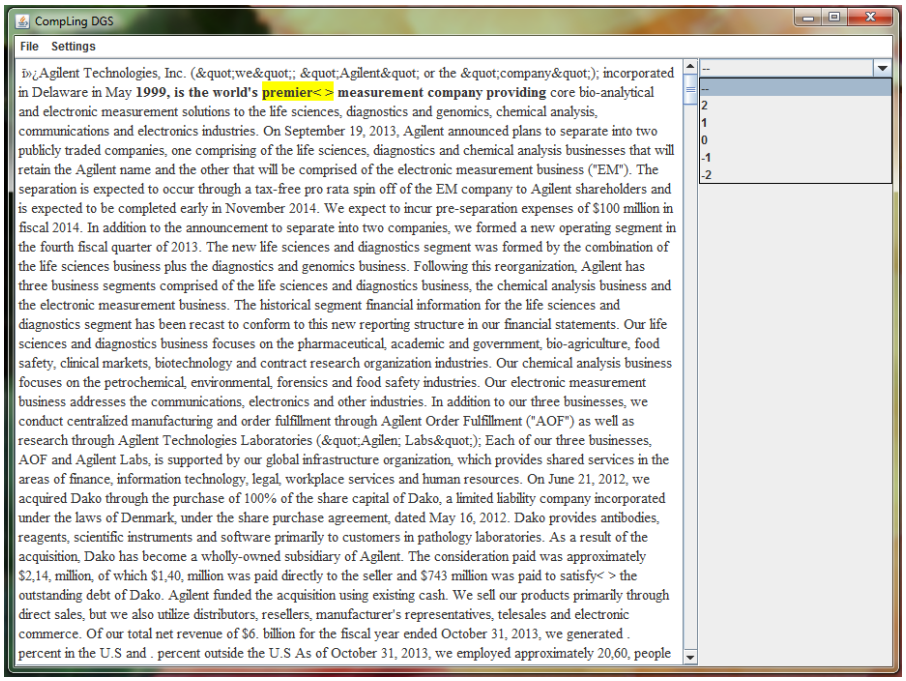
Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... & Patwardhan, S. (2005, October). OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations* (pp. 34-35). Association for Computational Linguistics.

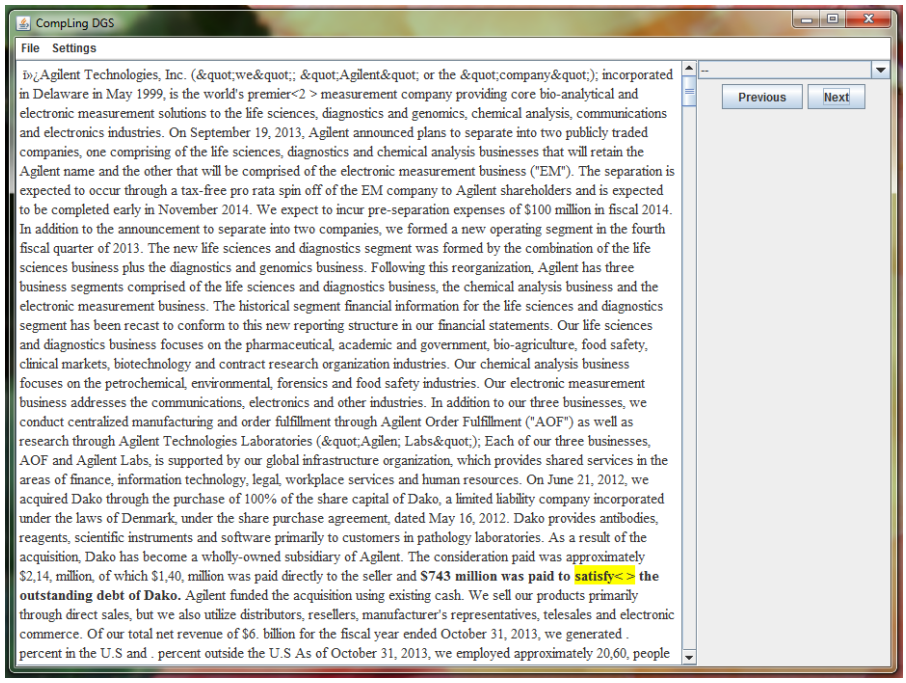
APPENDIX A

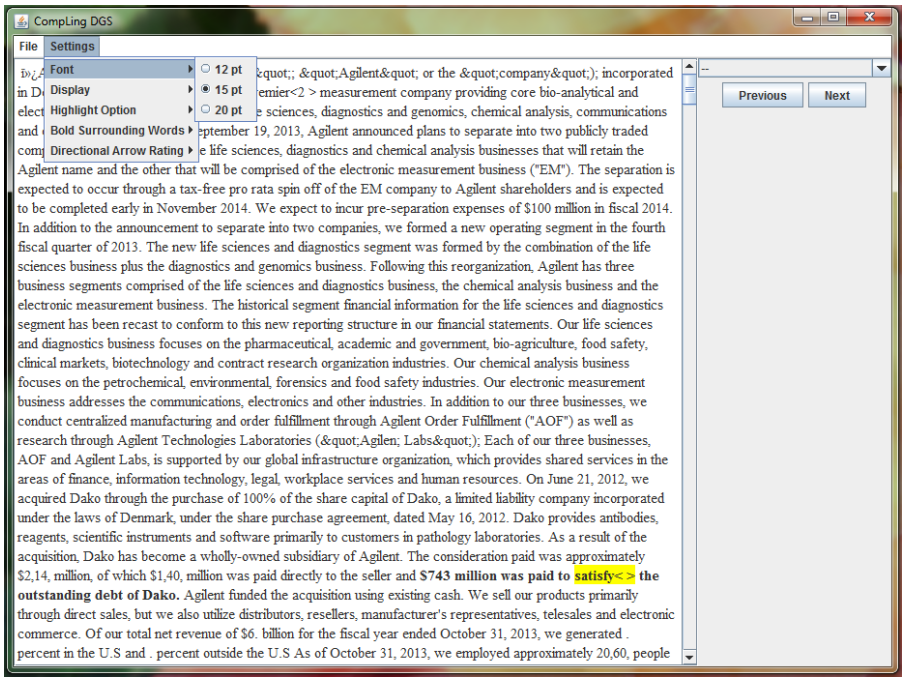
SCREENSHOTS FOR DATA GATHERING SOFTWARE

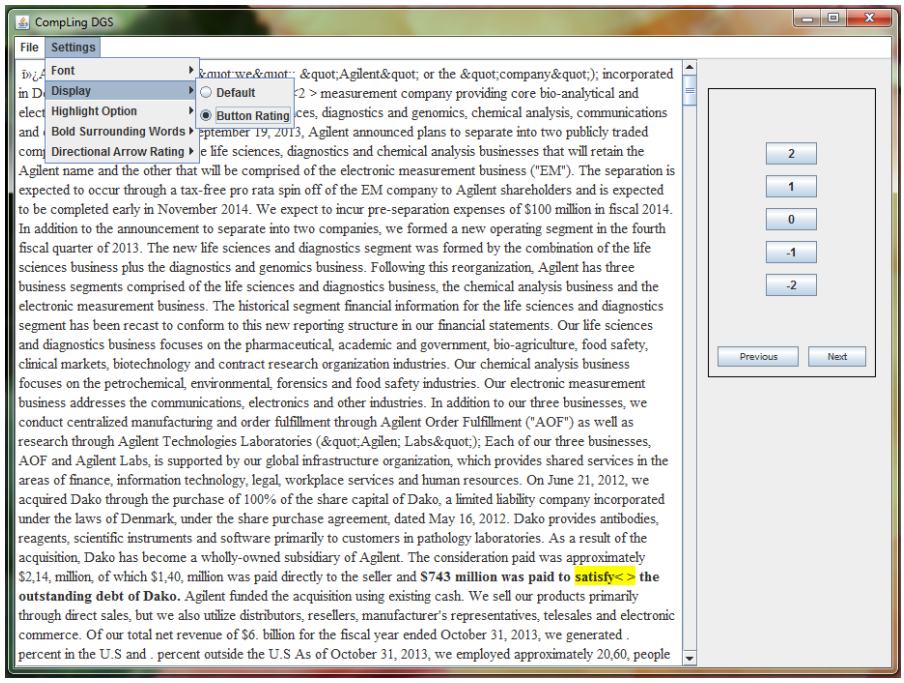


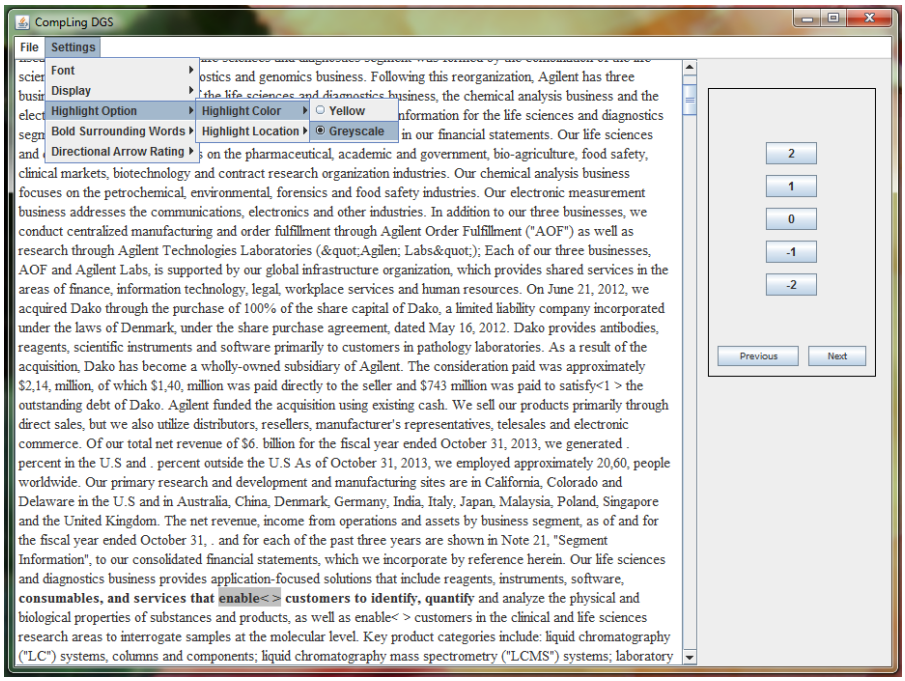


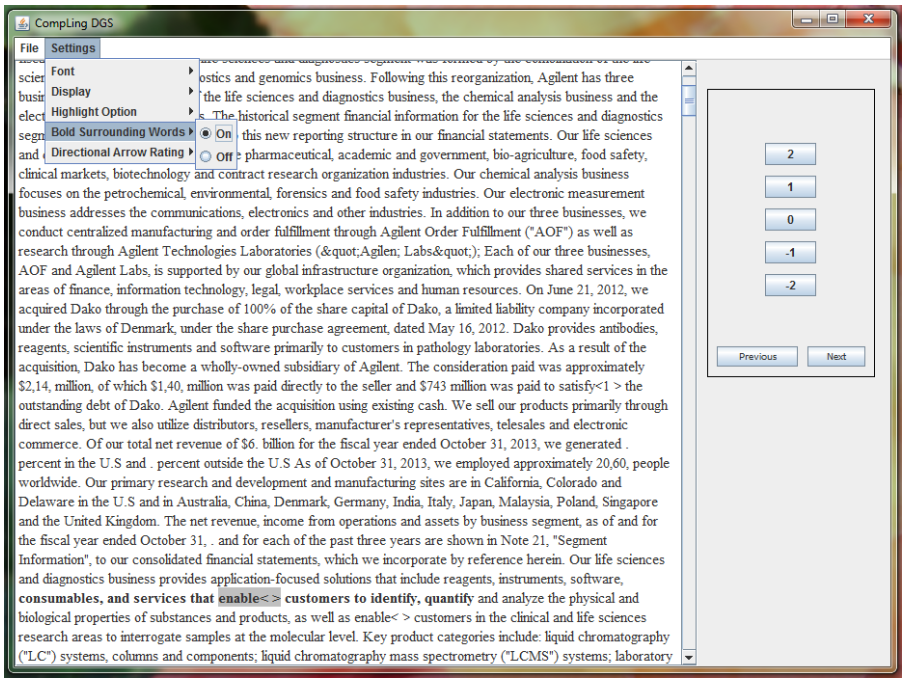

CompLing DGS
File Settings
ï»¿Agilent Technologies, Inc. ("we"; "Agilent" or the "company"); incorporated in Delaware in May 1999, is the world's premier< > measurement company providing core bio-analytical and electronic measurement solutions to the life sciences, diagnostics and genomics, chemical analysis, communications and electronics industries. On September 19, 2013, Agilent announced plans to separate into two publicly traded companies, one comprising of the life sciences, diagnostics and chemical analysis businesses that will retain the Agilent name and the other that will be comprised of the electronic measurement business ("EM"). The separation is expected to occur through a tax-free pro rata spin off of the EM company to Agilent shareholders and is expected to be completed early in November 2014. We expect to incur pre-separation expenses of $100 million in fiscal 2014. In addition to the announcement to separate into two companies, we formed a new operating segment in the fourth fiscal quarter of 2013. The new life sciences and diagnostics segment was formed by the combination of the life sciences business plus the diagnostics and genomics business. Following this reorganization, Agilent has three business segments comprised of the life sciences and diagnostics business, the chemical analysis business and the electronic measurement business. The historical segment financial information for the life sciences and diagnostics segment has been recast to conform to this new reporting structure in our financial statements. Our life sciences and diagnostics business focuses on the pharmaceutical, academic and government, bio-agriculture, food safety, clinical markets, biotechnology and contract research organization industries. Our chemical analysis business focuses on the petrochemical, environmental, forensics and food safety industries. Our electronic measurement business addresses the communications, electronics and other industries. In addition to our three businesses, we conduct centralized manufacturing and order fulfillment through Agilent Order Fulfillment ("AOF") as well as research through Agilent Technologies Laboratories ("Agilen; Labs"); Each of our three businesses, AOF and Agilent Labs, is supported by our global infrastructure organization, which provides shared services in the areas of finance, information technology, legal, workplace services and human resources. On June 21, 2012, we acquired Dako through the purchase of 100% of the share capital of Dako, a limited liability company incorporated under the laws of Denmark, under the share purchase agreement, dated May 16, 2012. Dako provides antibodies, reagents, scientific instruments and software primarily to customers in pathology laboratories. As a result of the acquisition, Dako has become a wholly-owned subsidiary of Agilent. The consideration paid was approximately $2,14, million, of which $1,40, million was paid directly to the seller and $743 million was paid to satisfy< > the outstanding debt of Dako. Agilent funded the acquisition using existing cash. We sell our products primarily through direct sales, but we also utilize distributors, resellers, manufacturer's representatives, telesales and electronic commerce. Of our total net revenue of $6. billion for the fiscal year ended October 31, 2013, we generated . percent in the U.S and . percent outside the U.S As of October 31, 2013, we employed approximately 20,60, people
--
--
2
1
0
-1
-2

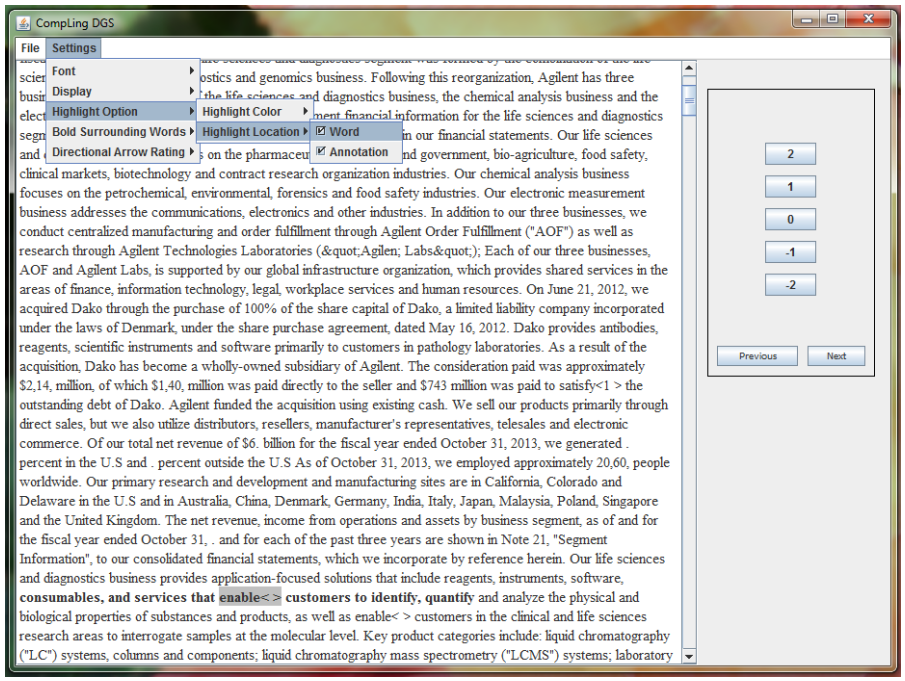

CompLing DGS
File Settings
ï»¿Agilent Technologies, Inc. ("we"; "Agilent" or the "company"); incorporated in Delaware in May 1999, is the world's premier<2 > measurement company providing core bio-analytical and electronic measurement solutions to the life sciences, diagnostics and genomics, chemical analysis, communications and electronics industries. On September 19, 2013, Agilent announced plans to separate into two publicly traded companies, one comprising of the life sciences, diagnostics and chemical analysis businesses that will retain the Agilent name and the other that will be comprised of the electronic measurement business ("EM"). The separation is expected to occur through a tax-free pro rata spin off of the EM company to Agilent shareholders and is expected to be completed early in November 2014. We expect to incur pre-separation expenses of $100 million in fiscal 2014. In addition to the announcement to separate into two companies, we formed a new operating segment in the fourth fiscal quarter of 2013. The new life sciences and diagnostics segment was formed by the combination of the life sciences business plus the diagnostics and genomics business. Following this reorganization, Agilent has three business segments comprised of the life sciences and diagnostics business, the chemical analysis business and the electronic measurement business. The historical segment financial information for the life sciences and diagnostics segment has been recast to conform to this new reporting structure in our financial statements. Our life sciences and diagnostics business focuses on the pharmaceutical, academic and government, bio-agriculture, food safety, clinical markets, biotechnology and contract research organization industries. Our chemical analysis business focuses on the petrochemical, environmental, forensics and food safety industries. Our electronic measurement business addresses the communications, electronics and other industries. In addition to our three businesses, we conduct centralized manufacturing and order fulfillment through Agilent Order Fulfillment ("AOF") as well as research through Agilent Technologies Laboratories ("Agilen; Labs"); Each of our three businesses, AOF and Agilent Labs, is supported by our global infrastructure organization, which provides shared services in the areas of finance, information technology, legal, workplace services and human resources. On June 21, 2012, we acquired Dako through the purchase of 100% of the share capital of Dako, a limited liability company incorporated under the laws of Denmark, under the share purchase agreement, dated May 16, 2012. Dako provides antibodies, reagents, scientific instruments and software primarily to customers in pathology laboratories. As a result of the acquisition, Dako has become a wholly-owned subsidiary of Agilent. The consideration paid was approximately $2,14, million, of which $1,40, million was paid directly to the seller and $743 million was paid to satisfy< > the outstanding debt of Dako. Agilent funded the acquisition using existing cash. We sell our products primarily through direct sales, but we also utilize distributors, resellers, manufacturer's representatives, telesales and electronic commerce. Of our total net revenue of $6. billion for the fiscal year ended October 31, 2013, we generated . percent in the U.S and . percent outside the U.S As of October 31, 2013, we employed approximately 20,60, people
--
Previous
Next

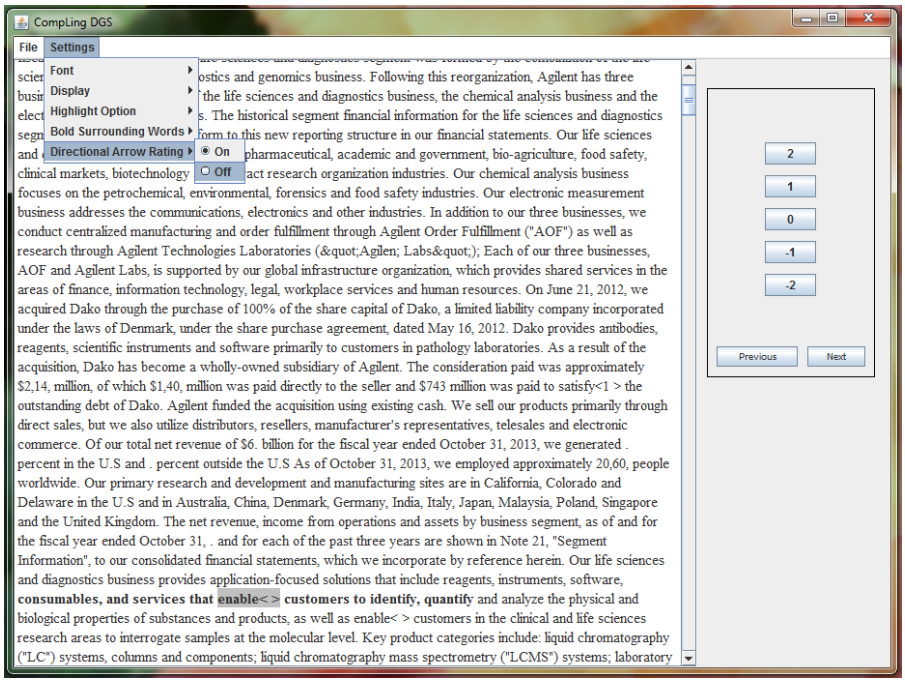

CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
12 pt
15 pt
20 pt
Previous
Next

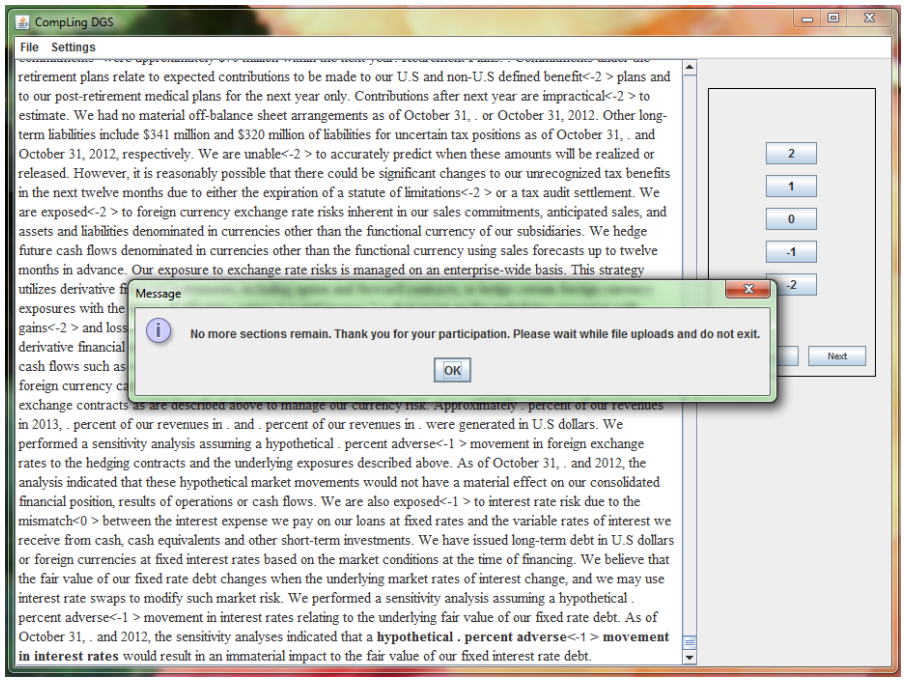

CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
Default
Button Rating
2
1
0
-1
-2
Previous
Next


CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
Highlight Color
Highlight Location
Yellow
Greyscale
2
1
0
-1
-2
Previous
Next


CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
On
Off
2
1
0
-1
-2
Previous
Next


CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
Highlight Color
Highlight Location
Word
Annotation
2
1
0
-1
-2
Previous
Next


CompLing DGS
File
Settings
Font
Display
Highlight Option
Bold Surrounding Words
Directional Arrow Rating
On
Off
2
1
0
-1
-2
Previous
Next


CompLing DGS
File Settings
retirement plans relate to expected contributions to be made to our U.S and non-U.S defined benefit<-2 > plans and to our post-retirement medical plans for the next year only. Contributions after next year are impractical<-2 > to estimate. We had no material off-balance sheet arrangements as of October 31, . or October 31, 2012. Other long-term liabilities include $341 million and $320 million of liabilities for uncertain tax positions as of October 31, . and October 31, 2012, respectively. We are unable<-2 > to accurately predict when these amounts will be realized or released. However, it is reasonably possible that there could be significant changes to our unrecognized tax benefits in the next twelve months due to either the expiration of a statute of limitations<-2 > or a tax audit settlement. We are exposed<-2 > to foreign currency exchange rate risks inherent in our sales commitments, anticipated sales, and assets and liabilities denominated in currencies other than the functional currency of our subsidiaries. We hedge future cash flows denominated in currencies other than the functional currency using sales forecasts up to twelve months in advance. Our exposure to exchange rate risks is managed on an enterprise-wide basis. This strategy
utilizes derivative fi
exposures with the
gains<-2 > and loss
derivative financial
cash flows such as
foreign currency ca
exchange contracts as are described above to manage our currency risk. Approximately . percent of our revenues in 2013, . percent of our revenues in . and . percent of our revenues in . were generated in U.S dollars. We performed a sensitivity analysis assuming a hypothetical . percent adverse<-1 > movement in foreign exchange rates to the hedging contracts and the underlying exposures described above. As of October 31, . and 2012, the analysis indicated that these hypothetical market movements would not have a material effect on our consolidated financial position, results of operations or cash flows. We are also exposed<-1 > to interest rate risk due to the mismatch<0 > between the interest expense we pay on our loans at fixed rates and the variable rates of interest we receive from cash, cash equivalents and other short-term investments. We have issued long-term debt in U.S dollars or foreign currencies at fixed interest rates based on the market conditions at the time of financing. We believe that the fair value of our fixed rate debt changes when the underlying market rates of interest change, and we may use interest rate swaps to modify such market risk. We performed a sensitivity analysis assuming a hypothetical . percent adverse<-1 > movement in interest rates relating to the underlying fair value of our fixed rate debt. As of October 31, . and 2012, the sensitivity analyses indicated that a **hypothetical . percent adverse**<-1 > **movement in interest rates** would result in an immaterial impact to the fair value of our fixed interest rate debt.
2
1
0
-1
-2
Next
Message
No more sections remain. Thank you for your participation. Please wait while file uploads and do not exit.
OK

APPENDIX B

DEMOGRAPHIC SURVEY

1 Experiment Code Number

2 Age in years

3 Gender
- Male
- Female
- Prefer not to answer

4 What is the ethnicity with which you most identify?
- White
- Latino
- Black
- Asian
- Native American
- Prefer not to answer
- Other_____

5 Citizenship:
- U.S.
- Other_____

6 Did you complete high school in the United States?
- Yes
- No

6a If Yes, what was the 5 digit zip code of your residence at the time?
6b If no, did you complete high school in a country where English is an official language?

7 Are you currently working in addition to going to school?
- Part-time
- Full-time
- Not working

8 Primary Major or Intended Major (if you have 2 or more, your primary major)
8a Secondary Major (if applicable)

9 In what year at UGA are you?

1
2
3
4
5+
Special student
Graduate student

10 Was your first language English?

Yes
No

10b If No, what was your first language?

11 Rank your fluency in English on a scale from 1 to 5:

1 Novice (Being at UGA is my first real-world experience using English)
2 Below average
3 Average for a U.S. college student
4 Above average for a U.S. college student
5 Expert (I could make my living as a writer if I wanted to)

12 Rank your knowledge of Finance on a scale from 0 to 4:

0 None (No formal college accounting course)
1 Took or am taking my first college finance course
2 Took or am taking my second college finance course
3 Took or am taking my third college finance course
4 Have completed 4 or more college finance courses

13 Rank your knowledge of Accounting on a scale from 0 to 4:

0 None (No formal college accounting course)
1 Took or am taking my first college accounting course
2 Took or am taking my second college accounting course
3 Took or am taking my third college accounting course
4 Have completed 4 or more college accounting courses

14 Rank your knowledge of Linguistics on a scale from 0 to 4:

0 None (No formal college linguistics course)
1 Took or am taking my first college linguistics course
2 Took or am taking my second college linguistics course
3 Took or am taking my third college linguistics course
4 Have completed 4 or more college linguistics courses

APPENDIX C

EXPERIMENT SCRIPT

"Welcome to our Computational Linguistics Financial Sentiment Dictionary Experiment. My name is _____, and I am part of a research team here at UGA. Our goal is to create a sentiment dictionary of words that are used in financial documents, specifically 10-K's. These are the annual reports that publicly traded corporations file with the United States Securities and Exchange Commission (SEC). You do not need any expertise in financial documents or language. If at any time you feel uncomfortable, please tell me. If you need any assistance or have a question, please ask me at any time. This experiment should take about an hour.

If you have any questions or concerns about this experiment, contact the lead researcher, Dr. Janine Elyse Aronson at 706.542.0991 or jaronson@uga.edu). In addition, her contact information is on the informed consent document.

A word's 'sentiment' is its emotional meaning in addition to its literal definition. Certain words in the English language are known to have sentiment, whether used in context (high risk, good taste) or by themselves (happy, sad). When thinking about words in context, one can see that the word risk might have a negative connotation, but the words 'low risk' together might have a positive connotation. We are studying individual words, so if you see something like 'low risk' and 'risk' is the word, it has a negative meaning. The word 'low' is a modifier, e.g., an adjective.

We are interested in a word's sentiment being positive or negative. We use a five (5) point scale from -2 being the most negative sentiment, -1 being somewhat negative, 0 being neither negative nor positive, +1 being somewhat positive, and +2 being the most positive. You will use

our custom Data Gathering Software that will present words to you in context, and let you easily enter your sentiment ratings. Try your best to use context, yet to ignore modifiers (like very, low, high, more, less, etc.)

This experiment involves four main steps:

1. You log into the Terry Network, access the Web to view and agree to the consent document or not, and to answer a short questionnaire.
2. I will demonstrate and we will train to use the easy-to-use Data Gathering Software. You will learn how to adjust the font and the way you enter your opinions about the sentiment of each highlighted word.
3. You will perform a short training run to gain familiarity and comfort with the software.
4. You will run the software on a random sample of text from a randomly selected 10-K annual report. You should take context into account by reading the bolded words around each sentiment word, but you are not expected to read the entire document.

When the experiment is concluded, your results are saved in a secure, encrypted file. Your identity will be protected. You may ask me questions at any time. If you need help with the software, or feel uncomfortable, please inform me.

Ready? Let's begin."

APPENDIX D

EXPERIMENT FEEDBACK

Subject Comment	Experiment Improvement
Without knowing what the motive behind the experiment is I cant (sic) really give much insight but the repetition of certain words caused me to move more rapidly through other words.	We wanted subjects' unbiased annotations and therefore chose not to give too much detail about the experiment's motives.
Entirely too long	All subjects were told the experiment would not last longer than 1 hour, and indeed many took less than 30 minutes to complete their annotations.
Whole sentence containing the word should be bold	We disagree. Sentences can range from just a couple of words (not enough context) to an entire paragraph (too much context). We wanted the subjects to rate the sentiment words themselves, not the sentences in which they appear.
The instructions were kinda (sic) confusing. Just state focusing on the word and not surrounding content.	A line has been added to the script, indicating that subjects need not read the entire 10-K, but should focus on the bolded context surrounding each highlighted word.
greyscale doesn't work	A line has been added to the DGS instructions indicating that subjects must click the "next" button, in order for the grayscale to take effect.
Might need break or percentage complete bar.	The scroll bar in the DGS serves as an indicator of progress.

APPENDIX E

HEDGING LEXICON

Item	Part of Speech	Purpose	Item	Part of Speech	Purpose
about	adjective	approximation	implied	verb	objectivity
almost	adverb	approximation	implies	verb	objectivity
anticipat	verb	prediction	imply	verb	objectivity
apparent	adjective	objectivity	imply	verb	objectivity
apparently	adverb	objectivity	indicat	verb	objectivity
appear	verb	objectivity	indicate	verb	objectivity
approximat	verb	approximation	indicated	verb	objectivity
approximate	adjective	approximation	indicating	verb	objectivity
approximately	adverb	approximation	indicative	adjective	objectivity
around	adverb	approximation	infer	verb	logic
assum	verb	subjectivity	intend	verb	intention
assumptive	adjective	subjectivity	likelier	adverb	probability
assur	verb	subjectivity	likeliest	adverb	probability
assurance	noun	subjectivity	likely	adverb	probability
belief	noun	subjectivity	many	adjective	approximation
believ	verb	subjectivity	may	modal	modality
calculat	verb	logic	maybe	adverb	prediction
claim	verb	subjectivity	might	modal	modality
conclud	verb	logic	most	adjective	approximation
connot	verb	subjectivity	mostly	adverb	degree
connotative	adjective	subjectivity	nearly	adverb	approximation
could	modal	modality	normally	adverb	frequency
deduc	verb	logic	occasionally	adverb	frequency
deductive	adjective	logic	often	adverb	frequency
essentially	adverb	degree	ought	modal	modality
estimat	verb	approximation	partially	adverb	degree
eventually	adverb	prediction	perhaps	adverb	prediction
expect	verb	prediction	plan	verb	intention
feel	verb	subjectivity	possibility	noun	probability
felt	verb	subjectivity	possible	adjective	probability
forecast	verb	prediction	possibly	adverb	probability
generally	adverb	frequency	potential	adjective	probability
guess	verb	subjectivity	potentially	adverb	probability
however	adverb	subjectivity	predict	verb	prediction

Item	Part of Speech	Purpose	Item	Part of Speech	Purpose
predicted	verb	prediction	should	modal	modality
predicting	verb	prediction	show	verb	objectivity
predictive	adjective	prediction	slightly	adverb	degree
predicts	verb	prediction	some	adverb	approximation
presum	verb	subjectivity	somehow	adverb	prediction
presumably	adverb	subjectivity	somewhat	adverb	degree
presumptive	adjective	subjectivity	soon	adverb	prediction
probability	noun	probability	speculat	verb	prediction
probable	adjective	probability	speculative	adjective	prediction
probably	adverb	probability	suggest	verb	objectivity
project	verb	prediction	suggestive	adjective	objectivity
propos	verb	intention	think	verb	subjectivity
quite	adverb	degree	thought	verb	subjectivity
rarely	adverb	frequency	unlikely	adverb	probability
reckon	verb	prediction	usually	adverb	frequency
relatively	adverb	degree	virtually	adverb	degree
seek	verb	intention	would	modal	modality
seem	verb	objectivity			