Jonathan L. Pulliza. An Analysis of Speculative Language in SEC 10-K Filings. A Master's Paper for the M.S. in I.S degree. May 2015 36 pages. Advisor: Stephanie W. Haas

This study applies sentiment analysis techniques to model the usage of speculation within a collection of financial documents. The model is trained on the MPQA corpus to extract features that correlate with speculative sentences and applied to a collection of SEC 10-K documents from a five year period. The documents with the highest amount of speculation contained a different concentration of terms compared to the entire collection, and the sentences mostly consisted of explaining potential risks concerning projects, taxes, and pensions.

Headings:

Sentiment Analysis

Text Mining

Machine Learning

Speculation Detection

# AN ANALYSIS OF SPECULATIVE LANGUAGE IN SEC 10-K FILINGS

by
Jonathan L. Pulliza

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

May 2015

Approved by

_______________________________________

Stephanie W. Haas

# Table of Contents

## 1. Introduction

The field of finance depends on information pertaining to the past to test and to choose the best investment hypotheses. Data from a variety of sources inform those choices, from overall economic trends to company specific reporting. Each piece of information contains a set of assumptions and biases that may obstruct the true performance of the target company. These underlying components may take the form of a stated estimate, such as an assumed year over year growth rate, or involve much more complex modeling.

An important source of information for those interested in the financial performance of a company is the data contained in publicly available filings required by the U.S. Securities and Exchange Commission (SEC) [1]. There is a great deal of structure and guidelines concerning the creation and presentation of these figures, summarized by the U.S. General Accepted Accounting Principles (GAAP), which allows for a more standard analysis and comparison within and across industries [2]. Along with this numerical data, these filings are also rich with textual information, but as with most large textual collections it is difficult to analyze the full depth and meaning of the information. The written text in these documents contains non-numerical auxiliary information, as well as elaboration on the reported figures. These statements have the potential to signal the presence of issues that may not be captured by financial accounting, as well as completely change one's view of the report.

Since these filings are crafted by the executives of a firm, there is a dual necessity to adequately and accurately represent the performance and potential risks of the firm, while also presenting the firm’s health and potential in the best possible light. Because of these potentially divergent needs, the writers of these statements may overemphasize possible strengths while understating potential risks. There is a great deal of regulatory pressure to include accurate and factual information, so writers often rely on adjusting the presentation and expression of a statement to affect its perception to a reader.

Speculation provides a way to express an idea that may not be factual, since by definition it is “a conclusion by abstract or hypothetical reasoning” [3]. This is often necessary and unavoidable when speaking about potential opportunities and risks, since they inevitably involve some estimation regarding the future. Since speculation involves communicating a degree of uncertainty, the expression includes some indication of the likelihood of the occurrence. For example, a writer can choose to state something “is very likely” versus “might be” true. Though both cases are speculative, there is an implicit higher chance of the first use compared to the other, creating some room for interpretation on the part of the reader.

In the case of regulatory filings, a writer can use speculative language to present information that may not be factual while still adhering to the proper guidelines. Outside of cases where a speculative statement is required, there is an opportunity for a writer to use speculation to mold the view of a company. One possible option is to adjust a statement to either over- or under- represent the possibility of a particular outcome. A document could include an overemphasis of positive opportunities while undermining the plausibility of possible risks, such as using the stronger “likely” versus the weaker “may.” There is also an opportunity for firms to add distance between pieces of related information, increasing the amount of cognitive effort to deduce truthfulness, such as projecting a large gain or loss in a table but communicating a decreased or increased likelihood of that figure in another section. Even without a motive for deception, tables cannot hold all of the information necessary to understand the figures, so analysts must find these speculative sentences somewhere in the text.

By keeping track of speculative statements readers can better understand if and how the company is distorting their report and allow for appropriate adjustments to be made. Analysts need to discern between what is factual and what is speculative in order to create an accurate picture of the present and future. Analysts then must be able to identify when an item is being presented in such a way that exaggerates strengths and assuages weaknesses.  Firms that rely heavily on speculation versus factual reporting may be presenting an overly distorted view of the financial health of the company, and investors should be made aware of any underlying issues being distorted. Aside from finance, the study of speculation is important to any field where the discernment of factual information is key, such as the automation of summaries of news events or legal discovery. This study will utilize machine learning techniques to measure speculation in the annual filings of a subset of publicly traded companies, and then categorize the underlying companies for any common characteristics.

## 1.1 Creating a Use Case for this system

The target users for this system are interested in finding potential risks underlying the statements within company filings. They could be very conservative investors who have a higher aversion to risk, or outside stakeholders that are looking for potential weaknesses to either strengthen or exploit. The decision as to whether some pieces of information are worthy of action or not is very dependent on the expertise, judgement and motivations of the individual. One investor may find a sentence indicative of a short term weakness and choose to sell the stock, while another, looking for long-term opportunities and challenges may decide to continue to hold the stock

Given the wide variance in user needs, even if the system achieves perfect precision and recall in respect to speculation, not every sentence that contains speculation will necessarily be a cause for concern for every user. In general these users are more concerned with finding all potential weaknesses than with losing time looking at a sentence that is not speculative. Since there could be a very high financial cost for missing a key piece of information. The system should then be focused on showing users more potential speculative sentences than on minimizing the amount of total sentences showed, and allowing users the opportunity to make the proper decisions. This study will create a system that identifies sentences that are believed to contain speculation.

## 1.2 What is Sentiment Analysis

Sentiment analysis, also called opinion mining, can be defined as the search for author intent and emotion conveyed through text [4]. The goal is to extract features from the text to categorize documents into corresponding labeled groupings. This approach is sometimes limited by the ability of a phenomenon to be recognized and agreed upon between human annotators, since it is difficult to evaluate the effectiveness of a system if humans cannot strictly define a success. Building an automated classification system consists of three main components: the collection of documents, the modeling of features and the selection of features.

## 1.3 What is Speculation

For the purpose of this study, speculation is defined as the modification of a possibly factual statement in order to express a degree of uncertainty. Speculation, hedging and uncertainty are very similar constructs that are often used interchangeably in text mining, and all point to some degree of truthfulness or ability to trace back to facts. A person's ability to sort statements into speculative and non-speculative statements allows for better decisions making. As stated by Vincze et al., "[d]etecting uncertain and negative assertions is essential in most Text Mining tasks where, in general, the aim is to derive factual knowledge from textual data." [5]

Wikipedia is an example of an organization that has a strong motivation to separate fact from speculation, as its content is meant to be encyclopedic and factual while being completely community based. Editors have created a subset of terms called weasels where "[a] word is considered to be a weasel word if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive or ambiguous." **[6]** Wikipedia considers these terms as indicative of authors' intent to emphasize their opinion over communicating facts, and editors can then use those terms to help identify potentially weak statements.

Much of the recent academic research revolves around separating the factual and the speculating pieces of research papers, moving from the scope of sentences down to phrases. These researchers examine and evaluate speculation patterns using the same tools utilized in sentiment analysis, though not directly defining speculation as a sentiment. Using these techniques, certain terms or cues have been found indicative of speculation, such as hedges which indicate an opinion that lacks corroborating evidence. [6] These cues can be grouped by part of speech, including "adjectives or adverbs (probable, likely, possible, unsure, etc.), auxiliaries (may, might, could, etc.), conjunctions (either. . . or, etc.), verbs of speculation (suggest, suspect, suppose, seem, etc.)." [7] This study will utilize sentiment analysis tools to continue to explore the use of speculation within a collection of natural language. Table 1 summarizes previous research that utilized sentiment analysis techniques on a collection of financial documents.

## 2. Survey of Previous Work

### Table 1: Summary of studies

| **Target Document** | **Study** | **Phenomenon** | **Scale** | **Train/Test Set** |
|---|---|---|---|---|
| IPO Prospectus | Hanley and Hoberg 2010 | Informativeness | word | N/A |
| IPO Prospectus, Complete 10-K | Loughran & McDonald | Style | word, sentence | N/A |
| Complete 10-K | Li 2008 | Readability | word, sentence | N/A |
| Complete 10-K | Loughran & McDonald 2014 | Readability | file, sentence, word | N/A |
| Complete 10-K | Chen 2013 | Optimism | multi-word expressions | MPQA |
| 10-K Section 1A | Huang 2011 | Risk Factors | sentence | Human annotated subset |
| 10-K Item 7 | Brown & Tucker 2011 | Content Change | word | N/A |
| 10-K Item 7 | Feldman 2010 | Tone Change | word | N/A |
| 10-K Item 7 | Davis & Sweet 2012 | Optimism | word | N/A |
| 10-K Item 7 | Li, Lundholm, Minnis 2013 | Competition | word | N/A |
| 10-K Item 7 | Pulliza 2015 (Current Study) | Speculation | sentence | MPQA |

## 2.1 Building Features from the Collection

Choosing a human annotated collection for sentiment analysis allows features to be selected from the collection by a co-occurrence statistic to the target subset to determine which features are descriptive of a phenomenon. A common type of collection utilized by researchers are review datasets, which consist of the textual reviews as well as some separate indication of the reviewers' sentiment toward the product such as a thumbs up or a selection on a scale. Pang et al. used a movie-review collection for positive-negative review classification, utilizing the star rating on IMDB reviews [8].

Researchers may also choose to build their own collection, which requires the development of a classification system for human annotations. Boiy and Moens manually annotated the sentences of blogs in English, Dutch and French as having positive, negative, or neutral sentiment in order to train their classification system [9]. This can be a difficult process, which requires a sufficient level of inter-annotator agreement concerning the target classifications. Building a model from an annotated collection allows for automatic classification, but researchers can also utilize a mix of automatic and manual approaches on an annotated collection. This approach is exemplified by Dasgupta and Ng, who used clustering to find the extreme or "easy" cases, then manually labeled the more difficult and ambiguous edge cases to create a more complete and robust training set [10].

## 2.2 Transplanting Feature Sets

While gathering features from the target collection can be effective, it relies on the collection to be pre-annotated in order to train and evaluate the model. Human annotation is a very intensive endeavor which requires strict guidelines and relies on a sufficient level of inter-annotator agreement to be credible. The process does not scale particularly well, so as a collection grows so must the amount of annotators and time needed.

Given this high cost, researchers could utilize a set of words and features that have previously been found by others to correlate with the target phenomenon. There are a few collections of words that have been crafted for general language use that can be used for classification in multiple systems, including WordNet [11]. WordNet is a large lexical database of nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms, each expressing a distinct concept, similar to a thesaurus. It can be used to extend a feature set, assuming that synonyms of positive words are positive and antonyms negative [12]. Aman extended a set of corpus based features using Roget's Thesaurus and WordNet to automatically categorize text to fit into Ekman's six basic emotion categories (happiness, sadness, anger, disgust, surprise, fear and no-emotion) [4]. There was a project to extend WordNet called SentiWordNet, which is an opinion lexicon derived from WordNet where each term is described with numerical scores indicating positive, negative or neutral sentiment information [13]. The polarity of a sentence can then be grouped as a function of the scores given by the SentiWordNet system, thus creating a categorization model independent of any particular corpus.

Another approach is to train models on an annotated set and to transplant the system onto another non-annotated collection. A main benefit of training on an annotated set is the freedom to identify complex features that may otherwise be unnoticed when using a more general lexicon. One who uses this technique must assume that the annotated and unannotated corpus share enough key characteristics that they would then also exhibit the target phenomenon in the same way. This may be as elementary as the documents being written in English versus French, or may include more subtle structural and stylistic choices, such as subject matter and reading level. The larger lexicons also deal with these issues, but are built for general purposes and not necessarily to fit the needs of any particular research objective. By selecting an appropriate annotated corpus, researchers assert control over these factors that affect the system, and can then adjust and optimize the system to accomplish their goals.

## 2.3 Issues with transplanting feature sets

The domain of the collection often determines the sentiment orientation of particular words or phrases. A term which may have a strong correlation to the positive class in one domain may have the opposite relationship in another. For example, the words “drama” and “intense” may have a positive correlation in film reviews, but may be correlated to the negative class when applied to hotel reviews. There has been a considerable amount of effort put into building repositories of sentiment terms, but the use of these terms across corpora have had some issues. There have been studies on the effectiveness of implementing models trained on one domain onto another, and often there is a strong overlap between similar domains (movies, books, product support, and knowledge base) [14]. Financial documents are a domain where the polarity of the most common use of a word is not necessarily applicable to the usage [15]. The terms “cancer” and “crude” are largely viewed as negative in most contexts, but are more likely identifiers for industry segment than a negative financial event (healthcare and energy respectively). Researchers concerned with this issue may choose to build domain specific classifiers into their model, where the system adjusts the classification of a document based on prior knowledge of its domain [16].

Beyond domain, there are also a few other issues with building a model for sentiment analysis. Unlike other fields of text mining, sentiment analysis has been found to be highly sensitive to the scale of the unit being analyzed, whether judging the sentiment of a given word or document. This problem goes beyond machine learning technology, as Nasukawa and Yi determined that inter-annotator agreement was consistently higher on small scale units of text compared to the whole document. [17] There is also the issue of negation, as machine learning is more effective with direct expression of a sentiment (using positive or negative words directly) than indirect expression (*not good*, *not very tall*) [18]. This may require a more sophisticated model then just term weighing, such as using a conditional random field model, which relaxes the feature independence assumption in other models [19]. Councill et al. used CRF with an English dependency parser on product reviews to create a system to detect the scope of negation [20].

## 2.4 Using the MPQA Opinion Corpus

MPQA Corpus is a collection of news articles manually annotated for private states, which are defined as statements “which can only be subjectively verified: i.e. states of mind, volition, attitude, etc.” [21] A breakdown of the documents in the collection is shown in Table 2.

**Table 2: Topics and descriptions of documents in the MPQA Corpus** [22]

| Topic | Description |
|---|---|
| Argentina | Economic collapse in Argentina |
| Axis of Evil | U.S. President’s State of the Union Address |
| Guantanamo | Detention of prisoners in Guantanamo Bay |
| Human Rights | U.S. State Department Human Rights Report |
| Kyoto | Kyoto Protocol ratification |
| Settlements | Israeli settlements in Gaza and the West Bank |
| Space | Space missions of various countries Relationship |
| Taiwan | Relationship between Taiwan and China |
| Venezuela | Presidential coup in Venezuela Presidential |
| Zimbabwe | Presidential election in Zimbabwe |

The system classifies privates states into five categories: sentiment, agreement, arguing, intention, and speculation [23]. Since this collection was built for the specific purpose of training models, there was a tremendous focus on setting up and training annotators to identify expressions of private states [22]. MPQA Corpus version 2.0 contains the attitude and target annotations of 344 documents (5,957 sentences). The target of an annotation is defined as the topic of the speech event or private state. The annotations also include their intensity, which are illustrated in Table 3 based on the target sentiment.

**Table 3: Measures of intensity for different attitude types** [22]

| Attitude Type | Measure of Intensity | Example |
|---|---|---|
| Sentiment Positive | degree of positiveness | like < love |
| Sentiment Negative | degree of negativeness | criticize < excoriate |
| Agreement Positive | degree of agreement | mostly agree < agree |
| Agreement Negative | degree of disagreement | mostly disagree < completely disagree |
| Arguing Positive | degree of certainty/strength of belief | critical < absolutely critical |
| Arguing Negative | degree of certainty/strength of belief | should not < really should not |
| Intention Positive | degree of determination | promise < promise with all my heart |
| Intention Negative | degree of determination | no intention < absolutely no intention |
| Speculation | degree of likelihood | might win < really might win |

For the purposes of this study the focus will be on the speculation annotations, with their intensity based on the degree of likelihood of the event, such as “might” being tagged as a lower intensity than “really might.” The examples in Table 4 are extracted from the corpus for each level of speculation intensity. A sentence that does not contain a speculation annotation was deemed by the annotators as not containing a speculative private state.

**Table 4: Speculative annotation intensity examples**

| **Speculative Sentence** | **Intensity** |
|---|---|
| However, it is extremely unlikely that the leak occurred during the flight from Paris to Memphis, Hakansson said. | High |
| Hugo Chavez is very likely to be soon return to the post of president of Venezuela. | Medium-high |
| Six degrees, the professor believes, is unlikely, since that kind of thing has not happened on the planet in the past 420,000 years. | Medium |
| And it is possible that, at some given time, OPEC may hand it the bill for not collaborating in the reduction of the oil supply. | Low-medium |
| Such tiny devices could one day fit into cells and supervise biological processes, or even synthesize drugs. | Low |

The corpus has been used as a training corpus for other machine learning systems. A key example relevant to this study is the work done by Chen et al., where the researchers utilized the MPQA corpus to train a model to measure optimism using multi-word expressions (MWEs) instead of single terms. This model was then transferred onto a set of SEC 10-K filings, where they found that managers used negative and positive MWE’s to mitigate the effects of declining earnings and to accentuate positive aspects [24]. Their research indicated that the MPQA is an appropriate training corpus for modeling the language in 10-K documents, so this study will continue under that assumption.

## 2.5 Defining the SEC 10-K

The U.S. Securities and Exchange Commission states its mission as the protection of investors and to maintain fair, orderly and efficient markets [25]. As part of its mission, it requires certain entities that access the capital markets to file certain public filings, including the Form 10-K, which serves as a comprehensive overview of the business' financial condition and audited financial statements. The form is made of up of over twenty sections, including sections dedicated to risk factors, legal proceedings, and the management's discussion and analysis of operations. Companies are required to describe these financials and along with key underlying assumptions, as well as address many of other business concerns in their filings. These filings must also be written in accordance to the SEC's "plain English rule," which is designed to ensure shareholders can read and understand any communications provided by management concerning the company [26].

## 2.6 Textual Analysis of 10-K's

Researchers have been looking at the text of 10-K's as a rich source of information, especially in respect to readability and financial health. Li created a system that measured readability of 10-K flings using a statistical measure that combines the numbers of words per sentence and syllables per word called the Fog Index. Firms that had higher Fog Index scores had lower earnings, and lower scores were correlated with persistent positive earnings [27].

Loughran and McDonald also studied the readability of annual reports, and utilized multiple readability measures including file size, sentence size, and complexity of words as measured by average word length. Their readability measures drew from guidance in the SEC Rule 421(d), which advises firms to comply with plain English principles, including short sentences, active voice, and no multiple negatives [26]. They found that file size was not only the simplest measure of readability but also had a higher correlation to other readability constructs such as average words per sentence and percentage of complex words compared to the Fog Index [28]**.** They also created a list of terms that are typically negative in a financial sense and added terms that were frequent in documents filed by firms under investigation for fraud under Rule 10b-5, and firms disclosing at least one material weakness in internal control. This list included terms such as loss, claims, impairment, and litigation. They highlighted negative words that appeared throughout the 10-K, as well specifically in the Item 7 section [15]**.**

## 2.7 Analyzing of Item 7 and Other Documents

There has also been some research into specific sections of the 10K, especially Item 7 Management Discussion and Analysis (MD&A). The SEC had some concern that many companies provide only boilerplate disclosures in MD&A instead of thoughtful and useful commentary [29]. Brown and Tucker investigated the extent of boilerplate discussion by comparing year to year changes in the section, and found that firms make larger changes to the section when they experience large economic shifts from the previous year [30]. Feldman classified all of the words in MD&A as positive/negative in 10-K’s and 10Qs, and found stock market changes are significantly related to changes in the amount use of one over the other [31]. Davis and Sweet in “Managers’ use of language across alternative disclosure outlets” found that managers whose incentives create increased sensitivity to current stock prices will be more likely to use pessimistic language in the MD&A section than in the earnings press release in an attempt to delay negative a stock price response. They used DICTION 5.0 to measure use of pessimistic and optimistic language [32].

Beyond financial performance, there are also some more complex relationships imbedded in the text. Li et al in “A Measure of Competition Based on 10-K Filings” focused on finding textual evidence of competition within the firm’s industry in the MD&A section. Their text based statistic found a similar level of competition compared to the Herfindahl Index, which is the generally accepted measure of competition based on the market share each firm holds within an industry  [33]. Another key section of the 10-K is the “Section 1A – Risk Factors,” and Huang used K nearest neighbor to categorize to 25 different risk factors present within a set of filings [34].

Other official financial documents outside of the 10-K have also been utilized for sentiment analysis. Hanley and Hoberg studied the information content of Form 424, the Initial Public Offering (IPO) prospectuses, and measured the residual content when taking into account previous IPO offerings from other firms as a measure of informativeness [35]. Loughran and McDonald analyzed the IPO prospectus as well as 10-K filings and found that firms are more likely to comply with the “Plain English” rule before an equity instance [26].

## 2.8 Defining this Study

Since speculation corresponds to the expression of something that may or may not be factual, financial documents that contain a high amount of speculation are reporting a disproportionate amount of non-factual information. This could correspond to some financial weakness within a firm that warrants further examination. This system will be trained and tested on the MPQA dataset, with the goal of maximizing the recall of statements containing speculation. It will then categorize sentences within Item 7 of the SEC 10-K filings of a group of firms as containing or not containing speculation, and filings with the highest percentage of speculative sentences will be grouped and analyzed based on word usage to identify any underlying themes that are particular to the subset.

# 3. Methods

## 3.1 Extracting the Training/Testing Corpus

The system will use sentences containing an annotation of speculation within the MPQA Opinion Corpus. Any sentence containing speculation is considered speculative, regardless of the intensity tagging. For the purposes of this study it is more important that the system identifies the usage of speculation versus the intensity of the private state, which as defined by the MPQA corresponds to the likelihood of the event occurring. All of the training and evaluation will be on the sentence level. The target documents are formally written and carefully crafted, so the sentence level offers a clear and complete unit of analysis, unlike another form of communication such as tweets, which do not follow common grammar rules and are difficult to parse in isolation.

The total number of speculation annotations is approximately 3% of the total annotations, totaling 309 unique sentences. The training set will contain half of these sentences, along with an equal number of sentences that do not contain speculation at all. This balanced training set will ensure features that distinguish between the two classes are captured instead of overweighing the majority class, in this case non-speculation. The remaining withheld sentences containing speculation will be combined with 2,597 sentences that do not contain speculation, which reflects the approximate distribution of the MPQA sentences in the corpus. This design is based on the assumption that this distribution is approximately the same as the most common use of language, including the 10-K's. The performance of the system on this test set will then serve as guidance for what to expect from the system on the unlabeled sentences in the 10-K collection.

**Table 5: Training and testing set breakdown**

| | Speculation Sentences | Non-Speculation Sentences | Total Sentences |
|---|---:|---:|---:|
| Training | 155 | 155 | 310 |
| Test | 154 | 2,597 | 2,751 |

## 3.2 Data Source for 10-K's

The target collection is available through the SEC EDGAR online database. It holds the full text of 10-Ks in a text file format. Each file is labeled by the firm's unique Central Index Key (CIK), which is the unique key given to all filers to the SEC [36]. These codes can also be used for industry groupings and for retrieving any other metadata, such as firm name, industry and size. The industry chosen for analysis is the Utility group, which consists of thirty companies, all of which were on the exchange for the entire 2009 to 2013 period. Selecting firms all from the same industry helps control for industry specific jargon which may skew the text analysis. It also allows for a more appropriate comparison across firm performance. Some of the filings indicated that the information that should be in Item 7 were moved to another section within the 10-K or into another document completely, so were removed from the sample, leaving 114 documents remaining.

**Table 6: Breakdown of industry filings 2009 – 2013**

| Filings | 2009 | 2010 | 2011 | 2012 | 2013 | Total |
|---|---|---|---|---|---|---|
| Not in Item 7 | 8 | 7 | 7 | 7 | 7 | 36 |
| Item 7 | 22 | 23 | 23 | 23 | 23 | 114 |
| **Total** | **30** | **30** | **30** | **30** | **30** | **150** |

The 10-K filings themselves are composed of both textual data and financial tables. These tables are rich with financial data but lack the textual content useful for this process. Like the 10-K subset created by Chen et al [24], for the purpose of this study those tables, along with many of the XML tags added on to the document by the firm for the interactive functionality on the EDGAR website, were removed using the Python library Beautiful Soup [37]. Only Item 7 is considered in this study, and the other sections are completely ignored. Table 7 shows the final sentence distribution.

**Table 7: Sentence distribution**

| Year | Max | Min | Median | Total |
|---|---|---|---|---|
| 2009 | 2,669 | 125 | 1,055 | 24,241 |
| 2010 | 2,505 | 126 | 976 | 23,252 |
| 2011 | 2,895 | 134 | 906 | 23,017 |
| 2012 | 4,571 | 132 | 831 | 25,089 |
| 2013 | 4,572 | 69 | 873 | 25,043 |

# 4. Results

## 4.1 Feature Extraction

The features were extracted from the training set using the program LightSide Researcher's Workbench [38]. All stop words were initially included in the generated features in order to ensure no modifying terms such as "could" or "would" were missing from the set. The final feature set included Word/Part of Speech pairs, with proper nouns and determiners (the, that, these) filtered out of the set since they would be unlikely to distinguish between classes across domains. The most discriminating features that correlated with the speculation class also coincided with the auxiliary words highlighted by Velldal et al. [7]

**Table 8: Top 10 features extracted by Pearson's r correlation to speculation**

| **Feature** | **Correlation** |
|---|---|
| may / MD | 30.55% |
| could / MD | 28.19% |
| might / MD | 25.55% |
| be / VB | 17.26% |
| more / JJR | 16.72% |
| have / VB | 16.55% |
| analysts / NNS | 15.20% |
| expected / VBN | 15.20% |
| than / IN | 14.48% |
| perhaps / RB | 14.05% |

## 4.2 Model Creation and Evaluation

With the features in place, a series of models were created and trained. The models were trained and tuned on the training set. The models utilized for this study were Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Decision Trees. Once the models were trained, they were evaluated on their performance on the test set, and combined into three different meta-classifiers: Two or more models agree (Majority Vote Classifier), unanimous vote, and at least one model predicts speculation (Max Classifier). The models and their evaluation metrics are in Table 9, ranked by recall. The Max Classifier performed significantly better on recall in respect to the speculation class, but also had the greatest false positive rate. Compared to the classifier with the next highest recall (Majority Vote), the user would see three extra sentences that did not contain speculation versus missing an extra ten out of one hundred.

**Table 9: Model metrics on the test set, ranked by recall (N=2,751)**

| | # of sentences (% of N) | # of sentences (% of N) | # of sentences (% of N) | # of sentences (% of N) | | |
|---|---|---|---|---|---|---|
| **Model** | **True Negatives** | **False Positives** | **True Positives** | **False Negatives** | **Precision** | **Recall** |
| Max Classifier | 1,404 (51%) | 1,193 (43%) | 132 (5%) | 22 (1%) | 9.96% | 85.71% |
| Majority Vote | 1,812 (66%) | 785 (29%) | 117 (4%) | 37 (1%) | 12.97% | 75.97% |
| Naïve Bayes | 1,729 (63%) | 868 (32%) | 113 (4%) | 41 (1%) | 11.52% | 73.38% |
| Logistic | 1,866 (68%) | 731 (32%) | 111 (4%) | 43 (1%) | 13.18% | 72.08% |
| SVM | 1,793 (65%) | 804 (29%) | 110 (4%) | 44 (2%) | 12.04% | 71.43% |
| Decision Tree | 2,264 (82%) | 333 (12%) | 92 (3%) | 62 (2%) | 21.65% | 59.74% |
| Unanimous Vote | 2,383 (87%) | 214 (8%) | 73 (3%) | 81 (3%) | 25.44% | 47.40% |

The generally small amount of false negatives seem to be caused by the use of some nominal form of speculation (“the speculation”, “the possibility”) instead of the verb or modifier verb structure (“may happen”, “expect”). The models seemed to have the most trouble with sentences marked as arguing instead of speculation, and there seems to be a large overlap between potential argument and speculation cues. For example, the phrase “I think” was sometimes annotated as arguing rather than speculating. It may be that the MPQA annotators chose arguing as the label for phrases that could be considered speculative, since many of the false positives follow the definition of speculation in this study.

**Table 10: Examples of false positives**

| | |
|---|---|
| While he **believed** the details of Clarkes account to be incorrect, President Bush acknowledged that he **might** well have spoken to Clarke at some point, asking him about Iraq. | Agreement |
| But Commons Leader Mr Cook said: If you look back over the past month there has been no situation in which we have put British troops into the ground civil war and I don’t myself **imagine** that’s going to change. | Arguing |
| Pentagon officials said some prisoners **might** also be sedated during the more-than-20-hour flight, but it was not clear whether that had happened. | Arguing |

## 4.3 Using the Model on the 10-Ks

For the purposes of this study we analyze the 10-K documents using the three meta-classifiers and the best performing single classifier, which was the logistic model. Using a single model is more computationally cost efficient than using a meta-classifier and may be the best options for some users.

Using the LightSide application, each of the four selected classifiers were used to categorize the sentences as either containing speculation or not. These data were used to create aggregated metrics based on the company, year, and percentage of total sentences that were labeled as speculative. Table 11 shows the number and percentage of speculative sentences identified by each classifier. As expected based on results of the test set, the Max Classifier labeled the most sentences as speculative, with over fifty percent of sentences in the corpus labeled as speculation.

**Table 11: Speculative sentence distribution by model (N = 120,642)**

| Model | 2009<br>N=24,241 | 2010<br>N=23,252 | 2011<br>N=23,017 | 2012<br>N=25,089 | 2013<br>N=25.043 | Average<br>N=23,900 | Standard<br>Deviation |
|---|---|---|---|---|---|---|---|
| Max | 12,201<br>(50.33%) | 12,150<br>(52.25%) | 12,176<br>(52.90%) | 13,499<br>(53.80%) | 13,336<br>(53.25%) | 12,672<br>(52.52%) | 9.72% |
| Unanimous | 4,344<br>(17.92%) | 4,231<br>(18.20%) | 4.585<br>(19.92%) | 5,015<br>(19.99%) | 4,956<br>(19.79%) | 4,626<br>(19.17%) | 5.03% |
| Vote | 9,288<br>(38.32%) | 9,172<br>(39.45%) | 9,467<br>(41.13%) | 10,408<br>(41.48%) | 10,260<br>(40.97%) | 9,719<br>(40.28%) | 8.24% |
| Logisitic | 7,967<br>(32.87%) | 7,982<br>(34.33%) | 8,190<br>(35.58%) | 8,954<br>(35.69%) | 8,797<br>(35.13%) | 8,378<br>(34.72%) | 7.19% |

## 4.4 Analyzing Filings with the highest amount of speculation

Each document is ranked by the percentage of speculative sentences present in each classifier, and then the average of those rankings is taken as the indicator of the overall speculation in the document. Table 11 includes the average rank across models, the percentage of speculative sentences per document from each model, and a ranking of the length of each document by number of sentences. Table 12 provides the same breakdown for documents with the lowest level of speculation by percentage of total sentences in the document.

**Table 12: Ranking top speculative documents**

| Average Rank | Max % | Unanimous % | Vote % | Logit % | Rank of # of Sentences |
|---:|---:|---:|---:|---:|---:|
| 3 | 71.63% | 31.70% | 57.01% | 49.04% | 34 |
| 4 | 69.51% | 30.23% | 55.98% | 49.12% | 35 |
| 4 | 68.40% | 32.24% | 56.01% | 48.45% | 42 |
| 6 | 68.38% | 28.96% | 55.20% | 48.45% | 13 |
| 7 | 67.80% | 31.13% | 58.07% | 45.32% | 39 |
| 7 | 68.44% | 30.76% | 57.78% | 44.04% | 61 |
| 10 | 64.48% | 25.69% | 54.26% | 48.75% | 72 |
| 11 | 65.73% | 24.87% | 54.70% | 48.25% | 74 |
| 11 | 65.24% | 23.82% | 53.86% | 51.29% | 96 |
| 11 | 64.97% | 26.88% | 51.16% | 46.62% | 43 |
| 11 | 66.94% | 23.79% | 52.42% | 49.60% | 92 |

**Table 13: Ranking bottom speculative documents**

| Average Rank | Max % | Unanimous % | Vote % | Logit % | Rank of # of Sentences |
|---|---|---|---|---|---|
| 106 | 43.34% | 13.22% | 30.74% | 26.95% | 54 |
| 106 | 41.99% | 13.65% | 32.16% | 26.06% | 55 |
| 106 | 20.63% | 18.25% | 19.05% | 19.05% | 117 |
| 108 | 44.79% | 11.66% | 29.06% | 26.43% | 5 |
| 109 | 41.40% | 13.52% | 30.64% | 25.04% | 6 |
| 109 | 43.74% | 13.50% | 28.90% | 25.71% | 7 |
| 111 | 43.72% | 12.97% | 28.25% | 24.18% | 8 |
| 111 | 38.17% | 12.37% | 28.23% | 28.76% | 114 |
| 112 | 40.48% | 12.88% | 29.33% | 25.96% | 50 |
| 114 | 38.96% | 12.04% | 30.33% | 24.27% | 46 |

The sentences in documents with high speculation were analyzed in order to find words that are highly indicative of the group versus the whole corpus. Stop-words were removed from the collection in order to highlight more informative terms. The words were stemmed before processing to capture the largest amount of count per term. The measurement used was the term frequency multiplied by the inverse document frequency. Below is the list of terms that appear relevant in the highly speculative subset compared to the collection as a whole. There is an added annotation if the word appears on the negative word subset created by Loughran and McDonald [15].

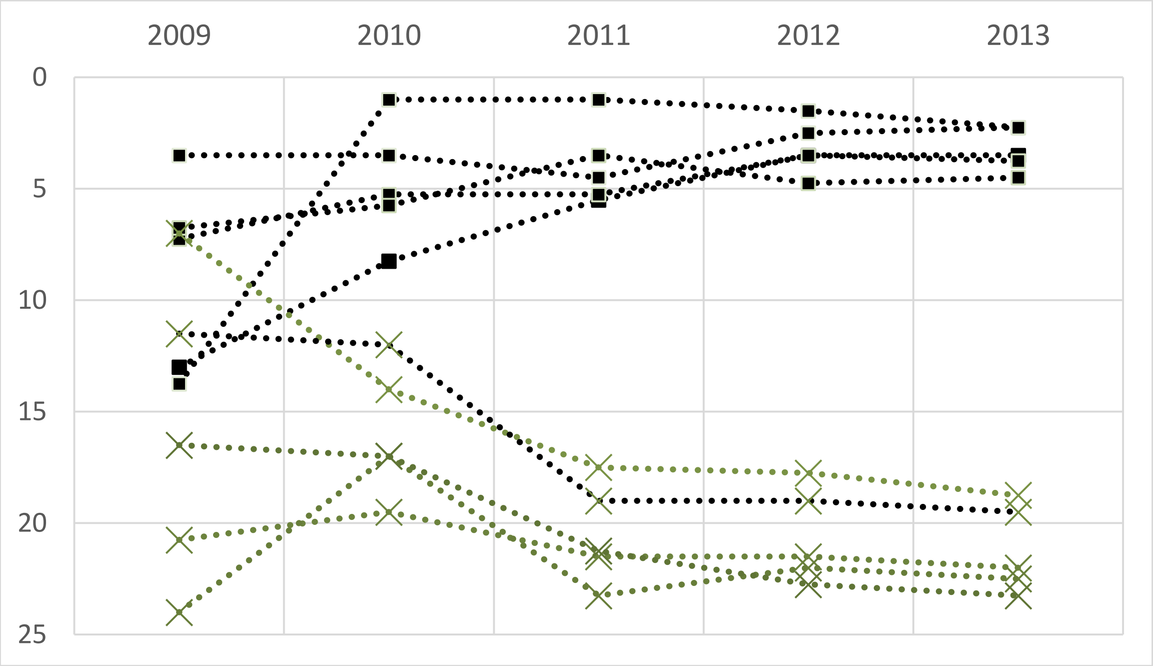
## Table 14: Top ranked terms by TF.IDF within High Speculation Documents versus the total collection

| Term | Rank of TF.IDF | Positive/(Negative) difference in rank compared to the collection | Negative Word List |
|---|---:|---:|---|
| oper | 1 | 0 | |
| decreas | 2 | 1 | X |
| increas | 3 | 1 | |
| rate | 4 | (2) | |
| result | 5 | 2 | |
| electr | 6 | 12 | |
| servic | 7 | 28 | |
| cost | 8 | 3 | X |
| total | 9 | 4 | |
| regul | 10 | 14 | |
| distribut | 11 | 56 | |
| energi | 12 | (6) | |
| expect | 13 | (3) | |
| effect | 14 | 11 | |
| other | 15 | 6 | |
| util | 16 | 13 | |
| tax | 17 | 3 | X |
| custom | 18 | 19 | |
| expens | 19 | (2) | X |
| chang | 20 | (4) | |
| due | 21 | 30 | |
| provid | 22 | (7) | |
| gener | 23 | (14) | |
| invest | 24 | 2 | |
| project | 25 | 7 | |
| offset | 26 | 18 | |
| credit | 27 | (13) | |
| plan | 28 | (9) | |
| addit | 29 | 1 | |
| purchas | 30 | 1 | |
| amount | 31 | 8 | |
| capit | 32 | 9 | |
| fund | 33 | 20 | |
| note | 34 | (12) | |
| include | 35 | (2) | |
| suppli | 36 | 61 | |

## 4.5 Ranking Firm filings over the sample period

Chart 1 highlights the top five and bottom five firms by average percentage of speculation sentences in the year 2013 and then traces their filings back to 2009. The two sets of firms seem to diverge after 2009 and flatten out over the rest of the period. It may be that 2009 was a year where there was a major change in the industry, or perhaps the industry shifted and resettled once the economic recession subsided after 2009 – 2010. Overall, the rankings contain some variability that points to actual changes in use of language and perhaps changes in the individual firm's financial operations.

**Chart 1: Average ranking for the top and bottom 2013 speculation percentage ranking, 2009 – 2013**



## 4.6 Sampling of documents highlighted as speculative

The top ten documents by percentage of speculative sentences were examined to find if there were any prevailing patterns or shared themes. The sentences within those documents that were marked as speculative consisted mostly of explaining a potential risk to future operations. It is expected that these documents state potential risks to the firm, but the model is highlighting a particularly high density of risk hedging statements within these documents. Common topics among these sentences included current projects, taxes, and pension funds.

**Table 15: Examples of sentences that were labeled as speculative**

| |
|---|
| “Although the generating capacity may be higher during the winter months, the facilities are used to meet summer peak loads that are generally higher than winter peak loads.” |
| “As a result, [firm] is exposed to the risk that it may not be able to renew these contracts or that the contract counterparties may fail to perform their obligations thereunder.” |
| “Failure to achieve forecasted taxable income or successfully implement tax planning strategies may affect the realization of deferred tax assets and the amount of any associated valuation allowance.” |
| “Even where collateral is provided, capital market disruptions, the lowered rating or insolvency of the issuer or guarantor, changes in 33the power supply market prices and other events may prevent a party from being able to meet its obligations or may degrade the value of collateral, letters of credit and guarantees, and the collateral, guarantee or other performance assurance provided may prove insufficient to protect against all losses that a party may ultimately suffer.” |
| "The closure of the plan to entrants after December 31, 2013 and the cessation of benefit accruals in 2023 are expected to further lessen the significance of pension costs and the criticality of the related estimates to the Company’s financial statements." |

## 5. Discussion

The model was trained using the MPQA corpus, which had a small percentage of speculation annotations compared to the size of the collection. It would be interesting to include in future research the likelihood notations, perhaps revealing a connection between the uses of speculation intensity and underlying operations, but this study was limited by the amount of speculation sentences available in the MPQA corpus. Transplanting the trained model onto financial documents assumed that speculation is used in the same way across domains and the resulting extracted sentences seem to corroborate that theory, though a full analysis by human annotators would be necessary to confirm.

The most discriminating features that correlated with the speculation class also coincided with the auxiliary words highlighted by Velldal et al. [7], which mostly consisted of modal terms such as *may*, *could* and *might*. The model was limited to extracting features sentence by sentence due to the structure of the corpus. There may be discourse-level features within each 10-K document that are strong indicators of speculation, such as references to previous passages or position within the document.

There was a difference in term usage when comparing the top speculative documents versus the entire collection. The increased usage of terms such as *regulation* and *capital* point to possible themes within these documents, such as dealing with government agencies and raising or investing funds into the firm. Deeper analysis is required from a financial expert in order to assess whether there are real structural challenges that are particular to the top speculative companies, but this system has at least shown that there are discernable differences in language usage for speculative documents.

The rankings of firm filings by year show that there is some variability in language usage across time, perhaps coinciding with the findings of Brown and Tucker that business make a significant amount of changes to filings when there are large economic shifts [30]. The model highlighted a concentration of speculative documents from a small group of firms in the top and bottom of the speculative rankings over the period. Having a high amount of speculation may not be a negative indicator of performance but rather a positive indicator of the quality of the document. It may be the case that all of these firms have the same issues but a small group chose to reveal and explain their weaknesses much more or much less compared to the rest of the industry. It is also interesting that the 2009 ranking is so different from the rest of the years, which could be a function of some industry shift or of the economic volatility of the time. Further research could focus on this pivotal time period, perhaps growing the corpus to include other recessions.

Future analysis could include some more structured groupings, including firms that have recently gone through a large merger, firms with shareholder lawsuits, and firms that have been flagged by their auditors as having poor internal controls, which were all highlighted by the work of Loughran and McDonald [15]. It would also be interesting to see how speculation is distributed across industries and if it correlates with other industry measures such as competition.

## 6. Conclusion

This study created a model for speculative language based on the MPQA corpus and applied the resulting model to a corpus of financial documents. The documents with the most speculative sentences contained a different concentration of terms compared to the complete collection on a TF.IDF basis. Specifically terms such as *regulation*, *fund*, and *supplier* were ranked much higher in the documents with the highest amount of speculative sentences. Upon further examination some of the sentences labeled as speculative contained discussion regarding potential risks to the firm, especially pertaining to projects, pensions and taxes. The model succeeded in labeling these particular groups of sentences that perhaps could be overlooked within these large documents by a human reader.

Speculation is a necessity when communicating possible future events, so there needs to be some level of speculation present in all financial documents. Firms deciding to overuse speculation may be cloaking risks, but firms that choose to underutilize speculation may not be fully expressing the possible opportunities and dangers to the firm. In this study the relative concentration of speculative sentences was taken as the barometer for the abnormality of a document, and therefore its worthiness of further inspection. That does not necessarily mean that any of these firms were using speculation inappropriately. The statements within these filings are very particular to the operations of the firm at a point in time, and one firm's choice to use more or less speculative language compared to the rest of the industry may be more indicative of its particular situation than any planned manipulation of the text for the purpose of obfuscating financial results.

Speculative sentences are often connected to some underlying risk that an event may or may not occur, and capturing the amount of speculation in a document could be critical to a field such as finance which is built around risk modeling. This study will hopefully improve the visibility of this issue within financial filings. If the SEC continues to enforce the "plain English" guidance, these document should reflect the actual underlying sentiment of the firm's management, and therefore serve as an important source of information beyond financial figures.

# Bibliography

[1] "SEC.gov | The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation," 2013. [Online]. Available: https://www.sec.gov/about/whatwedo.shtml#.VRWP5e7F854. [Accessed: 27-Mar-2015].

[2] "Authoritative Source of Guidance," *Federal Accounting Standards Advisory Board*. [Online]. Available: http://www.fasab.gov/accounting-standards/authoritative-source-of-gaap/#gaap. [Accessed: 27-Mar-2015].

[3] O. E. Dictionary, "'speculation, n.'. ." .

[4] S. Aman and S. Szpakowicz, "Using Roget ' s Thesaurus for Fine-grained Emotion Recognition," in *International Joint Conference on Natural Language Processing*, 2007, pp. 312–318.

[5] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, "The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes," *BMC Bioinformatics*, vol. 9, no. Suppl 11, pp. S9–S9, Nov. 2008.

[6] R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas, "The CoNLL-2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning --- Shared Task*, 2010, pp. 1–12.

[7] E. Velldal, L. Øvrelid, J. Read, and S. Oepen, "Speculation and Negation: Rules, Rankers, and the Role of Syntax," *Computational Linguistics*, vol. 38, no. 2. MIT Press, Cambridge, MA, USA, pp. 369–410, Jun-2012.

[8] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002, vol. 10.

[9] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr. Boston.*, vol. 12, no. 5, pp. 526–558, Sep. 2008.

[10] S. Dasgupta and V. Ng, “Mine the Easy , Classify the Hard : A Semi-Supervised Approach to Automatic Sentiment Classification,” in *ACL ’09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 701–709.

[11] “About WordNet,” *Princeton University*, 2010. [Online]. Available: http://wordnet.princeton.edu/. [Accessed: 29-Oct-2014].

[12] S. Kim, M. Rey, and E. Hovy, “Determining the Sentiment of Opinions,” in *COLING Conference*, 2004.

[13] “SentiWordNet.” [Online]. Available: http://sentiwordnet.isti.cnr.it/. [Accessed: 29-Oct-2014].

[14] A. Aue and M. Gamon, “Customizing Sentiment Classifiers to New Domains : a Case Study,” 2005.

[15] T. I. M. Loughran and B. Mcdonald, “When Is a Liability Not a Liability ? Textual Analysis , Dictionaries , and 10-Ks,” *J. Finance*, vol. LXVI, no. 1, pp. 35–65, 2011.

[16] Y. Choi, Y. Kim, and S.-H. Myaeng, “Domain-specific sentiment analysis using contextual feature generation,” *Proceeding 1st Int. CIKM Work. Top. Anal. mass Opin. - TSA ’09*, p. 37, 2009.

[17] T. Nasukawa and J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” *Proc. 2nd Int. Conf. Knowl. capture*, pp. 70–77, 2003.

[18] L. Qiu, W. Zhang, C. Hu, and K. Zhao, “SELC : A Self-Supervised Model for Sentiment Classification,” in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 929–936.

[19] J. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” *ICML ’01 Proc. Eighteenth Int. Conf. Mach. Learn.*, pp. 282–289, 2001.

[20] I. G. Councill, R. Mcdonald, and L. Velikovich, “What ’ s Great and What ’ s Not : Learning to Classify the Scope of Negation for Improved Sentiment Analysis,” in *NeSp-NLP ’10 Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 2010, no. July, pp. 51–59.

[21] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*, vol. 1. 1985.

[22] T. A. Wilson, “Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States,” University of Pittsburgh, 2008.

[23] “MPQA Opinion Corpus.” [Online]. Available: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/. [Accessed: 29-Oct-2014].

[24] C. Chen, C. Liu, Y. Chang, and H. Tsai, “Opinion Mining for Relating Subjective Expressions and Annual Earnings in US Financial Statements,” *J. Inf. Sci. Eng.*, vol. 29, no. 3, 2013.

[25] U. States and E. Commission, “SECURITIES AND EXCHANGE COMMISSION FORM 10-K ANNUAL REPORT PURSUANT TO SECTION 13 OR 15 ( d ) OF THE SECURITIES EXCHANGE ACT OF 1934,” 2015.

[26] T. Loughran and B. McDonald, “Regulation and financial disclosure: The impact of plain English,” *J. Regul. Econ.*, vol. 45, no. 1, pp. 94–113, Nov. 2013.

[27] F. Li, “Annual report readability, current earnings, and earnings persistence,” *J. Account. Econ.*, vol. 45, no. 2–3, pp. 221–247, Aug. 2008.

[28] T. Loughran and B. Mcdonald, “Measuring Readability in Financial Disclosures,” *J. Finance*, vol. 69, no. 4, pp. 1643–1671, Aug. 2014.

[29] R. Pozen, “Final Report of the Advisory Committee on Improvements to Financial Reporting to the United States Securities and Exchange Commission,” 2008.

[30] S. V. Brown and J. W. Tucker, “Large-Sample Evidence on Firms’ Year-over-Year MD&A Modifications,” *J. Account. Res.*, vol. 49, no. 2, pp. 309–346, May 2011.

[31] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, “Management’s tone change, post earnings announcement drift and accruals,” *Rev. Account. Stud.*, vol. 15, pp. 915–953, 2010.

[32] A. K. Davis and I. Tama-Sweet, “Managers’ Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A*,” *Contemp. Account. Res.*, vol. 29, no. 3, pp. 804–837, Sep. 2012.

[33] F. Li, R. Lundholm, and M. Minnis, “A Measure of Competition Based on 10-K Filings,” *J. Account. Res.*, vol. 51, no. 2, p. no–no, Jan. 2013.

[34] K.-W. Huang and Z. Li, “A multilabel text classification algorithm for labeling risk factors in SEC form 10-K,” *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 3, pp. 1–19, Oct. 2011.

[35] K. W. Hanley and G. Hoberg, “The information content of IPO prospectuses,” *Rev. Financ. Stud.*, vol. 23, pp. 2821–2864, 2010.

[36] “SEC.gov | EDGAR Company Filings | CIK Lookup.” [Online]. Available: http://www.sec.gov/edgar/searchedgar/cik.htm#.VFbFL5DF-Cs. [Accessed: 02-Nov-2014].

[37] L. Richardson, “Beautiful soup,” *Crummy Site*, 2013.

[38] E. Mayfield and C. P. Rosé, “LightSIDE: Open source machine learning for text accessible to non-experts,” *Invit. chapter Handb. Autom. Essay Grading*, 2012.