Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach

Article *in* Journal of Emerging Technologies in Accounting · October 2018

DOI: 10.2308/jeta-52312

Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach

Chi-Chun Chou*
Associate Professor of Accounting
College of Business
California State University, Monterey Bay
100 Campus Center
Seaside, CA 93955-8001
Tel: 831-582-5213
Fax: 831-582-4251
cchou@csumb.edu

C. Janie Chang
Vern Odmark Professor of Accountancy
The Charles W. Lamden School of Accountancy
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-8221

Chen-Lung Chin
Professor
Department of Accounting
National Chengchi University
Taipei, TAIWAN

Wei-Ta Chiang
Senior Auditor, KPMG
Taipei, TAIWAN

*Contact author

Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach

ABSTRACT

This study uses a design science approach to examine the consistency between quantitative financial ratios and qualitative narrative disclosures in the annual reports. To extract information on the tone of unstructured qualitative textual data, we first use the TFIDF (term frequency/inverse document frequency) text mining technique to classify each company's narrative disclosure as either "Positive" or "Negative." For the quantitative information, we use the K-means method to cluster each company's financial performance data into "Good" or "Poor" groups. Consistency is said to occur when the textual and numerical data form either a "Positive-Good" pair or a "Negative-Poor" pair. The design model is presented in a stepwise fashion and therefore is transparent for evaluation and validation. Our evaluation process demonstrates the feasibility of the design model. The evaluation was conducted using listed semiconductor companies in countries with different levels of market development. The results show that US firms are less likely to exaggerate in their narrative disclosures and are more likely to understate their performance in MD&As compared to companies in other markets such as China and Taiwan.

Keywords Text analytics/mining, K-means, MD&A, disclosure discipline

Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach

INTRODUCTION

Companies use their annual reports, which contain both quantitative and qualitative information about financial performance, to communicate messages and to market themselves to their stakeholders (Herreman and Ryans 1995). Quantitative information in these reports takes the form of structured data such as financial ratios, prices, and accounting items, which are easy to process and are considered hard information. Qualitative information takes the form of unstructured data such as textual descriptions and disclosures and is known as soft information (Petersen 2004). Qualitative information may be either required or voluntarily disclosed in annual reports, where it is used to explain, support, and expand on the numerical or financial data.

Prior studies have generally concluded that both quantitative and qualitative data are value relevant in financial reporting (Lev and Thiagarajan 1993; Amir and Lev 1996; Bryan 1997; Back, Toivonen, Vanharanta, and Visa 2001; Cole and Jones 2004; Petersen 2004; Engelberg 2008; Sun 2010; Van der Laan Smith, Adhikari, Tondkar, and Andrews 2010). Some studies have also confirmed that textual disclosures, especially the management discussion and analysis (MD&A), are crucial for informed investment decisions (Back et al. 2001) such as the prediction of future firm performance (Sun 2010; Li 2010).

Despite the common perception that both quantitative and qualitative information are useful, comparisons of the two types of information in US annual reports indicate that the relative usefulness of the two types of data might be affected by companies' tendency to exaggerate performance in the text (Back et al. 2001; Davis and Tama-Sweet 2012). The general preference for positive versus negative words is first addressed by Hildebrandt and Snyder's

(1981) Pollyanna Hypothesis[1] which uses the annual letter to the stockholders as the source of text. The preference is also supported by Kohut and Segars's (1992) study of the President's Letters, David's (2001) study of the annual report, Pava and Epstein's (1993) examination on the prospective information contained in MD&As and Huang, Teoh, and Zhang's (2014) study of tone management in earnings press releases.

Prior studies, as introduced above, have generated mixed findings on the consistency between textual and numerical information in financial reports. The mixed results highlight a measurement issue that a better indicator of inconsistency might be useful. Accordingly, this study develops a scientific method for measuring consistency that minimizes the influence of human biases. In specific, we develop a new model to measure consistency that combines text mining (text analytics) and data mining techniques. Using a design science methodology, we first use the text mining technology to extract information about the tone (or sentiment) of the unstructured qualitative data (such as business operations or MD&A sections) in an annual report. Then, we use a data mining technique to assess and classify the quantitative financial data (such as financial numbers or ratios) in the same annual report. The consistency index developed in our model is based on a simple matching of the results from the tone of qualitative information (Positive/Negative) and the financial performance highlighted by quantitative information (Good/Poor). A balanced message (such as "Positive-Good" pair or "Negative-Poor" pair) reflects a coherent reporting behavior, and the report's consistency is classified as "Fair." On the other hand, an unbalanced message (such as "Positive-Poor" pair or "Negative-Good" pair) indicates the existence of inconsistency.

[1] Pollyanna is a Walt Disney Productions feature film based upon the novel of the same name by Eleanor Porter. The story is about a cheerful, radically optimistic youngster who always focuses on the goodness of life, no matter how much difficulty she is in.

To validate these concepts and the methodology developed in our model, we conduct an evaluation using a cross-country sample. Using samples of semiconductor manufacturers located in the US, China, and Taiwan, we demonstrate the technical feasibility of our model in a stepwise manner. As studies using text analytics techniques to compare the quantitative and qualitative information released by Chinese companies are rather rare, this study provides a foundation for future research in this area.

The rest of the paper is organized as follows. The second section summarizes the role of text analytics in the study of financial reporting and develops our research question. The third section presents the methodology and research design in a stepwise fashion. The fourth section reports the findings of our cross-country comparison of the consistency of qualitative and quantitative financial information. The last section concludes the study and discusses future research directions.

PRIOR RESEARCH AND RESEARCH QUESTIONS

Using Text Analytics to Analyze Narrative Disclosure

Due to the high costs of processing large amounts of unstructured textual data, there are relatively few studies of the economic effect or usefulness of companies' narrative disclosures (Plumlee 2003; Engelberg 2008; Sun 2010). Abrahamson and Amir (1996) first defined MD&A information as "softer" information relative to reliable, audited financial information. Petersen (2004) further differentiates information into two forms: hard (numerical) and soft (textual). Hard information is quantitative, often communicated with numbers. It includes data such as accounting elements or financial ratios that are not only more objective and easily comparable, but also easier to store and process. In contrast, soft information, such as the text of an annual report describing the company's financial performance and the results of operations for the year, is often communicated in the form of text. It is costly to store, more subjective, and difficult to pass along without loss of information. For this reason, text-based, qualitative information (i.e., soft information), despite containing valuable and relevant information, is sometimes ignored by analysts. According to Engelberg (2008), complicated and superabundant textual information is likely to cause ineffective information processing, which may contribute to market anomalies such as post-earnings announcement drift.

To address the problem of processing costs, several accounting studies have conducted computerized content analysis using text mining/analytics technology. For example, Frazier et al. (1984) use WORDS—a computer application for content analysis—to analyze the association between the content of MD&A reports and market performance. Tennyson et al. (1990) use content analysis to examine the incremental usefulness of the President's Letter contained in annual reports for explaining bankruptcy, after controlling for relevant financial ratios. Previts, Bricker, Robinson, and Young (1994) analyze documents issued by financial analysts to determine what type of information is used in analysts' investment recommendations. They show that analysts use both the financial and nonfinancial information available in the MD&A and President's Letter. Abrahamson and Amir (1996) use the relative number of negative words to measure the amount of negativity expressed in President's Letters. They find the measure is negatively correlated with accounting-based performance measures in the six annual periods around the release of the President's Letter. All of these studies rely on word-count systems that use computerized software to extract word counts from electronic documents and then draw inferences from word frequencies (Beyer et al. 2010; Fisher et al. 2010; Berger 2011; Ernstberger and Grüning 2013).

Expanding the simple word count techniques, Brown and Tucker (2011) adopt term frequency/inverse document frequency (TFIDF) and vector space model (VSM) methods, which are used by Internet search engines to determine similarities between documents. They use a new similarity algorithm to measure year-to-year changes in MD&A disclosures and assign these changes a modification score. They conclude that the magnitude of stock price responses to 10-K filings is positively associated with the MD&A modification score, but analyst earnings forecast revisions are unassociated with the score. They also find that over their study period, the MD&A modification scores decline and the MD&A disclosures become longer. In addition, the price reaction to MD&A modification scores weakens, suggesting a possible decline in MD&A usefulness over the past decade.

Alternately, Li (2010) uses a Naive Bayesian machine-learning algorithm to examine the information content of forward-looking statements (FLS) in the MD&A section of 10-K and 10-Q filings. He finds that the average tone of the FLS is positively associated with future earnings. To justify the validity of this machine-learning algorithm, he tests a dictionary approach (such as Diction, General Inquirer, or Linguistic Inquiry and Word Count) to measure the tone in MD&A, but fails to identify a link between information content and tone using these methods. Therefore, he argues that the dictionary technique does not work well for corporate financial statements.

Echoing Li (2010) and Brown and Tucker (2011), Ernstberger and Grüning (2013) comments that a "simple word-count" measurement approach "restrict[s] the software application to merely assisting the manual coding process." Hence, they suggest using more sophisticated machine learning or natural language processing methods such as the VSM and TFIDF methods to calculate similarity and to extract the information content of MD&As. Building on the insights of Ernstberger and Grüning (2013), Li (2010), and Brown and Tucker

(2011), this study adopts advanced text analysis techniques, specifically TFIDF and VSM, to detect the tone embedded in texts.

Financial Performance Measurement

Instead of using a single financial number (e.g., earnings) or a single financial ratio, Back et al. (2001), Kloptchenko et al. (2004), and Qiu, Srinivasan, and Street (2006) use vector-based measurements of financial performance. Back et al. (2001) categorize 47 ratios into six groups and Kloptchenko et al. (2004) show that seven financial ratios[2] in a self-organizing map (SOM) of selected quantitative financial ratios can be used to indicate future financial performance. Qiu et al. (2006) use the return on equity (ROE) ratio as the single financial performance measure for each firm/year. They linearly partition ROE into three classes, neutral, positive, and negative,[3] without using the clustering method.

Consistency of Narrative Disclosures and Financial Performance Indicators

Hildebrandt and Snyder (1981) propose and test the Pollyanna Hypothesis by applying a positive-negative word-count method to a small sample (24 President's Letters). They find that regardless of a corporation's financial position, positive words occur more frequently than negative words in annual letters to stockholders, suggesting that "the annual letter is used to underplay the negative news, replacing it with positive conclusions." This finding implies the existence of inconsistency between a firm's narrative disclosure and its financial performance. In

[2] The seven ratios are operating margin, return on total assets, return on equity for profitability, current ratio for liquidity, equity to capital, interest coverage for solvency, and receivables turnover for efficiency.

[3] If the ROE ratio in year $t+1$ is within 5 percent of year t, then the ROE of year $t+1$ is classified as "Neutral." If the ROE ratio in year $t+1$ is greater than year t by more than 5 percent, then the ROE of year $t+1$ is classified as "Positive." If the ROE ratio in year $t+1$ is less than year t by more than 5 percent, then the ROE of year $t+1$ is classified as "Negative."

contrast, Li (2010) provides evidence that MD&A tone predicts future performance, suggesting that managers truthfully disclose information in MD&As.

Regulators, analysts, and investors are interested in the quality of MD&A disclosures because management has relatively higher flexibility in preparing the MD&A information (Schroeder and Gibson 1990; Clarkson et al. 1994; Abrahamson and Amir 1996; Roulstone 2011). For example, the SEC's report in 2006 on the state of MD&A disclosures (Glassman 2006) states that their review of 6,000 public company filings from the previous year showed that companies were making progress in explaining their financial statements in the MD&A. However, the SEC also believed there was room for improvement.

Applying the word-count method to a sample of 14 listed organizations in the Portuguese Stock Market Index for the 2005 to 2007 period, Dias and Matias-Fonseca (2010) use a correlation matrix to measure the correlation between the language contained in annual reports and financial results. Their comparison of quantitative data (financial ratios) and qualitative data (linguistic terms) proves that organizations are not indifferent to the communication of information in annual reports, and that performance affects the way organizations communicate their results. Their results also demonstrate that the textual content of annual reports not only reflects companies' performance in the current year, it also contains information about the events that led up to the performance, which could predict future changes.

Given the mixed findings on the consistency of narrative disclosure and financial performance, academics, regulators, analysts, and investors could benefit from a scientific, objective method for measuring the consistency of soft and hard information.

MEASURING PAIRWISE CONSISTENCY

Building on prior studies, we combine text mining and data mining techniques to develop a new approach to measure consistency. Specifically, when the pairwise qualitative and quantitative information describing a company's financial performance for the year forms either a "Positive-Good" pair or a "Negative-Poor" pair, the report's consistency is classified as "Fair." However, if the quantitative information indicates a "Poor" financial performance while the qualitative information uses a "Positive" tone to discuss the company's performance, then the consistency is classified as "Exaggerated." In contrast, if the quantitative information indicates a "Good" financial performance while the qualitative information discusses the performance in a "Negative" tone, then the consistency is classified as "Pessimistic" (see Table 1). [4] The measurement is consistent with prior studies that have suggested that companies may manipulate data presentation to influence the way information users make decisions and judgments (Hildebrandt and Snyder 1981; Kohut and Segars 1992; Pava and Epstein 1993; Back et al. 2001; David 2001).

[Please insert Table 1 about here]

Figure 1 depicts the three stages of this study's methodology: training, measurement, and matching. In the training stage, we perform the VSM/TFIDF text analytics using 70 percent of our MD&A sample documents to create two feature-based representational vectors: "Positive" and "Negative." Our comprehensive measurement design integrates the traditional word-count

[4] It is possible to develop matching matrices with higher degrees than our 2 x 2 matching matrix design (such as 3 x 3, 4 x 4, or 5 x 5), as both the clustering and the tone analysis allow us to divide samples into more groups. However, the ability to identify inconsistencies could be limited by too many possible combinations. For example, for a 3 x 3 matrix, there would be nine combinations, for a 4 x 4 matrix there would be 16, etc. In this study, we propose an extensible measurement approach instead of providing a deterministic design model. Therefore, we believe the instantiation of a 2 x 2 matrix, despite its simplicity, is an appropriate form for examining the issue of inconsistency.

method with advanced text mining methods such as VSM and TFIDF to measure the tone of the documents and to calculate the similarity in tone of MD&A disclosures. The training process not only addresses technical concerns, it is designed to fit our settings and to ensure construct validity. In the measurement stage, each document in the remaining 30 percent of the sample is "vectorized" using the same VSM/TFIDF method and compared to the two representational vectors to determine if it has a "Positive" or "Negative" tone.

[Please insert Figure 1 about here]

To measure general performance, we combine three of the most important financial ratios, return on equity, receivables turnover, and current ratio, which measure profitability, efficiency, and liquidity, respectively. The major reason for choosing three representative financial ratios rather than a single number (such as earnings) is to capture the broader definition of financial performance. As the aim is to measure overall performance, the three ratios are denoted as a three-dimensional vector. Mathematically, the three-dimensional proxy for financial performance used in our model cannot be presented as a single continuous variable. Therefore, to measure relative financial performance, we use the K-means clustering algorithm to cluster the three chosen ratios by their vector values. The K-means algorithm is used to divide a data sample into a collection of distinct groups based on their similarities. It is a common technique in data/text mining and is based on the idea that in a multi-dimensional space, objects are more similar to nearby objects than to objects farther away (Gupta and Lehal 2009). Thiprungsri and Vasarhelyi (2011) find K-means clustering to be a promising technique in the field of auditing, as it could help auditors flag transactions with suspicious characteristics that had not been identified by other methodologies. One advantage of the K-means approach is that when there are a large number of variables, K-means may be computationally faster than hierarchical clustering and may also produce tighter clusters (Gupta and Lehal 2009).

Specifically, we use K-means to evaluate and classify each firm's comprehensive financial performance (profitability, efficiency, and liquidity) as either "Good" or "Poor." In addition, since the MD&A's and the status of operations sections in annual reports disclose a company's financial conditions, changes in financial conditions, and the results of operations for the year, we calculate the changes in the three financial ratios instead of the ratio values.

In the matching stage, a pairwise comparison is conducted based on the ranking results of the first two stages to determine if the tone of the qualitative data is consistent with the quantitative performance. The procedures used in these three stages are described in detail as follows:

Training Stage

Miner et al. (2012) note that to use text mining technology to classify or categorize text, it is necessary that the "classifiers should be *trained* for a specific task." Training is required to set the numerical parameters (weights and threshold) of the classification model. In a related study, Li (2010) manually categorizes 30,000 sentences of randomly selected forward-looking statements extracted from MD&As and uses this dataset to train his Naive learning algorithm to categorize the tone (i.e., positive versus negative) and content (e.g., profitability operations, liquidity, etc.) of other FLS in 10-Q and 10-K filings.

This study proposes a more comprehensive strategy that incorporates both advanced text mining methods and training techniques. The text analysis method we develop is a combination of sentiment dictionary and the training procedure adopted by the VSM/TFIDF approach. In the training stage, we first use K-means method to determine the tone of 70 percent of our MD&A sample documents based on the percentage of positive/negative words to avoid the subjectivity bias that may occur from human involvement. Once they have been categorized by positive/negative word counts, all of the texts in the training sample are separately converted into two feature-based representational vectors: "Positive" and "Negative." The two large featured vectors can be considered the knowledge base in the form of two virtual reference documents: "Positive" and "Negative." The two virtual documents are then used to classify the remaining 30 percent of the sample documents into Positive and Negative groups using a VSM/TFIDF similarity calculation. As Brown and Tucker (2011) indicate, VSM and TFIDF are commonly used in search engines to improve their search quality and therefore the external validity of VSM and TFIDF can be assured. Appendix A provides a brief introduction to VSM and TFIDF technology.

We describe the four steps used in our study to construct the knowledge base below.

Step 1: Preprocessing

Before processing any qualitative information, it is necessary to apply a word segmentation program to the dataset, as previous studies have shown that phrase-based segmentation does not produce good results (Scott and Matwin 1999). For English documents, word-based segmentation is based on the spaces between each word. For traditional Chinese documents, the Chinese Knowledge and Information Processing (CKIP) system is the most powerful tool (Ma and Chen 2003). For the Simplified Chinese documents, we use the Stanford Chinese Word Segmenter (Tseng, Chang, Andrew, Jurafsky, and Manning 2005). The differences in the word segmentation process for English and Chinese documents are discussed in Appendix

B. We also consider stop words[5] to remove unnecessary characters from the text. For English words, stemming is performed to normalize words by removing prefixes and suffixes. Finally, each document is saved to an individual text file and imported into Excel for later analysis.

Step 2: Split the Data into Training And Testing Samples

In Step 2, the dataset is divided randomly into two partitions, 70 percent for training and 30 percent for testing. As discussed earlier, this ratio is often used in data mining and has been used in many text analytics studies.[6] The training process could be simplified by using the positive-negative percentage to measure the tones of MD&As. However, as "Positive" and "Negative" are relative concepts, we design an extra training process to create two more class vectors. For example, in a market prone to optimism, percentages slightly greater than one might have been considered "Neutral" or even "Pessimistic," and vice versa. Therefore, directly using percentages of word-counts might obscure the fact that tone is a relative concept rather than an absolute metric. Another advantage of adopting the training process is that researchers can learn patterns of "Positive"/"Negative" words and use this knowledge to generate representative positive and negative terms. Using patterns and terms rather than a word-count frequency based on a fixed dictionary to determine the tone of a test document ensures that we capture the differences in the characteristics of the research samples from different countries/markets. In addition, once patterns have been identified, we can use a variety of pattern recognition methods

---

5 "Stop words" are the common words in a language, such as *the*, *is*, *at*, *which*, and *on* in English, that are not significant in natural language processing. Usually, these words are filtered out from text analytics because they add unnecessary information to the dataset.

6 Qiu et al. (2006) randomly split their non-financial data into 2/3 for training and 1/3 for testing. The training sample is then used to select feature vectors based on the TFIDF weights, and the testing sample is used for financial performance prediction. However, most of data mining software uses 70/30 as default values during partitioning and Sporleder (2007), Bhattacharyya et al. (2008), Vanneschi et al. (2011), and Ville (2006) all split their training and testing data according to the 70/30 rule in their text analytics studies. "The choice of 70/30 split for training and testing is somewhat arbitrary, but is empirically a good practical ratio according to more detailed experiments" (Deng and Moore 1998).

to classify the documents. In this study, the VSM/TFIDF is chosen as the pattern recognition method.

Step 3: Create a TFIDF Table and Term Space

In the training stage, after each document has been imported into Excel, a term frequency (TF) is determined for every document. Inverse document frequency (IDF) is calculated based on all of the training data, and the product of TF and IDF is calculated, as described in Appendix A.

Step 4: Create "Positive" and "Negative" Class Vectors

For each training document, tone is captured by examining the total appearance of positive-to-negative words in each document. The word list is taken from a dictionary list of positive and negative words provided by Loughran and McDonald (2011), who caution that a majority of the words considered negative in general contexts or in a general dictionary do not actually represent negative information in a business context. Accordingly, K-means is used to cluster the reports in the training set into two groups, "Positive" and "Negative," according to the ratio of positive words to negative words. The positive and negative documents clustered by K-means through sentiment analysis is then used to create the class vectors "Positive" and "Negative."

Measurement Stage

As depicted in Figure 1, text analytics and clustering methods are adopted in this stage to determine the consistency between the tones retrieved from the qualitative data and the financial performance reported by the quantitative data. There are two steps in this stage: measuring the tone of the soft information and selecting the clusters to measure hard information.

Step 1: Measure the Tone of Soft Information

First, we conduct a tone analysis of the qualitative data (e.g., statements from MD&A). Each document in the testing sample is converted into its own TFIDF document vector to determine its similarity to the two class vectors. Individual documents are then classified as "Positive" if it is more similar to the positive-toned virtual document or "Negative" if it is more similar to the negative-toned virtual document.

A similarity measure is a function that computes the degree of similarity between vectors. A common method for computing the similarity of two documents is to use the cosine similarity measure. The inner product of the two vectors is divided by the product of their vector lengths, which normalizes the vectors to unit length and only the angle, more precisely the cosine of the angle, between the vectors is used to determine their similarity.

Given a specific class vector cv_k, and a specific document dv_i, the cosine similarity, *COS(θ)*, is represented using a dot product and magnitude as

$$sim(cv_k, dv_i) = COS(\theta) = \frac{V(cv_k) \cdot V(dv_i)}{|V(cv_k)||V(dv_i)|}.$$

As all of the vector attributes are positive, the resulting similarity ranges from 0 to 1, with 1 indicating that the two vectors are exactly identical, 0 indicating independence, and in-between values indicating intermediate similarity, i.e., the degree of similarity. Therefore, we calculate two *COS(θ)* values for each document: one represents the similarity between the document and the positive class vector, and the other represents the similarity between the document and the negative class vector. Then we choose the higher *COS(θ)* value to classify the document as having either a generally positive or generally negative tone.

Step 2: Select Clusters to Measure Hard Information

Second, we measure each company's financial performance based on the financial ratios that are computed based on the numbers in the financial statements. For company *i* between year *t* and year *t-1,* we calculate the changes for its return on equity (ΔROE_{it}), receivables turnover ($\Delta RECTO_{it}$), and current ratio ($\Delta CURRENTRATIO_{it}$), respectively. Then we conduct a cluster analysis using K-means to cluster the three chosen ratios into by their vector values. The general processes are as follows (Roiger and Geatz 2003).

1) Choose a value for K that is the total number of clusters to be determined.
2) Choose K instances (data points) within the dataset at random. These are the initial cluster centers.
3) Use Euclidean distance to assign to the remaining instances their closest cluster center.
4) Use the instances in each cluster to calculate a new mean for each cluster.
5) If the new mean values are identical to the mean values of the previous iteration, the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

As proposed in Table 1, this study specifies two clusters, which represent "Good" and "Poor" financial performance for the year of study. Extreme values are adjusted to prevent unnecessary negative biases in the clustering during later evaluation.

[Please insert Table 2 about here]

Matching Stage

The results from the TFIDF text analysis of the qualitative textual data are matched with the K-means clusters of the quantitative financial ratios to determine the consistency of a company's annual report disclosures. If the paired qualitative and quantitative information on a company's financial performance for the year are represented by either a "Positive-Good" pair or "Negative-Poor," implying that the company's reporting behavior is coherent, the report is classified as "Fair." The report is classified as "Exaggerated" if the result of the TFIDF analysis classifies the qualitative data as "Positive," but the quantitative data indicates a "Poor" performance. Conversely, the report is classified as "Pessimistic" if the opposite pattern occurs.

EVALUATION OF THE MODEL

In this section, we evaluate our model using the annual reports of selected US, Taiwanese, and Chinese publically listed companies from the semiconductor industry.[7]

Although the US is the biggest semiconductor producer in the world, this industry is also one of the most strategically important high-tech industries in Taiwan and China. Taiwan is the world's fourth largest semiconductor producer, trailing the US, Japan, and South Korea. In addition, the patents granted to the Taiwanese semiconductor industry account for 42 percent of the total patents in Taiwan, and 56 percent of all Taiwanese patents granted by the United States Patent and Trademark Office (Chin, Lee, Chi, and Anandarajan 2006).[8] China also considers the development of the semiconductor industry a top strategic priority (Simon and Rehn 1988; Zhang, Yu, and Liu 2013). The prominence and importance of the semiconductor industry in these three countries makes it a suitable environment for exploring the consistency between qualitative and quantitative information in financial reports.

Data Selection and Collection

We analyze the tone of the narrative information provided by listed companies in the US, Taiwan, and China in the semiconductor industry for the 2002 to 2010 period. We also calculate changes in each company's financial ratios (return on equity, receivables turnover, and current

[7] We do not use the whole dataset for the three markets because our aim is to demonstrate the methods used to measure inconsistency. At the evaluation stage of research, it is not necessary to use a large sample in the analysis.
[8] For a detailed overview of the Taiwanese semiconductor industry, please see Tung (2001) and Hung and Yang (2003).

ratio). YCharts (http://www.ycharts.com) is used to collect the historical financial data of companies in the US, the TEJ Smart Wizard offered by the Taiwan Economic Journal (TEJ) is used to collect historical financial data for companies in Taiwan, and SINA Finance is used (http://money.finance.sina.com.cn) to collect historical financial data for companies in China.

We collect annual reports from the following databases: (1) for the United States, MD&As in the 10-K filings in the SEC EDGAR database, (2) for Taiwan, the status of operations sections in annual reports from the Market Observation Post System (MOPS) offered by TWSE, and (3) for China, the Director's Report in the annual reports from SINA Finance. The qualitative data are extracted from the MD&As of the US 10-K Filings for US listed companies, from the Status of Operations of annual reports for listed companies in Taiwan, and from the Director's Report of annual reports for listed companies in China. For China, the dataset includes 225 reports from 30 companies (not all of the companies have complete data for the study period, as some were listed after 2002), 157 of which are used for training and 68 for testing. For the US, the dataset includes 290 reports from 33 companies, 203 of which are used for training and 87 for testing. For Taiwan, the dataset includes 311 reports from 39 companies, 217 of which are used for training and 94 for testing. Table 3 summarizes the general information about our samples.

[Please insert Table 3 about here]

Qualitative Data – Training and Measurement

As discussed above, the first step in the training stage is the preprocessing of the data. The textual data is extracted from the annual reports in three different languages: English, Traditional Chinese, and Simplified Chinese. First, word segmentation procedures are implemented for all three languages using the techniques discussed in Appendix B. Second, UltraEdit is used to remove unnecessary characters, separate words by lines, and import the data into Excel, where the TFIDF calculations are performed.

The data are randomly partitioned into training and testing sets according to the 70/30 split rule, and the TFIDF table and term space are created. After each document is imported into Excel, the term frequency is calculated for each document. The training documents are then combined to calculate the inverse document frequency.

Then, we use sentiment analysis to classify each training document as having either a “Positive” or “Negative” tone. First, we count the positive and negative words defined by Loughran and McDonald (2011) in each training document. In this study, the positive and negative word list for Traditional and Simplified Chinese are translated directly from Loughran and McDonald’s list[9]. We further calculate the percentage of positive and negative words in each document. Then, the K-means algorithm is used to cluster the documents into two clusters, positive-toned documents and negative-toned documents, based on the percentage of positive and negative words. The two class-featured vectors are thus created based on the clustered results. The term frequencies of all of the positive-toned and negative-toned documents are recounted separately. Finally, the class featured vectors and the TFIDF calculation table are created. In Figure 2, Columns A to E represent the words contained in the positive-toned class vectors, including IDF, TF, TFIDF, and the power of TFIDF. Columns F to J represent one sample document’s vector values. Column K represents the inner product of the testing sample and the positive-toned class vectors.

---

[9] We use two very well accepted English-Chinese dictionaries for the translation, the ACME Excellence English-Chinese Dictionary for Traditional Chinese and the Bilingual English-Chinese Dictionary for Simplified Chinese. Both dictionaries usually provide several synonyms in Chinese for each English vocabulary. When the situation occurs, we search selected synonyms in the Chinese annual reports to resolve the potential semantic-pragmatic difference due to translations.

[Please insert Figure 2 about here]

In the measurement stage, an individual testing document is classified as either “Positive” or “Negative” based on its vector-based similarity to the “Positive” or “Negative” tone class vectors using the VSM calculation, as illustrated in Figure 3.

[Please insert Figure 3 about here]

Quantitative Data - Measurement

Table 4 presents the numbers and percentages of documents with good and poor financials by country. For the US subsample, cluster 1 represents positive changes in the financial ratios compared to the previous year and is labeled “Good” (i.e., cluster 2 is labeled “Poor”). For the Taiwan subsample, cluster 1 represents negative changes in financial ratios compared to the previous year and is labeled “Poor” (i.e., cluster 2 is labeled “Good”). For the China subsample, cluster 1 represents positive changes in financial ratios compared to the previous year and is labeled “Good” (i.e., cluster 2 is labeled “Poor”). Then, the results of the TFIDF analysis are matched with the results of the K-means clustering (see Table 5).

[Please insert Tables 4 and 5 about here]

Quantitative and Qualitative Data - Matching

Table 6 provides the summary (percentages) of the cases based on the raw data from Table 5. Based on the narrative disclosures in the annual reports of listed companies in the semiconductor industry, companies in China are more likely to exaggerate or overstate their financial conditions than companies in the US or Taiwan. Roughly half of the narrative disclosures by US and Taiwanese companies are “Fair.” Chinese companies offer “Fair” statements on financial reports around 76.5 percent of the time. Interestingly, Chinese companies never produce “Pessimistic” reports. One-way, between-group ANOVA tests show significant differences between the three sample groups in both the overall data and separately for each of the three ratios.

[Please insert Table 6 about here]

Further Adjustments

We also test the correlations between the three quantitative financial ratios, return on equity, receivables turnover, and current ratio, in the three country subsamples. The three ratios are positively correlated with each other in the US and Taiwanese subsamples, but not in the Chinese subsample. As a result, we separately cluster each of the three financial ratios into “Good” and “Bad” categories and compare them to the results of the TFIDF text analytics using $\chi 2$ analysis. The results, shown in Table 7, suggest that US and Taiwanese companies have similar degrees of consistency in their reports, although companies in Taiwan tend to exaggerate more often than US companies ($\chi 2=0.156$, $p<0.014$). Companies in China tend to provide fair statements in their financial reports, but do not use a negative tone in narrative disclosures in either good or bad financial situations.

[Please insert Table 7 about here]

As another robustness test, we use the 50 training documents with the highest percentage of positive words and the 50 training documents with the highest percentage of negative words to create a class-featured vectors sample. There are two reasons for this adjustment. First, using training documents with a higher percentage of positive/negative words could improve the accuracy of the similarity and classification procedures by eliminating documents with potential neutral tones, that is documents that have a close or equal number of positive and negative claims (Tang, Fang, and Wang 2014). Second, the clustering of training documents based on the percentage of positive words is not evenly clustered. In some cases, the number of training documents classified as positive-toned significantly exceeds the number of negative-toned documents. Creating samples that have an equal number of positive and negative documents allows us to recalculate the inconsistency. The results for each country are similar to those produced by the full sample.

Overall, the results show that listed semiconductor companies in China have the greatest tendency to exaggerate in their narrative disclosures about their financial performance, as given in the Director's Report of annual reports. Listed semiconductor companies in the United States have the lowest tendency to exaggerate and are more likely to be pessimistic about their performance in the MD&As of the annual reports.

CONCLUSION

This study, using a design science approach, proposes a replicable model to extract useful information from unstructured qualitative textual data (such as MD&As) with the use of the TFIDF text mining technique. Our model is designed to help financial statement users measure and test the consistency of reports of financial performance presented in the form of quantitative financial ratios (i.e., hard information) and qualitative narrative disclosures (i.e., soft information) in annual reports. To evaluate the feasibility of the model, we examine cross-country variation between countries with different levels of market development. Our analysis of a sample of listed semiconductor companies' annual reports from the 2002 to 2010 period shows that companies in China are more likely to exaggerate and overstate their performance in their narrative disclosures. Their narrative disclosures are predominantly positive regardless of their financial state, as predicted by the Pollyanna Hypothesis (Hildebrandt and Snyder 1981). None of the Chinese companies make "conservative" disclosures when their financial performance is good. US companies are less likely to offer “exaggerated” (4.6%) narratives than their counterparts in China (23.53%) and Taiwan (20.21%). In addition, US companies are more “pessimistic” in reporting (45.98%) than their counterparts in China (0%) and Taiwan (37.23%). These results suggest that legal environments probably affect the tone of the narrative disclosures in the annual reports of companies in different jurisdictions. Future studies should look into possible factors that affect the consistency of soft and hard information in financial reports, such as national culture, legal environment, industry, etc.

This study has some limitations. First, the consistency score proposed in our model is not a continuous variable. The major reason is the selection of financial performance indicators. Instead of choosing one number, such as earning amount, we select three financial ratios. This choice limits the development of a linear metric for financial performance. The adopted clustering method creates a challenge for linearizing the calculation of financial performance.

Second, given the challenges in text analytics, we are not able to fully computerize the process. Although text analytics provide an automated method for extracting information from large volumes of unstructured textual data in a fast and cost-effective way, it is still difficult to depend on computers to accurately accomplish tasks normally performed by humans. For example, the extraction of the target sections from annual reports is conducted manually. In addition, it is difficult for a text analytics algorithm to know that the words “significant” and “increase” can explain either a significant increase in revenue and income (positive data) or in expenses and losses (negative data). Furthermore, humans can easily overcome obstacles that challenge computers such as slang, spelling variations, and contextual meaning. Hence, using a text analytics tool requires some loss of accuracy in exchange for efficiency.

In addition, we have focused our research on the textual information that compares the financial performance of the current year with that of the previous year. These sections of annual reports also contain information about future outlooks and risks. Future studies could examine the relationship between the narrative disclosures about future outlooks in $year_t$ and financial performance in $year_{t+1}$. The inclusion of more financial ratios, accounting items, and technical indicators in the quantitative features could also be considered in the future. Another study could examine the effect of company size on the tone of textual discussions of financial performance. Furthermore, future studies could develop an empirical model to test the explanatory power of a consistency score as a determinant of MD&A inconsistency. Possible explanatory variables might include firm performance, accruals, size, firm-level risk factors, corporate governance quality, major events, analysts following, and firm age.

REFERENCES

Abrahamson, E., and E. Amir. 1996. The information content of the president's letter to shareholders. *Journal of Business Finance and Accounting* 23 (8): 1157–1182.

Amir, E., and B. Lev. 1996. Value relevance of nonfinancial information: The wireless telecommunications industry. *Journal of Accounting and Economics* 22 (1-3): 3–30.

Back, B., J. Toivonen, H. Vanharanta, and A. Visa. 2001. Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems* 2 (4): 249–269.

Berger, P. G. 2011. Challenges and opportunities in disclosure research - A discussion of the financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics* 51 (1-2): 204–218.

Beyer, A., D. A. Cohen, T. Z. Lys, and B. R. Walther. 2010. The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics* 50 (2-3): 296–343.

Bhattacharyya, D., P. Das, D. Ganguly, K. Mitra, P. Das, S. K. Bandyopadhyay, and T. Kim. 2008. Unstructured document categorization: A study. *International Journal of Signal Processing, Image Processing and Pattern Recognition* 1 (1): 55–62.

Brown, S. V., and J. W. Tucker. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49 (2): 309–346.

Bryan, S. H. 1997. Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review* 72 (2): 285–301.

Chin, C. L., P. Lee, Y. H. Chi, and A. Anandarajan. 2006. Patent citation, R&D spillover, and Tobin's Q: Evidence from the Taiwan semiconductor industry. *Review of Quantitative Finance and Accounting* 26: 67–84.

Clarkson, P. M., J. Kao, and G. D. Richardson. 1994. The voluntary inclusion of forecasts in the MD&A section of annual reports. *Contemporary Accounting Research* 11 (1): 423–450.

Cole, C., and C. L. Jones. 2004. The usefulness of MD&A disclosures in the retail industry. *Journal of Accounting, Auditing, and Finance* 19: 361–388.

David, C. 2001. Mythmaking in annual reports. *Journal of Business and Technical Communication* 15 (2): 195–222.

Davis, A., and I. Tama-Sweet. 2012. Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research* 29: 804–837.

Deng, K., and A. W. Moore. 1998. *On the Greediness of Feature Selection Algorithms*. Paper presented at the *International Conference of Machine Learning (ICML) '98*. Madison, Wisconsin, USA. Available at: http://www.cs.cmu.edu/~kdeng/thesis/feature.pdf

Dias, W., and R. Matias-Fonseca. 2010. The language of annual reports as an indicator of the organizations' financial situation. *International Review of Business Research Papers* 6 (5): 206–215.

Engelberg, J. 2008. *Costly Information Processing: Evidence from Earnings Announcements*. Working paper, Rady School of Management, University of California San Diego, San Diego.

Ernstberger, J., and M. Grüning. 2013. How do firm- and country-level governance mechanisms affect firms' disclosure? *Journal of Accounting Public Policy* 32 (3): 50–67.

Fisher, I. E., M. R. Garnesey, S. Goel, and K. Tam. 2010. The role of text analytics and information retrieval in the accounting domain. *Journal of Emerging Technologies in Accounting* 7: 1–24.

Frazier, K. B., R. W. Ingram, and B. M. Tennyson. 1984. A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research* 22 (1): 318–331.

Glassman, C. A. 2006. MD&A Report Card. *Journal of Accountancy.* Blog post. Available at: http://www.journalofaccountancy.com/Issues/2006/Aug/MdAReportCard.htm

Gupta, V., and G. S. Lehal. 2009. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence* 1 (1): 60–76.

Herreman, I. M., and J. Ryans. 1995. The case for better measurement and reporting of marketing performance. *Business Horizons* 38 (5): 51–60.

Hildebrandt, H. W., and R. D. Snyder. 1981. The Pollyanna hypothesis in business writing: Initial results, suggestions for research. *Journal of Business Communication* 18 (1): 5–15.

Huang, X., S. H. Teoh, and Y. Zhang. 2014. Tone management. *Accounting Review* 89 (3): 1083–1113.

Hung, S. W., and C. Yang. 2003. The IC fables industry in Taiwan: Current Status and Future Challenges. *Technology in Society* 25: 385–402.

Kloptchenko, A., T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa. 2004. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management* 12 (1): 20–28.

Kohut, G. F., and A. H. Segars. 1992. The president's letter to stockholders: An examination of corporate communication strategy. *Journal of Business Communication* 29 (1): 7–21.

Lev, B., and S. R. Thiagarajan. 1993. Fundamental information analysis. *Journal of Accounting Research* 3 (2): 153–192.

Li, F. 2010. The information content of forward-looking statements incorporate filings - A naïve Bayesian learning approach. *Journal of Accounting Research* 48 (5): 1049–1102.

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

Ma, W. Y., and K. J. Chen. 2003. Introduction to CKIP Chinese word segmentation system for the First International Chinese Word Segmentation Bakeoff. Available at: http://www.aclweb.org/anthology/W03-1726

Miner, M., D. Delen, J. Elder, A. Fast, T. Hill, and B. Nisbet. 2012. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. 1st Edition. Waltham, MA: Academic Press.

Pava, M. L., and M. Epstein. 1993. MD&A as an investment tool: User beware!. *Journal of Accountancy* 75 (3): 51–53.

Petersen, M. A. 2004. *Information: Hard and Soft*. Working paper, Kellogg School of Management, Northwestern University.

Plumlee, M. 2003. The effect of information complexity on analysts' use of that information. *The Accounting Review* 78 (1): 275–296.

Previts, G., R. Bricker, T. Robinson, and S. Young. 1994. A content analysis of sell-side financial analyst company reports. *Accounting Horizons* 8 (2): 55–70.

Qiu, X. Y., P. Srinivasan, and N. Street. 2006. Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology* 3 (1): 1–15.

Roiger, R. J. and M. W. Geatz. 2003. *Data Mining: A Tutorial-based Primer.* International edition. Pearson.

Roulstone, D. 2011. Discussion of large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49 (2): 347–357.

Schroeder, N. and C. Gibson. 1990. Readabilities of management's discussion and analysis. *Accounting Horizons* (Dec): 78–87.

Scott, S., and S. Matwin. 1999. Feature engineering for text classification. *Proceedings of ICML-99, 16th. International Conference on Machine Learning,* Bled, Slovenia, 1999, 379–388.

Simon, D. F., and D. Rehn. 1988. *Technological Innovation in China: The Case of the Shanghai Semiconductor Industry*. Cambridge, MA: Ballinger Publishing Company.

Sporleder, C. 2007. *Text Mining for Cultural Heritage Data from Natural History Domain*. Working paper, Computational Linguistics Saarland University, Saarland University.

Sun, Y. 2010. Do MD&A disclosures help users interpret disproportionate inventory increases?. *The Accounting Review* 85 (4): 1411–1440.Tang, T., Fang, E., & Wang, F. (2014). Is neutral really neutral? The effects of neutral user-generated content on product sales. *Journal of Marketing 78* (4): 41-58

Tennyson, B. M., R. W. Ingram, and M. T. Dugan. 1990. Assessing the information content of narrative disclosure in explaining bankruptcy. *Journal of Business Finance and Accounting* 17 (3): 391–410.

Thiprungsri, S. and M. Vasarhelyi. 2011. Cluster analysis for anomaly detection in accounting data: An audit approach. *The International Journal of Digital Accounting Research* 11: 69–84.

Tseng, H., P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.* Jeju Island, Korea, 2005*,* Association for Computational Linguistics.

Tung, A. 2001. Taiwan's semiconductor industry: What the state did and did not. *Review of Development Economics* 5: 266–288.

Van der Laan Smith, J., A. Adhikari, R. H. Tondkar, and R. L. Andrews. 2010. The impact of corporate social disclosure on investment behavior: A cross-national study. *Journal of Accounting and Public Policy* 29: 177–192.

Vanneschi, L., A. Farinaccio, G. Mauri, M. Antoniotti, P. Provero, and M. Giacobini. 2011. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Mining* 4: 1–12.

Ville, B. D. 2006. Text mining with holographic decision tree ensembles. *Proceedings of the Thirty-first Annual SAS Users Group International Conference.* Barry de Ville, 2006, SAS Institute Inc., Cary, NC.

Zhang, Y., J. Yu, and Y. Liu. 2013. Evolutionary dynamics of high technology industry: Modeling of semiconductor sector in China. *Chinese Management Studies* 7 (2): 194–214.

TABLE 1 – Types of *Consistency*

Quantitative Data (Financial Performance)

Qualitative Data (Narrative Tone)

	Good	Poor
Positive	Fair	Exaggerated
Negative	Pessimistic	Fair

Table **TABLE 2 - Selected Financial Ratios and Definitions**

Financial Ratio	**Definition**
ΔROE_{it}	Change-in return on equity between *year t* and *year t-1* of company *i*. [*Net profit$_t$ / Net assets$_t$*] - [*Net profit$_{t-1}$ / Net assets$_{t-1}$*]
$\Delta RECTO_{it}$	Change-in receivables turnover between *year t* and *year t-1* of company *i*. [*Net salest$_t$ / Average accounts receivable$_t$*] - [*Net sales$_{t-1}$ / Average accounts receivable$_{t-1}$*]
$\Delta CURRENTRATIO_{it}$	Change-in current ratio between *year t* and *year t-1* of company *i*. [*Current assets$_t$ / Current liabilities$_t$*] - [Current *assets$_{t-1}$ / Current liabilities$_{t-1}$*]

TABLE 3 - Dataset

Country	No. of training reports	No. of testing reports	Total
United States	203	87	290
Taiwan	217	94	311
China	157	68	225

TABLE 4 – Number (Percentage) of Cases in Each Cluster by Sample Group

		US	China	Taiwan
Cluster	Good	144 (54.5%)	157 (59.5%)	130 (49.2%)
	Poor	120 (45.5%)	107 (40.5%)	134 (50.8%)
	Total	264 (100%)	264 (100%)	264 (100%)

TABLE 5 – Number of Cases of Soft and Hard Data by Sample Group

Panel A: US		Hard Data (Financial Ratios)	
		Good	Poor
Soft data (from text mining)	Positive	5	4
	Negative	40	38

Panel B: China		Hard Data (Financial Ratios)	
		Good	Poor
Soft data (from text mining)	Positive	52	16
	Negative	0	0

Panel C: Taiwan		Hard Data (Financial Ratios)	
		Good	Poor
Soft data (from text mining)	Positive	16	19
	Negative	35	24

TABLE 6 –
Consistency of Financial Reports

	Fair	Exaggerated	Pessimistic
United States	49.42%	4.60%	45.98%
China	76.47%	23.53%	0.00%
Taiwan	42.56%	20.21%	37.23%

TABLE 7 - χ2 Analysis by Scheffe

Panel A: Overall results

Group	US-China		China-Taiwan		US-Taiwan	
	χ2	P	χ2	P	χ2	P
Fair	0.270	**0.003**	0.339	**0.000**	-0.687	0.631
Exaggerated	0.189	**0.005**	0.332	0.843	0.156	**0.014**
Pessimistic	-0.460	**0.000**	-0.372	**0.000**	-0.087	0.492

Panel B: Return on Equity Ratio (ROE) comparison

Group	US-China		China-Taiwan		US-Taiwan	
	χ2	P	χ2	P	χ2	P
Fair	0.342	**0.000**	0.375	**0.000**	-0.033	0.900
Exaggerated	0.175	**0.013**	-0.013	0.973	0.188	**0.003**
Pessimistic	-0.517	**0.000**	-0.362	**0.000**	-0.156	0.047

Panel C: Receivables Turnover Ratio (RT) comparison

Group	US-China		China-Taiwan		US-Taiwan	
	χ2	P	χ2	P	χ2	P
Fair	0.100	0.462	0.160	0.132	-0.060	0.723
Exaggerated	0.325	**0.000**	0.202	**0.004**	0.123	0.089
Pessimistic	-0.425	**0.000**	-0.362	**0.000**	-0.064	0.593

Panel D: Current Ratio (CR) comparison

Group	US-China		China-Taiwan		US-Taiwan	
	χ2	P	χ2	P	χ2	P
Fair	0.017	0.978	0.213	0.965	-0.004	0.999
Exaggerated	0.466	**0.000**	0.330	**0.000**	0.136	**0.047**
Pessimistic	-0.483	**0.000**	-0.351	**0.000**	-0.132	0.107

FIGURE 1 – Measuring Pairwise Consistency

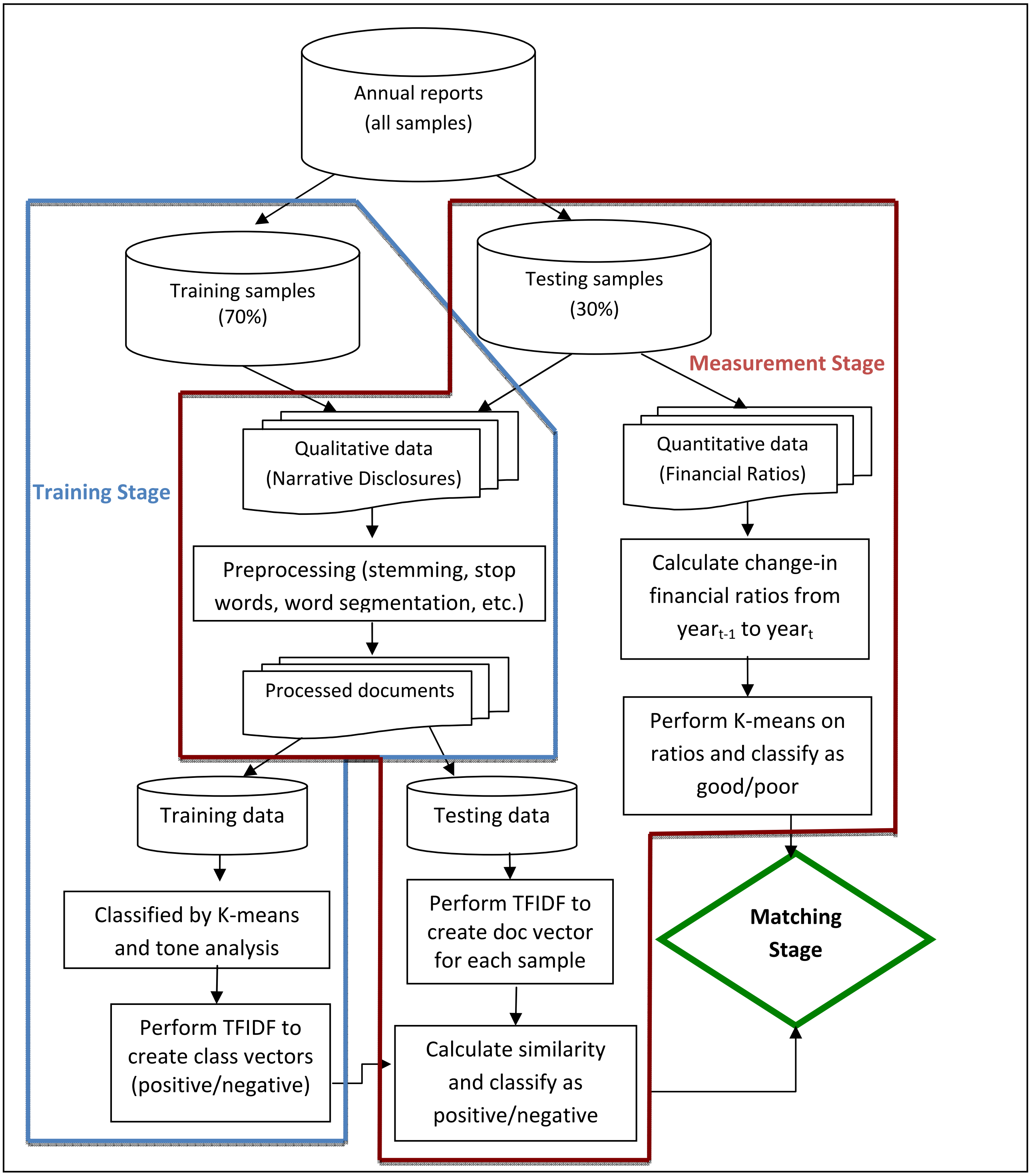

FIGURE 2 - Screenshot of a Positive Toned Class Vector and Document Vector

	A	B	C	D	E	F	G	H	I	J	K	L
1	TERM	IDF	TF	TFIDF	POWER	TERM	IDF	TF	TFIDF	POWER	PRODUCT	1170.415
2	of	0	27974	0	0	of	0	508	0	0	0	46.95495
3	the	0	29885	0	0	the	0	402	0	0	0	21573.33
4	and	0	22820	0	0	and	0	341	0	0	0	0.392551
5	in	0	20588	0	0	in	0	328	0	0	0	
6	to	0	17963	0	0	to	0	269	0	0	0	
7	million	0.011899	8949	106.4861	11339.3	million	0.011899	170	2.022868	4.091995	215.4074	
8	fiscal	0.041862	4469	187.0833	35000.15	fiscal	0.041862	163	6.823579	46.56123	1276.577	
9	our	0.039285	11516	452.4063	204671.5	our	0.039285	139	5.460618	29.81835	2470.418	
10	for	0	9110	0	0	for	0	118	0	0	0	
11	a	0	7944	0	0	a	0	103	0	0	0	
12	we	0.047064	7553	355.4717	126360.1	we	0.047064	98	4.612237	21.27273	1639.52	
13	as	0	5966	0	0	as	0	93	0	0	0	
14	net	0	3472	0	0	net	0	79	0	0	0	
15	increase	0	1897	0	0	increase	0	72	0	0	0	
16	on	0	4369	0	0	on	0	65	0	0	0	
17	cash	0	2802	0	0	cash	0	63	0	0	0	
18	from	0	4631	0	0	from	0	60	0	0	0	
19	revenue	0.036723	1948	71.53603	5117.403	revenue	0.036723	59	2.166646	4.694353	154.9932	
20	an	0	2310	0	0	an	0	57	0	0	0	
21	was	0	3142	0	0	was	0	54	0	0	0	
22	by	0	3742	0	0	by	0	53	0	0	0	
23	with	0	2933	0	0	with	0	52	0	0	0	

IDF | Good | Poor | 2003_APPLIEDMATERIALS | 2003_CIRRUS | 2003_CREE | 2003_DSP | 2003_EXAR | 2003_MA

FIGURE 3 - Screenshot of Process for Classifying Testing Documents Based on Similarity

ID	Classification	Positive-toned	Negative-toned
2010_INTEL	Poor	0.45994219	0.59099705
2005_INTEL	Good	0.485401425	0.286004309
2007_ADVANCEDMICRO	Poor	0.332072227	0.403926409
2006_ADVANCEDMICRO	Poor	0.290139907	0.383318744
2010_APPLIEDMATERIALS	Good	0.353326226	0.247177971
2007_APPLIEDMATERIALS	Good	0.349633996	0.254587695
2006_APPLIEDMATERIALS	Good	0.339853177	0.25163829
2003_APPLIEDMATERIALS	Good	0.284261758	0.242848846
2010_INTERSIL	Good	0.413589733	0.380749806
2009_INTERSIL	Good	0.425600577	0.399137258
2007_INTERSIL	Good	0.430947726	0.399988008
2007_SUPERTEX	Good	0.335180269	0.286754627
2005_INTEGRATEDDEVICE	Good	0.445714144	0.430812212
2004_EXAR	Good	0.311394085	0.282099215
2003_EXAR	Good	0.291904659	0.275798533
2009_CYPRESS	Poor	0.40336438	0.410234509

APPENDIX A

BRIEF INTRODUCTION TO VSM AND TFIDF

Vector Space Model

Term Weighting

Each document is represented as a vector of its distinctive terms and their weights. Weighting methods include (1) the term frequency (TF) method, (2) the inverse document frequency (IDF) method, and (3) the term frequency inverse document frequency (TFIDF) method.

1. TF measures how many times a term appears in a document, represented as

$$TF_{ij}=N_j/N_{all},$$

where

TF_{ij} is the frequency of term T_j in document D_i,

N_j is the number of times term T_j appears in document D_i, and

N_{all} is all of the terms in document D_i.

2. IDF is a logarithm of a fraction of the total number of documents in the text collection measured by the number of documents in which a term has appeared. It refers to the rarity of a term in a document collection and is represented as

$$IDF_j=log_2(N/DF_j),$$

where

IDF_j is the inverse document frequency of term T_j,

N is the total number of documents, and

DF_j is the document frequency of term T_j.

Note that when DF_j is N, i.e., when term T_j appears in all of the documents, IDF_j becomes 0. In this approach, all of the common words, such as "the," "of," and "a" in English, are automatically eliminated.

3. The TFIDF method is the most successful and widely used weighting scheme to estimate the usefulness of a term in a document. The implication is that the most descriptive and representative terms in a given document are those that occur very frequently in the given document and less frequently in other documents. Therefore, it assigns the highest weights to those terms that appear with a high frequency in an individual document, but are relatively rare in the text collection. The weight of TFIDF can be represented by this formula:

$$W_{ij} = TF_{ij} * IDF_j = (N_j/N_{all}) * (log_2(N/DF_j)),$$

where

W_{ij} is the weight value of term T_j in document D_i.

TFIDF and Vector Space Model

The TFIDF values can be used to create vector representations of documents. Each component of a vector corresponds to the TFIDF value of a particular term in the corpus. Terms that do not occur in a document are weighted zero. To compare analyzed results more conveniently, the information must have an established standard or an index value. The purpose of the index value is to give weights to each individual word based on their importance and representativeness of the document. When the documents are represented as vectors in high dimensional space corresponding to all of the keywords, relevance can be measured with an appropriate similarity measure defined over the vector space.

Vector space model representations are developed to represent each document in a collection as a point in space; points that are close together are semantically similar and points that are far apart are semantically distant. Consider a document space consisting of documents D_i, where i is the number of documents, each identified by one or more index terms (W_1, W_2, W_3, ... W_j), and j is the number of terms. For example, assume there is a typical three-dimensional index space with three weighted keywords (W_1, W_2, and W_3), and three documents as vectors (*Doc1*, *Doc2*, and *Doc3*). As each document has different index keywords, each document has a different position in space. The similarity of one document vector to another is the cosine of the angle between them. A collection of i documents and k terms can be represented in the vector space model by a term-document matrix. Generally, we use W_{ki} corresponding to the weight of a term k in document i.

APPENDIX B

DIFFERENCES IN TEXT SEGMENTATION FOR ENGLISH AND CHINESE TEXTS

Text Segmentation

Text in documents can be segmented into terms, phrases, or sentences, depending on your needs. In English, the simplest method is to use spaces to segment text into terms; this is known as the “bag of words” or “bag of tokens” method. This method has the advantages of easier computations and mature theories for term weighting that have emerged over the last couple of decades. However, the bag of words method loses the sequence of information in the text and suffers from the problems of homonymy, polysemy, synonymy, and hyponymy.

Scholars have argued that phrase-based or sentence-based approaches will perform better than term-based ones because phrases may carry more semantic information. However, this argument has not been supported by actual studies (Scott and Matwin 1999). Although phrases are less ambiguous and more discriminative than individual terms, studies using phrases have had disappointing results, as phrases occur at relatively low frequencies and there are many redundant and noisy phrases.

In Chinese, term-based segmentation is not as easy as in English, as there are no spaces between characters. To solve this problem, the most effective approach is to include all of the Chinese words in the program’s dictionary. For Traditional Chinese, the Chinese Knowledge and Information Processing (CKIP) system can automatically extract new words, create words, or conduct an online word-segmentation function. This system has a vocabulary of approximately 100,000 words and the ability to recognize parts of speech, word frequency, and the frequency of double-conjunction classes.

For Simplified Chinese, which is also written without spaces between words, the Stanford Chinese Word Segmenter, a Java implementation of the CRF-based Chinese Word Segmenter offered by The Stanford Natural Language Processing Group, has the ability to split the Simplified Chinese text into a sequence of words, defined according to some word-segmentation standard (Tseng et al. 2005).