Spring 2011

Applying text timing in corporate spin-off disclosure statement analysis: understanding the main concerns and recommendation of appropriate term weights

Aravindh Sekar

APPLYING TEXT TIMING IN CORPORATE SPIN-OFF DISCLOSURE STATEMENT ANALYSIS: UNDERSTANDING THE MAIN CONCERNS AND RECOMMENDATION OF APPROPRIATE TERM WEIGHTS

By

ARAVINDH SEKAR

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN INFORMATION SCIENCE AND TECHNOLOGY

2011

Approved by

Dr. Wen-Bin Yu, Advisor
Dr. Ying Chou Lin
Dr. Bhi-Ru Lea

ABSTRACT

Text mining helps in extracting knowledge and useful information from unstructured data. It detects and extracts information from mountains of documents and allowing in selecting data related to a particular data.

In this study, text mining is applied to the 10-12b filings done by the companies during Corporate Spin-off. The main purposes are (1) To investigate potential and/or major concerns found from these financial statements filed for corporate spin-off and (2) To identify appropriate methods in text mining which can be used to reveal these major concerns.

10-12b filings from thirty-four companies were taken and only the "Risk Factors" category was taken for analysis. Term weights such as Entropy, IDF, GF-IDF, Normal and None were applied on the input data and out of them Entropy and GF-IDF were found to be the appropriate term weights which provided acceptable results. These accepted term weights gave the results which was acceptable to human expert's expectations. The document distribution from these term weights created a pattern which reflected the mood or focus of the input documents.

In addition to the analysis, this study also provides a pilot study for future work in predictive text mining for the analysis of similar financial documents. For example, the descriptive terms found from this study provide a set of start word list which eliminates the try and error method of framing an initial start list.

ACKNOWLEDGMENTS

I would like to begin by extending my gratitude to my thesis advisor and mentor, Dr. Wen-Bin Yu for having confidence in me and providing me with the opportunity to work on this research. He has been my inspiration and guiding force throughout my stay in Rolla and he is someone who I will always look up to. Without him, this thesis would not have come to this stage and I owe everything to him.

In addition, I would like to thank my thesis committee member, Dr. Ying Chou Lin, of the Business and Information Technology Department who taught me the financial concepts and the importance of SEC filings. This thesis would have been impossible without his support, advice and encouragement.

I would also like to thank my other thesis committee member, Dr. Bhi-Ru Lea of the Business and Information Technology Department. She taught me the basics and importance of research and the dedication required to perform any task, which I will carry with me my entire life time.

Finally, A huge thanks to my parents, sister and friends for their constant support and encouragement. More than my dream it has been my parents and sister's dream that I pursue and complete my Master's degree and today I hope I have made one of my family's dream come true. I owe all my hard work and struggle to reach till this goal, to them.

Thank you everyone and I am nothing without you all.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

The purpose of this research is to apply text mining to the 10-12b filing s done by the companies during Corporate Spin-off. The main purposes are (1) To investigate potential and/or major concerns found from these financial statements filed for corporate spin-off and (2) To identify appropriate methods in text mining which can be used to reveal these major concerns.

Deep penetration of personal computers, data communication networks, and the Internet has created a massive platform for data collection, dissemination, storage, and retrieval. Every day, people engage in numerous online activities, including reading the news and product reviews, commenting on developing events, buying and selling stocks, and widening their social networks. This widespread engagement with online worlds has facilitated the creation of large amounts of textual data (Lu et al., 2007).

The common knowledge is that almost 80% of the corporate data is textual (Chen, 2001; Robb, 2004). These texts contain vast amounts of untapped data, which is very difficult to decipher because of its unstructured nature. Text mining is often used to aide in the extraction of knowledge and useful information from these textual documents. Text mining explores for data in text files to establish valuable patterns and rules that indicate trends and significant features about specific topics (Lau et al., 2005). Text mining extracts high-level knowledge and useful patterns from low-level textual data (Durfee et al., 2007). Text mining tools seek to automatically analyze and learn the meaning of implicitly unstructured information. The key to gaining knowledge from internal and external textual repositories is by exploiting computers for processing the vast amounts of textual data with text mining software using text clustering to discover intrinsic knowledge within documents. Low-level data is transformed to richer data by detecting meaningful themes implicitly present in the data (Leong et al., 2004).

Text mining uncovers the underlying themes contained in large document collections. Text mining applications have two phases: exploring the textual data for its content, and then using discovered information to improve the existing processes. Both phases are important and are often referred to as descriptive mining and predictive mining, respectively.

Descriptive mining involves discovering the themes and concepts that exist in a textual collection. For example, many companies collect customers' comments from sources that include the Web, e-mail, and call centers. In general, mining the textual comments includes providing detailed information about the terms, phrases, and other content by extracting meaningful information from the textual collection. Also, clustering the documents into meaningful groups and reporting the concepts that are discovered in the clusters are performed as a part of descriptive mining. Results from descriptive mining provide a better understanding of the textual collection.

On the other hand, predictive mining involves classifying the documents into categories and using the information that is implicit in the text for decision making. Predictive modeling involves examining past data to predict future results. Both of these aspects of text mining share some of the same requirements. Namely, textual documents that human beings can easily understand must first be represented in a form that can be mined by the software. The raw documents require processing before the patterns and relationships that they contain can be discovered. Although the human mind comprehends chapters, paragraphs, and sentences, computers require structured (quantitative or qualitative) data. As a result, an unstructured document must be converted into a structured form before it can be mined.

Text mining differs from data mining in different ways. Unlike data mining, text mining works with an unstructured or semi-structured collection of text documents (Lau et al., 2005). In general, texts present in large databases cannot be analyzed by normal data mining statistical methods but can be preprocessed by text mining technology, which extracts knowledge from very large amounts of textual data (Nasukawa & Nagano, 2001).

Text mining is used in medical and business applications, as well as the sports and insurance industries. The new trend emerging from text mining is its application in the financial industry. Various text mining tools are applied to analyze the financial performance of an industry, and also aide in making major decisions on the company as a whole. Financial applications cover a wide range of functions, including forecasting the stock market, currency exchange rates, bank bankruptcies, understanding and managing financial risks, trading futures, future trends of stocks in the market, credit ratings, loan

management, bank customer profiling, and money laundering analyses (Nakhaeizadeh et al., 2002).

Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks, and which stocks to purchase. Financial institutions produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Potential significant benefits of solving these problems motivated extensive research for years (Kovalerchuk, 2006).

Generally, finance-related textual content falls roughly into three categories. The first category includes forums, blogs, and wikis. A typical IT company forum has hundreds of new messages every day. Users actively share their investment strategies, new product information, perspectives and opinions. Information about the company's background, rumors, and news updates are also prevalent in many finance-related blogs and wiki sites (Lu et al., 2007).

The second category of finance-related content includes news and research reports. Newspaper articles are often accessible on news websites. Moreover, various finance portals provide intraday updates with contents from newswire services. Some portal sites also provide access to research reports generated by analysts (Lu et al., 2007).

The third category involves finance-related content generated by firms. Many firms maintain their own websites as a communication channel with consumers and investors (Lu et al., 2007).

This thesis is primarily about applying text mining techniques in the case of corporate spin-off's. The structure of this thesis begins with the literature review which defines text mining, its processes and methods, and its application to various industries like medical, business and finance. An explanation about corporate spin-offs and more detailed information about it are mentioned as a part of the financial application. This is followed by the Methodology section, which explains the method used in the study for data analysis. The data used for this research has been identified and the results analyzed. The methodology and the data section is followed by the conclusion of the analysis and how these conclusions can be used for future study is suggested as future works.

2. LITERATURE REVIEW

2.1. DEFINITION OF TEXT MINING

Text mining explores data in text files to establish valuable patterns and rules that indicate trends and significant features about specific topics (Lau et al., 2005). Text mining is also defined as a sub-specialty of knowledge discovery from data and as a process of utilizing computers to extract useful information from vast volumes of digital content. Low-level data is transformed to richer data by detecting meaningful themes implicitly present in the data (Leong et al., 2004).

Text mining is a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools (Feldman et al., 2007). To be more specific, the process is aimed to understand and interpret semi-structured and unstructured data (Sirmakessis, 2004) in order to discover and extract knowledge from them, unlike data mining, which discovers knowledge from structured text (Ananiadou & McNaught, 2006).

In a direct reference to text mining and its properties, Nasukawa and Nagano (2001) defined text mining as a method which detects and extracts relevant documents from mountains of documents based on select data related to specific topics of interest so that the amount of data to be handled is reduced without losing the required information. This makes it possible to discover patterns and trends semi-automatically from huge collections of unstructured text using technologies such as natural language processing, information retrieval, information extraction, and data mining (Uramoto et al., 2004).

It is not necessary that all text mining process follow the order of natural language processing, information retrieval, information extraction, and data mining, but most of the text mining processes follow one of these steps, thereby showing the importance of each step in the process of text mining.

2.2. TEXT MINING PROCESS

2.2.1. Information Retrieval. Much of business information is text and the information is subject to frequent changes. The use of efficient and effective mechanisms to retrieve required business information is a key to business success, and automated processing of text to extract key terms is an essential component of such an information retrieval (IR) system (Gao et al., 2005).

Nasukawa and Nagano (2001) mentioned that information retrieval is probably the most common technology to use when faced with a very large number of documents. They also stated that the term "Text Mining" (or Text Data Mining) detects and extracts documents wanted from mountains of documents, and allows selection of data related to some specific topics.

An Information Retrieval (IR) system performs a matching function between data (Miller, 2005). IR consists of two key processes: document storage and document retrieval (Gao et al., 2005). In the storage process, an IR system defines a collection of documents, specifies the manipulation of the documents, and represents the documents with an index. In the retrieval process, a user specifies an information requirement, which is then manipulated by the system and represented with a query in a certain format according to the retrieval strategy of the system. The query is then compared with the index to identify documents that are relevant to the query and relevant documents are retrieved and presented to the user.

Three key issues must be considered in Information Retrieval (Gao et al., 2005). Firstly, the choice of appropriate terms is a challenge for both index creation and query generation. A term can be a single word or a multi-word phrase. Most experiments show that using phrases in IR obtains consistent results (Koster & Seutter, 2003).

Secondly, a fundamental problem that hinders a successful retrieval is term mismatch or a vocabulary problem (Tseng, 2002). Frequently, terms used by users do not match those that represent the same or similar meanings in documents. A common solution to this problem is to create a thesaurus, which coordinates the usage of the query terms and index terms.

Thirdly, without detailed knowledge of the document collection and retrieval environment, users find it difficult to formulate appropriate queries. In some situations,

users do not know what they really want to search for (Zhou & Zhang, 2003). In other situations, they cannot specify their precise information requirement (Nakashima et al., 2003). However, most users can explain their requirements with reference to a specific example. Therefore, some IT systems incorporate the technique of case-based reasoning (CBR) (Gao et al., 2005), which formulates a query by analyzing examples of relevant documents. A major task of CBR is to extract terms from the examples for the generation of a query.

2.2.2. Information Extraction. Text mining looks for patterns in unstructured text. The related task of Information Extraction (IE) is about locating specific items in Natural-language documents (Kanya & Geetha, 2007). Companies are increasingly applying IE behind the scenes to improve information and knowledge management applications such as text search, text categorization, data mining and data visualization (Taylor, 2004).

The objective of IE is to extract certain pieces of information from text that are related to a prescribed set of related concepts, namely, an extraction scenario (Jordi et al., 2007).

Mooney and Bunescu (2007) mentioned that Information Extraction distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. IE systems can be used to directly extricate abstract knowledge from a text corpus or to extract concrete data from a set of documents, which can then be further analyzed with traditional data mining techniques to discover more general patterns.

Information retrieval (IR) and Information Extraction (IE) are two major areas of Text Based Intelligence systems (Jordi et al., 2007). IR techniques are used to select those documents from a collection that most closely conform to the restrictions of a query, commonly a list of keywords. As a consequence, IR techniques allow recovering relevant documents in response to the query.

IE technology involves a more in-depth understanding task. While in IR the answer to a query is simply a list of potentially relevant documents, in IE the relevant content of such documents has to be located and extracted from the text. This relevant

content, represented in a specific format, can be integrated into knowledge-based systems as well as used in IR in order to obtain more accurate responses.

IE can serve an as important technology for text mining (Mooney & Bunescu, 2007). If the knowledge to be discovered, is expressed directly in the documents to be mined, then IE alone can serve as an effective approach to text mining. However, if the documents contain concrete data in unstructured form rather than abstract knowledge, it may be useful to first use IE to transform unstructured data in the document corpus into a structured database, and then use traditional data mining tools to identify abstract patterns in this extracted data.

IE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from unstructured text (Mooney & Bunescu, 2007). One type of IE is entity recognition, which involves identifying references to particular kinds of objects such as names of people, companies, and locations.

Another application of IE is extracting structured data from unstructured or semi-structured web pages. When applied to semi-structured HTML, typically generated from an underlying database by a program on a web server, an IE system is typically called a wrapper, and the process is sometimes referred to as screen scraping.

IE has its own benefits which help an analyst in a great way. It ideally projects as a sole way for information extraction from huge sets of documents which an analyst can use to exploit the data. These aides an analyst in identifying the required information from the pile of data, thereby, speeding up the process for the analyst. Patterns and trends will also be easily identified, further assisting the analyst in simplifying the process in a more effective manner.

The most related research is document explorer (Feldman et al,; 1998) which uses automatic term extraction for discovering new knowledge from texts. However document explorer assumes semi-structured documents such as SGML text unlike DISCOTEX developed for natural language text. Similarly automatic text categorization has been used to map web documents to pre-defined concepts for further discovery of relationships among the identified concepts (Loh et al.; 2000). One of the limitations for these approaches is that they require a substantial amount of domain knowledge.

2.2.3. Data Mining. Data mining means extracting or "mining" knowledge from large amounts of data, which are also referred to as knowledge mining from data, knowledge extraction, pattern analysis, data dredging and knowledge discovery in a database. Data mining is a multidisciplinary field including database technology, AI (artificial intelligence), Machine Learning, Neural Networks, statistics and so on. Data cleaning and integration, data collection, data transformation, data mining, knowledge evaluation and presentation are the general processes in a project of data mining (Li & Zhang, 2009).

The function of data mining includes association analysis, classification and prediction, clustering analysis, outlier analysis, etc. (Olson et al., 2001).

Data mining is a process by which accurate and previously unknown information is extracted from large volumes of data. This information should be in a form that can be understood, acted upon, and used for improving decision processes (Apte, 2007).

Apte (2007) divided data analysis algorithms into three major categories based on the nature of their information extraction: predictive modeling (also called classification or supervised learning), clustering (also called segmentation or unsupervised learning) and frequent pattern extraction. In this thesis, classification and clustering are discussed below.

Classification. Classification or Predictive modeling is based on techniques used for classification and regression modeling. One field in the tabular data set is pre-identified as the response or class variable. These algorithms produce a model for that variable as a function of the other fields in the data set, pre-identified as the features or explanatory variables (Apte, 2007).

Classification is identified as the task of assigning the class label to the unclassified data objects as accurately as possible by building a model for target attribute as a function of predictive attributes based on the pre-classified dataset (Hongqi Li et al., 2008). On the other hand, Guyon and Elisseeff (2003) mentioned that feature selection for classification is the technique of choosing an optimal subset of features (also called attributes or variables) by removing the most irrelevant and redundant features from the dataset in order to enhance model generalization capability and simplify data mining results.

The classification process has two phases (Wah et al., 2001). The first phase is the learning process whereby training data is analyzed by a classification algorithm. The learned model or classifier is represented in the form of classification rules. The second phase is classification, and test data are used to estimate the accuracy of classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data.

Some of the techniques used for data classification are decision trees. The advantage of the decision tree technique is that it does not require any domain knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. The second technique is neural-network, which has a high tolerance of noisy data as well as the ability to classify patterns on which they have not been trained. It can be used when we have little knowledge of the relationship between attributes and classes. Next, the K-nearest-neighbor technique is an instance-based learning using distance metric to measure the similarity of instances.

Clustering: Clustering is one of the important techniques of data mining (Han et al., 2001; Qian & Dong 2004). Clustering can divide data objects into several classes or clusters based on data objects comparability (Zhifu et al., 2007). So the objects of the same cluster have high comparability but have a greater difference between the objects of different clusters.

Clustering analysis has been applied to various fields such as pattern identification, data analysis, and image processing and so on (Zhifu et al., 2007). K-Means clustering algorithm is a general and simple clustering algorithm and it can divide n objects into K classes or clusters by using a K parameter, resulting in a high comparability in one class and low comparability between different classes (Han et al., 2001).

The process of a common K-Means algorithm is as follows: First, randomly select K objects where each object originally expresses the average or center of one cluster. Then the remaining objects can be given to the nearest cluster according to the distances between the object and the center of every cluster. Next, compute the average of every cluster over again. This process repeats continually until the rule function is constringed.

Clustering determines the features which better describe objects in the set, intra-cluster similarity, while distinguishing objects in the set from the collection, inter-cluster dissimilarity (Yates & Neto, 1999). Intra-cluster similarity measures a raw frequency of a term k_i inside document d_j, aka the TF factor. Inter-cluster dissimilarity measures the inverse of the frequency of a term k_i among the documents in the collection, aka inverse document frequency or IDF factor. IDF weighting focuses on inter-cluster dissimilarity and tries to reduce the effect when the terms appearing in many documents are not useful for distinguishing documents. The product of TF and IDF (TFIDF) was proposed as a reasonable measure which tries to balance the two effects, intra-cluster similarity and inter-cluster dissimilarity.

2.3. TEXT MINING METHODS

Text mining is the base of several analysis and researchers have developed various text mining methods to analyze related issues. Durfee et al. (2007) proposed the use of a text clustering methodology known as the Prototype matching method as a text mining technique. Prototype matching method was implemented in a prototyping software package called GILTA-3, which seeks similarity between the document prototype and the closest-matching subject documents.

A new prototype of text mining called Text Analysis and knowledge mining (TAKMI) was developed by Nasukawa and Nagano (2001) which, when applied in the PC help centers, can automatically detect product failures and can also determine issues that have lead to an increase in problems and thereby help in analyzing them and identifying changes in customer behavior involving a particular product. Mooney and Bunescu (2005) proposed the use of Information Extraction as a methodology to directly extract knowledge from text and then discover knowledge by mining data previously extracted from an unstructured or semi-structured text. Mooney and Bunescu (2005) developed an IE method called Relational Markov Networks that captures dependencies between distinct candidate extractions in a document, whereas, (Hearst, 1999) used text compression as a key technology of text mining.

Dörre et al. (1999) developed IBM Intelligent Miner, which is a software development tool kit for building text mining applications. It addresses system integrators, solution providers, and application developers. The main work of the intelligent miner's tool is the feature extraction and mining in documents. Wang et al. (2004) proposed a text mining methodology using an associational approach. This method enables multiple classifications of a same set of high frequency words and achieves high performance even with unstructured text data in terms of retrieval efficiency and explanatory power of the final result.

Udoh and Rhoades (2006) explained a new method named Wordstat. It determines the dominant activities in a small enterprise such as a company or an institution. An analysis of document profiles is done, which is generated by extracting the frequencies of certain terms on the basis of repetitive occurrence and co-occurrence of those terms. The main conclusion drawn is Wordstat's ability in detecting patterns and similarities in documents.

2.3.1. Dimension Reduction. Dimension reduction refers to mapping points in a high dimensional space to a space with low dimensions while approximately preserving some property of the original points (Charikar & Sahai, 2002).

Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis (Fodor, 2002). Such datasets, in contrast with smaller, more traditional datasets that have been studied extensively in the past, present new challenges in data analysis. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation. The dimension of the data is the number of variables that are measured on each observation (Fodor, 2002).

High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments (Donoho, 2000). One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are "important" for understanding the underlying phenomena of

interest. While certain computationally expensive novel methods (Breiman, 2001) can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data.

The set of techniques that can be employed for dimension reduction can be partitioned in two important ways; they can be separated into techniques that apply to supervised or unsupervised learning, and into techniques that either entail feature selection or feature extraction (Cunningham, 2007).

2.3.1.1 Feature selection. Feature selection (FS) algorithms take an alternate approach, to dimension reduction by locating the best minimum subset of the original features, rather than transforming the data to an entirely new set of dimensions (Cunningham, 2007). For the purpose of knowledge discovery, interpreting the output of algorithms based on feature extraction can often prove to be problematic, as the transformed features may have no physical meaning to the domain expert. In contrast, the dimensions retained by a feature selection procedure can generally be directly interpreted.

Feature selection in the context of supervised learning is a reasonably well posed problem. The objective can be to identify features that are correlated with or predictive of the class label. Or more comprehensively, the objective may be to select features that will construct the most accurate classier. In unsupervised feature selection the object is less well posed and consequently it is a much less explored area (Cunningham, 2007).

In supervised learning, selection techniques typically incorporate a search strategy for exploring the space of feature subsets, including methods for determining a suitable starting point and generating successive candidate subsets, and an evaluation criterion to rate and compare the candidates, which serve to guide the search process. The evaluation schemes used in both supervised and unsupervised feature selection techniques can generally be divided into three broad categories (Jolliffe, 1972; Cardoso, 1984).

Filter. Filter approaches attempt to remove irrelevant features from the feature set prior to the application of the learning algorithm. Initially, the data is analyzed to identify those dimensions that are most relevant for describing its structure. The chosen feature subset is subsequently used to train the learning algorithm. Feedback regarding an algorithm's

performance is not required during the selection process, though it may be useful when attempting to gauge the effectiveness of the filter (Cunningham, 2007).

Wrapper. Wrapper methods for feature selection make use of the learning algorithm itself to choose a set of relevant features. The wrapper conducts a search through the feature space, evaluating candidate feature subsets by estimating the predictive accuracy of the classier built on that subset. The goal of the search is to find the subset that maximizes this criterion (Cunningham, 2007).

Embedded. Embedded approaches apply the feature selection process as an integral part of the learning algorithm. The most prominent example of this is the decision tree building algorithms such as Quinlan's C4.5 (Hialthouse, 1996). There are a number of neural network algorithms that also have this characteristic. Breiman (2007) has shown recently that Random Forests, an ensemble technique based on decision trees, can be used for scoring the importance of features. He shows that the increase in error due to perturbing feature values in a data set and then processing the data through the Random Forest is an effective measure of the relevance of a feature.

2.3.1.2 Feature extraction. Feature extraction involves the production of a new set of features from the original features in the data, through the application of some mapping. Well-known unsupervised feature extraction methods include Principal Component Analysis (PCA) and spectral clustering (Ng et al., 2001). The important corresponding supervised approach is Linear Discriminant Analysis (LDA) (Hyvarinen et al., 2001).

2.3.2. Term Document Frequency. Term Frequency is a weight and statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the term frequency weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query (Jones & Karen, 1972).

The document representation is one of the pre-processing processes that are used to reduce the complexity of the documents and make them easier to handle. The document is first transformed from the full text version to a document vector, an

important aspect in the document's categorization, which denotes the mapping of a document into a compact form of its content (Khan et al., 2010).

A text document is typically represented as a vector of term weights i.e. word features from a set of terms (dictionary), where each term occurs at least once in a certain minimum number of documents. A major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents. Feature selection is used to create vector space, which improves the scalability, effectiveness and accuracy of a text classifier. "A good feature selection method should consider domain and algorithm characteristics (Chen, 2009)".

The most common method for document representation is Vector Space Model (VSM) which is most widely used for document categorization. The VSM represents each document as a feature vector of the terms (words or phrases) in the document. Each feature vector contains term weights (usually term-frequencies) of the terms in the document. However, this method has some disadvantages, one which does not consider the dependency between the terms and also ignores the sequence and structure of the term in the documents (Khan et al., 2010).

A vector-space approach is commonly employed to convert qualitative representation of documents into a quantitative one since it is simple as well as has been proved to be superior or as good as the known alternatives (Baeza-Yates & Ribeiro-Neto, 1999). Coussement (2008) described the approach as "the mean that original documents are converted into a vector in a feature space based on the weighted term frequencies. Each vector component reflects the importance of the corresponding term by giving it a weight if the term is present or zero otherwise." The final vector is represented as a term-document frequency matrix.

In the first two steps, the most informative terms were selected. Thus, the current set of terms is ready to be converted. Base on the term assignment array of Salton and McGill (1983), the vector representation of documents can be represented as a term document frequency matrix as shown in Table 2.1. Terms are rows and documents are columns. Each cell contains a frequency value of the term in the document. In the matrix, $f_{i,j}$ is the number of times that term i appears in document j.

Table 2.1 Term-Document Frequency Matrix

Terms	Documents			
	D1	D2	…	Dn
T1	$f_{1,1}$	$f_{1,2}$	…	$f_{1,n}$
T2	$f_{2,1}$	$f_{2,2}$	…	$f_{2,n}$
…	…	…	…	…
Tm	$f_{m,1}$	$f_{m,2}$	…	$f_{m,n}$

Albright (2004) described this model in detail. The model ignores the context of the documents while providing their quantitative representation. The resulted matrix is generally sparse and will become much sparse when the size of document collection increases, since few terms are contained in any single document. Also, only hundreds of documents can yield thousands of terms. Huge computing time and space are required for the analysis. Therefore, reducing dimensions of the matrix can improve performance significantly.

In addition, another way to improve retrieval performance of the analysis is to apply weighting methods (Berry & Browne, 1999). According to Berry and Browne (1999), the performance refers to the ability to retrieve relevant information while dismissing irrelevant information. Each element of the matrix ($a_{i,j}$) can be applied to the weighting and represented as

$$a_{i,j} = l_{i,j} g_i d_j \text{ , where}$$

$l_{i,j}$ is the frequency weight for term i occurring in document j,

g_i is the term weight for term i in the collection, and

d_j is a document normalization factor indicating whether document j is normalized.

This equation was originally applied from information retrieval for search engines where longer documents have a better chance to contain terms matching the query than the shorter ones. Therefore, the document normalization factor was included to equalize the length of the document vectors from documents which vary in length (Salton & Buckley, 1988). Since this paper focused on text mining and the lengths of the documents in the collection were not varied, the third factor was unnecessary and ignored by replacing the variable with 1. Then, the final equation is

$$a_{i,j} = l_{i,j}$$

Defining the appropriate weighting depends on characteristics of the document collection. The frequency weights and term weights are popular weighting schemes which are described in more detail in the following subsections..

2.3.3. Frequency Weights. Frequency weight is used to adjust the frequencies in the term-by-document matrix to prevent high-frequency, commonly-occurring terms from dominating the analysis. Frequency weights are functions of how many times each term appears in a document (Chisholm & Kolda, 1999). Because unique, often rare terms can play a significant role in distinguishing between different types of documents, it is normal to try to adjust rare term frequencies with a weight factor to give them an opportunity to contribute more to the analysis. They are functions of the term frequency ($\mathbf{l_{i,j}}$). This factor measures the frequency of occurrence of the terms in the document by using a term frequency (TF). Common methods include binary and logarithm. Three common weighting schemes are shown below where $\mathbf{f_{i,j}}$ represents the original frequency of term i appears in document j.

Binary: $$l_{i,j} = \begin{cases} 1 \text{ if term i is in document j} \\ 0 \text{ otherwise} \end{cases}$$

Logarithm: $$l_{i,j} = \log_2(f_{i,j})$$

None or
Term Frequency: $l_{i,j} = f_{i,j}$

Sometimes, a term is repeated in a document for a lot of time; thus, it reflects high frequency in the document collection as a whole even though it appears in only one document. To reduce the effect from the repetitive terms, Binary and Logarithm can be applied to the term frequency. Binary formula gives every word that appears in a document equal relevance. This can be useful when the number of times a word appears is not considered important (Polettini, 2004).

The Binary method takes no repetitive effect into account while Logarithm reduces the effect, but still maintains it in some degree. Therefore, the Logarithm is a method in between Binary and None. Moreover, taking log of the raw term frequency reduces effects of large differences in frequencies (Dumais, 1991). Logarithms are used to adjust within-document frequency because a term that appears ten times in a document is not necessarily ten times as important as a term that appears once in that document. Logarithms formulas decrease the effects of large differences in term frequencies (Polettini, 2004).

According to Berry and Browne (1999), the selection of appropriate weighting methods depends on the vocabulary or word usage patterns for the collection. The simple term frequency or none weighting term frequency is sufficient for collection containing general vocabularies (e.g., popular magazines, encyclopedias) (Berry & Browne, 1999). Term frequency formula counts how many times the term occurs in a document. Term frequency is used alone and it works well involving common words and long documents. This formula gives more credit to words that appears more frequently, but often too much credit (Kolda, 1997).

If the collection spans general topics such as news feeds, magazine articles, etc., using term frequency would suffice. If the collection were small in nature with few terms in the vocabulary, then binary frequencies would be the best to use (Giles et al., 2003).

Logarithms are a way to de-emphasize the effect of frequency. Literature proposes log as the most used frequency weight (Kolda, 1997). Logarithms are used to adjust within-document frequency because a term that appears ten times in a document is

not necessarily ten times as important as a term that appears once in that document. Logarithms formulas decrease the effects of large differences in term frequencies (Polettini, 2004).The logarithm formulas offer a middle ground (Polettini, 2004).

2.3.4. Term Weights. Term weights are statistical measures used to evaluate how important a word is to a document in a collection or corpus. They take word count in the document into account. Term weights are functions of how many times each term appears in the entire document collection (Chisholm & Kolda, 1999).Common methods are

Entropy : $G_i = 1 + \sum_j \frac{p_{i,j} log_2(p_{i,j})}{log_2 n}$

GF-IDF : $G_i = (\sum_j f_{i,j}) / \sum_j X(f_{i,j})$

IDF : $G_i = \log (n / \sum_j X(f_{i,j})$

Normal : $G_i = 1 / \sqrt{\sum_j f_{i,j}^2}$

None : $G_i = 1$

Where, $f_{i,j}$ represents the original frequency of term i appears in document j, n is number of documents in the collection, as well as

$$p_i = f_{i,j} / \sum_j f_{i,j}$$

$$X(f_{i,j}) = \begin{cases} 1 \text{ if term i is in document j} \\ 0 \text{ otherwise} \end{cases}$$

In determining the term weights, the likelihood that the collection will change needs to be considered (Giles et al., 2003). The choice for an appropriate term weight depends on the state of the document collection, or how often the collection is likely to change (Berry & Browne, 1999). This weighting scheme responds to new vocabulary and accordingly affects all rows of the matrix. All of the formulas emphasize those words that occur in few documents whereas they give less weight to terms appearing frequently or in

many documents in the document collection. In general, the document collection will work well with some weighting schemes and poorly with others (Giles et al., 2003).

Term weights try to give a "discrimination value" to each term. Many schemes are based on the idea that the less frequently a term appears in the whole collection, the more discriminating it is (Salton & Buckley, 1988).

Entropy is based on information theoretic ideas and is the most sophisticated weighting scheme. It assigns weights between 0 and 1 for a term that appears in only one document. If a term appears once in every document, then that term is given a weight of zero. If a term appears once in one document, then that term is given a weight of one. Any other combination of frequencies will yield a weight somewhere between zero and one. Entropy is a useful weight because it gives higher weight for terms that appear fewer times in a small number of documents (Polettini, 2004). So this formula takes into account the distribution of terms over documents (Dumais, 1991).

The Inverse Document Frequency (IDF) is a popular measure of a word's importance (Polettini, 2004). IDF is the logarithm of the inverse of the probability that term i appears in a random document j. It awards high weight terms appearing in few documents in the collection and low weight for terms appearing in many documents in the collection (Chisholm & Kolda, 1999). GF (Global Frequency) - IDF assigns the smallest possible weight if a term appears once in every document or once in one document (Chisholm & Kolda, 1999).

According to Salton and Buckley (1988), one of the commonly used documents term weighting is obtained by the inner product operation of none or simple term frequency and the Inverse document frequency. In the analysis done on improving the retrieval of information from external sources, Dumais (1991) found that using IDF and Entropy term weight improved the performance by an average of 30% whereas when used with the combination of log and entropy, the performance improved by 40%.

Another popular method is combining Term frequency and IDF to form their product TF * IDF. Accordingly to Mittermayer (2004), when TF is used it is assumed that important terms occur in the document collection more often than unimportant ones. The application of IDF presupposes that the rarest terms in the document collection have the highest explanatory power. With the combined procedure TF*IDF the two measures

are aggregated into one variable. Tseng et al. (2007) used TF*IDF as a weighting method to analyze the data for pattern analysis. Loughran and McDonald (2009) used the TF * IDF method weighting scheme to perform an analysis on the financial texts and also on the 10-K filings for finding out alternative negative word list that better reflects the tone of the financial text than the already existing word list. They used this weighting scheme as TF represents the method used to account for the word frequency and normalization and IDF is used to adjust the impact across the entire collection.

2.3.5. Core Mining Process. The stage inherits analysis methods from data mining such as classification, decision trees, and clustering. Since the goal was to cluster the document filings into several clusters without pre-defined categories, this research only focuses on clustering. The clustering method being used in this research was Expectation- maximization (EM).

Expectation-Maximization (EM): The Expectation-Maximization (EM) algorithm is generally a framework for estimating the parameters of distribution of variables in data (Feldman & Sanger, 2007). It is adapted to the clustering problem as a probabilistic clustering technique which is not based on distance unlike the k-means method. According to Bradley et al. (1998), EM performed superior to other alternatives for statistical modeling purposes. It attempts to group items similar to each other together. In general, data is not distributed in the same pattern; thus, some combinations of attributes are more preferable than the others. The concept of density estimation is applied to EM, in order to identify the dense regions of the probability density of the data source. The goal of EM is to identify the parameters of each of k distribution that meet the probability of the given items belonging to the cluster. Initially, parameters of k distributions are randomly or externally selected. Then, the algorithm proceeds iteratively as described in the following steps (Feldman & Sanger, 2007).

- Expectation: Compute probability of the item belonging to the cluster by using the current parameters of the distributions, and then re-label all items accordingly to the probability.
- Maximization: Using current labels of the items, re-estimate the parameters of the distributions to maximize the likelihood of the items

- If the change in log-likelihood after each iteration becomes small, stop the process; otherwise, repeat the process again

Finally, clustering results are labels of the items, generated clusters, attached with estimated distributions.

After text mining process is done, a set of clusters is generated, along with assignments of each document to clusters.

2.4. TEXT MINING APPLICATIONS

Various methodologies have been developed for text mining to be applied in diverse organizations. Methodologies ranging from simple text mining tools to complex algorithms have been researched and used for the final benefit of the organization. Separate text mining technologies have been developed respective to a particular organization. Mooney and Bunescu (2005) developed a simple text mining technology using information extraction to directly extract knowledge from text and then to discover knowledge by mining data. Already existing software such as IBM Intelligent Miner, SAS text miner and SPSS text mining tools are also used by organizations for their text mining process. Adeva and Calvo (2006) mention a text mining tool called Pimiento, which can be used to track plagiarism in universities.

For the medical industry, Uramoto et al. (2004) from IBM have developed a tool named MedTAKMI, which is capable of running the entire biomedical database in an interactive manner. Text mining tools have also been developed for analyzing competitors' online persuasive themes in the hotel and auction industries, as well as e-mail bounce management.

For patent analysis, Xu (2009) discusses patent map, a text mining technology which is a process of gathering information and building a map, which also mines and analyses the patent documents.

NTM Agent, a text mining agent for net auction, was developed to resolve the problems customers faced in net auctions. The NTM agent primarily does the work of collecting web pages of the satisfied items for the user's search demand, extracts certain features of the items from the web pages and then makes a table which contains all the

captured features. The table can be used by the user to get details about the different items which he/she would like to buy (Kusumura et al.; 2003).

Text mining is also used in developing standardized descriptions of taxa in paleontology (Lea et al., 2006). A framework that uses text mining techniques was proposed which develops a taxon description recommendation system. This study provided insights on how text mining can be used to develop a descriptive model, as well as how the descriptive terms generated during the text mining process can be used to provide a basic set for a standard lexicon to develop a standardized taxon description recommendation.

A pilot study was performed by Katerattanakul (2010) on the application of text mining to find new information from a collection of survey comments evaluating the civil engineering learning system. Text mining helped in categorizing the comments into different groups in an attempt to identify "major" concerns from the users or students. This assisted the evaluators of the learning system to obtain the ideas from those summarized terms without the need of going through a potentially huge amount of data.

Various text mining tools have been used in the fields of customer relationship management, insurance industry, and archeological industry and also in the sports industry.

2.4.1. Text Mining in Medical Applications. The life science industry is an emerging market in which application spaces, such as drug discovery and development in the pharmaceutical sector and clinical record management in health care, have become areas of significant recent interest (Arlington et al., 2004).

Early papers mention the possibility of knowledge discovery from biomedical literature (Hearst, 1999).

Uramoto et al. (2005) presented a method to perform text mining on unstructured biomedical documents to facilitate knowledge discovery from the very large text databases using a tool named MedTAKMI. MedTAKMI tool is capable of running the entire biomedical database in an interactive manner. MedTAKMI was developed for a hierarchical category viewer because most biomedical entries are defined hierarchically. By mainly developing for medical purposes, it can scan millions of documents and

retrieve information. This shows the flexibility of text mining tools in terms of unique business applications.

With the advances in medical technology and wider adoption of electronic medical record systems, large amounts of medical text data are produced in hospitals and other health institutions daily. These medical texts include the patient's medical history, medical encounters, orders, progress notes, test results, etc. Although these text data contain valuable information, most are just filed and not referred to again. These are valuable data that are not used to their full advantage. Gong et al (2008) mentions mining in radiology reports due to the availability of rich information like describing a radiologist's observation on the patient's medical conditions.

Gong et al. (2008) proposed a text mining system which extracts and uses information in radiology reports. This system consists of three main modules: "a medical finding extractor, a report and image retriever, and a text assisted image feature extraction". To conduct research using the given text mining approach, large amounts of textual data produced in hospitals and other health institutions were taken as input. These medical texts include the patient's medical history, medical encounters, orders, progress notes, test results etc. This paper has proposed a text mining system which extracts and uses the information present in radiology reports. The structuring of free text reports bridges the gap between users and report database, making the information contained in the reports readily accessible. It also serves as an immediate result to other components of the system.

The ability to automatically identify relationships between cancer diseases and external factors from medical records for supporting cancer diagnosis would be a valuable contribution in public health fields. Lee et al. (2007) proposed a prototype for automating the extraction of relationships between cancer diseases and potential factors from clinical records. The methodology proposed here is of three stages which describes the framework for discovery of the relationships between cancer diseases and potential factors from clinical medical records. This paper discusses the text mining processes which extract patterns from clinical records. The three stages of the methodology begin with utilizing the cancer ontology thesaurus for extracting the key terms from clinical records, after which the algorithm was used to extract relationship between cancer

diseases and potential factors from medical records, and finally the SVM method was used to support the relatedness between the text and clinical records.

Tasha et al. (2006) made comparisons between the use of traditional text mining and natural language processing techniques and how these techniques can be integrated for future biomedical ontology and a user development interface. They described a common vocabulary that can be used to describe age related macular degeneration (AMD) through certain methods, and one among them is text mining. In the text mining methodology, a collection of documents known as a "corpus" is used as an input for all text mining algorithms. "The unstructured text in the corpus becomes a structured data object via the creation of a term-by-document frequency matrix". The research of this paper has found that human expert results were the best. But they have worked on how "text mining methods and Natural Language processing methods will enhance the analysis and generation of future descriptions".

2.4.2. Text Mining in Business Applications. Text mining is used in various business applications. Lau et al. (2005) proposed text mining as a means of information management which the hoteliers can use to develop competitive and strategic intelligence. Application of text mining in the hotel Industry is relatively new (Lau et al., 2005). The authors used the "online Text Mining" method to search through the internet to get vast amounts of business information from customer forums, their expectations about rooms and their prices, which can help the managers to better understand the customers and their business as a whole thereby keeping off the competitors.

Leong et al. (2004) used text mining to analyze competitors' online promotional text messages by taking the sites of top educational sites in the USA and analyzing their position with respect to the competitors by using a text mining tool from Megaputer called "Text Analyst". Text Analyst summarizes the text and identifies the key concepts. It sums up the frequency of occurrences of each concept and assigns a numeric semantic weight to each concept in relation to its importance in the document.

In addition to other applications of text mining, it is also used in the context of news. Kroha et al. (2006) framed a methodology to cleanse and classify business news and then investigate the similarity between the good news and bad news, by framing two types of templates. The first template was "to analyze the relative frequency of the given

words" and the second template was to "analyze the probabilistic profile of news (frequency of positive and negative news)".

Kusumura et al. (2003) proposed a text mining agent named NTM Agent to resolve the problems faced during Net auctions. This text mining tool will help in supporting bidders on net auctions by automatically generating a table containing the features of some items for comparison. Grieser et al. (2009) used text mining for more day-to-day activities in e-mail bounce management. They proposed a model which predicts the possibility of deliverability to an e-mail address using decision trees, targeting on improvement of addressability and mentioned turnovers.

Text mining with classification techniques such as Naïve Bayes, Linear Regression, and Rule Induction were used as part of a methodology that used two text mining applications. They were Text Miner Software Kit and Rule Induction Kit for Text (Ticom et al., 2007). This methodology is subsequently integrated with an Expert System. The objective is to reduce the amount of different words to be treated. The aim of stemming is not reach the basic rules of language's linguistics but to improve the performance of the application of text mining.

In addition to above business applications, Text Mining is also applied in the insurance industry. Ellingsworth and Sullivan (2003) proposed a case study of how text mining was used in the field of insurance by Fireman's Fund Insurance Company to understand rising homeowner claims and suspicious auto claims. Arora and Purushotham (2005) also applied text mining in the field of sports where they demonstrated the benefits of combining classification and clustering techniques which will help in grouping articles which are very similar. They also mentioned the use of cluster hypothesis which helps in speeding up the retrieval process.

Chang et al. (2009) applied text mining by applying the data warehouse and data mining technologies to analyze customer behavior in order to form the correct customer profiles and its growth model under Internet and e-commerce environments. Godbole and Roy (2008) explained a text mining solution in the services industry which is mainly used in contact centers. They proposed a methodology using an application named C-Sat. This application primarily takes data from customers interacting with contact centers. The C-Sat analysis is integrated with the business intelligence solution and an interactive

document labeling interface named IBM technology to Automate Customer Satisfaction Analysis (I-TACS).

Yu et al. (2007) developed a framework which analyses news articles and helps to measure the social importance of many events, providing an understanding about current interests. A theoretical framework of a text mining enhanced approach is proposed to accommodate short-term variations caused by special events, such as severe weather conditions. A sentiment analysis approach for extracting sentiments associated with positive or negative polarities from a series of news reports is utilized to illustrate the impact on energy demand from a special event. The magnitudes of the sentiments from the series of news articles are used to compose a time-series pattern to represent the events that are translated into the causes of short-term demand or price variation.

2.4.3. Text Mining in Financial Applications. Extracting and mining relevant information from vast amount of text is a daunting task due to the lack of formal structure in the documents. Mining information from financial data can become even more complex because of the alphanumeric characteristics and other formulae involved as a part of the financial information.

Text mining has been used in various forms of financial data. Various research authors have analyzed the advantage of using text mining for identifying hidden information from financial news, analyzing performance indicators from financial statements, Analyzing stock market trends and other applications. In this section, more details on the application of text mining on financial data and the data analysis on it.

2.4.3.1 Text mining in predicting stock market trends from news. Large amounts of financial news are continuously posted on the web. For people following stock markets or market movements closely there is a need to organize this information and keep track of its development (Ingvaldsen et al., 2006). Online financial news from different sources is widely available on the Internet. In order to decide the best investment strategy, financial analysts have to catch up with the latest information provided by the online news agencies (Cheung et al., 2004). The behavior of the market is dictated by contemporary local and global events, which are not captured in the structured data. Text mining is expected to play an important role in designing strategies for prediction of market behavior since it can be employed successfully to analyze

financial news articles and reports in conjunction with time-series market data (Mahajan et al., 2004). This mined knowledge can assist financial analysts in making investment decisions in the shortest amount of time. They can catch up with or monitor the latest financial activities easily through the system. An information overflow problem can be significantly reduced as well (Cheung et al., 2004).

The financial news is mainly analyzed on predicting the performance of stocks during various time periods. Ingvaldsen et al. (2006) describes a framework that investigates the applicability of text mining operations as a means to manage and extract structures from financial news streams. The framework consists of following modules: Article Fetcher, Part of Speech Tagger, Named Entity Fetcher, Feature Extractor, and Vector Comparator. The Article Fetcher listens periodically for newly distributed news articles by extracting the titles and ingresses of the articles. The pre-trained part of speech taggers are available online in English. The named entity fetcher utilizes static lists of organizations and persons. According to Ingvaldsen et al. (2006) this framework shows how elements from information retrieval, information extraction and natural language processing can be applied to extract named entities from financial news streams and represent these as temporal and spatial vectors.

Fawcett and Provost (1999) proposed a framework which would issue an alarm on a specific company when a stock shifts at least 10%. They define activity monitoring as discovering market changes in a time series. A similar system developed by Lavrenko and Allan (2000) also monitors unusual trends in the time series with alarms in the form of recommended stories. They make a news recommendation by estimating the importance of a story to the stock market. A language model is learned based on trend types of the financial market. Their focus is on the influence of the news story on the market trend rather than the mining of related financial activities.

Cheung et al. (2004) developed a financial knowledge management system, known as FAM (Financial Activity Mining), which is able to digest online news and conduct financial activity mining. The online news can come from various news agencies from the Web or subscribed newsfeed services. These information sources provide real-time international, political and economic news, citations from worldwide bankers and politicians as well as recommendations from different financial analysts. FAM can fetch

the news articles from the above sources automatically. The whole mining process is conducted via an unsupervised learning algorithm. As a result, financial analysts can digest and monitor the latest financial activities produced by the mining results. The key difference between the FAM by Cheung et al. (2004) and other systems developed by Fawcett and Provost (1999) and Lavrenko and Allan (2000) is that FAM is able to present news on specific companies or activities. The system is particularly helpful for tracking the stock performance of a targeted company or an event, with all the related news collected in the form of clusters. Clear presentation of relationships between related activities provides a convenient environment for users to monitor the financial market.

Kaya and Karsligil (2010) also developed a model where they predict stock prices using financial news articles. A prediction model which finds and analyzes the correlation between contents of news articles and stock prices and then makes predictions for future prices was developed. The financial news articles published in the previous year are retrieved, and the prices for the same period are taken. All articles are labeled positive or negative according to their effects on stock price, so price changes are used to label the articles. While analyzing textual data, word couples consisting of a noun and a verb as are used instead of single words. Afterwards, a support vector machine classifier is trained with labeled train articles. Finally, classes of test articles are predicted using the model results from the train phase.

There are other substantial works completed on the prediction of stock prices. These works are basically text categorization systems targeted to predict stock price movement by classifying financial news articles as positive or negative. Since the problem is converted to a text categorization problem, several feature selection and classification methods are used in these works. In the frameworks of Mittermayer (2004) and Wuthrich et al. (1999), term frequency – inverse document frequency technique is used as a feature selection method. Falinouss (2007) use chi-square statistics feature selection method. Support vector machines, k nearest neighbor and naive bayes are most widely used methods for classification. In the classification phase several researchers (Koppel and Shtrimberg, 2004; Fung et al., 2005; Mittermayer and Falinouss, 2007), support vector machines method. While some researchers (Gidofalvi , 2001 ; Kroha and Baeza-Yates , 2004) use naive bayes. Other researches (Wuthrich et al., 1998; Y.-C. Wu,

2007) use k-nearest neighbor method for classification. The accuracy rates of these works are mostly below 60%. These relatively low success rates are caused by the nature of stock price movements, which are a result of decisions of investors, since it is hard to predict human behavior.

2.4.3.2 Text mining for predicting impact on stock market. Financial data analysis has traditionally dealt with large volumes of structured data reflecting economic performance. However the behavior of the market is dictated by contemporary local and global events, which are not captured in the structured data. Text mining is expected to play an important role in designing strategies for prediction of market behavior, since it can be employed successfully to analyze financial news articles and reports in conjunction with time-series market data. Text-mining can be employed to extract information about related contemporary events from financial news reports, and also explain the causes for poor performance or a sudden upturn in the market. In the recent past, the use of text-mining has been reported for predicting individual company's' stock prices. Information extracted from various sources is used to design strategies to help potential investors. However these systems did not attempt to identify the factors that affect the market as a whole.

Mahajan et al. (2008) proposed a text-mining system that analyzes market news about the Indian stock market and correlates it with the actual stock market behavior. The aim is to identify the major events that have impacts on the stock market, and characterize them in order to design strategies for predicting the market. Kloptchenko et al. (2002) presented a mining technique that analyzed quantitative and qualitative data from annual financial reports in order to see if the textual part of the report contains some indication about future financial performance. Seo et al. (2004) explained a multi-agent Portfolio management system that evaluates the risks associated with the individual companies in a portfolio. Wutrich et al. (1998) predicts the movement of five major global stock indices based on current news. Ingvaldsen et al. (2006) addressed the problem of extracting, analyzing and synthesizing valuable information from continuous text streams covering financial information. Lavrenko et al. (2000) presented an approach to identify news stories that influence the behavior of financial markets, by correlating contents of news articles to trends in financial market. Mittermayer(2004) proposed the

NewsCats system to categorize financial news articles into pre-defined categories and then derive appropriate trading strategies based on these categories. Thomas and Sycara (2004) described a system that can learn profitable trading rules using data from stock chat boards.

2.4.4. Text Mining on Financial Reports. A huge amount of electronic information concerning company's financial performance is available in organizational databases and on the Internet today. Numeric financial information is important for many stakeholders and is extensively analyzed with advanced computational methods. Textual financial information in form of reports and news contain not only the factual description of events, but also explain why they have happened (Kloptchenko, 2004). Exploiting finance and business related textual information in addition to numeric financial information should increase the quality of decision-making. Constantly updated text collections have grown so large that there is not enough time to read and analyze them manually. Additionally, the ambiguous structure of texts makes their analysis rather complicated. Researchers are searching for elegant and computationally feasible tools that would be able to handle sophisticated text-related tasks without thorough linguistic preprogramming (Kloptchenko, 2004).

The message, stylistic focus, language and readability of financial reports are good indications about the perspectives and developments of any company. These indications can guide companies' decision makers to more efficient acts on the market. Although, financial experts and experienced readers can detect those indications and make more precise financial decisions, the manual analysis of textual reports requires a lot of time, and time is a costly asset in a financial community. Text mining methods aim to offer an automatic way for analyzing and discovering previously unknown patterns in text (Hearst, 1999). Therefore, less expensive computer-based solutions for mining financial texts for hidden indications of companies' perspectives are needed.

Annual reports, while being important documents to stockholders and financial communities are controversial. They generate disagreement regarding audience, objectives and credibility (Thomas, 1997). As a genre, annual reports resemble quarterly reports closely. The same writers produce quarterly and annual reports for the same readers within the same community. The reports have a similar structure, conventions,

basic functions and communicative purposes but the time spans are different. The study of the linguistic contents of quarterly reports has nevertheless been overlooked in favor of the study of the language of annual reports (Kloptchenko, 2004). In the short-term perspective quarterly reports are informative and important means for companies in appraising past performance and projecting future opportunities to the readers, who primarily consist of investors and analysts. Typically the beginning of every report, known as the manager's/president letter/message to stockholders, contains management's strategy, summary of the financial performance for the year and an attempt to put in perspective the success or failure of the various initiatives of the company (Thomas, 1997).

Various researchers have analyzed the annual reports and the advantage of using text mining. Thomas (1997) concentrated on transitivity, thematic structure, context, cohesion and condensation in the language used in the reports. The researcher studied the annual reports of a machine tool manufacturer during a period, which began with prosperity and ended with severe losses. During the time frame of the analysis, the structure of the language used in the reports had changed. According to Thomas' study, an increase in the use of passive constructions can be seen as the profits decrease. There is also an increase in verbs that present the actor (i.e. the company) as "being" rather than as "doing". This indicates that management is trying to present itself as a victim of unfortunate circumstances. This creates an impression of objectivity for the reader, as if the management was presenting plain facts on recent events. On the other hand, when the company was making more profit, it presented itself as aggressive and forward moving through the use of active voice and verbs with both an actor and a goal. A close look at the language structure in the letters to stockholders made by Thomas (1997) showed that the structure of the financial reports might reveal some things that the company may not wish to announce directly to its outside audience.

Kendal (1993) introduced the concept of drama when she noticed a similar opposition between the actions of the company and circumstances created by nonhuman agents. Kendall has classified the words and phrases describing actors and objects in the drama into two groups, God terms and Devil terms. Some examples of god terms are growth, increased sales and competitive position. These words represent concepts that are

unquestionably good in the eyes of the company. Devil terms, on the other hand, are terms like losses, decline in sales and regulations.

Other studies have been made with a focus on the relationship between the readability of the annual reports and the financial performance of a company (Subramanian et al., 1993). The annual reports of the companies that performed well were easier to read than those that originated from companies that did not perform well. Studies have also shown that writers of annual reports see the message they put in the report as their personal representation (Winsor, 1993). The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. Thus, the communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens (Kohut & Segars, 1992). After performing computer-aided content analysis of more than four hundred president's letter to shareholders and examining empirical linkages between themes in annual reports and companies' performances, Osborn et al. (2001) conclude that the text in annual reports reflects the strategic thinking of the management of a company.

Attempts to semi-automatically analyze a company's performance by examining quantitative and qualitative parts from annual reports have been done by Back et al. (2001) and Kloptchenko et al. (2002). Back et al. (2001) indicated that there are differences in qualitative and quantitative data clustering results due to a slight tendency to exaggerate the performance in the text. Kloptchenko et al. (2002) attempted to explain this tendency using quantitative analysis by means of self-organizing maps for financial ratio clustering, and qualitative analysis by means of the prototype matching for quarterly report text clustering. In both studies the researchers noticed that the combination of two mining techniques for two different types of data describing the same phenomena could bring additional knowledge to a decision maker. While annual/quarterly reports explicitly state information about a company's past performance, they also contain some indications of future performance, i.e. the tables with financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text may tell what a company intends to do. The study has shown that the sophisticated semi-

automatic analysis of the style and content of the financial reports help to reveal insiders' moods and anticipations about the future performance of their company.

2.5. CORPORATE SPIN – OFF'S

Corporate spin-offs play a key role for industrial dynamics, innovations and national competitiveness in developed countries. Spin-offs are a major determinant of the formation of new firms in high and low technology industries (Pickerodt & Stieglitz, 2004). Since Hite et al., (1983), the empirical literature has repeatedly documented that parent company stockholders gain during spin-off announcement period (for example, Allen et al., 1985; Krishnaswami and Subramaniam, 1999).

Recent empirical evidence goes beyond showing the positive announcement effects of spin-offs on stock price. Cusatis et al. (1993) show that, in addition to the positive abnormal stock returns for parent firms on the announcement date, both spin-offs and their parents experience significantly positive abnormal returns for up to three years beyond the spin-offs' announcement date. Further, both spin-offs and their parents experience significantly more takeovers than do control groups of similar firms. Cusatis et al. (1993) also show that spin-off/parent combinations not reporting takeover activity within three years do not have positive long-term abnormal stock returns.

Many explanations are provided in literature to explain why shareholders gain during spinoffs. Miles and Rosenfeld (1983) and Daley et al. (1997) verify a correlation between announcement return and investment policy of the parent company, interpreting the spin-off as the chance to eliminate "negative synergies" generated by a management unable to replicate the role of financial markets. Burch and Nanda (2002) showed that an increase in corporate focus partly explains the increase in the value of the firm. Aron (1991) argued that spin-offs benefit the firm since, after the spin-off; the equity values of the securities traded provide a much cleaner signal of managerial productivity than when the two divisions were part of a combined firm. The argument is that this enables the firm to provide better incentives for firm management based on the stock price of the individual firms. However, this argument requires the somewhat strong assumption that

equivalent incentive contracts cannot be written based on the profitability of the individual divisions when they are part of a combined firm.

Habib et al. (1997) argued that spin-offs improve the quality of the information managers and uninformed investors can infer from the prices of the firm's traded securities, therefore leading to an increase in the expected price of the firm's equity. Nanda and Narayanan (1999) suggested that the firm may be undervalued if the market cannot observe the cash flows of each individual division in that firm. Therefore, the firm that needs external financing could resort to divestures such as spin-offs in order to raise capital at a fair market price after the divesture. Krishnaswami and Subramaniam (1999) tested the hypothesis that such positive market reactions to spin-offs are due to a reduction in the information asymmetry existing in the market for the equity of the parent firm.

Chemmanura and Yanb (2003) develops a new rationale for the performance and value improvements arising from spin-offs, which is consistent with this recent (as well as earlier) empirical evidence. The authors developed a theoretical analysis which demonstrates how spin-offs can increase the probability of a takeover by the right kind of (value-improving) management team. The authors showed how such spin-offs can enhance the level of firm performance even in the absence of such a value-improving takeover by serving to discipline firm management. Finally, the analysis demonstrates that, while a spin-off will lead to positive abnormal stock-price returns on the announcement day, it will also lead to increases in operating performance and to abnormal stock price performance (on average) in the period following the spin-off for certain categories of firms.

A different body of literature focused on the effects of investment decisions on stock prices. A number of studies explored the effects of investment choices on stock returns (McConnell & Muscarella, 1985; Fazzari et al., 1988; Morck et al., 1990). In these studies, correlation between stock prices and investment policy has been documented in two ways: on the one hand, firms tend to invest more following increases in their stock prices (Fazzari et al., 1988; Morck et al., 1990), on the other hand, it is also the case that stock prices tend to respond favorably to announcements of major capital investments (McConnell & Muscarella, 1985). Furthermore, a significant positive

relationship between the magnitude of the stock market reaction to capital investment announcements and the level of new investment has been documented (Blose & Shieh, 1997).

In contrast with the main findings of the above studies, Titman et al. (2003) registered an inverse relationship between increase in capital investments and stock returns. Adopting Jensen's approach, the authors accept that managers can be "empire builders", and invest for their own benefits rather than for the benefits of the firm's shareholders (Jensen, 1986), with negative consequences on stock prices. The authors show that firms that increase capital investments the most tend to underperform their benchmarks over the following five years.

Investment policies seem to be an important factor in determining the stock performance. With a more general perspective, investment choices are documented to play an important role in explaining returns of all the listed companies. Titman et al. (2003) provide a significant contribution to the debate on capital investments and stock returns. Differently from what previously documented (McConnell and Muscarella, 1985; Blose and Shieh, 1997), the authors verify the existence of a negative relation between increase in capital investments and subsequent excess returns, measured through the Fama-French-Cahart α (Cahart, 1997). This negative relationship is shown to be stronger for firms with greater investment discretion (firms with higher cash flows and lower leverage ratios).

Rovetta (2005) analyzed excess returns related to corporate spin-off with respect to changes in investment policies of the spun off companies (subsidiaries). Spun off companies gain substantial excess returns on the three years following the spin-off and, at the same time, they show a general decrease in the level of capital investments. Moreover, following the spin-off, a substantial increase in investment efficiency can be documented for the well performing companies. Investment in low-Q subsidiaries strongly decreases and investment in high-Q tends to increase or remain substantially unchanged. Results provide evidence on the existence of a direct relationship between the size of the change in the level of investment, the Tobin's Q, and the dimension of the excess return.

Linking together the financial literature on the changes in investment policies after corporate spinoffs and on the effects of investment decisions on stock prices (Titman et al., 2003), this article provides evidence on the relationship between the dimension of the excess returns subsequent to the spinoffs, measured through the Fama and French alpha (Fama and French, 1993), and the changes in investment behavior in the spun-off companies.

Corporate spin-offs could relax financial constraints at the origin of investment inefficiency in two different ways. Spin-offs help divisions to adopt specific financial policies that allow them to define their capital structure in a more efficient way, using an amount of debt that fits the segment growth opportunities (Ofek & Stulz, 1996). On the same lines, Gertner et al. (2002), Ahn and Denis (2003), and Dittmar and Shivdasani (2003) explore corporate spin-offs with the objective to verify a relationship between market values and investment policies. Gertner et al. (2002) show that changes in the investment behavior of the spun off companies explain the gains on the financial market. Ahn and Denis (2003) provide evidence that the reduction in the diversification discount is positively related to changes in measures of investment efficiency for spinoffs. Dittmar and Shivdasani (2000) examine the effects of divestitures of specific business segments on the investment policy of the parent company. Over a sample of 278 divestitures (15 of which are pure spin-off) completed by 235 firms from 1983 to 1994, the authors verify a correlation between the decline in the diversification discount around the divestiture and the change in the investment policy of the firms' remaining segments. The level of investment in segments that under-invest relative to single segment firms increases after the divestiture, while the level of investment in segments that overinvest declines (Rovetta, 2005).

Researchers (Desai and Jain, 1999 ; Daley et al., 1997), showed that both the market reaction to spin-off announcements and the long-term abnormal returns and operating performance are significantly greater in unrelated spin-offs (where the spun-off subsidiary operates in an industry unrelated to the parent firm) than in related spin-offs. Other studies show that the magnitude of the market reaction to spin-off announcements is increasing in the size of the spun-off division as a fraction of the combined firm existing prior to the spin-off.

3. RESEARCH OBJECTIVES AND METHODOLOGY

In this study, text mining is applied to the 10-12b filings done by the companies during corporate spin-off. The main purposes are

(1) To investigate potential and/or major concerns found from these financial statements filed for corporate spin-off, and;

(2) To identify appropriate methods in text mining can be used to reveal these major concerns.

A spin- off is used to separate two businesses that have become incompatible or whose collective business success has become subdued by the common ownership. This will help a corporation in getting the investors and lenders provide capital to one but not all operations. Also since a spin-off would result in two separate entities, compensation in the form of stock ownership could be given to employees in the specific business for which they are responsible. Spinning off a separate business can establish a separate identity and operating history that makes both the spun-off company and the distributing company more readily marketable to a buyer in the future.

3.1. DATA

In this research, the financial documents (10-12b filings) from corporate spin-off are collected. The analysis is focused on the risk factors part of the financial documents, in particular, the risk factors for the company or the risk factors for the business environment are used.

10-12b is a filing with Securities and Exchange Commission (SEC) which is required when a public company issues a new class of stock through spin-off. SEC Form 10-12b contains information about the original shared issued, the new shares affected and the information about how and on which exchange the new shares will trade.

For this study, the 10-12b filings done by 34 companies were taken for analysis. The number of employees in these companies varies between 500 and 50,000. Table 3.1 shows the list of 34 companies used for analysis.

Table 3.1 Companies Data for Analysis

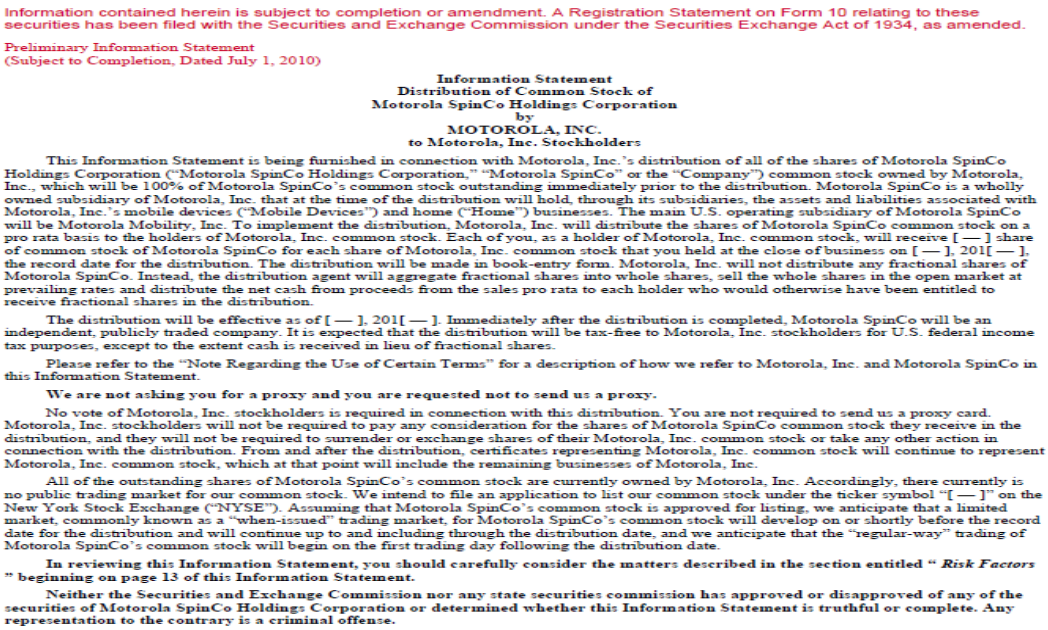
ID	Company	Year of Filing	SIC Code	Industry
1	A.H.BELO CORP	2007	2711	Newspapers: Publishing, or Publishing and Printing
2	ACUITY BRANDS INC	2001	3640	Electric Lighting And Wiring Equipment
3	ALBERTOCULVER CO	2006	5990	Retail Stores, Not Elsewhere Classified
4	ALLEGIANCE CORP	1996	8093	Specialty Outpatient Facilities, Not Elsewhere Classified
5	ALTISOURCE PORTFOLIO SOLU	2009	7380	Miscellaneous Business Services
6	AMC NETWORKS INC	2011		
7	AMETEK_INC	1997	3621	Motors and Generators
8	AOL_INC	2009	7374	Computer Processing and Data Preparation and Processing Services
9	ARCH_CHEMICALS_INC	1998	2800	Chemicals & Allied Products
10	BABCOCK_&_WILCOX_CO	2010	3510	Engines And Turbines
11	BRINK'S_HOME_SECURITY_HOL	2008	7380	Miscellaneous Business Services
12	CERIDIAN_CORP_DE	2000	8742	Management Consulting Services
13	CIMAREX_ENERGY_CO	2002	1311	Crude Petroleum and Natural Gas
14	CIRCOR_INTERNATIONAL_INC	1999	3490	Miscellaneous Fabricated Metal Products
15	CIT_GROUP_INC	2002	6172	Finance Lessors
16	CONSOLIDATED_FREIGHTWAYS	1996	4213	Trucking, Except Local
17	DELTA_APPAREL,_INC_	1999	5130	Apparel, Piece Goods, And Notions
18	DELTIC_TIMBER_CORP	1996	2421	Sawmills and Planning Mills, General
19	DUN_&_BRADSTREET_CORPNW	2000	7320	Consumer Credit Reporting Agencies, Mercantile
20	HANESBRANDS_INC.	2006	5600	Retail-Apparel & Accessory Stores
21	HOSPIRA_INC	2003	2834	Pharmaceutical Preparations
22	INTERMEC,_INC	1997	3577	Computer Peripheral Equipment, Not Elsewhere Classified

Table 3.1 Companies Data for Analysis (Cont'd)

ID	Company	Year of Filing	SIC Code	Industry
23	JOHN_BEAN_TECHNOLOGIES_CO	2008	3550	Special Industry Machinery, Except Metalworking
24	LUCENT_TECHNOLOGIES_INC	1996	3661	Telephone and Telegraph Apparatus
25	MARINE_PRODUCTS_CORP	2000	3730	Ship And Boat Building And Repairing
26	MARRIOTT_INTERNATIONAL_IN	1998	7011	Hotels and Motels
27	NCR_CORP	1996	3578	Calculating and Accounting Machines, Except Electronic Computers
28	NEENAH_PAPER_INC	2004	2621	Paper Mills
29	NORTEK_INC	2010	3634	Electric House wares and Fans
30	PHILIP_MORRIS_INTERNATION	2007	2111	Cigarettes
31	TELEDYNE_TECHNOLOGIES_INC	1999	8711	Engineering Services
32	WYNDHAM_WORLDWIDE_CORP	2006	7011	Hotels and Motels
33	MEDCO_HEALTH_SOLUTIONS_IN	2003	5912	Drug Stores and Proprietary Stores
34	MOTOROLA_SPIN OF	2010	3663	Radio and Television Broadcasting and Communications Equipment

Also in Table 3.1, additional information such as the year the companies had performed the spin – off along SIC code of the companies is shown. SIC code is a coding system developed by United States government for classifying industries and it is a four digit coding system. It is a number used to specify what industry a particular company belongs to. Some companies append two or four additional digits to the standard SIC code to form a six or eight digit SIC code, allowing more specific business classification.

From Table 3.1, A.H Belo Corp Company has a SIC code of 2711. This means that that company belongs to the Newspapers (Publishing, or Publishing and Printing) Industry. Each of the thirty – four companies used for analysis have been classified based on their SIC code, type of industry and also the year the filing was done by the companies. Figure 3.1 shows a screenshot of 10-12b filing filed by Motorola Inc



Information contained herein is subject to completion or amendment. A Registration Statement on Form 10 relating to these securities has been filed with the Securities and Exchange Commission under the Securities Exchange Act of 1934, as amended.

Preliminary Information Statement
(Subject to Completion, Dated July 1, 2010)

Information Statement
Distribution of Common Stock of
Motorola SpinCo Holdings Corporation
by
MOTOROLA, INC.
to Motorola, Inc. Stockholders

This Information Statement is being furnished in connection with Motorola, Inc.'s distribution of all of the shares of Motorola SpinCo Holdings Corporation ("Motorola SpinCo Holdings Corporation," "Motorola SpinCo" or the "Company") common stock owned by Motorola, Inc., which will be 100% of Motorola SpinCo's common stock outstanding immediately prior to the distribution. Motorola SpinCo is a wholly owned subsidiary of Motorola, Inc. that at the time of the distribution will hold, through its subsidiaries, the assets and liabilities associated with Motorola, Inc.'s mobile devices ("Mobile Devices") and home ("Home") businesses. The main U.S. operating subsidiary of Motorola SpinCo will be Motorola Mobility, Inc. To implement the distribution, Motorola, Inc. will distribute the shares of Motorola SpinCo common stock on a pro rata basis to the holders of Motorola, Inc. common stock. Each of you, as a holder of Motorola, Inc. common stock, will receive [—] share of common stock of Motorola SpinCo for each share of Motorola, Inc. common stock that you held at the close of business on [—], 201[—], the record date for the distribution. The distribution will be made in book-entry form. Motorola, Inc. will not distribute any fractional shares of Motorola SpinCo. Instead, the distribution agent will aggregate fractional shares into whole shares, sell the whole shares in the open market at prevailing rates and distribute the net cash from proceeds from the sales pro rata to each holder who would otherwise have been entitled to receive fractional shares in the distribution.

The distribution will be effective as of [—], 201[—]. Immediately after the distribution is completed, Motorola SpinCo will be an independent, publicly traded company. It is expected that the distribution will be tax-free to Motorola, Inc. stockholders for U.S. federal income tax purposes, except to the extent cash is received in lieu of fractional shares.

Please refer to the "Note Regarding the Use of Certain Terms" for a description of how we refer to Motorola, Inc. and Motorola SpinCo in this Information Statement.

We are not asking you for a proxy and you are requested not to send us a proxy.

No vote of Motorola, Inc. stockholders is required in connection with this distribution. You are not required to send us a proxy card. Motorola, Inc. stockholders will not be required to pay any consideration for the shares of Motorola SpinCo common stock they receive in the distribution, and they will not be required to surrender or exchange shares of their Motorola, Inc. common stock or take any other action in connection with the distribution. From and after the distribution, certificates representing Motorola, Inc. common stock will continue to represent Motorola, Inc. common stock, which at that point will include the remaining businesses of Motorola, Inc.

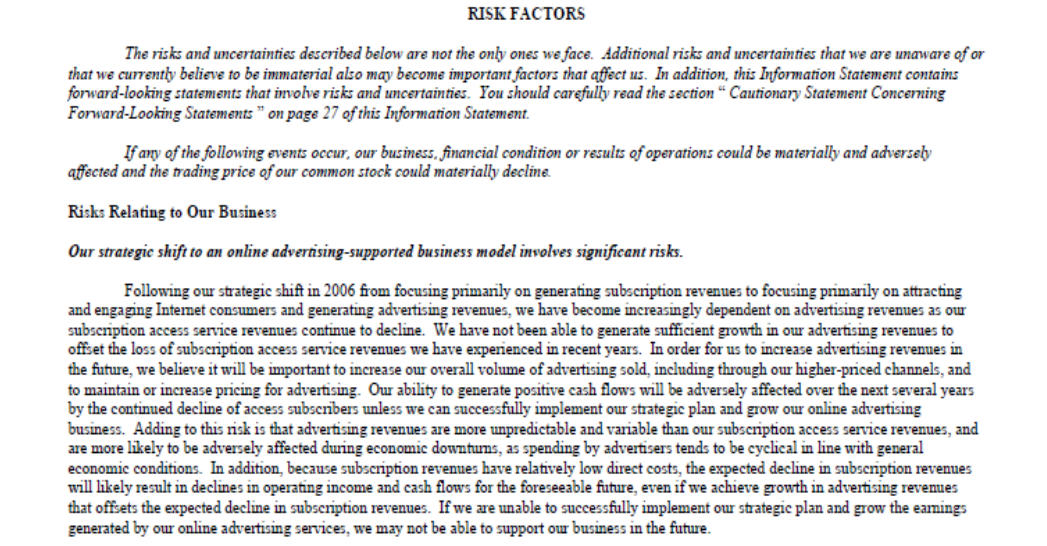
All of the outstanding shares of Motorola SpinCo's common stock are currently owned by Motorola, Inc. Accordingly, there currently is no public trading market for our common stock. We intend to file an application to list our common stock under the ticker symbol "[—]" on the New York Stock Exchange ("NYSE"). Assuming that Motorola SpinCo's common stock is approved for listing, we anticipate that a limited market, commonly known as a "when-issued" trading market, for Motorola SpinCo's common stock will develop on or shortly before the record date for the distribution and will continue up to and including through the distribution date, and we anticipate that the "regular-way" trading of Motorola SpinCo's common stock will begin on the first trading day following the distribution date.

In reviewing this Information Statement, you should carefully consider the matters described in the section entitled " *Risk Factors* " beginning on page 13 of this Information Statement.

Neither the Securities and Exchange Commission nor any state securities commission has approved or disapproved of any of the securities of Motorola SpinCo Holdings Corporation or determined whether this Information Statement is truthful or complete. Any representation to the contrary is a criminal offense.

Figure 3.1 Sample 10-12b Filing

Figure 3.2 shows the screenshot of the Risk Factors mentioned in the 10-12b filing of AOL, Inc.



RISK FACTORS

The risks and uncertainties described below are not the only ones we face. Additional risks and uncertainties that we are unaware of or that we currently believe to be immaterial also may become important factors that affect us. In addition, this Information Statement contains forward-looking statements that involve risks and uncertainties. You should carefully read the section " Cautionary Statement Concerning Forward-Looking Statements " on page 27 of this Information Statement.

If any of the following events occur, our business, financial condition or results of operations could be materially and adversely affected and the trading price of our common stock could materially decline.

Risks Relating to Our Business

Our strategic shift to an online advertising-supported business model involves significant risks.

Following our strategic shift in 2006 from focusing primarily on generating subscription revenues to focusing primarily on attracting and engaging Internet consumers and generating advertising revenues, we have become increasingly dependent on advertising revenues as our subscription access service revenues continue to decline. We have not been able to generate sufficient growth in our advertising revenues to offset the loss of subscription access service revenues we have experienced in recent years. In order for us to increase advertising revenues in the future, we believe it will be important to increase our overall volume of advertising sold, including through our higher-priced channels, and to maintain or increase pricing for advertising. Our ability to generate positive cash flows will be adversely affected over the next several years by the continued decline of access subscribers unless we can successfully implement our strategic plan and grow our online advertising business. Adding to this risk is that advertising revenues are more unpredictable and variable than our subscription access service revenues, and are more likely to be adversely affected during economic downturns, as spending by advertisers tends to be cyclical in line with general economic conditions. In addition, because subscription revenues have relatively low direct costs, the expected decline in subscription revenues will likely result in declines in operating income and cash flows for the foreseeable future, even if we achieve growth in advertising revenues that offsets the expected decline in subscription revenues. If we are unable to successfully implement our strategic plan and grow the earnings generated by our online advertising services, we may not be able to support our business in the future.

Figure 3.2 Sample Risk Factors in a 10-12b Filing

3.2. IMPLEMENTATION PLATFORM

This study applies SAS Enterprise Miner as the tool to analyze 10-12b filings done by the companies during corporate spin-off. SAS Text Miner 4.1 is a plug-in for the SAS Enterprise Miner 6.1 environment. SAS Enterprise Miner provides various data mining tools that facilitate the prediction aspect of text mining. Text Miner encompasses the parsing and exploration aspects of text mining and prepares data for predictive mining and further exploration. Also the Text Miner enables to choose from a variety of parsing options to parse documents for detailed information about the terms, phrases, and other entities in the collection. Documents can be clustered into meaningful groups and report concepts that can be discovered in the clusters.

The following processing steps were conducted in this study with the help of the software.

File preprocessing: This step creates a single data set from the document collection.

Text parsing: This step decomposes textual data and generates a quantitative representation suitable for data mining purposes.

Transformation (dimension reduction): During this step, transformation of the quantitative representation into a compact and informative format is performed.

Document analysis: This step performs clustering analysis on the document collection.

The outcome of clustering is presented by the descriptive terms of each cluster. It is necessary to try a different combination of frequency weights and term weights options before a satisfied result is available. Some terms may be eliminated from the document collection. Different number of clusters will be examined. Finally, Human expert opinion is used to verify the satisfactory of the text mining data.

3.3. TEXT MINING PROCESS

This study focuses on applying Frequency weights and term weights methods, such as Entropy, GF-IDF, IDF, Normal and None, on the 10-12b filings done by companies. The purpose is to find out which method(s) present the results would meet human expert's expectation.

For this study, analysis was performed on the “Risk factors” category mentioned in each filing by the Spin-off companies. Figure 3.3 shows the input data fed into the enterprise miner

ID	TEXT	TARGET
1	If A. H. Belo is unable to respond to evolving industry trends and changes in technology	1
2	equipment business's sales are made to customers in the new construction and renovation industries.	1
3	could have an adverse impact on its business, financial condition and results of operations. New Alberto-	1
4	meaning of the Private Securities Litigation Reform Act of 1995. Such forward looking statements	1
5	adversely affect us. Our business is dependent on our ability to grow which is affected by our ability to	1
6	distributors and our viewers. Our business depends in part upon viewer preferences and audience	1
7	forth below, as well as other inforation contained in this Information Statement. For purposes of this	1
8	Following our strategic shift in 2006 from focusing primarily on generating subscription revenues to	1
9	will own and operate the Specialty Chemical Businesses. These businesses have no operating history as	1
10	industries, with demand for our products and services depending on capital expenditures in these	1
11	our use of the Brink’s brand following the spin-off. As a business unit of Brink’s, we have marketed our	1
12	payroll business, and cannot assure you that its efforts and the amount it invests in these plans will	1
13	information statement. Some of the following risks relate principally to the spin-off and the merger.	1
14	which may affect CIRCOR's financial condition or results of operations and/or the value of its common	1
15	business, financial condition and results of operations may be affected by various economic factors,	1
16	certain risk factors, including those described below and elsewhere in this Information Statement,	1
17	The following discussion contains various "forward-looking statements". Please refer to "Forward-	1
18	affected by the cyclical nature of the forest products industry. Prices and demand for logs and	1
19	District Court for the Southern District of New York, naming as defendants the corporation then known	1
20	placing us at a product cost disadvantage to our competitors who have a higher percentage of their	1
21	customers may not buy our products and our revenue and profitability may decline. Demand for our	1

Figure 3.3 Input Data Fed into Enterprise Miner

The risk factors from each spin-off filing was entered into one cell of an excel sheet. Each cell was assigned an ID to keep track of the identification of the documents. A target value was induced to be default “1”, as this study focuses only on one category which is the “Risk Factors”. So, for the analysis, the input file contained 34 lines with each line representing the risk factors of the companies.

After the input data file was prepared, the model for text mining was created as a diagram in SAS Enterprise Miner, displayed in Figure 3.2. The left node is titled Input Data node into which the data file was imported, and the right node titled Text Miner, in which the text mining process would be performed to explore information in the document collection. Both nodes were connected via a line. The direction of the arrow represents the data flow. The input data was fed into the text mining process. SAS Enterprise Miner automatically processes the files based on parameter settings. Figure 3.4 shows the text mining model used for the initial data analysis.

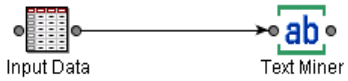


Figure 3.4 Text Mining Model

To improve performance, the dimensional reduction technique was applied; thus, the “Compute SVD” was set to “Yes”.

Singular Value Decomposition (SVD) is a popular approach which was also used in this research. SVD resolution determines the number of SVD (dimensions) extracted. For a “High” SVD dimension value more information is kept but it requires longer computation time. For a “Low” SVD dimension, less information is kept but it takes less computation time. Since the data size used in this study is small, the computation time is not an issue. Hence SVD dimension value of “High” was used for analysis.

The higher resolution yields more SVD dimensions, which summarizes the data set more efficiently although it requires more computing resources. The number of SVD dimensions should be large enough to prevent loss of concepts small enough to limit noise. Dumais (1991) performed information retrieval and found that performance increased over the first 100 dimensions, hitting the maximum, and then falling off slowly. Thus, 100 seemed to be a good start for the maximum number of SVD dimensions. Table

3.2 shows the key parameter settings used for the analysis. None, Log and Binary frequency weight was used for analysis and all the five term weights – Entropy, GF-IDF, IDF, Normal and None , was used for analysis.

Table 3.2 Parameter Settings for Term-Document Frequency Matrix Conversion Stage

Property	Value
Compute SVD	Yes
SVD Resolution	High
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	None
Term Weight	Entropy

Core Mining Processing**.** Clustering technique was applied to the risk factors taken from the 10-12b filing into clusters. Table 3.3 shows some key parameter settings for this clustering process. The fixed set of clustering was set. The exact number was set to 5 since the document collection was small and 5 clusters should be sufficient to cover all ideas. Expectation-maximization (EM) clustering technique was being used. The number of descriptive terms was set to 10. This number is reasonable for the size of data as it would help in identifying a cluster more easily. Clustering worked on the term-frequency matrix after dimensional reduction (i.e., SVD) had been applied.

Table 3.3 Parameter Settings for Clustering of Core Mining Processing

Property	Value
Automatically Cluster	No
Exact or Maximum Number	Exact
Number of Clusters	5
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Descriptive Terms	10
What to Cluster	SVD Dimensions

4. ANALYSIS AND RESULTS

4.1. FIRST ANALYSIS

The analysis was started without any pre-defined parameter setting. Table 4.1 shows the parameter setting performed for the first analysis.

Table 4.1 First Analysis Parameter Setting

Parameter	Value
SVD Dimension	High
Number of Cluster	Maximum 7
Descriptive Terms	5

The parameters started with a minimum descriptive term of 5 and the system was allowed to decide on the number of clusters, by keeping the setting as maximum.

For this analysis, the data set was analyzed with all the frequency terms and all the term weights. For example, an analysis was done first with the frequency weight of None, Log and Binary with the term weight Entropy. It was followed by the combination of None, log and Binary frequency weight with GF-IDF. This step was repeated till all the five term weights were covered with all the three frequency weights. Table 4.2 shows the analysis done with the combination of None/log/Binary with the term weight Entropy.

This analysis did not provide any reasonable outcome out of the clusters. The system generated 2 clusters due to this setting. Usage of different term frequencies like None, Binary and Log with the term weights didn't make any difference at all. For example, in Table 4.2, the descriptive terms generated out of None/Entropy is exactly the same as the descriptive terms generated for Log/Entropy and Binary/Entropy. Different frequency terms did not provide different descriptive terms for the clusters generated. Hence a decision was made to focus on the default frequency term which is None.

Table 4.2 Descriptive Terms from None/Log/Binary with Entropy

Frequency Weight	None			
Term weight	Entropy			
ID	Descriptive Terms	Freq	Percentage	RMS Std
1	may, + risk, + factor, new, + affect	24	0.7058824	0.181636
2	significant, economic, portion, historical, including	10	0.2941177	0.176612
Frequency Weight	Binary			
Term weight	Entropy			
ID	Descriptive Terms	Freq	Percentage	RMS Std
1	may, + risk, + factor, new, + affect	24	0.7058824	0.181636
2	significant, economic, portion, historical, including	10	0.2941177	0.176612
Frequency Weight	Log			
Term weight	Entropy			
ID	Descriptive Terms	Freq	Percentage	RMS Std
1	may, + risk, + factor, new, + affect	24	0.7058824	0.181636
2	significant, economic, portion, historical, including	10	0.2941177	0.176612

4.2. SECOND ANALYSIS

For the second analysis the parameter settings were slightly modified. Table 4.3 shows the parameter settings used for the second analysis.

Table 4.3 Second Analysis Parameter Setting

Parameter	Value
SVD Resolution	High
Frequency Weight	None
Number of clusters	5
Number of Terms	10

During the second analysis, fixed number of clusters and an increase in number of terms was performed. Increasing the number of descriptive terms to 10 will not change the outcome but will provide more information about the clusters. The "None" frequency weight was chosen to be the only frequency weight since the first analysis did not show any difference among various frequency weights. In this analysis, analysis was done using only one frequency weight on all the five term weights. Table 4.4 shows the cluster output using None/Entropy.

Table 4.4 Descriptive Terms from None with Entropy (Second Analysis)

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	information, + statement, statement, information, following, carefully, below, consider, forward-looking, in addition to	11	0.323529412	0.175388
2	economic, portion, significant, including, + operation, + business, + condition, + affect, + result	6	0.176470588	0.165882
3	+ year, + loss, december, historical, + result	4	0.117647059	0.169672
4	competition, compete, attract, effectively, intense, + industry, adversely, ability, + affect, + business	6	0.176470588	0.173274
5	history, operating, products, continue, company, + business, + affect, + operation	7	0.205882353	0.17859

According to human expert, this analysis resulted in clusters which were too noisy. Lots of "non – important" and "meaningless" terms such as information, statement, including, year, month (like december), historical/history, operation, in addition to, and U.S were found in the cluster. This resulted in the need to rerun the analysis by removing the unwanted terms. The number of clusters and the number of terms were unchanged.

4.3. THIRD ANALYSIS

For the third analysis, the parameters were kept the same. The only difference when compared with the second and third analysis is the removal of unwanted terms from the descriptive terms of the clusters. Also in this analysis, the distribution of documents in the cluster was taken for more detailed analysis. Table 4.5 shows the cluster outcome with None/Entropy

Table 4.5 Descriptive Terms from None with Entropy (Third Analysis)

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	portion, significant, + revenue, + product, + industry, + business	5	0.14705882	0.173003
2	value, + loss, financial, + result, continue, subject, substantial, + condition, + factor, may	9	0.26470588	0.17308
3	stock, distribution, common, below, + certain risk factor, + describe, + own, + risk, + shareholder, + involve	8	0.23529412	0.168023
4	+ constitute	3	0.08823529	0.170131
5	competition, adversely, ability, attract, compete, effectively, intense, introduce, manner, profitability	9	0.26470588	0.175318

This analysis resulted in the outcome of more meaningful information from the clusters. Cluster # 1, cluster # 2 and Cluster # 5 had good information to be analyzed and there were no repetitions of any term in the cluster due to the removal of unwanted words before running the analysis. Analysis was also performed on the clustering of the documents in different clusters. The document distribution was taken from the SAS Enterprise miner and analysis was performed on which documents went to which cluster. Also the appearance of patter in these document distributions was noticed. Table 4.6 shows the document distribution of Table 4.5.

Table 4.6 Document Distribution in None/Entropy

Cluster	Document ID
1	10,2,26,8,20
2	34,15,28,23,14,27,18,30,17
3	16,6,24,7,29,9,19,22
4	13,25,4
5	31,32,11,5,12,3,21,1,33

More detailed description of the third analysis is given under the Results section.

4.4. FINAL ANALYSIS

In the final analysis, an effort was made to reduce the number of clusters and to find better information from the clusters. Except the cluster settings, all the other parameter settings remained the same. Table 4.7 shows the parameter settings used for the final analysis.

Table 4.7 Final Analysis Parameter Setting

Parameter	Value
SVD Resolution	High
Frequency Weight	None
Number of clusters	4 and 3
Number of Terms	10

This analysis did not provide the result as expected. The 4-cluster and 3-cluster analysis were not as meaningful as the 5-cluster analysis. Table 4.8 shows the descriptive terms formed from None/Entropy method

Table 4.8 Descriptive Terms from None with Entropy (4 Clusters)

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	businesses, manufacturing, portion, products, + company, + constitute, significant, continue, company, + industry	12	0.35294118	0.182761
2	+ involve, stock, common, consider, set, + certain risk factor, + contain, + describe, + shareholder, + risk	7	0.20588235	0.172841
3	competition, adversely, ability, attract, financial condition, intense, introduce, manner, part, profitability	13	0.38235294	0.171909
4	+ loss	2	0.05882353	0.076692

Table 4.9 shows the descriptive terms formed by using three clusters and using the frequency weight of None with term weight Entropy.

Table 4.9 Descriptive Terms from None with Entropy (3 Clusters)

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ involve, stock, distribution, common, consider, set, + certain risk factor, + describe, + own, + shareholder	9	0.26470588	0.177409
2	continue, substantial, portion, spinco, + loss, + industry, + product, significant, + business, new	12	0.35294118	0.184773
3	financial, competition, ability, attract, financial condition, intense, introduce, manner, timely, value	13	0.38235294	0.169627

4.5. RESULTS

The third analysis was chosen as the final analysis. According to human expert's opinion, more useful and considerable meaningful information was found from Entropy

method and GF-IDF method of the 5-cluster analysis. Table 4.10 shows the descriptive terms formed out of None/GF-IDF method and Table 4.11 shows its corresponding document distribution in the clusters formed. Remaining analysis from other methods with None frequency weight and its corresponding document distribution are showed in Appendix.

Table 4.10 Descriptive Terms from None/GF-IDF

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ loss	2	0.05882353	0.070571
2	+ risk, + involve, stock, distribution, common, below, businesses, consider, set, + certain risk factor	13	0.38235294	0.176889
3	financial, financial condition, introduce, manner, profitability, timely, value, + result, + condition, + affect	12	0.35294118	0.158539
4	+ constitute	3	0.08823529	0.169985
5	+ product, substantial, continue, significant, + industry	4	0.11764706	0.166162

Table 4.11 Document Distribution for None/GD-IDF

Cluster	Document ID
1	34,27
2	13,32,16,2,6,24,7,8,12,9,1,33,22
3	31,15,11,28,23,5,29,3,14,21,19,17
4	26,25,4
5	10,18,30,20

From the results, sentences which show the descriptions for each cluster were taken from the input documents. Table 4.12 shows the descriptive terms and its corresponding input text sentences. The analysts have to use their domain knowledge in

finding the “real sentences” which matches most of the descriptive terms from each cluster.

In the None/Entropy method, cluster # 2, 3 and 5 provide good details about each cluster, whereas clusters 1 and 4 are considered to be outliers. The outliers could be useful when performing further analysis combined with target variables. However, for clustering analysis purpose, these outlier clusters are not considered.

Table 4.12 Descriptive Terms and Input Text with None/Entropy Method

ID	Descriptive Terms	Input Text/Example
2	value, + loss, financial, + result, continue, subject, substantial, + condition, + factor, may	Substantial operating losses in each of the last three years and may continue to incur financial losses Changes in food consumption patterns may negatively affect our business, financial condition, results of operations and cash flows Not be able to realize the entire book value of goodwill and other intangible assets If we are unable to develop, preserve and protect our intellectual property assets, our business, financial condition, results of operations and cash flows may be negatively affected.
3	stock, distribution, common, below, + certain risk factor, + describe, + own, + risk, + shareholder, + involve	The combined trading price of Western Atlas Common Stock and Company Common Stock held by shareholders after the Distribution may be less than, equal to or greater than the trading price of Western Atlas Common Stock prior to the Distribution Substantially all of the shares of Company Common Stock will be eligible for immediate resale in the public market after the Distribution. Trading in the Company Common Stock to be distributed may commence on a "when issued" basis prior to the Distribution Date.
5	competition, adversely, ability, attract, compete, effectively	Increasing Competition could reduce the demand for our products and services. Having no operating history as an independent company makes it difficult to predict out profitability as a stand-alone company

Table 4.12 Descriptive Terms and Input Text with None/Entropy Method (Cont'd)

ID	Descriptive Terms	Input Text/Example
5	, intense, introduce, manner, profitability	We operate in a competitive business environment, and if we are unable to compete effectively, our results of operations and financial condition may be adversely affected. There can be no assurance that we will be able to compete successfully against current or future competitors or that competitive pressures we face in the markets in which we operate will not materially adversely affect our business, financial condition and results of operations

Table 4.13 shows the descriptive terms and its corresponding input texts from the input documents with the method None/GF-IDF

Table 4.13 Descriptive Terms and Input Text with None/GF-IDF Method

ID	Descriptive Terms	Input Text/Example
		Shareholders of CFI should be aware that the Distribution and ownership of the Common Stock involves certain risk factors, including those described below and elsewhere in this Information Statement, which could adversely affect the value of their holdings. Until the Company Common Stock is fully distributed and an orderly market develops, the prices at which such stock trades may fluctuate significantly and may be lower than prices that would be expected for a fully distributed issue The aggregate market values of Olin Common Stock and Company Common Stock after the Distribution may be less than, equal to, or greater than the market value of Olin Common Stock prior to the Distribution.

Table 4.13 Descriptive Terms and Input Text with None/GF-IDF Method (Cont'd)

ID	Descriptive Terms	Input Text/Example
2	+ risk, + involve, stock, distribution, common, below, businesses, consider, set, + certain risk factor	Trading in the Company Common Stock to be distributed may commence on a "when issued" basis prior to the Distribution Date. Substantially all of the shares of Company Common Stock will be eligible for immediate resale in the public market after the Distribution.
3	financial, financial condition, introduce, manner, profitability, timely, value, + result, + condition, + affect	Performance under government contracts has certain inherent risks that could have a material effect on our business, results of operations and financial condition. Delays or further declines in U.S. military expenditures could adversely affect our business, results of operations and financial condition, depending upon the programs affected, the timing and size of the changes and our ability to offset the impact with new business or cost reductions. Acquisitions involve inherent risks that may adversely affect our operating results and financial condition Our operating results will depend in part on our ability to introduce new and enhanced products on a timely basis The distribution and ownership of our common stock involve a number of risks and uncertainties

The sentences taken from the original data set, which are related to the descriptive terms from both the Entropy and GF-IDF, provide some sense of important concepts reviewed under the "Risk Factors" category in each document filing.

In the next two sections, the two accepted methods which produced acceptable results according to the human expert's expectations are analyzed. The important information from each term method as well as the pattern formed out of the document distribution in the clusters, are explained.

4.5.1. Result Analysis from None/Entropy. Using the None/Entropy method, from cluster # 2, important information such as "Continuation of substantial operating losses", "Change in pattern which might affect the business, financial condition, results of operations and cash flows", "loss of goodwill and other intangible assets" and also the requirement to "develop, preserve and protect the intellectual property assets", are obtained. These are some of the most important risk factors which any organization needs to take care**.**

From cluster # 3, important information related to the stocks can be derived such as "the trading price of the parent and the spin-off company after the spin-off may be less or more or equal to the trading price of the parent company before a spin-off". Other information which the shareholder might find useful are the "Eligibility to resell the company's common stock after the distribution" and also the information that "trading can commence on a "when issued" basis even prior to the distribution date. This information will really be useful from a shareholder point of view.

Cluster # 5 gives out details like how 'increasing competition" can affect the company's products and services. It also gives out important information where the new company with "no operating history" makes it difficult to for that organization to predict the profitability as a stand along company. Also due to the competitive environment, if the new companies are unable to compete effectively, then the operation results and financial condition will be severely affected. These details will give an idea to the shareholder and stockholder to decide on buying shares and stocks from the new company.

Lau et al. (2005) considered text mining as exploring for data in text files to establish valuable patterns and rules that indicate trends and significant features about specific topics. Identical to this, a pattern has been noticed on the document distribution based on the clusters formed under the none/Entropy

From the results, a pattern is formed based on the mood or focus of the risk factors given in the input documents. For example, consider the cluster # 2 from the method None/Entropy. The documents formed under this cluster are document ID's 34, 15, 28, 23, 14, 27, 18, 30, and 17. All these documents belong to different industrial spin - off's. But there is one common point upon which a pattern is formed. They are give

primary importance to the "financial condition" of the company after a spin-off. These documents also mention about the products and common stock, but the focus of the risk factors in these documents has been primarily towards the financial condition of these companies.

Similarly cluster # 4 under the None/Entropy method has the documents 16, 6, 24, 7, 29, 9, 19, and 22 distributed. These documents have "common stock distribution" as a pattern which has made all these documents to appear under the same cluster. The risk factors in all these documents predominantly mention about the distribution of stock after the spin-off and how the shareholder will be affected due to this spin-off.

Cluster # 5 has the documents 31, 32, 11, 5, 12, 3, 21, 1, and 33 distributed in it. In this cluster, the pattern points towards the "competitiveness" mentioned in the risk factors in each financial document. These documents mostly emphasize the risks on how the new company will face the competition from the competitors in the open market as an independent entity.

4.5.2. Result Analysis from None/GF-IDF. From the None/GF-IDF method, relatable information such as "Distribution and ownership of the Common Stock involves certain risk factors", "fluctuating of company Common Stock prices before and after a spin-off", "trading of common stock prior to the distribution date" and also the "ability to resell the common stock after spin-off". These are very useful information to a shareholder in getting the right details out of a company and also in deciding the future course of the spin-off company.

Cluster # 5 from None/GF-IDF mentions more about the financial outcome of the spin-off companies related to the input data. From the cluster, information on how the delay in military expenditures could adversely affect the business, results of operations and financial condition of the company. All these sentences provide a gist of the important concepts mentioned in the spin-off filing reports under the risk factors category.

Both these clusters provide the information regarding the stocks of the parent as well as the spin-off company.

Text mining explores data in text files to establish valuable patterns and rules that indicate trends and significant features about specific topics (Lau et al., 2005). Identical

to this, a pattern has been noticed on the document distribution based on the clusters formed under None/GF-IDF method.

Consider the cluster # 2 under the None/GF-IDF method. Documents 13, 32, 16, 2, 6, 24, 7, 8, 12, 9, 1, 33, and 22 are distributed under this cluster. These documents predominantly focus on the risks regarding the common stock distribution among its shareholders. These input documents also mention about the loss that will be incurred and also the products that will be affected due to the loss of good will by the company, but the primary focus has been towards the common stock distribution and how this is going to be risky to its shareholders and stockholders.

Under cluster # 3 documents 31, 15, 11, 28, 23, 5, 29, 3, 14, 21, 19, 17 are distribution. These documents have stressed the importance of the financial condition of the company after a spin-off. Thus most of the documents emphasizing this point have been distributed under the same cluster.

Even though None/Entropy and None/GF-IDF provided different outcomes, it meets the human expert's expectations on the results.

5. CONCLUSION AND FUTURE WORK

5.1. CONCLUSION

Text mining provides an idea of what the final output will look like. It basically helps in the data analysis thereby helping to find related sentences based on the descriptive terms generated in the clusters.

In this study, analysis is done on the 10-12b filings done by the companies during corporate spin-off. Text mining was applied to the 10-12b filings to investigate potential and/or major concerns found from these financial statements filed for corporate spin-off and also to identify appropriate methods in text mining which can be used to reveal these major concerns.

10-12b filings from thirty-four companies were taken and only the "Risk Factors" category was taken for analysis.

The most important thing in any analysis is the data gathered. Getting the right data is the key in getting some reasonable results. Spin-off's done by thirty-four companies were taken for analysis. All these spin-off filings were done between 1996 and 2011 and none of these companies belonged to the same industry. This resulted in the data containing companies filed during various periods and belonging to various industries.

The first analysis was performed using all three frequency weights such as None, Log and Binary in a combination with all the five term weights such as Entropy, GF-IDF, IDF, Normal and None. Even though the formula used by the frequency weights were different, they all provided the same outcome. This resulted in the narrowing down the analysis on using only the default "None" frequency term.

Among all the analysis performed in this study, the third analysis in which 5-clusters were used and the unwanted words were removed provided better results. From the cluster # 2, 3 and 5 produced in the 5-cluster analysis under the None/Entropy method, related sentences from the input texts were formed. The remaining clusters 1 and 4 are outliers. The same was done with the cluster # 2 and 3 from the None/GF-IDF method. These sentences provided important concepts which were reviewed under the risk factors category in each filing. In this method, clusters 1, 4 and 5 are outliers.

In this study, clusters formed from the methods Entropy and GF-IDF, produced better results. This confirms with prior literature, that Entropy and GF-IDF are two methods that generally produce better results (Dumais, 1991; Chisholm & Kolda, 1999; Jarman, & Berndt, 2010; Katerattanakul, 2010).

Meaningful clusters were formed even from the final analysis where the number of cluster was taken as three and four, but it didn't keep an extra cluster to host the outliers. This actually made the three clusters from the None/Entropy method and the two clusters from the None/GF-IDF, easier to understand. Forcing down the analysis to three clusters actually made the descriptive terms set (from the final analysis) confusing. This is due to the inclusion of outliers in one of the meaningful clusters thereby reducing the purity of the clusters.

Text mining creates patterns based on the input data, taken for analysis. With the input data used for analysis, a pattern was formed based on the mood or focus of the documents. Documents which predominantly emphasized on the effects of the spin-off on the financial condition of the company were grouped together in a cluster. Similarly, documents which emphasized more on the effects of the spin-off on the common stock of the company were grouped together.

Analysis was also performed to identify any pattern which was formed based in the year of filing the 10-12b and also based on the SIC code. But no significant pattern was found out of the clusters.

For this analysis, the appropriate term weights were chosen as Entropy and GF-IDF. The better of the two term weights cannot be identified due to the unavailability of the target variable. Hence the evaluation is based on human judgment.

Also for any analysis, the analyst needs to have full understanding of the domain. This limits the application of text miner by an analyst without considerable understanding on the domain details.

Irrespective of these limitations, text mining makes it easier by taking sizable amounts of data and finding the best descriptive terms which in turn benefits in identifying typical sentences represented by these descriptive terms from the input texts.

5.2. FUTURE WORK

Future works include performing a study to identify how the outline clusters formed during an analysis can be used in finding clues related to unsuccessful spin-off. This study might require a model building with classification.

With the availability of more data, a pattern can be found which can be associated with the SIC code or the year of filing. Individual categorized data can further be categorized in order to deepen the analysis. Further clustering can be performed on the existing cluster for better analysis and results.

In this analysis, two potential useful set of outcomes have been proposed. Identifying the best out of it is something which can be done in future. In order to perform this, result of the spin-off is essential, as it gives target variables of the spin-off of a particular company. A predictive modeling using target variables would be able to actually evaluate the two clustering outputs.

Generally, a 10-12b filing has huge information about the Risk Factors associated with any spin-off. The filing has minimum four pages of documents which explain all the unknown and known risks. But in one cell of an excel sheet, the maximum number of characters which can be entered is 32767. Hence due to the restriction on the character limit in the input file, this study focused on only the “Risks Relating to the Company” and “Risks Relating to the Business”. Future study can be performed to overcome this restriction in order to perform complete analysis of a particular category and can be applied to all categories in the filing.

Further, other categories from the Spin – off filing such as “Forward-looking statements”, “Distribution Details”, “Dividend Policy”, “Compensation Discussion and Analysis” can be undertaken and then the results from each category can be combined which will in turn form the basic precautions which any organization can take into account when any spin – off occurs. This could help the analyst obtain a broader view of the spin-off documents.

This research work serves as a pilot study of next levels of similar studies. A set of start word list can be developed which will eliminate the try and error method of framing an initial list. This work will help researchers in saving time associated in the cumbersome process.

Also more detailed analysis can be performed on the financial data by working with human experts. This will help in identifying any previous ignored or missed data which can be used for future analysis.

APPENDIX: ANALYSIS RESULTS

RESULTS FROM 5-CLUSTER ANALYSIS

Entropy Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	portion, significant, + revenue, + product, + industry, + business	5	0.14705882	0.173003
2	value, + loss, financial, + result, continue, subject, substantial, + condition, + factor, may	9	0.26470588	0.17308
3	stock, distribution, common, below, + certain risk factor, + describe, + own, + risk, + shareholder, + involve	8	0.23529412	0.168023
4	+ constitute	3	0.08823529	0.170131
5	competition, adversely, ability, attract, compete, effectively, intense, introduce, manner, profitability	9	0.26470588	0.175318

Document Distribution

Cluster	Document ID
1	10,2,26,8,20
2	34,15,28,23,14,27,18,30,17
3	16,6,24,7,29,9,19,22
4	13,25,4
5	31,32,11,5,12,3,21,1,33

GF-IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ loss	2	0.05882353	0.070571
2	+ risk, + involve, stock, distribution, common, below, businesses, consider, set, + certain risk factor	13	0.38235294	0.176889
3	financial, financial condition, introduce, manner, profitability, timely, value, + result, + condition, + affect	12	0.35294118	0.158539
4	+ constitute	3	0.08823529	0.169985
5	+ product, substantial, continue, significant, + industry	4	0.11764706	0.166162

Document Distribution

Cluster	Document ID
1	34,27
2	13,32,16,2,6,24,7,8,12,9,1,33,22
3	31,15,11,28,23,5,29,3,14,21,19,17
4	26,25,4
5	10,18,30,20

IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	significant, portion, substantial, + revenue, + product, + industry, + business	6	0.1764706	0.170895
2	value, + loss, financial, + result, + condition, + factor, continue, may, + affect	8	0.2352941	0.168234
3	stock, distribution, common, below, + certain risk factor, + describe, + own, + risk, + shareholder, + involve	8	0.2352941	0.167535
4	+ constitute	3	0.0882353	0.169359
5	competition, adversely, ability, attract, compete, effectively, intense, introduce, manner, profitability	9	0.2647059	0.173958

Document Distribution

Cluster	Document ID
1	10,2,26,8,30,20
2	34,15,28,23,14,27,18,17
3	16,6,24,7,29,9,19,22
4	25,4
5	31,32,11,5,12,3,21,1,33

Normal Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	portion, + competitor, significant, substantial, + industry, + product	6	0.1764706	0.176993
2	+ relate, set, consider, + risk, carefully, + business	4	0.1176471	0.171749
3	operating, distribution, compete, effectively, introduce, manner, part, profitability, new, company	16	0.4705882	0.178784
4	+ constitute	2	0.0588235	0.115472
5	value, financial, + result, + condition, economic, + factor, + affect, may, + business	6	0.1764706	0.161414

Document Distribution

Cluster	Document ID
1	10,2,26,33,30
2	13,32,7,8
3	34,31,16,6,24,11,5,12,3,9,21,1,27,18,19,22
4	25,4

None Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	significant, portion, continue, + product, + industry, new, + business	7	0.2058824	0.17001
2	+ loss, + result, financial	5	0.1470588	0.1524
3	stock, distribution, common, carefully, below, consider, part, set, + certain risk factor, + factor	12	0.3529412	0.17619
4	+ constitute	4	0.1176471	0.18099
5	competition, compete, attract, effectively, intense, + industry, adversely, ability, + affect, new	6	0.1764706	0.17139

Document Distribution

Cluster	Document ID
1	2,26,8,12,18,30,20
2	34,28,23,27,19
3	13,31,15,16,6,24,7,11,29,9,14,22
4	25,21,4,17
5	10,32,5,3,1,33

RESULTS FROM 4-CLUSTER ANALYSIS

Entropy Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	businesses, manufacturing, portion, products, + company, + constitute, significant, continue, company, + industry	12	0.35294118	0.182761
2	+ involve, stock, common, consider, set, + certain risk factor, + contain, + describe, + shareholder, + risk	7	0.20588235	0.172841
3	competition, adversely, ability, attract, financial condition, intense, introduce, manner, part, profitability	13	0.38235294	0.171909
4	+ loss	2	0.05882353	0.076692

Document Distribution

Cluster	Document ID
1	10,2,26,28,12,25,9,1,18,30,4,20
2	13,16,24,7,8,22,17
3	31,15,32,6,11,23,5,29,3,14,21,19,33
4	34,27

GF-IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	effectively, compete, businesses, part, unable, + depend, + business, new, performance, competition	10	0.29411765	0.163013
2	+ factor, value, + certain risk factor, + contain, + describe, financial, + condition, + result, + affect, carefully	11	0.32352941	0.161086
3	+ competitor, demand, + product, substantial, + revenue, + industry, continue, significant	6	0.17647059	0.170276
4	+ loss, + constitute	7	0.20588235	0.18528

Document Distribution

Cluster	Document ID
1	31,32,2,6,8,11,12,3,9,1
2	15,16,7,28,23,5,29,14,19,22,17
3	10,21,18,33,30,20
4	13,34,26,24,25,27,4

IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	businesses, manufacturing, portion, products, + constitute, significant, company, continue, + industry, substantial	11	0.32352941	0.181136
2	+ involve, stock, common, consider, set, + certain risk factor, + describe, + shareholder, + risk, carefully	6	0.17647059	0.166166
3	competition, adversely, ability, attract, financial condition, intense, introduce, manner, part, profitability	13	0.38235294	0.170714
4	+ loss, + result, financial	4	0.11764706	0.163969

Document Distribution

Cluster	Document ID
1	10,2,26,12,25,9,1,18,30,4,20
2	13,16,24,7,7,8,22
3	31,15,32,6,11,5,29,3,14,21,19,33,17
4	34,28,23,27

Normal Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	substantial, manner, introduce, demand, part, portion, spinco, timely, + competitor, + depend	12	0.35294118	0.176358
2	+ involve, + describe, + certain risk factor, set, consider, + risk, stock, common, carefully, + factor	5	0.14705882	0.16472
3	+ affect, financial, performance, adversely, compete, effectively, financial condition, value, + result, + condition	14	0.41176471	0.178999
4	+ constitute	3	0.08823529	0.169809

Document Distribution

Cluster	Document ID
1	34,10,31,2,26,6,24,21,19,33,30,20
2	13,16,7,8,22
3	15,32,11,28,23,5,29,12,3,9,14,1,18,17
4	25,27,4

None Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ factor, carefully, consider, set, value, + certain risk factor, + contain, + describe, financial, + risk	12	0.35294118	0.164319
2	competition, businesses, compete, effectively, intense, part, products, unable, + depend, + business	13	0.38235294	0.17054
3	demand, significant, + product, substantial, + revenue	5	0.14705882	0.169544
4	+ loss	4	0.11764706	0.178683

Document Distribution

Cluster	Document ID
1	13,15,16,7,28,23,5,29,14,19,22,17
2	31,32,2,6,8,11,12,3,25,9,1,18,33
3	10,26,21,30,20
4	34,24,27,4

RESULTS FROM 3-CLUSTER ANALYSIS

Entropy Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ involve, stock, distribution, common, consider, set, + certain risk factor, + describe, + own, + shareholder	9	0.26470588	0.177409
2	continue, substantial, portion, spinco, + loss, + industry, + product, significant, + business, new	12	0.35294118	0.184773
3	financial, competition, ability, attract, financial condition, intense, introduce, manner, timely, value	13	0.38235294	0.169627

Document Distribution

Cluster	Document ID
1	13,16,26,24,7,8,25,9,22
2	34,10,2,6,11,12,1,27,18,20,4,20
3	31,15,32,28,23,5,29,3,14,21,19,33,17

GF-IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	substantial, ability, attract, demand, introduce, manner, timely, value, + company, financial	15	0.44117647	0.1731
2	stock, common, consider, set, + certain risk factor, + describe, + shareholder, + risk, carefully, + involve	7	0.20588235	0.176683
3	businesses, compete, effectively, products, + business, company, + industry, adversely, competition, performance	12	0.35294118	0.163592

Document Distribution

Cluster	Document ID
1	34,10,31,15,28,23,5,14,21,27,19,33,30,17,20
2	13,16,26,24,7,4,22
3	32,2,6,8,11,29,12,3,25,9,1,18

IDF Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	financial, economic, compete, businesses, effectively, financial condition, portion, value, + affect, + result	17	0.5	0.179281
2	+ risk, + involve, stock, common, consider, set, + certain risk factor, + describe, + relate, + shareholder	7	0.20588235	0.171056
3	substantial, manner, introduce, demand, timely, + constitute, + loss, + product, ability, + revenue	10	0.29411765	0.17437

Document Distribution

Cluster	Document ID
1	15,2,26,6,11,28,23,5,29,12,3,9,14,1,18,17,20
2	13,32,16,24,7,8,22
3	34,10,31,25,21,27,19,33,30,4

Normal Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	+ customer, stock, competition, common, ability, attract, compete, consider, demand, + risk	18	0.52941176	0.180616
2	value, + condition, financial, + result, economic, + affect, + factor, + business	8	0.23529412	0.177482
3	products, + constitute, + loss, continue, significant, substantial, company, + result, + business	8	0.23529412	0.180713

Document Distribution

Cluster	Document ID
1	13,10,31,32,16,2,6,24,7,28,5,3,21,1,19,33,22,20
2	15,11,23,29,12,9,14,17
3	34,26,8,25,27,18,30,4

None Analysis

ID	Descriptive Terms	Freq	Percentage	RMS STD
1	carefully, consider, set, value, + certain risk factor, + contain, + describe, + loss, + factor, financial	12	0.35294118	0.170932
2	+ competitor, demand, + product, + revenue, significant, + industry	5	0.14705882	0.175184
3	+ business, adversely, businesses, compete, effectively, financial condition, part, + affect, new, continue	17	0.5	0.17195

Document Distribution

Cluster	Document ID
1	13,34,16,7,28,23,14,27,19,4,22,17
2	10,26,21,33,20
3	31,15,32,2,6,24,8,11,5,29,12,3,25,9,1,18,30

BIBLIOGRAPHY

ADRAIG CUNNINGHAM, P., (2007): "Dimension Reduction," University College Dublin Technical Report UCD-CSI-2007-7.

ALBRIGHT, R., (2004). Taming Text with the SVD. Cary: SAS Institute Inc.

ALLEN, J.W., LUMMER, S.L., MCCONNELL, J., and REED, D.K., (1985). Can takeover losses explain spin-off gains? Journal of Financial and Quantitative Analysis 30, 465-485.

APTE, C., (1997). Data mining: an industrial research perspective. Computational Science & Engineering, IEEE, 4(2). 6 – 9.

ARLINGTON, S.J., BARNETT, S., HUGHES, S., and PALO, J., (2004). Pharma 2010-The Threshold of Innovation, IBM Corporation.

ARREIRA-PERPINAN, M.A.; (1997): A review of dimension reduction techniques, Technical report CS-96-09, Department of Computer Science, University of Sheffield, 1997.

BACK, B., TOIVONEN, J., VANHARANTA, H., and VISA, A. (2001): "Comparing Numerical Data and Text Information from Annual Reports using Self-organizing Maps," International Journal of Accounting Information Systems, 2(4): 249-269.

BAEZA-YATES, R., and RIBEIRO-NETO, B., (1999). Modern Information Retrieval. New York: Addison Wesley.

BALAKRISHNAN, S., CHU, V., HERNÁNDEZ, M.A., HO, H; KRISHNAMURTHY, R., LIU, S.X., PIEPER, J.H., PIERCE, J.S., POPA, L., ROBSON, C.M., SHI, L., STANOI, I.R., TING, E.L., VAITHYANATHAN, S; and YANG, H.; (2010). Midas: integrating public financial data. International conference on Management of data.

BERRY, M. W., and BROWNE, M., (1999). Understanding Search Engines: Mathematical Modeling and Text Retrieval. Philadelphia: Society for Industrial and Applied Mathematics.

BLEI, D., NG, A., and JORDAN, M., (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022.

BLOSE, L., AND J. SHIEH (1997), Tobin's Q Ratio and Market Reaction to Capital Investment Announcements., Financial Review 32, 449 – 476.

BRADLEY, P. S., FAYYAD, U., and REINA, C., (1998). Scaling EM (Expectation-Maximization) Clustering to Large Databases. Redmond, WA: Microsoft Corporation.

BREIMAN, L. (2001): Random forests, Technical report Department of Statistics, University of California.

CARDOSO, J.F.; (1984): ICA website of J.-F. Cardoso. 27 Feb. 2011. http://www.tsi.enst.fr/~cardoso/icacentral/.

CARHART, M. (1997), On Persistence in Mutual Fund Performance., Journal of Finance 52, 57 – 83

CARROLL, J., and LEE, T.Y., (2008).A genetic algorithm for segmentation and information retrieval of SEC regulatory filings.

CHARIKAR, M., and SAHAI, A., (2002): Dimension reduction in the l1 norm, To the proceedings of the 43rd Annual IEEE Symposium, 551 – 560.

CHEN, J. HOUKUAN, TIAN, S, and YOULI, QU, (2009). Feature Selection for Text Classification with Naïve Bayes", Expert system with applications, pp 5432-5435.

CHEUNG, P.S., HUANG, R., and LAM, W.;(2004). Financial activity mining from online multilingual news. Information Technology: Coding and Computing. 267 - 271 Vol.1

CHISHOLM, E., and KOLDA, T.G., (1999). New term weighing Formulas for the vector space method in information retrieval. Computer Science and Mathematics divison

CHUTTUR, M.Y., and BHURTUN, C., (2005). Monitoring financial market using French written textual data.Computational Cybernetics. 239 – 242.

COUSSEMENT, K. (2008): "Employing SAS Text Miner Methodology to Become a Customer Genius in Customer Churn Prediction and Complaint E-mail Management," SAS Global Forum. San Antonio: SAS Institute Inc.

CUSATIS, P.J., MILES, J.A., and WOOLRIDGE, R.J., (1993). Restructuring through spin-offs. Journal of Financial Economics 33, 293-311.

DALEY, L., MEHROTRA, V., and SIVAKUMAR R., (1997). Corporate focus and value creation: Evidence from spinoffs. Journal of Financial Economics 45, 257-281.

DESAI, H., and JAIN, P., (1999). Firm performance and focus: Long-run stock market performance following spin-offs. Journal of Financial Economics 54, 75-101

CHENGMIN, D., and PING, C., (2006). "Mining Executive Compensation Data from SEC Filings". International Conference on Data Engineering.

DONOHO, D.L., (2000): “High-dimensional data analysis: The curses and blessings of dimensionality,” Mathematical Challenges of the 21st Century - conference of The American Math’s Society.

DUMAIS, S. T., (1991). Improving the Retrieval of Information from External Sources. Behavior Research Methods, Instruments, and Computers , 23 (2), 229-236.

FALINOUSS, P., (2007). Stock trend prediction using news articles: a text mining approach.

FAMA, E., AND K. FRENCH (1993), Common Risk Factors in the Return on Stocks and Bonds., Journal of Financial Economics 33, 3 – 56

FAWCETT, T., and PROVOST, F., (1999). Activity Monitoring: Noticing Interesting Changes in Behavior. In Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, pages 53–62, 1999.

FELDMAN, R., and SANGER, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press.

FUNG, G.P.C., YU, J.X., and LU, H., (2005). The predicting power of textual information on financial markets. IEEE Intelligent Informatics Bulletin, 5(1):1-10, 2005.

GIDOFALVI, G; (2001). Using news articles to predict stock price movements.

GILES, J; WO, L; and BERRY, M.; (2003). GTP (general text parser) software for text mining. Statistical data mining and knowledge discovery, CRC Press, Boca Raton, FL (2003), pp. 455–471.

GUYON, I., and ELISSEEFF, A., (2003). An Introduction to Variable and Feature Selection”. Journal of Machine Learning Research. Vol.3, No.7-8, pp.1157-1182.

HIALTHOUSE, E., (1996): “Some theoretical results on nonlinear principal component”.

HABIB, M.A., JOHNSEN, B.D., and NAIK, N.Y., (1997). Spin-offs and information. Journal of Financial Intermediation 6, 153-177.

HAN, J., and KAMBER, M., (2008). Data Mining Concepts and Techniques. China Machine Press.

HEARST, M., (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computer Linguistics. 3–10.

HITE, G., and OWERS, J.E., (1983). Security price reactions around corporate spin-off announcements. Journal of Financial Economics 12, 409-436.

HOTELLING, H., (1933): “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, 24:417 – 441.

HYVARINEN, A., J. KARHUNEN, J., and OJA, E., (2001): “Independent Component Analysis,” John Wiley & Sons, Inc, 2001.

INGVALDSEN, J.E., GULLA, J.A., LAEGREID, T.; and SANDAL, P.C., (2006). Financial News Mining: Monitoring Continuous Streams of Text. Web Intelligence, 321 - 324

JARMAN, J; and BERNDT, J.D., (2010). Throw the Bath Water Out, Keep the Baby: Keeping Medically-Relevant Terms for Text Mining. AMIA Annu Symp Proc. 2010.

JIANG, Y., and ZHOU, Z. H., (2006). A Text Classification Method Based on Term Frequency Classifier Ensemble. Journal of Computer Research and Development. 2006.43(10):1681- 1687.

JOACHIMS, T., (1999). Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 11:41-54.

JOLLIFFE, I.T., (1972): “Discarding variables in principal component analysis I: artificial data”.

JONES, S. K., (1972). A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation 28 (1): 11–21.

KATERATTANAKUL, N., (2010). A pilot study in an application of text mining to learning system evaluation. Missouri University of Science and Technology.

KAYA, M.I.Y., and KARSLIGIL, M.E., (2010). Stock price prediction using financial news articles. Information and Financial Engineering (ICIFE). 478 - 482

KENDAL, J., (1993): “Good and Evil in Chairmen's “boiler plate”: an Analysis,” Organization Studies, 14: 571-592.

KHAN, A., BAHARUDIN, B., and KHAN, K., (2010). Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification. Second International Conference on Computer Engineering and Applications. Page(s): 398 - 403

KLOPTCHENKO A., EKLUND T., BACK B., KARLSSON J., VANHARANTA, H., and VISA A.,(2002). Combining data and text mining techniques for analyzing financial reports, Proc. Eighth Americas Conference on Information Systems.

KLOPTCHENKO, A., MAGNUSSON, C., BACK, B., VISA, A., and VANHARANTA, H., (2004). Mining Textual Contents of Financial Reports.

KOHUT, G., and SEGARS, A., (1992): "The President's Letter to Stockholders: An Examination of Corporate Communication Strategy," Journal of Business Communication, 29(1): 7-21.

KOLDA, T.G., (1997). Limited-Memory Matrix Methods with Applications. Applied Mathematics Program. University of Maryland at College Park,

KOPPEL, M., and SHTRIMBERG, I., (2004). Good news or bad news? Let the market decide. In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications,2004.

KOSTER and SEUTTER, (2003). Taming wild phrases. Proceedings of the 25th European Conference on IR Research (ECIR 2003), Pisa (pp161-176).

KRISHNASWAMI, S., and SUBRAMANIAM, V., (1999). Information asymmetry, valuation, and the corporate spin-off decision. Journal of Financial Economics 53, 73-112.

KROHA, P., and BAEZA-YATES, R., (2004). Classification of stock exchange news.

KUMAR, V., and ZAKI, M., (2000). Knowledge Discovery & Data Mining. Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.

KUSUMURA, Y., HIJIKATA, Y., and NISHIDA, S., (2003). NTM-Agent: text mining agent for net auction.Applications and the Internet, 2003. Proceedings. 2003 Symposium on 27-31 Jan. 2003 Page(s):356 – 359

LAVRENKO, V., SCHMILL, M., LAWRIE, D., and OGILVIE, P., (2000). Language Models for Financial News Recommendation. In Proceedings of the 9th International Conference on Information and Knowledge Management, pages 389–396.

LEA, B., OBOH-IKUENOBE, F., and YU, W., (2006).Application of text mining in developing standardized descriptions of taxa in paleontology: A framework.U.S. Geological Survey, Information Services

LI, AIHUA, and ZHANG, LINGLING, (2009). A Study of the Gap from Data Mining to Its Application with Cases. Business Intelligence and Financial Engineering International Conference. 464 – 467.

LI, HONGQI, GUO, HAIFENG, GUO, HAIMIN, and MENG, ZHAOXU,(2008). Data Mining Techniques for Complex Formation Evaluation in Petroleum Exploration and Production: A Comparison of Feature Selection and Classification Methods. 37 - 43

MAHAJAN, A., DEY, L., and HAQUE, S.M., (2008). Mining Financial News for Major Events and Their Impacts on the Market. International Conference on Web Intelligence and Intelligent Agent Technology.

MCCONNELL, J., AND C. MUSCARELLA (1985), Corporate Capital Investment Decisions and the Market Value of the Firms., Journal of Financial Economics 14, 399 – 422.

MCNEIL, C., and MOORE W. (2001). Spinoff Wealth Effects and the Dismantling of Internal Capital Markets., Working Paper, University of South Carolina.

MILES, J.A., and ROSENFELD, J.D., (1983). The effect of voluntary spin-off announcements on shareholder wealth. Journal of Finance 38, 1597-1606.

MILLER, T.W. (2005). Data and Text Mining: A Business Applications Approach. New Jersey: Pearson Prentice Hall.

MITTERMAYER, M.A., (2004). Forecasting intraday stock price trends with text mining techniques. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS).

MOULTON, L., (2004). Understanding Taxonomies & Search for Corporate Applications. The Gilbane Report, 12(4):2-12.

NAKASHIMA, M., SATO, K., QU, Y., and ITO, T., (2003). Browsing-Based Conceptual Information Retrieval Incorporating Dictionary Term Relations, Keyword Association, and a User's Interest, Journal of the American Society for Information Science and Technology, 54(1):16-28.

NANDA, V., and NARAYANAN, M.P., (1999). Disentangling value: Financing needs, firm scope, and divestures. Journal of Financial Intermediation 8, 174-204.

NG, A., JORDAN. M, and WEISS, Y., (2001): "On spectral clustering: Analysis and an algorithm", In Proc. Advances in Neural Information Processing.

OLSON, D and SHI, Y. (2005). Introduction to Business Data Mining. McGraw-Hill/Irwin

OSBORN, J. D., STUBBART, C. I., and RAMAPRASAD, A., (2001), "Strategic Groups and Competitive Enactment: A Study of Dynamic Relationships between Mental Models and Performance," Strategic Management Journal, 22: 435-454.

PAPINENI, K; (2001). Why Inverse Document Frequency?. IBM T.J. Watson Research Center Yorktown Heights, New York.

PICKERODT, S., and STIEGLITZ, N., (2004). Innovation and development through corporate spin-offs? A theoretical appraisal. Paper presented at the DRUID Summer Conference 2004, Copenhagen

POLETTINI, N., (2004). The Vector Space Model in Information Retrieval- Term Weighting Problem. Department of Information and Communication Technology. University of Trento.

PUI CHEONG FUNG, G.; XU YU, J., and LAM, W.; (2003). Stock prediction: Integrating text mining approach using real-time news. Computational Intelligence for Financial Engineering, 2003. Proceedings. 395 - 402

QIAN, J., and DONG, Y., (2004). A Broad First Search Neighbors (BFSN) clustering algorithm. Journal of Dongnan university. 109-112

RAYMOND J., and MOONEY, R. B., (2005). Mining knowledge from text using information extraction. SIGKDD Explorations Newsletter, ACM.

ROVETTA, B., (2005). Investment Policies and Excess Returns in Corporate Spinoffs: Evidence from the U.S. Market

SALTON, G., and MCGILL, M. J., (1983): "Introduction to Modern Information Retrieval," New York: McGraw-Hill.

SALTON, G., and BUCKLEY, C., (1988). Term weighting approaches in automatic text retrieval. Information Provevssing and Management, 24(5):513 – 523

SCHIPPER, K., and SMITH, A., (1983). Effects of recontracting on shareholder wealth: The case of voluntary spinoffs. Journal of Financial Economics 12, 437-467.

SEO, Y-W., GIAMPAPA, J.A., and SYCARA, K., (2004). Financial News Analysis for Intelligent Portfolio Management. Robotics Institute, Carnegie. Mellon University.

SHI, GUOLIANG, KONG, YANQING, (2009). Advances in Theories and Applications of Text Mining. Information Science and Engineering (ICISE). 4167 - 4170

SHIHAVUDDIN, A.S.M., AMBIA, M.N., AREFIN, M.M.N., HOSSAIN, M., and ANWAR, A., (2010). Prediction of stock price analyzing the online financial news using Naive Bayes classifier and local economic trends. Advanced Computer Theory and Engineering.V4-22 - V4-26

SPARCKJONES, K., (1972). A Statistical interpretation of term specificity and its application in retrieval. J. Documentation, 28(1): 1 21.

SUBRAMANIAN, R., ISLEY, R., and BLACKWELL, R., (1993): "Performance and Readability: A Comparison of Annual Reports of Profitable and Unprofitable Corporations," Journal of Business Communication, 30: 50-61.

TAYLOR, S.M., (2004). Deciphering human language [information extraction]. 28 - 34

THOMAS, J., (1997): "Discourse in the Marketplace: The Making Meaning of Annual Reports," Journal of Business Communication, 34: 47-66.

THOMAS J. and CHEMMANURA, A., (2003). A theory of corporate spin-offs.

TITMAN, S., K.WEI, AND F. XIE (2003), Capital Investments and Stock Returns., Working Paper, National Bureau of Economic Research.

TSENG, Y., (2002). Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology, 53(13):1130-1138.

TSENG, Y., LIN, C-J., and LIN, Y-I., (2007). Text mining techniques for patent analysis. Information Processing and Management 43 (2007) 1216–1247

TURMO, J., AGENO, A., and CATALÀ, N., (2006). Adaptive information extraction. Computing Surveys (CSUR) , ACM Publisher.

VAPNIK, V.N., (1995). The Nature of Statistical Learning Theory.

WINSOR, D., (1993): "Owning corporate texts," Journal of Business and Technical Communication, 7(2): 179-195.

WUTHRICH, B., CHO, V., LEUNG, S., PERMUNETILLEK, D., ZHANG, J., and LAM, W., (1998). Daily stock market forecast from textual web data. IEEE International Conference on Systems, Man, and Cybernetics.

WU, Y.C.; (2007). Predicting the trend of Taiwan Weighted Stock Index with text mining techniques

XIANGZHU GAO, MURUGESAN, S. and LO, B., (2005). Extraction of keyterms by simple text mining for business information retrieval, 332 - 339

YING WAH THE, ABU BAKAR ZAITUN, and SAI PECK LEE;(2001). Data mining using classification techniques in query processing strategies.Computer Systems and Applications, ACS/IEEE International Conference.200 - 202

YU, W., LEA, B., and GURUSWAMY, B., (2007). A theoretic framework integrating text mining and energy demand forecasting.Electronic Business Management Society

ZHIFU, Y., FEI, Y., and JINGQING, L, (2007). A Multi-parameter Synthetic Signal Sorting Algorithm Based on Clustering. Electronic Measurement and Instruments. ICEMI '07. 8th International Conference. 2-363 - 2-366

ZHOU, L. and ZHANG, D., (2003). NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval. Journal of the American Society for Information Science and Technology, 54(2):115-123.

VITA

Aravindh Sekar was born in Pondicherry, India on June 22, 1985. He received his Bachelor's in Technology degree in Computer Science from the National Institute of Technology, Rourkela in 2006. After his graduation he worked with Accenture Services Pvt Ltd from May 2006 till July 2009 before joining Missouri University of Science and Technology in August 2009 for a Degree program in Information Science and Technology. From August 2009 till May 2010, he served as the Graduate Research Assistant in the Business and Information Technology Department. He completed his degree and earned a Master's degree in Information Science and Technology from the Missouri University of Science and Technology.