9/10 2019-10 pages- 2021.11.01
→ poorly written
→ summary table is very useful ref. ← collection of relevant papers
59 studies 1960 - 2017 (*)



# Text Mining in Financial Information

**Article** · January 2019

CITATIONS
3

READS
5,333

**1 author:**

Nida Turegun
Ozyegin University
**26** PUBLICATIONS **53** CITATIONS

SEE PROFILE


# Text Mining in Financial Information

Nida Türegün*

*School of Applied Sciences, Ozyegin University, Nişantepe Mah. Orman Sok, 34794 Istanbul, Turkey.*

**Abstract:**

Financial information mainly lies in financial statements, but it should be noted that financial information can be textual as well as numerals. There are many data sources (footnotes, sustainability reports, executive letters etc.) which are different from the financial statements that provide useful valuable information for decision makers. As a form of communication, text data allow researchers to understand managers' behavioral approaches and business behavior. The analysis of these data requires text mining methods. Text mining is a large data analysis used in the analysis of semi-structural and non-structural data. Within the scope of this information, it is aimed to raise awareness on the use of text mining in non-structural data analysis in the field of financial information. In addition, the aim of this study is to provide a review of previous studies and to guide the researchers about its applications in the field of financial information.

**Publication History:** Received: 22 October 2018 | Revised: 11 December 2018 | Accepted: 24 December 2018

**Keywords:**
Text mining, financial information, data mining, data analysis, unstructured data, text analysis.

**JEL:**
M40, C38.

## 1. INTRODUCTION

Data mining is the extraction of information that is previously unknown and hidden within the data. It uses various statistical techniques to obtain valuable information from large-scale data for related research (Larose, 2005). Data mining is interested in analyzing the numerical data in the databases with various statistical and analytical methods, and interpreting the obtained results. However, various studies should carry out their analysis of the data in the text form. This need for analyzing the data that is not in numerical form has formed the text mining field. Text mining is a very new technique that is used in contemporary studies.

The purpose of text mining is to process unstructured information, to extract meaningful numerical contents of the text, and thus to access information contained in the text for various data mining algorithms (Allahyari et al., 2017). The information can be derived from the summaries of the words found in the documents. Thus, the words used in documents, word sets, etc. can be analyzed. Moreover, documents can be analyzed and similarities can be identified. Text mining is the process of converting the text into numbers or converting to meaningful content. It can be combined with other analyses such as in predictive data mining and unstructured learning methods.

The pioneering works of text mining has been conducted in 1960s. Readability studies are considered as the first steps in the process of the evolution of text mining in financial information. However, the limitations of the financial statements started to be questioned in 1990s, correspondingly text mining studies increased concurrently.

The numerical data structure of the financial information was not satisfying the users of the financial statements, when they were taking their decisions. They were looking for additional sources of information to help them to increase the quality of the decisions they take.

The statements are analyzed using the financial statement analyses, such as ratio analysis, trend analysis, and comparative

*Address correspondence to this author at School of Applied Sciences, Ozyegin University, Nişantepe Mah. Orman Sok, 34794 Istanbul, Turkey;
Tel: +90 2165649826; E-mail: nida.turegun@ozyegin.edu.tr

**Mesford Publisher Inc**
Office Address: Suite 2205, 350 Webb Drive, Mississauga, ON L5B3W4, Canada; T: +1 (647) 7109849 | E: caef@mesford.ca, contact@mesford.ca, https://mesford.ca/journals/caef/


**Fig. (1).** Text Mining Process.

table analysis. However, financial reports provide many qualitative and non-structured data, such as text and illustrations as well. Therefore, the non-structural data can be analyzed with text mining analysis methods. Thus, more useful information can be provided to stakeholders.

Within the scope of this information, this study is conducted with the intention to raise awareness to the use of text mining on non-structural data in the field of financial information. In addition, the aim of this study is to provide a review of previous literature and guide the researchers about its applications in the field of financial information.

## 2. TEXT MINING

Similar to data mining, text mining investigates the data in text files, sets up the rules, and builds important features about specific topics. Unlike data mining, text mining works with unstructured or semi-structured assemblies of text documents. Text mining starts with keyword selection amongst the document heaps. Access machines recognize thousands of keywords and expressions, but they do not analyze the content behind the text (Bothma, 2004). It is necessary to establish a dictionary to be used as a source of information. The dictionary is then used to translate meaningful signs and the contents of the text that was not organized. With text access results, it can make further analysis and transform it into a successful database.

Text mining automatically extracts information from different written sources and reveals unknown information in the text. The purpose of text mining is to find the overall tendency in textual data and to identify new or not presented information. Text mining can offer flexible approaches to accessing, analyzing and managing information (Taparia, 2017). Thus, text mining can expand the scope of data mining with the ability to refer to textual materials. Text mining has not emerged from an academic gap, but it has evolved from similar technologies such as probability theory, statistics and artificial intelligence.

Beyond traditional surveys, focus interviews and other research methods to understand what people think, it is the basic function of text mining to address written texts and analyze them as a large database and convert them into numerical values.

### 2.1. Steps of Text Mining Process

There are five main steps in the text mining process. Fig. (**1**) shows the steps of the process. It starts with gathering. In this step, data must be collected from various resources such as document files, websites, emails or comments. This gathering process can be automated or directed by the user.

The process continues with preprocessing. In this step, content identification and extraction of descriptive characteristics takes place. The most important techniques can be counted as text cleanup, tokenization, and feature extraction. Text clean up technique eliminates the unwanted or unnecessary information. Tokenization technique separates the text into meaningful units like words or sentences. Feature extraction characterizes the text to have a set of measurable dimensions like frequency of words. In the third step, an indexed must be created for particular terms. Moreover, it indexes the location and number of the particular term. This process gives structure and allows rapid access to the data (Chang et al., 2018).

The process continues with the fourth step, which is mining. In this step, special data exploration methods are used to disclose new information. This step helps to reveal specific terms, their relationship between other terms and linkage with the dictionary (Chang et al., 2018). However, this step may encounter challenges and complexities. The most common challenge is the structure of the documents, but the main challenge is natural language processing. In various languages, one word may have more than one meaning or different words may have the same meaning which would lead to complexity. There are many studies to find a solution for this complexity, but it is very difficult to generalize the results since semantics may differ from user to user (Gorvankolla & Rekha, 2017).

The last step is related to the user. The analysis step takes the raw results from the fourth step of mining and this raw data has to be evaluated and visualized by the user, so that the user can make the interpretations (Chang et al., 2018).

### 2.2. Applications

Text mining can be applied in many areas such as national and company security applications, legal and law cases, corporate finance, patent analysis, public relations, and web page comparison. In a survey study, questions can be prepared as open-ended in order for respondents to express their opinions without restricting them in a specific response format. Thus, previously unexplored customer views and ideas can be obtained. An internet page can be scanned. The list of terms and documents on the site can be automatically removed, and the important features or terms can be specified.

For the purpose of the market research, published documents, press releases and web pages can be searched and tracked for the measurement of the market impact. Text mining can be

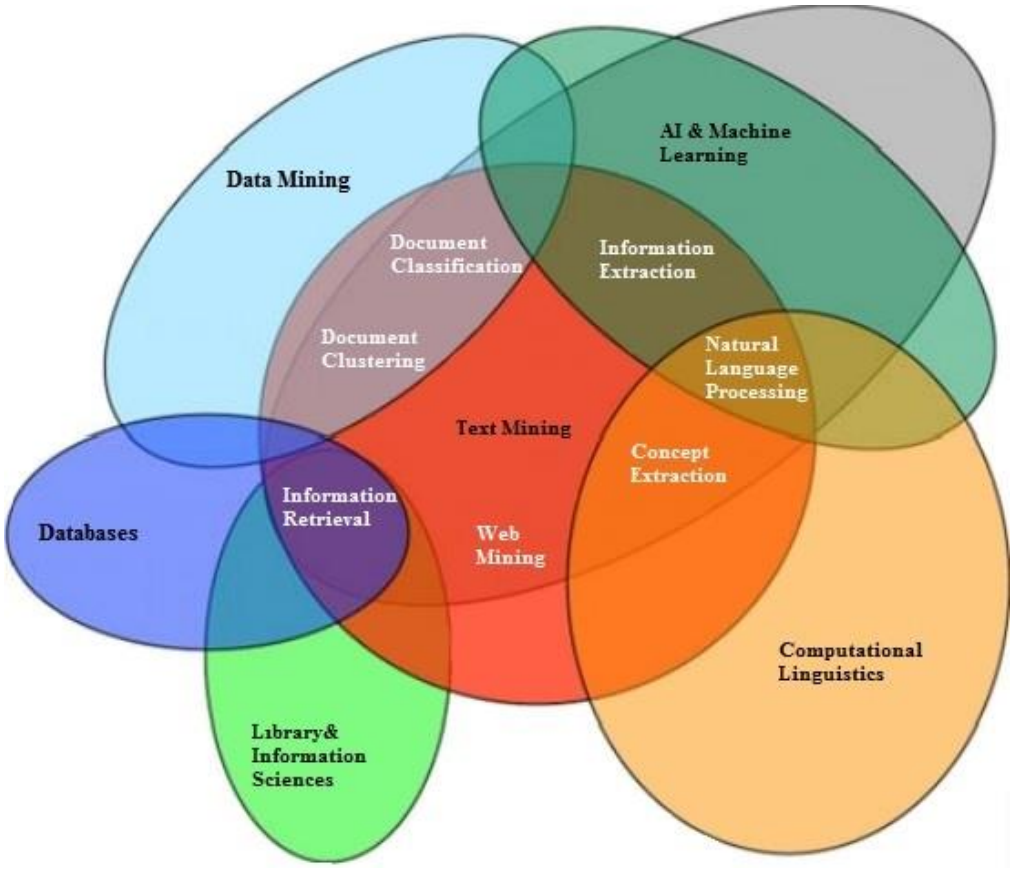


**Fig. (2).** Venn Diagram of Text Mining.

used with quantitative methods to evaluate open-ended survey questions and interviews. Moreover, it can be used in CRM (Customer Relationship Management). Quality information is extracted from text information obtained from access points such as emails, transaction data, call center data and survey results of all customers. This qualified information is used to estimate the customer's abandonment and cross sales.

There are many software resources for text mining. The few examples are Oracle Data Mining, Megaputer's Text Analyst, R-Language Programming text mining, STATISTICA Text Miner, SAS-Text Mining and SPSS-Text Mining and Text Analysis for Surveys.

Fig. (**2**) shows the Venn Diagram for text mining. This diagram is created by Miner et al. in 2012. It shows the five practice areas (databases, data mining, artificial intelligence and machine learning, computational linguistics, library and information sciences) as the key intersections and seven main additional areas that contribute to text mining. Although they have distinctive characteristics, they are very much interrelated with each other. An usual text mining task would necessitate techniques from multiple areas. The seven practice areas are search and information retrieval (IR), document clustering, document classification, web mining, information extraction (IE), natural language processing (NLP) and concept extraction (Miner et al., 2012).

Fig. (**2**). Venn Diagram. Adapted from Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, by Miner, G., Elder IV, J., Hill, T. 2012, UK: Academic Press.

The IR area contains the storage and recovery of documents, web search engines and keyword search. The document clustering area groups and categorize terms, documents etc. by applying clustering methods. The document classification area groups and categorize terms, documents etc. by applying classification methods. Web mining contains data and text mining on the Internet. The IE area identifies and extracts the correlated evidence and relations from unstructured data. The NLP area contains processing of low-level language and understanding of the tasks. The concept extraction area groups words and phrases them into semantically alike sets (Miner et al., 2012).

## 3. TEXT MINING IN FINANCIAL INFORMATION

In recent decades, numerical data structure of the financial statements started to become insufficient in supporting business decisions of the stakeholders. Thereby, studies on text mining of financial information have begun to raise with the increase in the text data. Understanding the text data contained in the financial reports which are provided by the enterprises, is getting more and more important for financial information research. The data in form of texts can give information about the content of the financial data (Heidari, 2015). The methods in which managers choose to communicate, can provide an understanding of corporate decisions due to their particular administrative characteristics. As a form of communication, the

text data allow researchers to understand managers' administrative styles, tendencies and approaches as well as the reflections of those on the business applications.

As stated before, text mining started with readability studies in financial information analysis. The documents used in financial information analysis can be exemplified as activity and quarterly reports, financial statement footnotes, sustainability reports, executive letters, SEC documents, 10-K executive discussions and analyzes section, analyst reports, and newspaper and social media news. As can be seen above, there are many data sources which are different from the financial statements both in structure and content that provide useful and valuable information to decision makers.

Text mining can compute the text data and calculate the word frequencies. Once the input document is sorted out and the first word frequencies are calculated, several additional transformations can be made to summarize and combine the extracted information. There are many methods for text mining analysis, but the log, binary and inverse document frequency methods are the basic and widely used examples of computing word frequencies (Munkova et al., 2013).

In Log frequencies, the word or term frequencies emphasize how a word in each document is significant or important (Miner et al., 2012). A general conversion to calculate the raw word frequency (wf) is as follows;

$f(wf) = 1 + \log(wf)$, for $wf > 0$

In binary frequency, a simpler transformation is used to enumerate whether a term is often used in a document (Miner et al., 2012). The conversion is as follows;

$f(wf) = 1$, for $wf > 0$

Inverse frequency shows relative document frequencies of different words (Miner et al., 2012). In addition to the word frequencies of their visibility, the available and very useful conversions of document frequencies are called inverse document frequency.

$$idf(i, j) = \begin{cases} 0 & \text{if } wf_{ij} = 0 \\ \left(1 + \log\left(wf_{ij}\right)\right) logdf_i & \text{if } wf_{ij} \geq 1 \end{cases} \quad N$$

In this formula, N is the total number of documents and dfi is the document frequency for the first word. Therefore, in this formula, both the reduction of the simple word frequencies via the log function and giving the value 0 (log(N/N=1)=0) when the word is seen in the whole document, and a word seen in a single document (log(N/1)=log(N)) contains a weight factor that gives the maximum value. It can easily be seen how this information both reflects relative frequencies of the visibility of the word and creates indices of the semantic characteristics of the documents present in the analysis.

## 4. METHODOLOGY

In this study, it is aimed to give information about the text mining applications on financial information. EBSCOhost, Google Scholar and Science Direct databases were scanned between 1960 to 2017 with 'text mining/analytic and financial information/reporting' keywords and content analysis was performed.

## 5. TEXT MINING WORKS ON FINANCIAL INFORMATION

The aim of the article is to systematically examine the contents of the past studies. For this reason, full-text available studies from above mentioned databases were included in the scope. After the through search a total of 59 studies have been found with the content of text mining on financial information. Table 1 shows the summary for text mining articles in financial reporting. According to the Table **1**, the most commonly used dataset for the text mining analyses are annual reports, quarterly reports, sustainability reports, 10-K reports, financial reports, financial news and SEC documents. The research area is clustered under six topics, which are business performance (25 studies), audit (10 studies), bankruptcy (2 studies), sustainability (5 studies), corporate governance (2 studies) and others (15 studies). Text mining is mostly used in estimation of business performance.

**Table 1. Summary of Text Mining Works in Financial Reporting.**

| Author | Article | Year | Area | Dataset | Model |
|---|---|---|---|---|---|
| Soper & Dolphin | Readability and Corporate Annual Reports | 1964 | Other | Annual Reports | Readability |
| Smith & Smith | Readability: A Measure of the Performance of the Communication Function of Financial Reporting | 1971 | Other | Annual Reports | Flesch ve Dale-Chall Readability Models |
| Frazier *et al.* | A Methodology for the Analysis of Narrative Accounting Disclosures | 1984 | Other | Annual Reports | WORDS |
| Abrahamson & Park | Concealment of Negative Organizational Outcomes: An Agency Theory Perspective | 1994 | Performance | Annual Reports | Computer aided text analysis |
| Abrahamson & Amir | The Information Content of the President's Letter to Shareholders | 1996 | Performance | Annual Reports | Computer aided text analysis |
| Clatworthy & Jones | Financial Reporting of Good News and Bad News: Evidence from Accounting Narratives | 2003 | Performance | Annual Reports | Natural Language Processing |

| | | | | | |
|---|---|---|---|---|---|
| Kloptchenko *et al.* | Mining Textual Contents of Financial Reports | 2004 | Other | Quarterly Reports | Computer aided text analysis |
| Antweiler *et al.* | Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards | 2004 | Other | Internet Stock Message Boards | Computer aided text analysis |
| Baginski *et al.* | Why Do Managers Explain Their Earnings Forecasts? | 2004 | Performance | Management Forecasting | Natural Language Processing |
| Kloptchenko *et al.* | Combining Data and Text Mining Techniques for Analysing Financial Reports | 2004 | Performance | Financial Reports & Tables | Self-Organizing Maps & Text Clustering Analysis |
| Cecchini | Quantifying the Risk of Financial Events Using Kernel Methods and Information Retrieval | 2005 | Audit | 10-K Reports | Kernel methods |
| Clatworthy & Jones | Differential Patterns of Textual Characteristics and Company Performance in the Chairman's Statement | 2006 | Performance | Annual Reports | Natural Language Processing |
| Schumaker & Chen | Textual Analysis of Stock Market Prediction Using Financial News Articles | 2006 | Performance | Financial News Articles | Support Vector Machine |
| Li | Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? | 2006 | Performance | Annual Reports | Natural Language Processing |
| Nelson & Pritchard | Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language | 2007 | Performance | Management Disclosures | Bag of words |
| Tettlock | Giving Content to Investor Sentiment: The Role of Media in the Stock Market | 2007 | Performance | Financial News | Sentiment Analysis |
| Henry | Are Investors Influenced by How Earnings Press Releases are Written? | 2008 | Performance | Earning Release | Sentiment Analysis |
| Shirata & Sakagami | An Analysis of the "Going Concern Assumption": Text Mining from Japanese Financial Reports | 2008 | Bankruptcy | Financial Reports | Bag of words |
| Li | Annual Report Readability, Current Earnings and Earnings Persistence | 2008 | Performance | Annual Reports | Readability |
| Tettlock *et al.* | More Than Words: Quantifying Language to Measure Firms' Fundamentals | 2008 | Performance | Financial News | Sentiment Analysis |
| Demers & Vega | Soft Information in Earnings Announcements: News or Noise? | 2008 | Other | Earnings Announcements | DICTION |
| Wuthrich *et al.* | Daily Stock Market Forecast from Textual Web Data | 2008 | Performance | Internet News | Knn Machine Learning Algorithm |
| Barkemeyer *et al.* | What the Papers Say: Trends in Sustainability | 2009 | Sustainability | Newspapers | Bag of words |
| Henry & Leone | Measuring Qualitative Information in Capital Markets Research | 2009 | Performance | Earning Release | Sentiment Analysis |
| Kothari *et al.* | The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: | 2009 | Performance | Reports of Management, Analysts, Press | Natural Language Processing |
| Feldmen *et al.* | Management's Tone Change, Post Earnings Announcement Drift and Accruals | 2010 | Performance | Management Disclosures | Sentiment Analysis |
| Balakrishnan *et al.* | On the predictive ability of narrative disclosures in annual reports | 2010 | Performance | 10-K Reports | Text classification techniques |
| Li | The Determinants and Information Content of the Forward-Looking Statements in Corporate Filings | 2010 | Performance | 10-K Reports | Naive Bayes |

| | | | | | |
|---|---|---|---|---|---|
| Goel *et al.* | Can Linguistic Predictors Detect Fraudulent Financial Filings? | 2010 | Audit | Annual Reports | Natural Language Processing & Bag of words |
| Cecchini *et al.* | Making Words Work: Using Financial Text as a Predictor of Financial Events | 2010 | Audit | 10-K Reports | Support Vector Machine |
| Aral *et al.* | Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice | 2011 | Performance | Stock Recommendations | Latent Dirichlet Allocation |
| Huang & Li | A Multilabel Text Classification Algorithm. For Labeling Risk Factors in SEC form 10-K | 2011 | Performance | 10-K Reports | Multilabel categorical K nearest neighbor |
| Rogers *et al.* | Disclosure Tone and Shareholder Litigation | 2011 | Performance | Earnings Announcements | Sentiment Analysis |
| Loughran & Mc Donald | When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks | 2011 | Other | Dictionaries & 10-K Reports | Sentiment Analysis |
| Shirata *et al.* | Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining | 2011 | Bankruptcy | Annual Reports | Bag of words |
| Glancy & Yadav | A Computational Model for Financial Reporting Fraud Detection | 2011 | Audit | Annual Reports | Computer aided text analysis |
| Humpherys *et al.* | Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis | 2011 | Audit | Annual Reports | Naive Bayes & C.4.5 algorithm |
| Davis *et al.* | Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language | 2012 | Performance | Press News | DICTION |
| Goel & Gangolly | Beyond the Numbers: Mining the Annual Reports for Hidden Cues Indicative of Financial Statement Fraud | 2012 | Audit | Annual Reports | Bag of words |
| Gupta & Gill | Financial Statement Fraud Detection Using Text Mining | 2012 | Audit | Text on Financial Tables | Bag of words & Support vector machine |
| Zheng & Zhou | An Intelligent Text Mining System Applied to SEC Documents | 2012 | Corporate Governance | SEC Reports | Intelligent Corporate Analysis &Rating System |
| Zaki & Theodoulidis | Analyzing Financial Fraud Cases Using a Linguistics-Based Text Mining Approach | 2013 | Audit | SEC Trail Reports | Natural Language Processing |
| Chakraborty *et al.* | Automatic Classification of Accounting Literature | 2014 | Other | Literature Documents | Decision Tree Algorithm |
| Liew *et al.* | Sustainability Trends in the Process Industries: A Text Mining-Based Analysis | 2014 | Sustainability | Sustainability Reports | Bag of words |
| Rivera *et al.* | A Text Mining Framework for Advancing Sustainability Indicators | 2014 | Sustainability | Newspapers | Bag of words |
| Shahi *et al.* | Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring | 2014 | Sustainability | Sustainability Reports | Naive Bayes & Decision Tree Algorithm |
| Campbell *et al.* | The information content of mandatory risk factor disclosures in corporate filings | 2014 | Performance | 10-K Reports | Sentiment Analysis |
| Purda & Skillicorn | Accounting Variables, Deception, and A Bag of Words: Assessing the Tools of Fraud Detection | 2015 | Audit | Annual & Quarterly Reports | Support Vector Machine |
| Kamaruddin *et al.* | A Text Mining System for Deviation Detection in Financial Documents | 2015 | Other | Financial Reports | Semantic analysis & rigorous textual comparison |

| | | | | | |
|---|---|---|---|---|---|
| Bala *et al.* | Tracking "Real Time Corporate Sustainability Signals Using Cognitive Computing | 2015 | Sustainability | Sustainability & Financial Reports | Natural Language Processing / Cognitive Computing |
| Heidari & Felden | Financial Footnote Analysis: Developing a Text Mining Approach | 2015 | Other | Financial Report Footnotes | Naive Bayes |
| Gemar & Jimenez-Quintero | Text Mining Social Media for Competitive Analysis | 2015 | Performance | Social Media News | Competitive Intelligence |
| Matthies & Coners | Computer-Aided Text Analysis of Corporate Disclosures- Demonstration and Evaluation of Two Approaches | 2015 | Other | Corporate Disclosures | Computer aided text analysis |
| Liu & Moffitt | Text Mining to Uncover the Intensity of SEC Comment Letters and Its Association with the Probability of 10-K Restatement | 2016 | Other | SEC letters & 10-K Reports | Bag of words |
| Rich *et al.* | Linguistic Tone of Municipal Management Discussion and Analysis Disclosures and Future Financial Reporting Delays | 2016 | Other | Annual Reports | Sentiment Analysis |
| Pencle & Mălăescu | What's in the Words? Development and Validation of a Multidimensional Dictionary for CSR and Application Using Prospectuses | 2016 | Social Responsibility | Initial Public Offering Prospectuses | Computer aided text analysis |
| Hajek & Henriques | Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud | 2017 | Audit | Annual Reports | Bayesian Belief Network & Naive Bayes |
| Aureli | A Comparison of Content Analysis Usage and Text Mining in CSR Corporate Disclosure | 2017 | Other | Corporate Disclosures | Content Analysis |
| Tsai & Vang | On the Risk Prediction and Analysis of Soft Information in Finance Reports | 2017 | Other | Financial Reports | Bag of words & Sentiment Analysis |

The limitations of the financial statements started to be questioned in 1990s, correspondingly number of text mining studies increased concurrently. Readability studies are considered as the first step in the process of text mining development. Therefore, the first studies in the field of financial reporting are readability studies.

## CONCLUSION

Text mining is an evolving field in research and many different computer programs are needed to be developed to serve various requirements of its applications. These computer programs will help to analyze the textual data sources conveniently and will enable researchers to identify new horizons of its usefulness.

Financial information has an important function in providing communication between the internal and external stakeholders of businesses. Financial information mainly lies within the boundaries of the financial statements, but it should be noted that financial information can be textual as well as numerical. The factors such as changing environmental conditions and technology have deeply influenced the conceptual basis phenomena in the definition of assets and the needs of the stakeholders. Along with the technology, changes in the asset structure of the enterprises have occurred and the intangible assets have gained weight rather than the tangible assets.

Society is closely following the environmental and social events. Also, give more importance to these issues for future generations. This phenomenon led managers to disclose more thorough and detailed reports to the general public beyond the numerical information they provide. Basically, these reasons have made the traditional financial reporting framework one of the topics discussed in the last 20 years. In this area, financial reporting frameworks that aim to provide more data for the enterprises have begun to emerge with international and national regulations. Companies have begun to share their information with the stakeholders through various intermediaries such as activity reports, sustainability reports, external audits and their internet sites.

The current state in financial reporting systems is working towards an integrated reporting which will cover all the sources of information mentioned above. However this integrated system is in its early stages of the evolution. Accounting systems are trying to reduce the information asymmetry between business and stakeholders with expanding the scope of the reporting frameworks. However, so much data that are targeted be presented in a single reporting, cannot be converted into useful information easily, because of the inclusion of texts and images.

The analysis of these data requires new methods. In this study, one of the new methods of text mining is discussed. Text mining is a large data analysis used in the analysis of semi-

structural and non-structural data. Text mining, within the expanding financial reporting framework, has been used progressively in financial information in the last two decades. There are many data sources which are different from the financial statements that provide useful information and contain valuable information for decision makers. As a result of the literature review, it is observed that text mining is frequently used in the field of financial information. The studies show that the data that cannot be presented in the financial statements as numerical data can be analyzed and provide more useful information to the stakeholders.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## REFERENCES


[1]. Abrahamson, E., Park, C.: Concealment of negative organizational outcomes: an agency theory perspective. Acad Manag J 37: 1302-1334 (1994).

[2]. Abrahamson, E., Amir, E.: The information content of the President's letter to shareholders. J Bus Finan Account 23: 1157-1182 (1996).

[3]. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K.: A brief survey of text mining: classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919 (2017).

[4]. Antweiler, W., Frank M. Z.: Is all that talk just noise? The information content of internet stock message boards. J Finance 59: 1259-1294 (2004).

[5]. Aral, S., Ipeirotis, P., Taylor, S.: Content and context: identifying the impact of qualitative information on consumer choice. 32nd International Conference on Information Systems Proceedings, Shanghai, China, 1-9 (2011).

[6]. Aureli, S. A comparison of content analysis usage and text mining in CSR corporate disclosure. Int J Dig Account Res 17: 1-32 (2017).

[7]. Baginski, S.P., Hassell, J. M., Kimbrough, M.D.: Why do managers explain their earnings forecasts? J Account Res 42: 1-29 (2004).

[8]. Bala, G., Bartel, H., Hawley, J. P., Lee, Y. J.: Tracking "real- time" corporate sustainability signals using cognitive computing. J Appl Corp Finan 27: 95-102 (2015).

[9]. Balakrishnan, R., Qiu, X.Y., Srinivasan, P.: On the predictive ability of narrative disclosures in annual reports. Eur J Operat Res 202: 789-801 (2010).

[10]. Barkemeyer, R., Figge, F., Holt, D., Hahn, T.: What the papers say: trends in sustainability: a comparative analysis of 115 leading national newspapers worldwide. J Corp Citizenship 33: 69-86 (2009).

[11]. Bothma, T.J.D.: Differentiating between data-mining and text-mining terminology. South Afr J Inform Manag 6: 1-9 (2004).

[12]. Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H.M., Steele, L. B.: The information content of mandatory risk factor disclosures in corporate filings. Rev Account Studies 19: 396-455 (2014).

[13]. Cecchini, M.: Quantifying the risk of financial events using kernel methods and information retrieval. Unpublished Doctoral Dissertation, University of Florida, (2005).

[14]. Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P.: Making words work: using financial text as a predictor of financial events. Decision Support Syst 50: 164-175 (2010).

[15]. Chakraborty, V., Chiu, V., Vasarhelyi, M.: Automatic classification of accounting literature. Int J Account Inform Syst 15: 122-148 (2014).

[16]. Chang, J., O'Reilly, C., Pontika, N., Haug, K., Owen, G., Oudenhoven, M. What is text mining, how does it work and why is it useful? https://www.fosteropenscience.eu/content/text-mining-101 (2018). Accessed 17 October 2018.

[17]. Clatworthy, M.A., Jones, M.J.: Financial reporting of good news and bad news: evidence from accounting narratives. Account Bus Res 33: 171-185 (2003).

[18]. Clatworthy, M.A., Jones, M.J.: Differential patterns of textual characteristics and company performance in the chairman's statement. Account Audit Accountability J 19: 493-511 (2006).

[19]. Davis, A.K., Piger, J.M., Sedor, L.M.: Beyond the numbers: measuring the information content of earnings press release language. Contemp Account Res 29: 845-868 (2012).

[20]. Demers, E., Vega, C.: Soft information in earnings announcements: news or noise?. Board of Governors of the Federal Reserve System International Finance Discussion Papers, No:951 (2008).

[21]. Feldman, R., Govindaraj, S., Livnat, J., Segal, B.: Management's tone change, post earnings announcement drift and accruals. Rev Account Studies 15: 915-953 (2010).

[22]. Fraizer, K.B, Ingram, R.W., Tennyson B.M.: A methodology for the analysis of narrative accounting disclosures. J Account Res 22: 318-331 (1984).

[23]. Gemar, G., Jimenez-Quintero, J.A.: Text mining social media for competitive analysis. Tourism Manag Studies 11: 84-90 (2015).

[24]. Glancy, F.H., Yadav, S.B.: A computational model for financial reporting fraud detection. Decision Support Syst 50: 595-601 (2011).

[25]. Goel, S., Gangolly, J., Faerman, S.R., Uzuner, O.: Can linguistic predictors detect fraudulent financial filings? J Emerging Technol Account 7: 25-46 (2010).

[26]. Goel, S., Gangolly, J.: Beyond the numbers: mining the annual reports for hidden cues indicative of financial statement fraud. Intell Syst Account Finan Manag 19: 75-89 (2012).

[27]. Gorvankolla, A.K. Rekha B.S.: Application of text mining in effective document analysis: advantages, challanges, techniques and tools. Int J Eng Res Technol 6: 60-64 (2017).

[28]. Gupta, R., Gill, N.S.: Financial statement fraud detection using text mining. 3: 189-191 (2012).

[29]. Hajek, P., Henriques, R.: Mining corporate annual reports for intelligent detection of financial statement fraud - a comparative study of machine learning methods. Knowledge Based Systems 128: 139-52 (2017).

[30]. Heidari, M., Felden, C.: Financial footnote analysis: developing a text mining approach. In Proceedings of International Conference on Data Mining (DMIN), 10-16 (2015).

[31]. Henry, E.: Are investors influenced by how earnings press releases are written? J Bus Commun 45: 363-407 (2008).

[32]. Henry, E., Leone, A.J.: Measuring qualitative information in capital markets research. (2009). Available at https://ssrn.com/abstract=1470807

[33]. Huang, K.W., Li, Z.: A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. ACM Transact Manag Inform Syst 2: 1-18: 19 (2011).

[34]. Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K., Felix, W.F.: Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems 50: 585-594 (2011).

[35]. Kamaruddin, S.S., Bakar, A.A., Hamdan, A.R., et al.: A text mining system for deviation detection in financial documents. Intell Data Anal 19: 19-44 (2015).

[36]. Kloptchenko, A., Magnusson, C., Back, B.: Mining textual contents of financial reports. Int J Digital Account Res 4: 1-29 (2004).

[37]. Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., Visa, A.: Combining data and text mining techniques for analysing financial reports. Intell Syst Account Finan& Manag Int J 12: 29-41 (2004).

[38]. Kothari, S.P., Li, X., Short, J.E. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. Account Rev 84: 1639-1670 (2009).

[39]. Larose, D.: Discovering knowledge in data an introduction to data mining. John Wiley & Sons, Inc, New Jersey, (2005).

[40]. Li, F.: Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan, (2006).

[41]. Li, F.: Annual report readability, current earnings, and earnings persistence. J Account Econ 45: 221-247 (2008).

[42]. Li, F.: Textual analysis of corporate disclosures: a survey of the literature. J Account Literature 29: 143-165 (2010).

[43]. Liew, W., Adhitya, A., Srinivasan, R.: Sustainability trends in the process industries: a text mining-based analysis. Computers Industry 65: 393-400 (2014).

[44]. Liu, Y., Moffitt, K.C.: Text mining to uncover the intensity of SEC comment letters and its association with the probability of 10-K restatement. J Emerg Technol Account 13: 85-94 (2016).

[45]. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10- Ks. J Finan 66: 35-65 (2011).

[46]. Matthies, B., Coners, A.: Computer-aided text analysis of corporate disclosures-demonstration and evaluation of two approaches. Int J Digital Account Res 15: 69-98 (2015).

[47]. Miner, G., Elder J. IV, Hill, T.: Practical text mining and statistical analysis for non-structured text data applications. Academic Press, (2012).

[48]. Munková, D., Munk, M., Vozár, M.: Data pre-processing evaluation for text mining: transaction/sequence model. Procedia Computer Sci 18: 1198-1207 (2013).

[49]. Nelson, K.K., Pritchard, A.C.: Litigation risk and voluntary disclosure: the use of meaningful cautionary language. Working paper, (2007). Available at http://ssrn.com/abstract=998590

[50]. Pencle, N., Mălăescu, I.: What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. J Emerg Technol Account 13: 109-127 (2016).

[51]. Purda, L., Skillicorn, D.: Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. Contemp Account Res 32: 1193-1223 (2015).

[52]. Rich, K.T., Roberts, B.L., Zhang, J.X.: Linguistic tone of municipal management discussion and analysis disclosures and future financial reporting delays. J Emerg Technol Account 13: 93-107 (2016).

[53]. Rimerman, T.W.: The changing significance of financial statements. J Accountancy 169: 79 (1990).

[54]. Rivera, S.J., Minsker, B.S., Work, D.B., Roth, D.: A text mining framework for advancing sustainability indicators. Environ Model Software 62: 128-138 (2014).

[55]. Rogers, J.L., Van Buskirk, A., Zechman, S.L.: Disclosure tone and shareholder litigation. Account Rev 86: 2155-2183 (2011).

[56]. Shahi, A. M., Issac, B., Modapothala, J. R.: Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring. International Journal of Computational Intelligence and Applications, 13(1), 1-27 (2014).

[57]. Shirata, C.Y., Sakagami, M.: An analysis of the "going concern assumption": text mining from Japanese financial reports. J Emerg Technol Account 5: 1-16 (2008).

[58]. Shirata, C.Y., Takeuchi, H., Ogino, S., Watanabe, H.: Extracting key phrases as predictors of corporate bankruptcy: empirical analysis of annual reports by text mining. J Emerg Technol Account 8: 31-44 (2011).

[59]. Schumaker, R., Chen, H.: Textual analysis of stock market prediction using financial news articles. AMCIS 2006 Proceedings, 1432-1440 (2006).

[60]. Smith, J.E., Smith, N.P.: Readability: a measure of the performance of the communication function of financial reporting. Account Rev 46: 552-561 (1971).

[61]. Soper, F.J., Dolphin, R.: Readability and corporate annual reports. Account Rev 39: 358-362 (1964).

[62]. Taparia, R.: Text mining as a better solution for analyzing unstructured data. https://www.projectguru.in/publications/text-mining-analyzing-unstructured-data/ (2017) Accessed 18 October 2018.

[63]. Tetlock, P.C.: Giving content to investor sentiment: the role of media in the stock market. J Finan 62: 1139-1168 (2007).

[64]. Tetlock, P.C., Saar- Tsechansky, M., Macskassy, S.: More than words: quantifying language to measure firms' fundamentals. J Finan 63: 1437-1467 (2008).

[65]. Tsai, M.F., Wang, C.J.: On the risk prediction and analysis of soft information in finance reports. Eur J Operat Res 257: 243-250 (2017).

[66]. Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J.: Daily stock market forecast from textual web data. In SMC'98 Conference Proceedings, 2720-2725 (2008).

[67]. Zaki, M., Theodoulidis, B.: Analyzing financial fraud cases using a linguistics-based text mining approach, (2013). Available at https://ssrn.com/abstract=2353834

[68]. Zheng, Y., Zhou, H.: An intelligent text mining system applied to SEC documents. IEEE/ACIS 11th International Conference, 155-160 (2012).