6/10 2015 - 10 pages - 2021.11.01
- not very relevant, could still refer to it
- very basic classification income tax categories

Financial Footnote Analysis: Developing a Text Mining Approach

Maryam Heidari[1] **and Carsten Felden**[1]
[1]Information system department, University of Freiberg, Freiberg, Saxony, Germany

Abstract - *Financial footnotes analysis provides an opportunity to communicate with stakeholders beyond the numbers in the main body of financial statements. The combination of values in financial reports and their disclosure in footnote parts supports financial decisions in a wisely manner. Nevertheless, the unstructured nature of footnotes poses a barrier for an accurate, automatic, and real-time financial analysis. To address this issue, this paper implements a text classification procedure to evaluate the benefits of text mining deployment to react to the manual financial footnote analysis. This supports the classification of textual parts of financial footnotes automatically into related financial categories, which are relevant for financial analysts, in order to avoid reading entire textual parts manually. This research provides useful insights about the impact of using text mining for an automatic financial footnote analysis in terms of time saving and increasing accuracy.*

Keywords: financial footnotes, text mining, text classification, income tax

1 Introduction

Analyzing financial disclosures is a key mechanism, which facilitates communication of financial analysts, auditors, internal and external decision makers, and other stakeholders gaining benefits from analyzing financial reports [1]. Using solely financial statements in this context does not represent the comprehensive financial story of a company. Financial footnotes provide useful information about company's financial performance [19], [25], [28]. However, the unstructured format of financial footnotes makes it difficult to analyze them automatically. A manual analysis is still a time-consuming issue, restricted in terms of accuracy and real time analysis. Based on 45 financial analysis research papers, Heidari and Felden show that there is no identified method to integrate financial analysis methods of structured values with unstructured footnotes automatically [14]. Furthermore, the widely usage of XBRL as a standard platform for financial data exchange, even with the detailed tagging process, has no impact on an automatic financial footnotes analysis [14], [31]. To overcome the identified bottleneck and to facilitate financial footnote analysis, our study suggests a text-mining approach and applies text classification procedures. The paper's goal is to assess to what extent and how text mining application in footnotes can support financial analytical tasks.

Existing approaches in financial analysis literature shows the variety usage of text mining in financial market prediction such as economic crises, stock price prediction, and risk management [5], [13], [12], and [20]. However, no identified article addresses a text mining application for financial footnotes analysis. Most related research in this area suffers from the automatic solution for financial footnotes and rely on manual information extraction [14]. There exist a number of attempts based on coding schemas to analyze financial disclosures [3]. But it seems to be obvious that a manual code assignment has no advantage for an automatic footnotes analysis [16]. This study tries to make a contribution regarding financial analysts and any stakeholders who benefit from financial analysis by applying a text mining approach based on pre-defined classes in order to assist them in required financial information extraction from footnotes in a real time and automatic fashion.

The paper proceeds as follows: Section 2 discusses related work in terms of existing solutions in literature to deal with unstructured financial information. Section 3 presents the details of the applied research method and the proposed text mining framework. Section 4 demonstrates the empirical implementation of text classification techniques and the results of applying the developed framework including the validation results. Section 5 concludes with implications of the research results and further research directions.

2 State of the art

Today, due to existing various financial analysis software, which provides companies with the financial information to make better decisions, an automatic analysis of financial figures is straightforward. However, it is difficult to analyze textual parts of financial reports automatically to explicate company's financial behaviors and to identify valuable information about the current and future financial status of a company [17], [18].

In the financial footnotes analysis domain, a literature review by Heidari and Felden identified the importance and the effect of financial footnote information on financial analysis processes [14]. However, existing methods for financial footnotes analysis rely on manual processes and there is a strong need to develop an appropriate solution to support footnotes analysis automatically within an integrated financial analysis process [14].

Since financial footnotes have textual format and consist of soft financial information, we reviewed existing literature

regarding text mining in financial analysis area as well. This means, according to Miner et al., particular characteristics and the purpose of text mining, information retrieval, concept extraction, clustering, and classification as typical text mining approaches [21]. As an example in terms of financial analysis, Beattie et al. introduced a comprehensive four-dimensional framework for content analysis of accounting narratives [3]. It uses a coding system based on four attributes in order to give structure to accounting texts. They performed this framework by qualitative research software in order to support the coding procedures via an index system. The advantage of this coding procedure is that it focuses not only on the topic, but also considers the different kinds of attributes. However, this does not have benefit for analyzing financial footnotes automatically, because of manual code assignment to the texts [16].

Among different researches in the financial analysis area, some articles concentrate on text classification techniques, which can be trained for recognizing and differentiating significant categories in documents automatically. Most related research in this area comprises text mining processes, which translate unstructured financial information from numerous text references into a machine readable format for predictive purposes. For example, Brent and Atkisson built a coding scheme by training pre-defined categories assigning code to text documents automatically in order to analyze economic crises through newspaper articles [5]. Neumann et al. designed a text classification approach to process financial news to automate stock price prediction [13]. It bases on three main text processing step: Dataset as a basis for the classification, feature processing to extract different features and generating machine readable information, and finally the machine learning step using a subset of data to train a classification algorithm to be able to response to the stock market trends. Another related study performed by Li, who employs several pre-defined dictionaries to predict stock feedbacks based on US corporate filings [20]. In another attempt, Groth et al. focus on German announcement to evaluate stock price effects [12]. It can be recognized that in the most related research in financial analysis area text mining approaches are used to classify financial news into positive and negative categories to have verifiable market trend [13].

Regarding reviewed articles, it can be mentioned that text mining methods are widely used in financial market trend prediction. Nevertheless, in financial footnote analysis area, the most used method bases on manual procedures. Due to human interference and probability of neglecting relevant content, the manual procedures is extremely time-consuming and error-prone. Therefore, it seems to be appropriate to identify an automatic-based solution to overcome this gap and to compose financial analysis processes reliable as well as accurate. Besides, text mining research shows that using classification techniques is an appropriate solution for analyzing information in textual formats. However, it is not enough to apply just a text classification algorithm, but there is a need to use the classified results regarding the main body of financial reports.

Regarding the identified gap, we demonstrate the utility of a text mining application in automatic financial footnote analysis through implementing a text classification workflow in the next section.

3 Research method

Concerning text mining application for financial footnotes analysis, we applied the Cross-Industry Standard Process (CRISP) process flow to define a complete lifecycle of the text mining workflow [7]. CRISP methodology is based on six phases providing a comprehensive coverage of all activities involved in data/text mining projects [21]. Fig. 1 illustrates the cyclic form of the CRISP process flow.

Concerning reviewed literature and regarding the purpose of this research, Fig. 2 shows the application of the proposed financial footnote analysis. After data preparation, the financial patterns in the narrative part of financial reports are recognized. Later on, well-known classification algorithms are applied and evaluated according to accuracy performance and other obtained results. We assess to what extent text classification method assists and facilitates financial footnote analysis considerably.

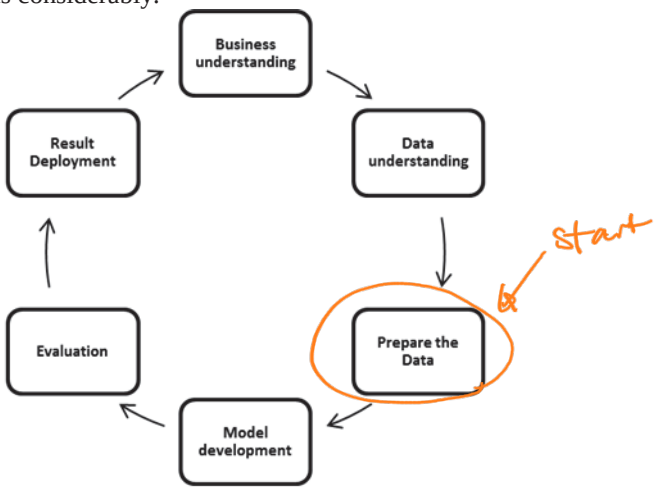

Fig. 1. CRISP process flow [7]

3.1 Data collection

The preliminary step is the collection of financial footnotes. We focused on one footnote item in financial reports: the income tax footnote, which is a material component of most financial statements. Income tax accounting requires the use of estimates, judgments, and other subjective information that cannot be fully discovered in the financial statement reports [11]. According to Graham et al. the income tax footnote can enable users to gain a better understanding of the income tax status of a company [11]. The used database to get financial footnotes of companies' filings was Edgar online of U.S. SEC[1]. This database provides free public access to corporate

[1] U.S. Securities and Exchange Commission

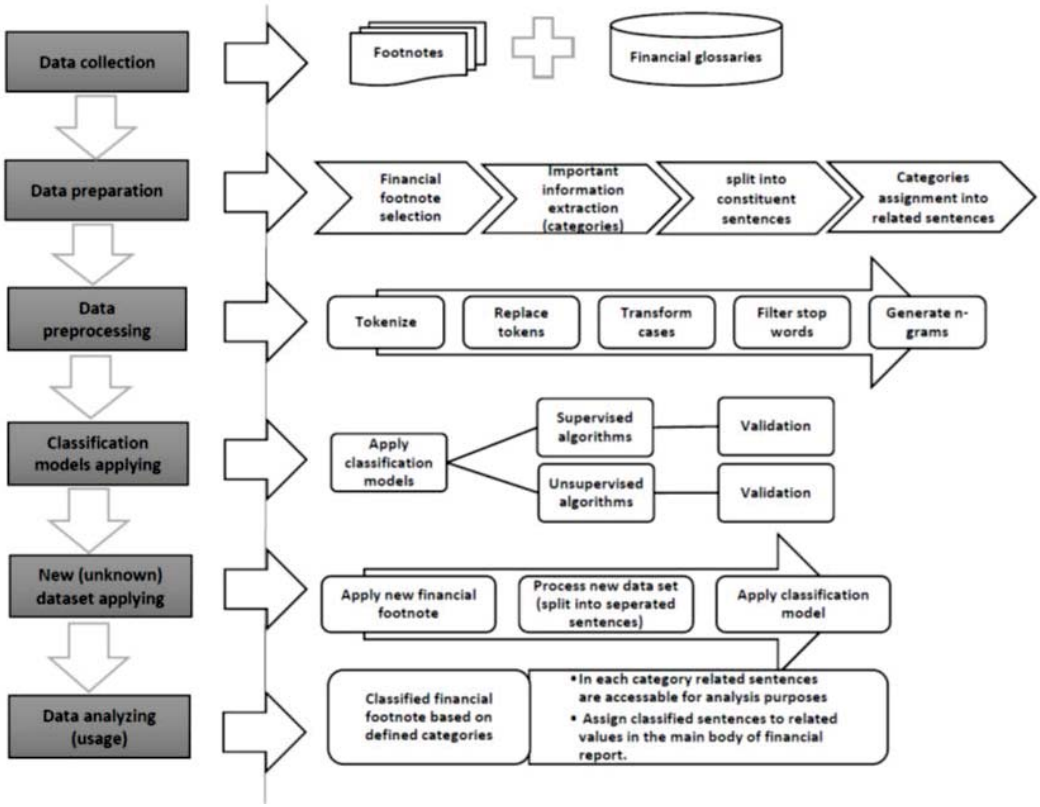

Fig. 2. Proposed text classification procedures for analyzing financial footnotes

information, allowing quick research on a company's financial information by reviewing registration statements and periodic reports filed on forms like 10-K (annual reports) and 10-Q (quarterly reports). Furthermore, existing financial glossaries and references are required to recognize important terms and issues not appearing in financial statement reports and which can be hidden in footnote lines.

3.2 Data preparation

We observed income tax footnotes from 120 different companies in different industries. First, we read some of the extracted income tax footnotes and compared them with existing financial glossaries and financial and accounting audit references in order to recognize financial terms which cannot be found in the main body of financial reports. The six identified categories related to income tax footnote are demonstrated in Table 1.

Table 1. Identified categories in income tax footnote

Income tax footnote	Deferred tax
	Effective tax rate
	Net operating loss
	Unrecognized tax benefit
	Taxation authority
	Valuation allowance

Afterwards, we split extracted footnotes into constituent sentences in order to assign all sentences to their related categories. This process supports not only recognizing rates and trends in footnotes' line, but also identifying soft information through sentences and their relations to financial figures in the main report body. Accordingly, we have extracted totally 1,290 sentences. Each sentence belongs to one particular pre-defined income tax category. It should be noted that alongside these six categories, other financial issues might appear in income tax footnotes, which can be categorized under one of these main categories. As an example, financial terms such as permanent and temporary differences, litigation charge, and net deferred tax asset are some income tax issues which can be classified under the deferred tax category.

3.3 Data pre-processing

The next step encompasses ordinary text mining preprocesses such as lower case transformation, stop words filtering, and tokenization. During the tokenization process, each sentence splits into a sequence of tokens in order to build a word vector for each sentence. The replace tokens operator allows replacing substrings within each token. To that end, the user can specify arbitrary patterns in the replace dictionary parameter. Table 2 shows some replacement examples, which have been applied to income tax footnotes analysis process.

Table 2. Sample replacement list in income tax footnote

Replace what	Replace by
jurisdiction	authority
exemption	deductibility
tax gain	tax benefit
unremitted	undistributed
uncertain tax position	unrecognized tax benefit
U.S.	united states
…	…

Afterwards, English stop words are removed from a document by checking every token that equals a stop word from the built-in stop word list. Finally, in the last pre-processing step, n-grams of tokens in a document are generated. A term n-gram is defined as a series of consecutive tokens of length n.; we defined n=3, because some financial terms consist of three relative words.

3.4 Apply classification model

Due to two main existing training concepts in the area of text classification methods, we applied both supervised and unsupervised machine learning algorithms. Although, in both methods, textual parts of financial footnotes are classified based on particular similarity criteria; in supervised algorithms we utilize training data sets (last two phases) including pre-defined classes in order to train the classification model and to classify new and unlabeled documents. As opposed to supervised algorithms, in unsupervised algorithms similar clusters are discovered without using training or labeled documents [9], [22], [29], and [30]. Both algorithms should be validated to observe the performance results, which assist users to employ more appropriate techniques for financial footnote analysis [15].

3.5 Apply new dataset

In this phase, the identified algorithm will be used for new financial footnotes. We processed this phase by splitting the financial text into constituent sentences. As an example, if our new income tax footnote is a document text file including 50 sentences, it will be split into 50 separated text files. Each one consists of one sentence and serves as input data set to our process. Later on, the approved classification model will be applied in order to classify income tax footnotes based on the recognized categories.

3.6 Data analyzing (usage)

It should be mentioned that the final purpose behind text classification processes is to facilitate text analyzing in order to recognize hidden textual patterns with reduced manual efforts [4]. To do this, we focus on utilization of classified financial footnotes in the last phase of the research, which consists of representation of classified output in each category. Related sentences in each category are accessible for further analysis, which helps analysts to extract required terms and related sentences without reading the whole footnote. Another usage of financial footnotes classification is the assignment of classified sentences to related values in the main body of financial statements like balance sheet or income statements.

4 Implementation and results

The implementation has been done by using the tool Rapid Miner[2], which is an open source tool for data mining and predictive analytics. All the above-discussed steps have been performed in this tool. As mentioned earlier, in order to obtain the best results, we applied both supervised and unsupervised algorithms to recognize the more appropriate one for our research goal.

The basic objective of all supervised classifiers is to recognize the degree of similarity between pre-classified training data and a new, unlabeled data set [6], [10], [23]. To do this, we preprocessed our training documents and trained the supervised model based on defined steps. We tested various supervised classification algorithms such as K-NN, Naïve Bayes, Support Vector Machine (SVM), and decision tree. After checking performance measures like accuracy, run time, absolute error, and Root Mean Square Deviation (RMSE)[3], we ended up to Naïve Bayes as the most appropriate supervised classifier for financial footnote analysis with 82.86% accuracy and the most minimum run time (Table 3).

Table 3. Results of supervised algorithm

Supervised algorithm	Run time	Accuracy	Absolute error	RMSE
K-NN	7s	81.82%	0.183	0.362
Naïve Bayes	4s	82.86%	0.171	0.414
SVM	28s	79.22%	0.784	0.786
Decision tree	1m45s	90.65%	0.136	0.280

In terms of unsupervised algorithms, we performed K-means as a clustering technique, which is used for extracting information from unlabeled data. According to an experimental study by Steinbach et al., among clustering algorithms, K-means method has better performance for text clustering. They compared two main approaches in clustering techniques: agglomerative hierarchical clustering and different types of K-means [27]. They argued that the hierarchical clustering does not work well because most of the time (like our case) it cannot be fixed by the hierarchical scheme. In contrast, K-means groups objects together that are similar to each other and dissimilar to the objects belonging to other clusters [22]. For this method of clustering we start by

[2] http://www.rapidminer.com

[3] Root Mean Square Error is a frequently used measure of the differences between value predicted by a model or an estimator and the values actually observed

deciding how many clusters (K) we would like to form based on our data. We set K to 6 based on our classes in supervised classification. Different performance criteria such as run time, Davies-Bouldin and average within centroid distance are supervised (Table 4).

Table 4. Results of k-means unsupervised algorithm

Unsupervised algorithm	K	Run time	Ave. within centroid distance	Davies-Bouldin
K-means	6	7min 20s	-0.827	-4.763

According to the obtain results, applying unsupervised algorithms can avoid relatively large amount of supervision and manual tasks [9], [29]. Nonetheless, regarding to the financial analysis purposes where searching particular financial terms in specified footnotes is significant, classification through supervised algorithms yield more reliable results. Furthermore, applying unsupervised methods are more appropriate n case of working with not very clean texts such as large amount of text data created by dynamic applications such as social networks [2]. Fig. 3 summarizes strengths and weaknesses of each machine learning algorithm based on this research criteria.

	Strengths	**Weaknesses**
Supervised algorithm	Classification based on pre-defined and specific required categories High accuracy Short run time	Complex data training process
Unsupervised algorithm	Not required training process Acceptable accuracy in related clusters	Clustering process happens not based on desired categories Long run time

Fig. 3. Comparison between supervised and unsupervised algorithms in this research

Regarding the analyzing step in terms of facilitating the usage of financial footnotes, two main benefits are notable. The first contribution of performing the text classification process is that analysts or any stakeholders can access financial footnotes

Table 5. Extracted sentences related to unrecognized tax benefit from income tax footnote of a company

Income tax footnote	Number of sentences	Related sentence
Unrecognized tax benefit	1	Differences between tax positions taken or expected to be taken in a tax return and the net benefit recognized and measured pursuant to the interpretation are referred to as "unrecognized benefits.
	2	A liability is recognized for unrecognized tax benefit because it represents an enterprise's potential future obligation to the taxing authority for a tax position that was not recognized as a result of applying the provisions of ASC 740.
	3	If applicable, interest costs related to the unrecognized tax benefits are required to be calculated and would be classified as "Other expenses – Interest" in the statement of operations.
	4	As of October 31, 2014 and October 31, 2013, no liability for unrecognized tax benefits was required to be reported.
	5	The Company does not expect any significant changes in its unrecognized tax benefits in the next year.

in a regular and ordered fashion. Table 5 shows an example of the output report for one of the major income tax category so called unrecognized tax benefit which consists of extracted sentences from income tax footnote of a company. Another benefit of the text mining approach to facilitate financial footnote analysis is the mapping of extracted sentences to related values in the main body of financial statements in order to perform a fully automatic process. In our case, all extracted sentences are related to "deferred income tax" (sometimes appears separately as deferred tax asset and deferred tax liability), which is one of the balance sheet report's term. As a result, the proposed text mining approach can be determined as an appropriate semi-automatic solution to overcome time-consuming manual analysis of financial footnotes by classifying footnotes based on pre-defined categories automatically. Thus, analysts can directly access required sentences of related categories instead of reading the entire text document in a real-time. The proposed text classification method can be seen as a starting point for further research to evaluate practically the influence of this solution on analysts' workflow in financial analysis process in a real world. We acknowledge that proposed text mining approach could still be improved by identifying other relevant footnotes items and by mapping extracted sentences to values in the main body of financial statements in order to implement fully automatic process.

5 Conclusion

Despite significant developments in fields of financial analysis e.g. based on XBRL-formatted data, the textual parts of financial reports, which are critical for comprehensive financial analysis, are still dependent on time-consuming manual procedures.

We addressed this challenge in this paper by proposing a text mining approach to be able to automate financial footnotes analysis and facilitate the usage of footnotes information by applying text classification methods. We have chosen income tax as a footnote example. We classified it sentence by sentence with supervised and unsupervised classification algorithms and implemented the proposed solution using an analytics tool. In terms of accuracy and run time, a supervised algorithm gains better results, however it requires a precise and careful data training process.

Comparing to existing approaches in terms of financial footnote analysis, our preliminary results show that text mining could be an appropriate semi-automatic solution to facilitate manual analysis of unstructured parts of financial reports. As a matter of fact, the text mining approach helps users to access required soft information as a separate sentence based on each financial pre-defined category.

It is also of interest in this research to develop this solution by adding more capabilities thereby to map extracted sentences into related figures in financial statements.

However, due to some limitations, it is not practically implemented, yet. One of the limitations is that footnote parts are normally received by analysts or auditors as separate documents and are not attached to main financial statements. This makes it difficult to map financial sentences into figures in financial statements. Another limitation is that some financial footnote sentences carry only some informative and explanatory data about financial terms and status of the organization and therefore cannot be connected to any financial values in the main body.

There is still room for an improvement concerning the automatic financial footnote analysis by evaluating this approach thorough different case studies and expert interviews to demonstrate the usefulness of this approach in accordance to analysts' processes. This is one of the future research areas of this research.

6 Reference

[1] Abahoonie, E. et al. (2013, December) Tax accounting services: Income tax disclosure, available at: http://www.pwc.com

[2] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." Mining Text Data. Springer US, 2012. 77-128.

[3] Beattie, V., McInnes, B. & Fearnley, S., 2004. A methodology for analyzing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. Accounting Forum, 28(3), pp.205–236.

[4] Botzenhardt, A., Witt, A., & Maedche, A. (2011). A Text Mining Application for Exploring the Voice of the Customer. AMCIS 2011 Proceedings.

[5] Brent, E.; Atkisson, C., 2011 " A standard-based automated coding program for unstructured text" Veyor ® Survey presentation, University of Surrey, USA

[6] Chaovalit, Pimwadee, and Lina Zhou. (2005) "Movie review mining: A comparison between supervised and unsupervised classification approaches." System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference.

[7] Chapman, P.; Clinton, J.; Kerber, R.; Khabanza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000) " CRISP-DM- step by step data mining guide." SPSS, Chicago, IL.

[8] Crammer, K.; Dredze, M.; Ganchev, K.; Talukdar, P. P.; Carroll, S., 2007, "Automatic code assignment to medical text" In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pp. 129-136. Association for Computational Linguistics.

[9] Dharmadhikari, S. C., & Kulkarni, P. (2011). Empirical Studies on Machine Learning Based Text Classification Algorithms. Advanced Computing: An International Journal, 2(6), 161–169.

[10] Ghosh, S., Roy, S. & Bandyopadhyay, P.S.K. (2012) A tutorial review on Text Mining Algorithms. , 1(4), pp.223–233.

[11] Graham, John R. & Raedy, Jana S. & Shackelford, Douglas A., 2012. "Research in accounting for income taxes," Journal of Accounting and Economics, Elsevier, vol. 53(1), pages 412-434

[12] Groth, S.S.; Muntermann, J., Supporting investment management processes with machine learning techniques, in: H.R. Hansen, D. Karagiannis, H.-G. Fill (Eds.), Proceedings of the 9. Internationale Tagung Wirtschaftsinformatik, Österreichische Computer Gesellschaft, Wien, Austria, 2009.

[13] Hagenau, M.; Liebmann, M.; Neumann, D.: Automated news reading: Stock price prediction based on financial news using context-specific features. System Science (HICSS), 2012 45th Hawaii International Conference on IEEE. (2012)

[14] Heidari, M.; Felden, C.: Toward Supporting Analytical Tasks in Financial Footnotes Analysis- A State of the Art, In: Multikonferenz Wirtschaftsinformatik MKWI 2014, Paderborn, Deutschland, 26-28, Februar, 2014.
[15] Hotho, A., Andreas, N., Paaß, G., & Augustin, S. (2005). A Brief Survey of Text Mining, 1–37.
[16] Hussainey, K. S. M. (2004). A study of the ability of (partially) automated disclosure scores to explain the information content of annual report narratives for future earnings, Doctoral dissertation, University of Manchester.
[17] Kloptchenko, Antonina; (2004) "Toward Automatic Analysis of Financial Reports- Readability of Quarterly reports & company's financial performance", AMCIS 2004 Proceedings, paper 412.
[18] Kloptchenko, Antonina; Eklund, Tomas; Back, Barbro; Karlsson, Jonas; Vanharanta, Hannu; and Visa, Ari (2002) "COMBINING DATA AND TEXT MINING TECHNIQUES FOR ANALYZING FINANCIAL REPORTS", AMCIS 2002 Proceedings. Paper 4.
[19] Leder, Michele; (2003), "Financial Fine Print: Uncovering a Company's True Value", John Wiely & Sons Inc., New Jersey, 17p.
[20] Li F., The information content of forward-looking statements in corporate filings — a Naïve Bayesian machine learning approach, Journal of Accounting Research 48 (5) (2010) 49–102.
[21] Miner, G.; Elder, J.; Nisbet, B.; Delen, D.; Fast, A.; Hill, T. (2012) Practical text mining and statistical analysis for non-structured text data applications. Massachusetts, USA: Elsevier
[22] Ozgür, Arzucan. (2004) Supervised and unsupervised machine learning techniques for text document categorization. Diss. Bogaziçi University.
[23] Padhye, Apurva. (2006) Comparing Supervised and Unsupervised Classification of Messages in the Enron Email Corpus. Diss. UNIVERSITY OF MINNESOTA, 2006.
[24] Pakhomov, Serguei VS, James D. Buntrock, and Christopher G. Chute. (2006) "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." Journal of the American Medical Informatics Association 13(5), pp. 516-525.
[25] Putra, Lie Dharma (2008) Understanding footnotes to financial statements, http://accounting-financial-tax.com/2008/08/understanding-footnotes-to-financial-statement
[26] Rapid miner, http://www.rapidminer.com
[27] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." KDD workshop on text mining. Vol. 400. No. 1. (2000)
[28] Tergesen, Anne (2002) "Getting to the bottom of a company's Debt", Business Week, 10/14/2002, Issue 3803, p156-158.
[29] Tsarev, Dmitry, Mikhail Petrovskiy, and Igor Mashechkin. (2013) "Supervised and Unsupervised Text Classification via Generic Summarization."
[30] Wagstaff, Kiri Lou. (2002) intelligent clustering with instance-level constraints. Diss. Cornell University.
[31] Weglarz, Geoffrey (2004), "Two worlds of data: unstructured and structured", DM Review, September2004