Visualizing Word Categorization of Corporate Annual Reports Using Self-Organizing Maps

Petr HÁJEK, **Vladimír OLEJ**
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice
Pardubice, 532 10, Czech Republic
petr.hajek@upce.cz, vladimir.olej@upce.cz

ABSTRACT

An increasing proportion of corporate information takes the form of unstructured or semi-structured text. Annual reports are one of the most important external documents that discuss corporate performance and managerial priorities. This paper is aimed at visualizing textual information in corporate annual reports. Several word categorization schemes are used to extract the tone of text. Self-organizing maps are employed to map the word categories. The results indicate that the annual reports of U.S. companies differ in terms of the tone emphasized. In addition, we show that the tone reflects both current financial performance and future change in the performance.

Keywords: Annual Report, Word Categories, Textual Analysis, Self-organizing Map.

1. INTRODUCTION

Corporate annual reports are semi-structured text documents that offer a detailed picture of a company's business, risks, operating and financial results and their drivers. Management uses the reports to discuss the perspective on the business results. Thus, the reports provide important information to external and internal users. The information comprises both quantitative business indicators and textual content. Actually, about 80% of the information takes the latter form. Automatic tools for the textual analysis of corporate documents are therefore necessary to provide additional support to stakeholder' decision-making process [6,12,18]. In addition, the textual analysis has become a central issue in financial analysis [24], significantly enhancing our understanding of agents' behavior in financial markets.

Two general approaches have been reported in the literature, word categorization (bag of words) [20] and statistical methods [23]. While the former approach requires appropriate dictionaries of terms for the categories, the latter one needs the likelihood ratios of subjective tone classification. We decided to use the word categorization approach mainly because several word categorization schemes are available which are specifically designed for financial domain.

The categories can be, however, interrelated and it is their combination that provides the comprehensive picture of textual content. However, the research to date has been limited to univariate analyses of word categories. The aim of this paper is to examine and visualize the relations between word categories. Self-organizing maps (SOM) offer a convenient tool for mapping this high-dimensional word categorization onto a low-dimensional space that can be easily visualized. Recent studies using SOMs in various domains [13,14] have shown that SOMs are well suited for the representation of nonlinear relationships between input attributes. In addition, current and future financial performance indicators can be mapped onto the constructed two-dimensional grid. In this paper we argue that the overall tone of the annual report may reveal the anticipation of financial performance development from the managerial point of view. We also hypothesize that current weak financial performance may be hidden by using specific word categories.

The rest of this paper is organized as follows. First, we review previous literature on textual analysis of corporate annual reports. Next, we briefly present the process of text pre-processing and methods used for the visualization. Finally, the results of experiments are provided and discussed.

2. TEXTUAL ANALYSIS OF CORPORATE ANNUAL REPORTS – A LITERATURE OVERVIEW

Kearney and Liu [20] reviewed the literature on the use of textual analysis to extract sentiment from sources such as corporate disclosures, media articles and internet postings. This review suggests that annual reports represent a valuable source of internal knowledge. However, managers may be tempted to manipulate investors' judgments. Thus, this source is particular relevant to examining the relationships between textual content and individual firm financial performance.

As highlighted above, word categorization and statistical (machine learning) methods are mostly used for the textual analyses of corporate annual reports. The mostly used general dictionary categories used in the formed approach included: General Inquirer [28], Harvard IV-4 [3,19] and Diction 5.0 [5,27]. However, this approach has been criticized due to the context-sensitivity of the dictionaries. Therefore, dictionaries specific for financial domain have been developed to address this issue [24].

Alternatively, Qiu et al. [26] employed machine learning methods to predict next year's earnings per share. Huang et al. [18] found that investors react more strongly to negative than to positive text. Magnusson et al. [25] employed (SOM) to visualize the changes in the writing style of the annual reports of telecommunication companies. The results suggested that for a well-performing company, the tone of the report was positive, optimistic and active. On the contrary, in the case of weak performers, the tone of financial report was negative, less optimistic and conservative. Similarly, Chen et al. [2] reported that decreasing earnings are often accompanied by ambiguous and mild statements in the reporting year and that increasing earnings are stated in an assertive and positive way. Li [23] showed that firms with better current performance, lower accruals, smaller size, lower market-to-book ratio, less return volatility, lower MD&A Fog index, and longer history tend to have a more positive statement tone.

Feldman et al. [4] and Hajek et al. [8] suggested that stock market significantly reacts to the change in the positive/negative sentiment. Hajek and Olej [7] reported that credit rating prediction can be improved with sentiment information extracted from annual reports. Huang et al. [16] estimated abnormal positive tone and showed that this tone has a positive stock return effect at the earnings announcement and a delayed negative reaction in the one and two quarters afterward. Hajek and Olej [11] used the WordNet ontology and singular value decomposition to extract concepts from annual reports and related the concepts to future financial indicators such as stock market risk and profitability. However, far too little attention has been paid to examining the role of sentiment information in financial distress prediction, including both credit rating evaluation [9] and bankruptcy/non-bankruptcy classification [10].

3. METHODOLOGY

Our study was designed and implemented in two broader steps, (1) textual data collection and pre-processing and (2) data visualization using SOM.

3.1 Textual Data Collection and Pre-Processing

First, we collected annual reports (10-Ks documents to be specific) of 611 US companies (year 2008) from U.S. Securities and Exchange Commission EDGAR System and pre-processed them linguistically. This step included tokenization and lemmatization. Second, the tagged lemmas (term candidates) were compared with the following dictionaries: (1) financial dictionary developed by [24]; and (2) Diction 5.0 [16]. The former one considered the following categories of terms:

- negative,
- positive,
- uncertainty,
- litigious,
- modal strong, and
- modal weak.

The frequency of net positive words was calculated as the positive term count minus the count for negation. The latter dictionary provided 5 additional categories:

- certainty (resoluteness, inflexibility, and completeness),
- activity (movement, change, etc.),
- optimism (highlighting the positive entailments of persons or events),
- realism (tangible, immediate, recognizable matters that affect people's everyday lives), and
- commonality (the agreed-upon values of a group).

More specifically, these categories were calculated as:

certainty = (tenacity + leveling + collectives + insistence) – (numerical + ambivalence + self-reference + variety), (1)

optimism = (praise + satisfaction + inspiration) – (blame + hardship + denial), (2)

realism = (familiarity + spatial awareness + temporal awareness + present concern + human interest + concreteness) – (past concern + complexity), (3)

activity = (aggression + accomplishment + communication + motion) – (cognitive + passivity + embellishment), (4)

commonality = (centrality + cooperation + rapport) – (diversity + exclusion + liberation). (5)

We expected that financially well-performing companies used generally more positive and optimistic tone of language in the annual reports. On the other hand, a more active language was anticipated in the case of poorly performing companies in need of actions aimed to improve their financial situation.

In the next stage, an average weight was calculated for each category using the *tf.idf* term weighting scheme:

$$w_{i,j} = \begin{cases} \dfrac{(1+\log(tf_{i,j}))}{(1+\log(a))}\log\dfrac{N}{df_i} & \text{if } tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where N represents the total number of documents in the sample, df_i denotes the number of documents with at least one occurrence of the i-th term, $tf_{i,j}$ is the frequency of the i-th term in the j-th document, and a denotes the average term count in the document. For example, Table 1 shows the terms with the highest weights $p_{i,j}$ in two selected categories – positive and negative.

Table 1. Terms with the highest weights – positive and negative categories.

positive term	$p_{i,j}$	negative term	$p_{i,j}$
allianc	0.723	plaintiff	1.010
innov	0.692	complaint	0.947
leadership	0.618	dismiss	0.918
stabl	0.617	infring	0.851
reward	0.615	deplet	0.824
resolv	0.613	divestitur	0.806
collabor	0.612	closur	0.796
superior	0.593	accid	0.786
satisfact	0.588	downgrad	0.769
attain	0.587	deni	0.760
transpar	0.573	foreclosur	0.737
integr	0.571	deficit	0.736
great	0.570	nonperform	0.734
excel	0.566	abandon	0.733
attract	0.554	antitrust	0.730
stronger	0.553	hazard	0.720
premier	0.534	catastroph	0.708
popular	0.529	purport	0.705
stabil	0.528	unfund	0.693
accomplish	0.520	defect	0.691

3.2 Visualizing Word Categorization using SOM

The weights $p_{i,j}$ of the sentiment categories (negative, positive, … , realism, commonality) were used as inputs to the SOM. The SOMs are based on competitive learning strategy. The input layer serves the distribution of the input patterns $\boldsymbol{p}_\mathrm{i}$. The neurons in the competitive layer serve as representatives (Codebook Vectors), and they are organized into topological structure (most often as a two-dimensional grid) which designates the neighboring network neurons. First, the distances d_j are computed between pattern $\boldsymbol{p}_\mathrm{i}$ and synapse weights $\boldsymbol{w}_\mathrm{i,j}$. The winning neuron j^* (Best Matching Unit, BMU), for which the distance d_j from the given pattern $\boldsymbol{p}_\mathrm{i}$ is minimum, is chosen. The output of this neuron is active, while the outputs of other neurons are inactive. When the representatives $\boldsymbol{w}_\mathrm{i,j}$ are identified, the representative $\boldsymbol{w}_{\mathrm{i,j}*}$ of the BMU is assigned to each vector $\boldsymbol{p}_i$.

In the learning process of the SOM, it is necessary to define the concept of neighborhood function, which determines the range of cooperation among the neurons, i.e. how many representatives $\boldsymbol{w}_\mathrm{i,j}$ in the neighborhood of the BMU will be adapted, and to what degree. Activity of the neurons and neighborhood are described in [22]. After the BMUs are found, the adaptation of synapse weights $\boldsymbol{w}_\mathrm{i,j}$ follows.

We employed a sequential training algorithm to adapt the weights of the two-dimensional SOM. The parameters of the sequential learning algorithm were set in the following manner: learning rate (initial = 0.5, final = 0.05), Gaussian neighborhood function with radius (initial = 3, final = 1), and the number of iterations = 25.

In the final step, we mapped the financial indicator Altman's Z-score (from the year 2010) [1] on the trained SOM. The Z value is in range -4 to +8, where a higher value denotes better financial situation of a company. Specifically, $Z \geq 3$ is called "safe zone"; $1.80 \leq Z \leq 2.99$ "grey zone"; and $Z < 1.80$ "distress zone" (serious financial problems).

4. EXPERIMENTAL RESULTS

Previous studies [15] have shown that various industries have specific determinants of financial performance and, in addition, current economic and technological situation in the industry is reflected across the companies in this industry. Thus, different language tone may be used in the annual reports of different industries. Therefore, we examined only two industries, manufacturing (SIC codes 20-39) and services (SIC codes 70-89). Finance and mining industries were excluded to prevent problems with different financial performance evaluation and the remaining industries were discarded from the data due to low frequencies. As a result, two datasets were available with 262 companies for manufacturing industry and 105 for services, respectively. On average, annual reports of service companies were more negative, less positive and uncertain and used more weak modal words (Fig. 1). At the same time, companies in this industry utilized more certain, realistic and active vocabulary (Fig. 2).

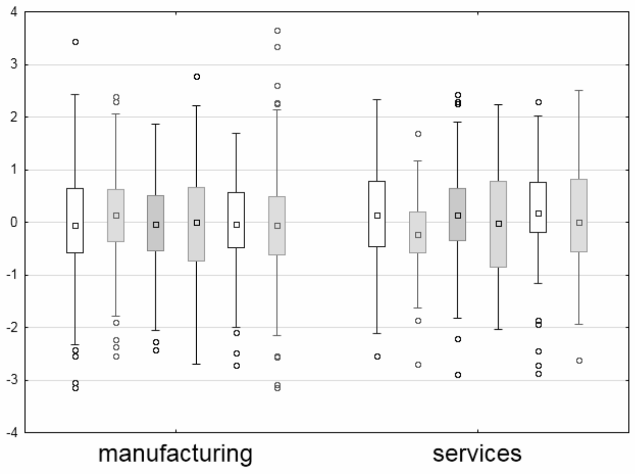

Fig. 1. Standardized values of word categories negative, positive, uncertainty, litigious, weak modal and strong modal. The box plots depict median, lower and upper quartile and extreme values.

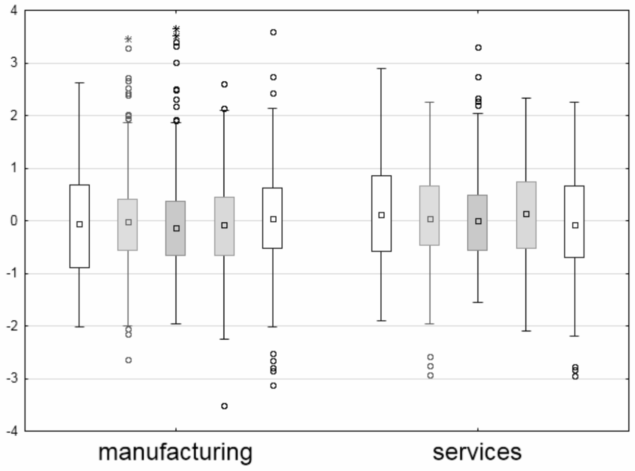

Fig. 2. Standardized values of word categories certainty, optimism, realism, activity and commonality. The box plots depict median, lower and upper quartile and extreme values.

We tested several structures of the SOM in order to both minimize quantization error and preserve topology [21]. The experiments with the SOM structures suggested that the best topology preservation (with a satisfactory quantization error) was achieved using the 10 × 6 sheet topology for both industries. The sheet topology enabled further utilization of SOM visualization capabilities, namely U-matrixes and component planes.

The Euclidean distances in the trained SOMs (**U**-matrixes) are depicted in Fig. 3. In the **U**-matrixes, each component denotes a distance between two adjacent neurons. These manufacturing companies made more compact clusters of representatives (with lower distances – dark color).

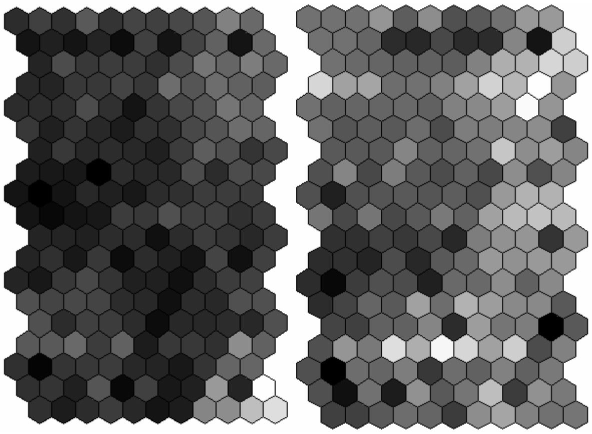

Fig. 3. **U**-matrixes for manufacturing (left) and service companies (right), where light colors indicate higher distances.

Fig. 4 and Fig. 5 depict component planes, where each neuron is colored according with the value of synapse weight. Thus, emerging data patterns and relations between word categories can be detected. These figures show that manufacturing companies in distress zone used: (1) both negative and positive vocabulary, and (2) active and common words. This suggests that managers attempted to balance the anticipated negative financial situation by using more positive tone and propose active policies. On the other hand, the bottom right distressed representative used vocabulary which was completely different from all the other companies. They used only few negative, positive, uncertain and weak modal words. At the same time, their vocabulary was more optimistic.

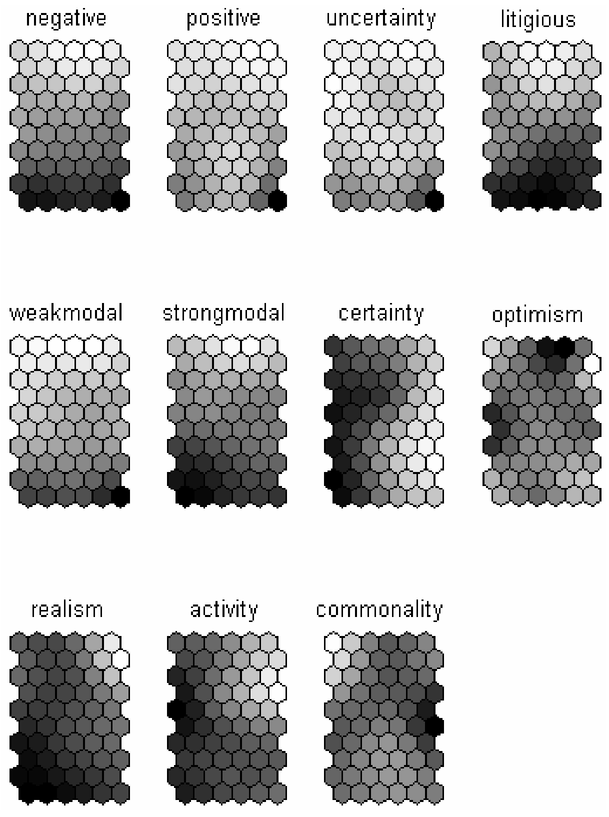

Fig. 4. Values of word categories for manufacturing companies' representatives.

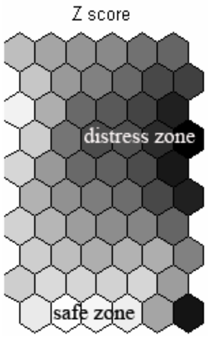

Fig. 5. Financial performance mapped onto trained SOM – manufacturing companies.

Manufacturing companies in the safe zone used also less negative words, but their annual reports were more positive, less litigious and rather realistic and active, respectively.

Distressed service companies, on the hand, used more litigious, certain and active words (Fig. 6 and Fig. 7). Two different clusters of well performing companies were observed in the map, the top left and the bottom left, respectively. The top left companies' sentiment was more negative and positive at the same time. They were also more uncertain and less optimistic about their situation. On the contrary, the bottom left ones used both less negative and less positive vocabulary. They were also more optimistic and used more realistic and active tone. These results differ from some published studies [25]. These differences can be explained by a larger dataset used in our study and rather specific industry used by [25].

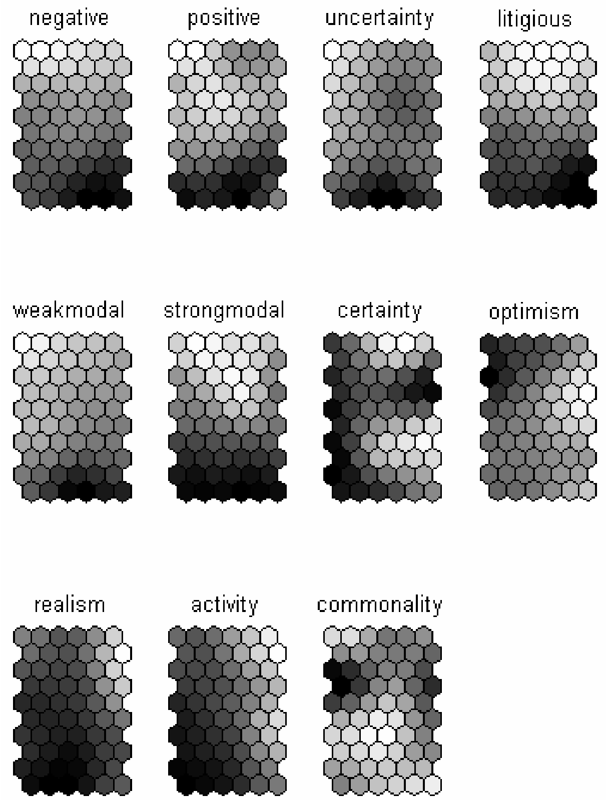

Fig. 6. Values of word categories for service companies' representatives.

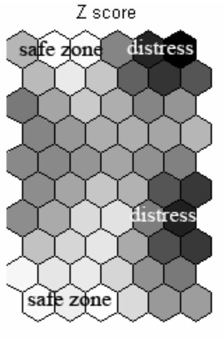

Fig. 7. Financial performance mapped onto trained SOM – service companies.

5. CONCLUSIONS

To sum up, our findings suggest that the word categories used in annual reports can be easily visualized using SOM and, in addition, the results can be compared with future financial performance.

Returning to the hypotheses posed at the beginning of this study, it is now possible to state that there weak performing companies use specific word categories depending on industry convention. For manufacturing companies, it was rather the balanced relation between positive and negative tone that differentiated the weak performers from well performing companies. On the other hand, our findings suggest that weak performers use more litigious, less realistic and less active word categories. For service companies, the situation was partially different. On one hand, a more litigious language was also observed for weak performing service companies. On the other hand, these companies used more active tone when compared with distressed manufacturing companies.

Taken together, the most obvious finding to emerge from this study is that annual reports using both realistic and active tone along with less litigious words show on well performing companies in both industries. These findings add substantially to our understanding of the role of qualitative information in financial distress forecasting. This paper has also underlined the importance of using multiple word categorizations when attempting to anticipate future financial performance.

In future, we consider mapping additional financial distress indicators on the SOM. Furthermore, the current categorizations are mostly limited by assigning equal weights to each word within category. Therefore, a further study could assess the role of the most important words in each category.

The experiments in this study were carried out in Statistica 10 (text pre-processing) and SOM Toolbox Matlab 7 (SOM) in MS Windows 7 operation system.

ACKNOWLEDGMENT

This work was supported by the scientific research project of the Czech Sciences Foundation Grant No: 13-10331S.

6. REFERENCES

[1] E.I. Altman, **Predicting Financial Distress of Companies: Revisiting the Z-Score and Zeta Models**, New York: New York University, 2000.

[2] C.L. Chen, C.L. Liu, Y.C. Chang, H.P. Tsai, "Opinion Mining for Relating Subjective Expressions and Annual Earnings in US Financial Statements", **Journal of Information Science and Engineering**, Vol. 29, No. 3, 2013.

[3] J.E. Engelberg, A.V. Reed, M.C. Ringgenberg, "How Are Shorts Informed?: Short Sellers, News, and Information Processing", **Journal of Financial Economics**, Vol. 105, No. 2, 2012, pp. 260-278.

[4] R. Feldman, S. Govindaraj, J. Livnar, B. Segal, "Management's Tone Change, Post Earnings Announcement Drift and Accruals", **Review of Accounting Studies**, Vol. 15, 2010, pp. 915-953.

[5] S.P. Ferris, G.Q. Hao, M. Liao, "The Effect of Issuer Conservatism on IPO Pricing and Performance", **Review of Finance**, Vol. 7, No. 3, 2013, pp. 993-1027.

[6] H. Gomes, M. de Castro Neto, R. Henriques, "Text Mining: Sentiment Analysis on News Classification", **8th IEEE Iberian Conference on Information Systems and Technologies** (CISTI), 2013, pp. 1-6.

[7] P. Hajek, V. Olej, "Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines", **Communications in Computer and Information Science**, Vol. 384, 2013, pp. 1-10.

[8] P. Hajek, V. Olej, R. Myskova, "Forecasting Stock Prices using Sentiment Information in Annual Reports - A Neural Network and Support Vector Regression Approach", **WSEAS Transactions on Systems**, Vol. 10, No. 4, 2013, pp. 293-305.

[9] P. Hajek, V. Olej, "Predicting Firms' Credit Ratings Using Ensembles of Artificial Immune Systems and Machine Learning–An Over-Sampling Approach", **IFIP Advances in Information and Communication Technology**, Vol. 436, 2014, pp. 29-38.

[10] P. Hajek, V. Olej, "Defuzzification Methods in Intuitionistic Fuzzy Inference Systems of Takagi-Sugeno Type. The Case of Corporate Pankruptcy Prediction", **11th International Conference on Fuzzy Systems and Knowledge Discovery** (FSKD'14), Xiamen, China, 2014, pp. 240-244.

[11] P. Hajek, V. Olej, "Comparing Corporate Financial Performance and Qualitative Information from Annual Reports using Self-organizing Maps", **10th International Conference on Natural Computation** (ICNC'14), Xiamen, China, 2014, pp. 93-98.

[12] P. Hajek, V. Olej, R. Myskova, "Forecasting Corporate Financial Performance using Sentiment in Annual Reports for Stakeholders' Decision-Making", **Technological and Economic Development of Economy**, 2014, in press.

[13] P. Hajek, V. Olej, "Modeling of Relationships between Economic Performance and Environmental Quality by SOM and Growing Hierarchical SOM - The Case of the Czech Republic Regions", **WSEAS Transactions on Environment & Development**, Vol. 9, No. 3, 2013, pp. 220-229.

[14] P. Hajek, R. Henriques, V. Hajkova, "Visualising Components of Regional Innovation Systems using Self-Organizing Maps - Evidence from European Regions",

Technological Forecasting and Social Change, Vol. 84, 2014, pp. 197-214.

[15] P. Hajek, "Credit Rating Analysis using Adaptive Fuzzy Rule-Based Systems: An Industry-Specific Approach", **Central European Journal of Operations Research**, Vol. 20, No. 3, 2012, pp. 421-434.

[16] R.P. Hart, "Redeveloping DICTION: Theoretical considerations", M.D. West (Ed.), **Theory, Method, and Practice in Computer Content Analysis**, 2001, pp. 43-60.

[17] X. Huang, S.H. Teoh, Y. Zhang, "Tone Management", **The Accounting Review**, Vol. 89, No. 3, 2014, pp. 1083-1113.

[18] A. Huang, A. Zang, R. Zheng, "Evidence on the Information Content of Text in Analyst Reports", **The Accounting Review**, in press, 2014, doi: http://dx.doi.org/10.2308/accr-50833.

[19] N. Jegadeesh, D. Wu, "Word Power: A New Approach for Content Analysis", **Journal of Financial Economics**, Vol. 110, No. 3, 2013, pp. 712-729.

[20] C. Kearney, S. Liu, "Textual Sentiment in Finance: A Survey of Methods and Models", **International Review of Financial Analysis**, Vol. 33, 2014, pp. 171-185.

[21] K. Kiviluoto, "Topology Preservation in Self-Organizing Maps", **IEEE International Conference on Neural Networks**, 1996, pp. 294-299.

[22] T. Kohonen, **Self-organizing Maps**, Berlin: Springer Verlag, 2001.

[23] F. Li, "The Information Content of Forward-Looking Statements in Corporate Filings - A Naïve Bayesian Machine Learning Approach", **Journal of Accounting Research**, Vol. 48, No. 5, 2010, pp. 1049-1102.

[24] T. Loughran, B. McDonald, "When is a Liability not a Liability? Textual Analysis, Dictonaries, and 10-Ks", **The Journal of Finance**, Vol. 66, No. 1, 2011, pp. 35-65.

[25] C. Magnusson, A. Arppe, T. Eklund, B. Back, H. Vanharanta, A. Visa, "The Language of Quarterly Reports as an Indicator of Change in the Company's Financial Status", **Information and Management**, Vol. 42, No. 4, 2005, pp. 561-574.

[26] X.Y. Qiu, P. Srinivasan, Y. Hu, "Supervised Learning Models to Predict Firm Performance with Annual Reports: An Empirical Study", **Journal of the Association for Information Science and Technology**, Vol. 65, No. 2, 2014, pp. 400-413.

[27] J.L. Rogers, A. Van Buskirk, S.L. Zechman, "Disclosure Tone and Shareholder Litigation", **The Accounting Review**, Vol. 86, No. 6, 2011, pp. 2155-2183.

[28] B. Twedt, L. Rees, "Reading between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports", **Journal of Accounting and Public Policy**, Vol. 31, No. 1, 2012, pp. 1-21.