Analyzing Customer Complaints: A Web Text Mining Application

Esra Kahya Özyirmidokuz[1,a,*] and Mustafa Hakan Özyirmidokuz[2,b]

[1] Kayseri Vocational College, Erciyes University, Kayseri, Turkey
[2]Bosch Thermotechnic, Ankara, Turkey
[a]esrakahya@erciyes.edu.tr, [b]hakan.ozyirmidokuz@tr.bosch.com
*Corresponding author

Keywords: Information management, Data mining, Text mining, Web mining, Heating systems firms, Managing customer complaints

Abstract. The amount and the complexity of Web pages have seen dramatic increases, as has the information contained within Web pages. In today's world, firms' Web data must be analyzed to gain a competitive advantage in the topic sector. Recently, Web text mining (TM) has gained much importance because of its increasing use in business applications for understanding and predicting valuable information. It plays a key role in organizing the huge amount of Web unstructured (textual) data and condensing it into valuable knowledge.

Customer complaints give businesses valuable information about how they need to improve. This paper addresses a Web TM application to extract useful, interesting and hidden knowledge for heating systems firms to implement in competition.

The top seven heating systems firms in Turkey are analyzed in terms of customers' complaints. Data are collected from a complaint Website with the RapidMiner Web Mining Tool. Then the data are transformed to a collection of documents by generating a document for each record. Every complaint is transformed to a document. These documents, which were collected over the period between December 2012 and October 2013 are analyzed with TM techniques. Summarization, tokenization, stemming, and filtering are also used. In addition, the similarities of firms on this subject are determined. Not only do we extract knowledge about the customers but also the firms in the sector.

1. Introduction

The world has turned into a small village. In addition, nowadays, the price advantage prevails in the world of sales. The only way to stand out from the competition lies in the after-market. In order to satisfy the customer we must analyze all kinds of data. This is because social media and Internet media have put the consumer in a stronger position. Customers are really the kings in the big data world.

Customer satisfaction is not an absolute scenario, but very much depends on interactions, feedback, praise, and complaints. Complaints have to be looked at in a constructive, positive and professional perspective [1]:

ISBN: 978-605-64453-0-9

- They are a way of receiving feedback from customers and therefore a necessary means for putting improvement plans into action.
- They are a tool for preventing complacency and harnessing internal competencies for optimizing products and services.
- They are a useful way of measuring performance and allocating resources to deal with the deficient areas of the business.
- They are a useful "mirror" for gauging internal performance against the competition and the best in class organizations.
- They are a useful exercise for getting closer to customers and understanding them better.

Understanding the complaint process is very important. Fig.1. [2] maps out the different stages of customer experience for the simple case of newspaper subscription customers. These customers basically have the following types of interactions [2]:

- Starting the subscription via some channel
- Changing the product (weekday to seven-day, weekend to seven-day, seven-day to weekday, seven-day to weekend)
- Suspending delivery (typically for a vacation)
- Complaining
- Stopping the subscription (either voluntarily or involuntarily)

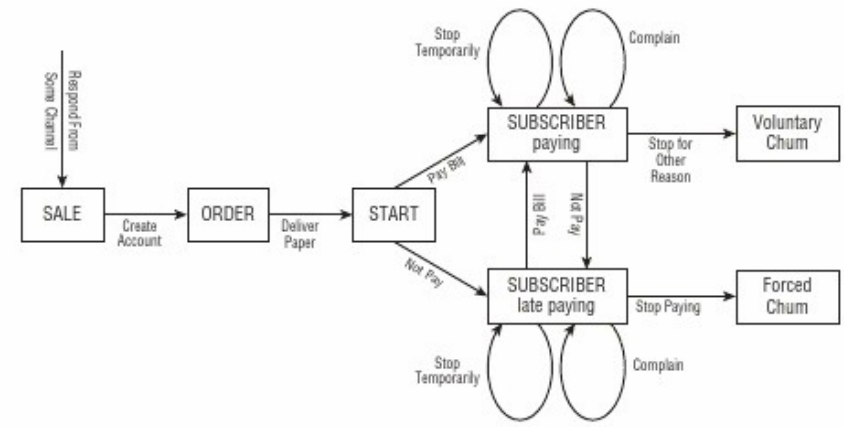


Fig.1. Different stages of the customer experience

Firms normally consider consumer complaints of any kind to be indispensable indicators of unsatisfactory performance. Without consumers' feedback, they will be unaware of their problems and retain their customers. Lau and Ng [3] found that dissatisfied consumers who complained had a higher level of repurchase intention than those who did not complain [3]. However, previous studies have also shown that many unsatisfied consumers prefer to change brands or suppliers and tell friends or families about their bad purchase experience than to voice their dissatisfaction to the companies concerned. For these reasons, it is clearly evident that Customer Complaints Management needs serious attention [1].

A firm must analyze consumers' complaints to pinpoint factors that are behind low satisfaction levels. Low satisfaction can be a result of a consumer's dissatisfaction with factors ranging from product quality to price. These data can also keep some factors in which the consumer is highly satisfied. Additionally, firms can develop marketing strategies to meet the consumer's needs.

In today's competitive business environment, the keyword 'future' is becoming more important because it can be directly connected with the identification of promising business opportunities for formulating long-term businesses. Various methods for identifying future business opportunities

ISBN: 978-605-64453-0-9

range from customary approaches including brainstorming, voice-of-customer analysis and data envelopment analysis to specific approaches such as system evolution patterns, disruptive innovation theory, weak signal analysis and customized patent mining methods [4].

Product reviews possess critical information regarding customers' concerns and their experience with the product. Such information is considered essential to firms' business intelligence and can be utilized for the purpose of conceptual design, personalization, product recommendation, better customer understanding, and finally to attract more loyal customers. Previous studies on deriving useful information from customer reviews have focused mainly on numerical and categorical data. Textual data have been somewhat ignored although they are deemed valuable. Existing methods of opinion mining in processing customer reviews concentrate on counting the positive and negative comments of review writers, but this is not enough to cover all the important topics and concerns across different review articles [5].

DM (Data Mining) can play a role in understanding whether or not customers are moving through the process the way they should be—or what characteristics cause a customer to fail during the activation stage. These results can help improve operational processes. They can also provide guidance during acquisition, by high lighting strategies that bring in sales that are not converted to paid subscriptions [2].

DM is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6] stored in structured databases, where the data are organized in records structured by categorical, ordinal and continuous variables. However, a vast majority of business data are stored in documents that are virtually unstructured. According to a recent study, 85–90% of all corporate data are stored in some sort of unstructured form (i.e., as text) [7]. This is where the TM fits into the picture. TM is the process of discovering new, previously unknown, potentially useful information from a variety of unstructured data sources including business documents, customer comments, Web pages, and XML files [8].

The Web is a highly dynamic information source. It contains a rich collection of data. Web TM is used to analyze Web data via TM. There are numerous advantages of Web TM for a firm. Firstly, Web TM provides additional traffic to the Web pages of a firm's site. In addition, Web TM is useful to improve the productive uses of mining for businesses, Web designers, and search engine operations. Firms can also use Web TM to improve marketing of their Web sites as well as the products they offer.

A large part of corporate information, approximately 80%, is available in textual data formats [9]. TM processes unstructured information, extracts meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to various DM algorithms including statistical and machine learning. Businesses use DM and TM to analyze customer and competitor data to improve competitiveness. The benefits of TM are obvious in areas where a large number of textual data are collected from business transactions. For example, the free-form text of customer interactions allows trending during time in the areas of complaint (and praise), warranty claims and error tracking, all of which are clearly input for product development and service allocation [8].

The paper is organized as follows: Section 2 provides the literature, Section 3 presents the preprocessing process and the Web TM application, Section 4 ends the paper with a brief conclusion.

2. Literature

DM techniques have long been applied to business [10,11]. TM has become an important research area in business in recent years [12]. Chang, Lin and Wang [13] aimed to apply the data warehouse and DM technologies to analyze customers' behavior in order to form the right customers' profile and a growth model in an Internet and e-commerce environment.

ISBN: 978-605-64453-0-9

TM, which is an interdisciplinary technique, analyzes unstructured data via DM. Although the technological developments underpinning TM are relatively recent, there already exists a lot of important studies in the literature.

In TM applications, determination of conversation topic is one of the important study areas. Most of the studies made in this area are conducted on the classification of news texts. Other studies in this area are related to the determination of the text writer's characteristics [14]. Weng and Liu [15] proposed a template for e-mails with multiple questions. Therefore, using multiple concepts to display the document topic is definitely a clearer way of extracting the information that a document wants to convey when the vector of similar documents is used. Zhan, Loh, and Liu [5] discovered and extracted salient topics from a set of online reviews and further ranked these topics. Özyurt and Köse [14] analyzed chat conversations to determine the characteristics of conversations via machine learning and data mining methods. Thorleuchter, Van den Poel and Prinzie [16] introduced idea mining as process of extracting new and useful ideas from unstructured text. They used an idea definition from technique philosophy and focused on ideas that can be used to solve technological problems. Fuller, Biros and Delen [17] reported on the promising results of a research study where data and TM methods along with a sample of real-world data from a high-stakes situation were used to detect deception.

Tsai and Kwee [18] explored the feasibility and performance of novelty mining and database optimization of business blogs. Gopal, Marsden, and Vanthienen [19] summarized the state of data and TM. Taking a very broad view, they used the term "information mining" to refer to the organization and analysis of structured or unstructured data that can be quantitative, textual, and/or pictorial in nature.

Sunikka and Bragge [20] combined a TM approach for profiling personalization and customization research with a traditional literature review in order to distinguish the main characteristics of these two research streams. Onishi and Manchanda [21] assembled a unique data set from Japan that contains market outcomes (sales) for new products, new media (blogs) and traditional media (TV advertising) in the movie category. Armentano, Godoy and Amandi [22] aimed to determine the impact of different profiling strategies based on the text analysis of micro-blogs as well as several factors that allow the identification of users acting as good information sources.

Thorleuchter and Van den Poel [23] analyzed the impact of textual information from e-commerce companies' Websites on their commercial success. Thorleuchter, Van den Poel and Prinzie [24] used Web TM. They analyzed the customers of a large German business-to-business mail-order company. Ur-Rahman and Harding [9] focused on the use of hybrid applications of TM or textual DM techniques to classify textual data into two different classes. Hao [25] compared the k-medoids algorithm and k-medoids social evolutionary programming in clustering documents.

He, Zha, and Li [26] increased competitive advantage and effectively assessed the competitive environment of businesses. Companies need to monitor and analyze not only the customer-generated content on their own social media sites, but also the textual information on their competitors' social media sites. They described an in-depth case study which applies TM to analyze unstructured text content on the Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino's Pizza and Papa John's Pizza. Kahya-Ozyirmidokuz [12] used TM to analyze online Turkish social shopping firms. Text preprocessing techniques (tokenization, term filtering methods, Euclidean distance measure etc.) were used. The relationships are discovered via a Web TM model.

In this research, we explored customer complaint data patterns associated with TM using Web TM. Unlike other studies our research aims to find relationships from heating system firms' customer complaints unstructured data. We use Web TM methods to extract hidden patterns.

3. Experimental analysis

ISBN: 978-605-64453-0-9

The top seven heating system firms in Turkey are analyzed in terms of customers' complaints. Data are collected from a complaint Website with the RapidMiner Web Mining Tool. Then, data are transformed to a collection of documents by generating a document for each record. Every complaint is transformed to a document. These documents which were collected for the period between December 2012 and October 2013 are analyzed with TM techniques. Summarization, tokenization, stemming, and filtering are also used.

3.1. Preprocessing process

Fig.2 shows the number of complaints submitted by customers and consumers about seven companies to a Web site. Two thousand documents are used in the analysis.

Fig.2. Pie chart of number of posted complaints

Fig.3 presents the preprocessing process of extracted web documents. The aim of preprocessing is to represent each document as a feature vector, that is, to separate the text into individual words. The preprocessing process starts by removing all the HTML tags and only preserving the actual content. Then tokenization, which is the exploration of the words in a sentence, is applied. We eliminate stopwords because they are not necessary for TM applications. Stemming, which is a technique for the reduction of words into their roots, is applied.

A numerical static TF-IDF (Term Frequency- Inverse Document Frequency), which reflects how important a word is to a document in a collection, is used for vector creation in the document processing step. In conclusion, we select the significant keywords that carry the meaning, and discard the words that do not.

ISBN: 978-605-64453-0-9

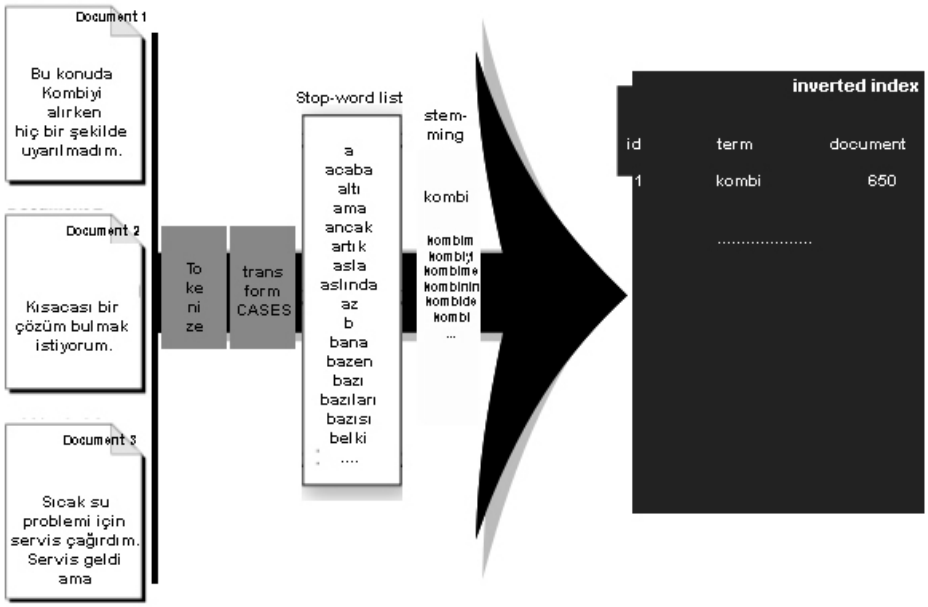


Fig.3. Preprocessing unstructured web complaint data.

The output from the preprocessing techniques consists of 1) a word list and 2) a document vector. TF-IDF scores are achieved, with attribute name, total occurrences and document occurrences. The outputs' datasets are the process documents example set's attributes *{row no, text, link, URL, response-code, response-message, content-type, content-length, date, last-modified, expires, title, language, description, keywords, robots, id, and words which are used in documents*}, and the process documents wordlist's attributes *{word, attribute name, total occurrences, document occurrences}*. One of the outputs of the preprocessing process is the example set which has 2,000 examples with 15 special attributes and 3,824 regular attributes.

"Şikayet" is the most frequently used word in the documents. The maximum number of total occurences of the attribute is 1,226. "Üslup, yapmak, taraf, uygulama, yanlış, seçmek, adil, beğenmek, hakkaniyet" are the other attributes that are most frequently used. The extracted document vector is used in clustering.

3.3. Model

Similarity analysis, which calculates the similarity among all examples of the dataset, is used. The histogram of similarity analysis is given in Fig.4. Fig. 5 presents the similarity graphs.

ISBN: 978-605-64453-0-9

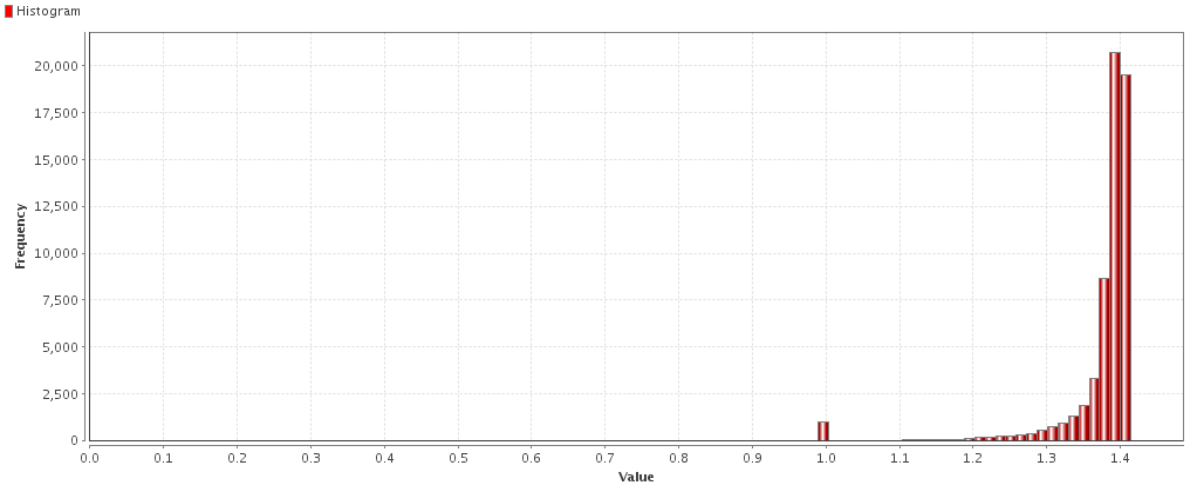
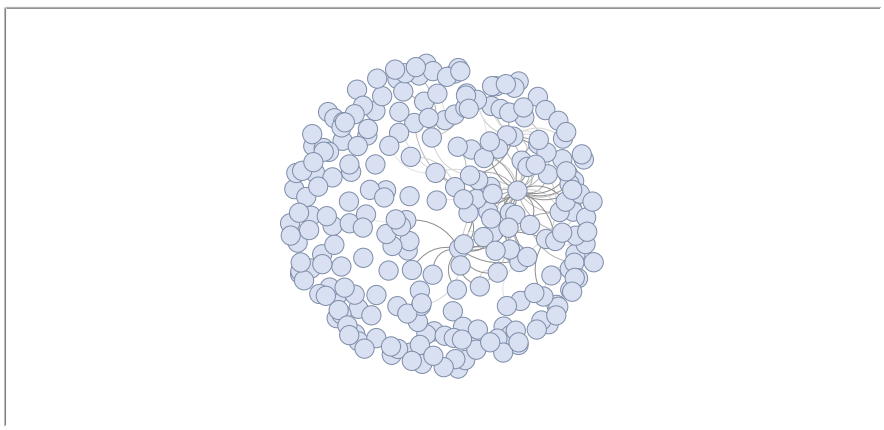


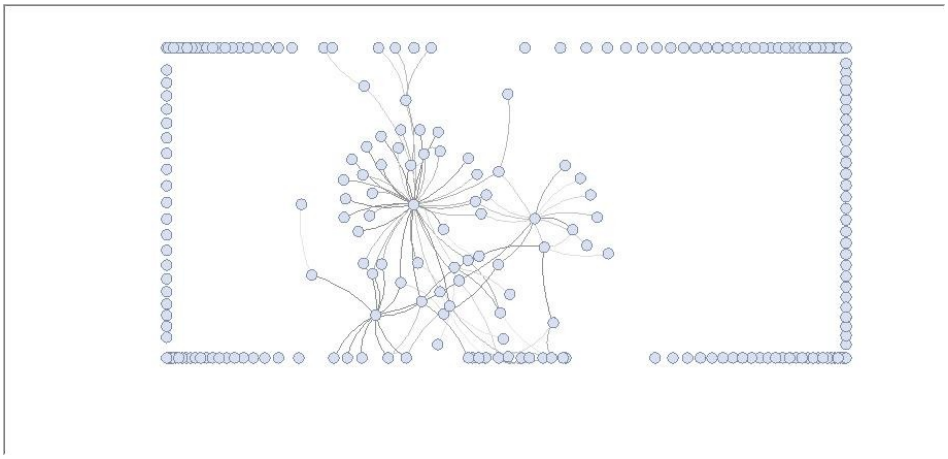
Fig.3. Histogram of similarity analysis.

The k-medoids clustering algorithm [27], which reduces the distance between all objects in a cluster and the most centrally located object in the cluster is applied to the preprocessed data. It is similar to the k-means algorithm except that the mean of each cluster is the object that is nearest to the "center" of the cluster. There are two clusters: cluster 0 and cluster 1 with 1,110 and 890 items respectively. A centroid table of the clustering model is achieved. The clustering outputs clearly show the top keywords from each of the documents. We can also discriminate attributes via vectors. The plot view of the cluster model is shown in Fig.5.



(a)

ISBN: 978-605-64453-0-9



(b)

Fig. 4 Similarity graphs with RapidMiner.

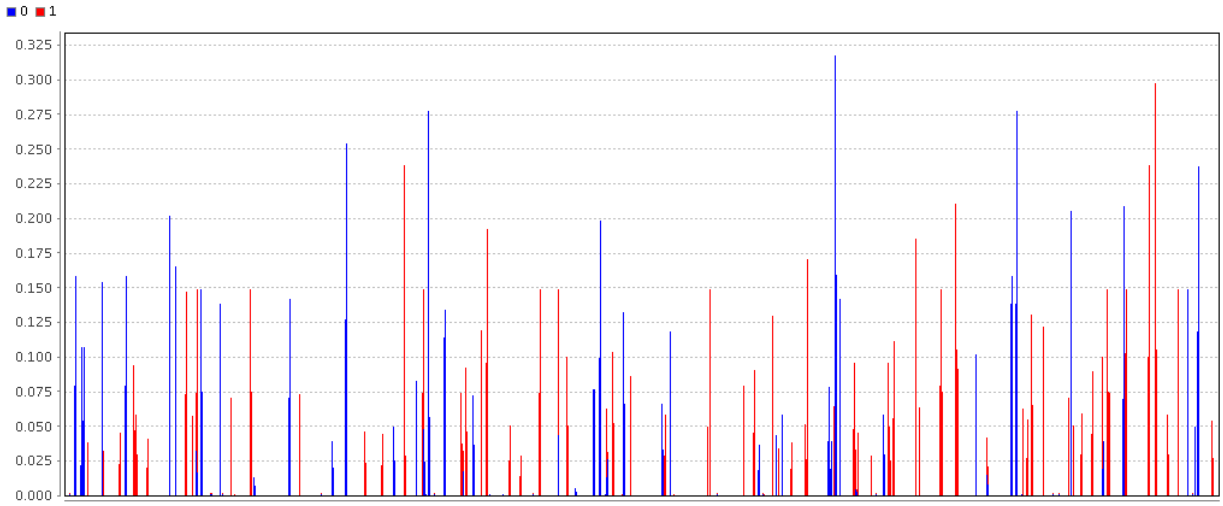


Fig.5. k-medoids clustering model centroid plot view.

A performance operator, which can be used to derive a performance measure (in the form of a performance vector) from the dataset, is used. The performance vector of the model's cluster number index is 0.992.

3. Conclusions

Analyzing customer complaints, is a part of being in business. Customers make buying decisions based on the price, the quality, and the service they receive. Successful businesses use customer information to truly evaluate feedback. However, firms generally ignore textual data and they often use categorical and numerical data. This situation causes lack of information, confidence and bad decision making because there is important hidden knowledge in textual databases. The amount of customer complaints' data is increasing at a higher rate on the Web. Traditional techniques cannot analyze these unstructured data.

ISBN: 978-605-64453-0-9

Textual data have been somewhat ignored, although they are deemed valuable. In this research customer complaints were analyzed via web TM to achieve useful knowledge from the customer complaints' documents of heating systems. Similarity analysis was used to determine similar documents. The similarities of firms about the subject were determined. We not only grouped the customer complaints in the heating sector but also the firms to which complaints were made.

Documents were clustered. Graphs and tables were obtained. Further work could be done as follows. A scenario could be improved to indicate the importance of this type of clustering. Solutions could be produced by these clusters. Responses could be added to the solution database. Thus, every cluster has similar solution documents. In conclusion, similar complaints could be answered by similar response mails.

4. References

[1] M. Taleghani, M. S.Largani, S. Gilaninia, and S. J. Mousavian, The role of customer complaints management in consumers satisfaction for new industrial enterprises of Iran, *International Journal of Business Administration*, vol. 2, no. 3, pp. 140-147, August 2011.

[2] M. J. A. Berry and G. S. Linoff, *Data mining techniques for marketing, sales, and customer relationship management,* 3rd edn., Wiley, New York, 2011.

[3] G.T lau and S. Ng, Individual and situational factors influencing negative word of mouth behaviour, *Revue Canadienne des Sciences de 1 'Administration*, vol. 18, No. 3, pp. 163 – 178, 2001.

[4] J. Yoon, Detecting weak signals for long-term business opportunities using TM of Web news, *Expert Systems with Applications,* vol. 39, pp. 12543–12550, 2012.

[5] J. Zhan, H. T. Loh and Y. Liu, Gather customer concerns from online product reviews – A text summarization approach, *Expert Systems with Applications,* vol. 36 pp. 2107–2115, 2009.

[6] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From DM to knowledge discovery in databases, *American Association for Artificial Intelligence*, vol.17, no.3, pp.37-54, 1996.

[7] W. McKnight, Text DM in business intelligence, *Information Management Magazine*, http://www.information-management.com/issues/20050101/1016487-1.html (accessed Dec 10, 2013), 2005.

[8] D. Delen and M. D. Crossland, Seeding the survey and analysis of research literature with TM, *Expert Systems with Applications,* vol. 34, pp. 1707–1720, 2008.

[9] N. Ur-Rahman and J.A. Harding, Textual DM for industrial knowledge management and text classification: A business oriented approach, *Expert Systems with Applications* vol. 39, pp. 4729–4739, 2012.

[10] C. Çiflikli and E. Kahya-Özyirmidokuz, Implementing a DM solution for enhancing carpet manufacturing productivity, *Knowledge Based Systems*, vol.23, pp.783-788, 2010.

[11] C. Çiflikli and E. Kahya-Özyirmidokuz, Enhancing product quality of a process, *Industrial Management & Data Systems*, vol.112, no. 8, pp.1181-1200, 2012.

[12] E. Kahya-Özyirmidokuz, Text mining of online social shopping, The Global Reach of Industrial Engineering, Turkey, Istanbul, pp.172-173, 26-28 June 2013.

ISBN: 978-605-64453-0-9

[13] C.-W. Chang, C.-T. Lin and L.-Q. Wang, Mining the text information to optimizing the customer relationship management, *Expert Systems with Applications,* vol. 36, pp. 1433–1443, 2009.

[14] Ö. Özyurt and C. Köse, Chat mining: Automatically determination of chat conversations' topic in Turkish text based chat mediums, *Expert Systems with Applications,* vol. 37, pp. 8705–8710, 2010 .

[15] S.-S. Weng and C.-K. Liu, Using text classification and multiple concepts to answer e-mails, *Expert Systems with Applications,* vol.26, pp. 529–543, 2004.

[16] D. Thorleuchter, D. Van den Poel and A. Prinzie, Mining ideas from textual information, *Expert Systems with Applications,* vol.37, pp. 7182–7188, 2010.

[17] C. M. Fuller, D. P. Biros and D. Delen, An investigation of data and TM methods for real world deception detection, *Expert Systems with Applications,* vol. 38, pp. 8392–8398, 2011.

[18] F. S. Tsai and A. T. Kwee, Database optimization for novelty mining of business blogs, *Expert Systems with Applications,* vol. 38, pp. 11040–11047, 2011

[19] R.D.Gopal, J.R. Marsden and J.Vanthienen, Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining, *Decision Support Systems,* vol.51, pp. 727–731, 2011.

[20] A. Sunikka and J. Bragge, Applying text-mining to personalization and customization research literature – Who, what and where?, *Expert Systems with Applications*, vol.39, pp. 10049–10058, 2012.

[21] H. Onishi and P. Manchanda, Marketing activity, blogging and sales, *Intern. J. of Research in Marketing,* vol. 29, pp.221–234, 2012.

[22] M.G. Armentano, D. Godoy and A.A. Amandi, Followee recommendation based on text analysis of micro-blogging activity, *Information Systems*, vol.38, pp. 1116-1127, 2013.

[23] D. Thorleuchter and D. Van den Poel, Predicting e-commerce company success by mining the text of its publicly-accessible Website, *Expert Systems with Applications,* vol.39, pp. 13026–13034, 2012.

[24] D. Thorleuchter D. Van den Poel and A. Prinzie, Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing, *Expert Systems with Applications*, vol. 39, pp. 2597–2605, 2012.

[25] Z.-G. Hao, A New Text Clustering Method Based on KSEP, *Journal of Software*, vol. 7, no. 6, pp. 1421-1425, 2012.

[26] W. He, S. Zha and L. Li, Social media competitive analysis and TM: A case study in the pizza industry, *International Journal of Information Management,* vol.33, no.3, pp. 464–472, 2013.

[27] T. Velmurugan and T. Santhanam, Computational complexity between k-means and k-medoids clustering Algorithms, *Journal of Computer Science, vol.*6, no.3, pp. 363-368, 2010.

ISBN: 978-605-64453-0-9