

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/260074667>

Analyzing customer complaints : A Web text mining application

Conference Paper · February 2014

CITATIONS

2

READS

775

2 authors:



[Esra Kahya-Özyirmidokuz](#)

Erciyes Üniversitesi

27 PUBLICATIONS 117 CITATIONS

[SEE PROFILE](#)



[Mustafa Hakan Özyirmidokuz](#)

1 PUBLICATION 2 CITATIONS

[SEE PROFILE](#)

All content following this page was uploaded by [Esra Kahya-Özyirmidokuz](#) on 20 October 2015.

The user has requested enhancement of the downloaded file. All in-text references [underlined in blue](#) are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

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 Termoteknik, Ankara, Turkey

^aesrakahya@erciyes.edu.tr, ^bhakan.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 been increasing explosively, as has the information contained within Web pages. In today's world, firms' Web data must be analyzed to gain competitive advantage in the topic sector. Web text mining (TM) is gaining a lot of importance because of being used increasingly in business applications for understanding and predicting valuable information. It plays a key role in organizing 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 to implement heating systems firms' in competition.

Top seven heating systems firms are analyzed about customers' complaints in Turkey. Data are collected from a complaint Website with RapidMiner Web Mining Tool. Then data is transformed to a collection of documents by generating a document for each record. Every complaint is transformed to a document. These documents which are collected for one years' time between 2012 december and 2013 october are analyzed with TM techniques. Summarization, tokenization, stemming, and filtering are also used. In addition, the similarities of firms about the subject are determined. Not only have we analyzed the customers in the sector but also the firms about the complaints.

1. Introduction

The world has turned into a small village. In addition, the price advantage prevails in the world of sales nowadays. The only way to stand out in the competition lies in the aftermarket. In order to satisfy the customer we must analyze all kinds of data. Because social media and Internet media put consumer on a stronger location. Customers are really 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]:

- They are a way of receiving feedback from customers and therefore necessary means for

putting into action improvement plans.

- 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 competition and best in class organizations.
- They are a useful exercise for getting closer to the customers and understanding them better.

Understanding the complaining process is very important. Fig.1. [2] maps out different stages of the 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)

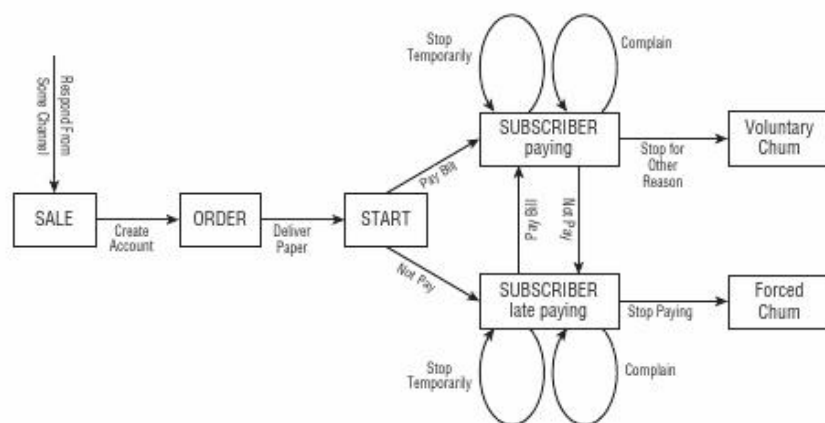


Fig.1. Different stages of the customer experience

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 companies. 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 driving low satisfaction levels. Low satisfaction can be a result of a consumer's dissatisfaction 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 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 which can be utilized for the purpose of conceptual design, personalization, product recommendation, better customer understanding, and finally attract more loyal customers. Previous studies of deriving useful information from customer reviews 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 concentrates on counting positive and negative comments of review writers, which is not enough to cover all 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 uses 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 engines 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 the various DM algorithms including statistical and machine learning. Businesses use DM and TM to analyse customer and competitor data to improve competitiveness. Benefits of TM obvious in the 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 is clearly input to product development and service allocation [8].

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

2. Literature

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

TM which is an interdisciplinary technique analyzes unstructured data via DM. Although the technological developments underpinning TM are relatively recent, there exists a lot of important researches 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 classification of news texts. Other studies on this area are related to the determination of text writer's characteristics [14]. Weng and Liu [15] propose a template for e-mails of multiple questions. Therefore, using multiple concepts to display the document topic is definitely a clearer way of extracting information that a document wants to convey when the vector of similar documents is used. Zhan, Loh, and Liu [5] discover and extract salient topics from a set of online reviews and further ranks these topics. Özyurt and Köse [14] analyze chat conversations to determine characteristics of conversations via methods of machine learning and data mining. Thorleuchter, Van den Poel and Prinzie [16] introduce idea mining as process of extracting new and useful ideas from unstructured text. They use an idea definition from technique philosophy and we focus on ideas that can be used to solve technological problems. Fuller, Biros and Delen [17] report 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 is used to detect deception.

Tsai and Kwee [18] explore the feasibility and performance of novelty mining and database optimization of business blogs. Gopal, Marsden, and Vanthienen [19] summarize the state of data and TM. Taking a very broad view, they use 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] combine 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] assemble 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 allows the identification of users acting as good information sources.

Thorleuchter and Van den Poel [23] analyze 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 analyze the customers of a large German business-to-business mail-order company. Ur-Rahman and Harding [9] focus on the use of hybrid applications of TM or textual DM techniques to classify textual data into two different classes. ZhanGang Hao [25] compare k-medoids algorithm and k-medoids social evolutionary programming in clustering documents.

He, Zha, Li [26] increase competitive advantage and effectively assess 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 describe an in-depth case study which applies TM to analyze unstructured text content on Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino's Pizza and Papa John's Pizza. Esra Kahya-Ozyirmidokuz [12] use TM to analyze online Turkish social shopping firms. Text preprocessing techniques (tokenization, term filtering methods, Euclidean distance measure etc.) are used. The relationships are discovered via a Web TM model.

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

3. Experimental analysis

Top seven heating systems firms are analyzed about customers' complaints in Turkey. Data are collected from a complaint Website with RapidMiner Web Mining Tool. Then data is transformed to a collection of documents by generating a document for each record. Every complaint is transformed to a document. These documents which are collected for one years' time between 2012 december and 2013 October are analyzed with TM techniques. Summarization, tokenization, stemming, and filtering are also used.

3.1. Preprocessing process

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

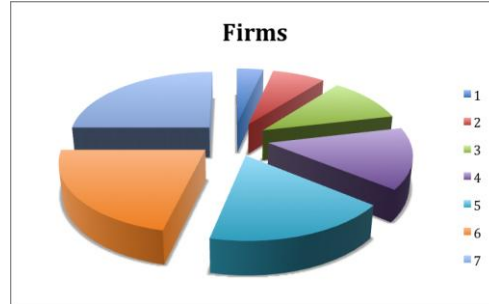


Fig.2. Pie chart of number of posted complaints

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 with 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 (lemmatisation) 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.

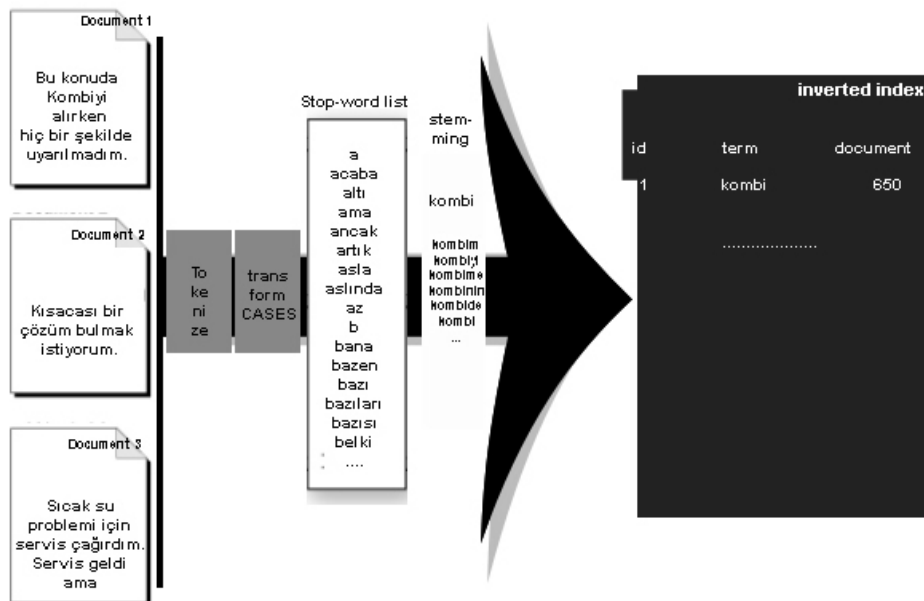


Fig.3. Preprocessing unstructured web complaint data.

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 preprocessing process is the example set which have 2000 examples with 15 special attributes and 3824 regular attributes.

“Şikayet” is the most used word in the documents. The maximum number of total occurrence of the attribute is 1226. “Üslup, yapmak, taraf, uygulama, yanlış, seçmek, adil, beğenmek, hakkaniyet” are the other attributes 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.

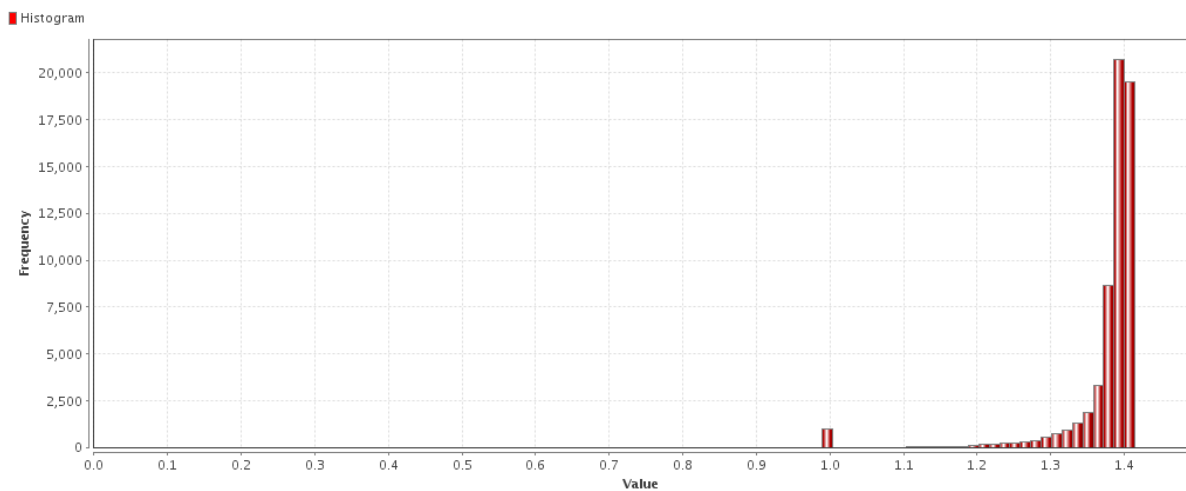
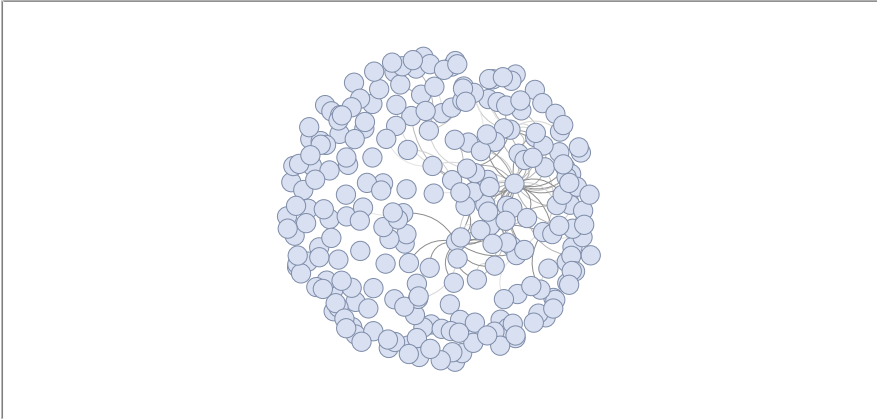
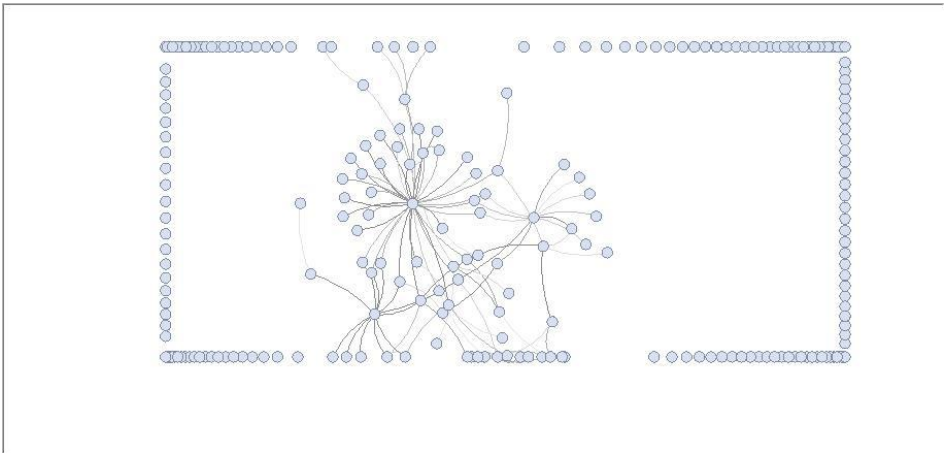


Fig.3. Histogram of similarity analysis.

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 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 1110 and 890 items respectively. Centroid table of 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 cluster model is shown in Fig.5.



(a)



(b)

Fig. 4 Similarity graphs with RapidMiner.

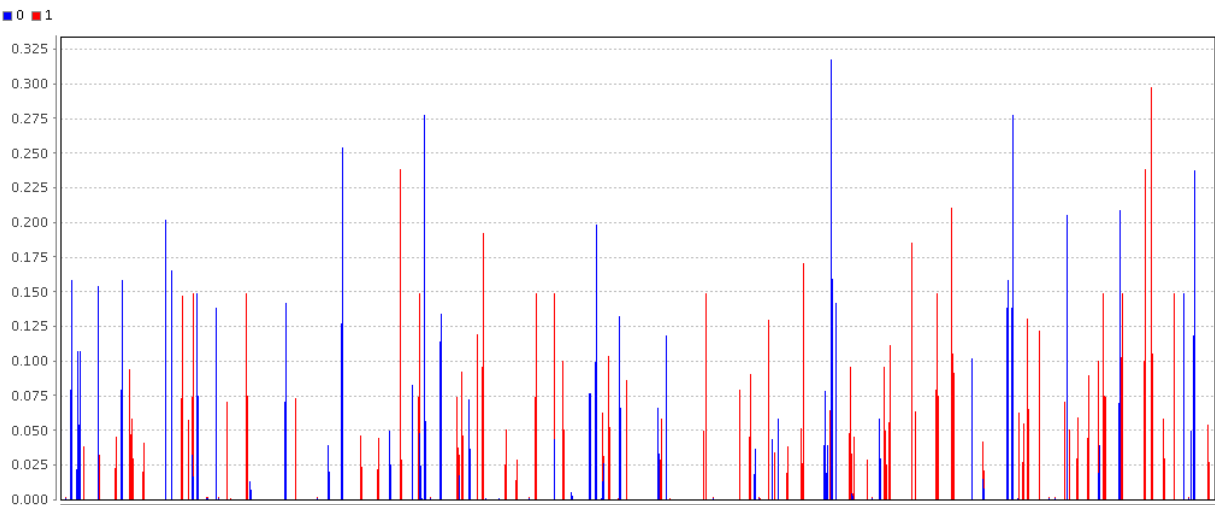


Fig.5. k-medoids clustering model centroid plot view.

A performance operator, which can be used to derive a performance measure (in 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 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 hidden important knowledge in textual databases. The amount of customer complaints data are increasing at a higher rate on Web. Traditional techniques cannot analyze these unstructured data.

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

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

4. References

- [1] M. Taleghani, M. S.Largani and S. Gilaninia, 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 l'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, Extual 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.
- [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.](#)