Mining Customer Feedback Documents

Eduard Alexandru Stoica and Esra Kahya Özyirmidokuz

***Abstract*—Managing customer feedback data has become a necessity for firms in order for them to gain competitive advantage in the sector. Analyzing customer complaints' data to find useful information that's hidden is an important step in understanding customers. This important, hidden knowledge must be extracted automatically to allow firms to gain a better understanding of the general market and of their own and their competitors' customers. A firm can learn the needs of customers and show how its products and services satisfy these needs by analyzing these documents.**

The aim of this research is to summarize and extract data from unstructured customer feedback documents which are about ignoring subscriptions to a telecommunication firm in Turkey. The data are transformed to a collection of documents by generating a document for each record. Text processing techniques are applied. Cosine similarity analysis is used to classify documents into relevant categories. Clusters are determined.

***Index Terms*—Data mining, text mining, customer feedback data, natural language processing**

I. Introduction

Nowadays one of the biggest needs of a firm is to extract knowledge by analyzing these unstructured data in the strategic decision making process. It is also important for a firm to make decisions by using these obtained patterns.

Collecting and analyzing customer feedback is important because it allows organizations to learn in a continuous manner to adapt their offerings to customer preferences. Increasingly, customers use multiple communication channels to provide feedback, making it cumbersome for organizations to develop efficient and effective processes to collect and analyze all the information [1]. Firms must manage customer complaints to improve the quality of their service. After collecting data correctly, data must be analyzed without loss of information in order to ensure an effective complaint management system.

A firm must analyze consumers' complaints to pinpoint the 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 show some factors in which the consumer is highly satisfied. Additionally, firms can develop marketing strategies to meet the consumer's needs [2].

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 [3].

Huge amount of data are available in textual data formats. It is impossible managers to read all big unstructured complaint data which form an unstructured free data type and analyze them without loss of information. Consequently, analysists need to use TM in addition to data mining (DM) techniques. Extracting meaningful information from the textual data is critical to the success of a business. Text mining, which processes unstructured information, will help managers to analyze customer feedback documents automatically. Consequently, knowledge is discovered that would have been extremely difficult to find, even if it had been possible to read all the documents.

TM extracts meaningful numeric indices from the text, and, thus, makes the information contained in the text accessible to various DM algorithms including statistical, machine learning and artificial intelligence. Words, clusters of words used in documents, etc., or documents can be analyzed and similarities between them can be determined. DM, machine learning, and artificial intelligence techniques can be applied to these digitized documents.

In this research, TM is used to analyze natural language documents. The customer complaint documents of a big telecommunication firm about ignoring their subscriptions are used in this research. The aim of this study is to summarize and group these customer feedback data. 841 customer complaint documents were collected in 2013. The paper is organized as follows: Section II provides the literature, Section III presents the experimental analysis, and Section IV ends the paper with a brief conclusion.

II. Related Research

DM, which is the process of identifying patterns in large sets of data, has long been studied [4], [5]. TM has become an important research area in business in recent years [6]. Chang, Lin and Wang [7] aimed at applying 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.

Gamon [8] demonstrated that it was possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. Gamon used natural language processing techniques and linear support vector machines that achieved high classification accuracy on data that present classification challenges even for a human annotator.

Gamon *et al*. [9] presented a prototype system, code-named Pulse, for mining topics and sentiment orientation jointly

Manuscript received August 1, 2014; revised December 17, 2014.

E. Kahya-Özyirmidokuz is with the Computer Technologies Department, Erciyes University, Kayseri, Turkey (e-mail: esrakahya@erciyes.edu.tr).

E. A. Stoica is with the Department of Information Economics, Lucian Blaga University, Sibiu, Romania (e-mail: eduard.stoica@ulbsibiu.ro).

DOI: 10.7763/IJKE.2015.V1.12

from free text customer feedback. They described the application of the prototype system to a database of car reviews; a simple but effective technique for clustering sentences, the application of a bootstrapping approach to sentiment classification, and a novel user-interface. Ittoo, Zhang and Jiao [10] proposed a TM-based recommendation system which assists customers in their decision making in online product customization. The proposed system allowed customers to describe their interests in textual format, and thus to capture customers' preferences to generate accurate recommendations.

Coussement, and Van den Poel [11] introduced a methodology to improve complaint-handling strategies through an automatic email-classification system that distinguishes complaints from non-complaints. The study focused on how a company could optimize its complaint-handling strategies through an automatic email-classification system. Natural processing techniques with TF-IDF were used.

Weng and Liu [12] 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 [3] discovered and extracted salient topics from a set of online reviews and further ranked these topics. Özyurt and Köse [13] analyzed chat conversations to determine the characteristics of conversations via machine learning and DM methods. Thorleuchter, Van den Poel and Prinzie [14] introduced idea mining as a 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 [15] 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 [16] explored the feasibility and performance of novelty mining and database optimization of business blogs. Gopal, Marsden, and Vanthienen [17] 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 [18] 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 [19] 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 [20] 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 [21] analyzed the impact of textual information from e-commerce companies' Websites on their commercial success. Thorleuchter, Van den Poel and Prinzie [22] used Web TM. They analyzed the customers of a large German business-to-business mail-order company. Ur-Rahman and Harding [23] focused on the use of hybrid applications of TM or textual DM techniques to classify textual data into two different classes. Hao [24] compared the k-medoids algorithm and k-medoids social evolutionary programming in clustering documents.

He Zha, and Li [25] 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.

Kahya Özyirmidokuz [6] used TM to analyze online Turkish social shopping firms. The relationships are discovered via a Web TM model. Two hundred popular Turkish companies' web URLs are used. Web TM through natural language processing techniques is examined. Kahya Özyirmidokuz and Özyirmidokuz [2] analyzed the top seven heating systems firms' customer complaint documents in Turkey via web TM. Ordenes *et al*. [1] used linguistics-based TM modeling to help in the process of developing an improved framework. The proposed framework incorporated important elements of customer experience, service methodologies, and theories such as cocreation processes, interactions, and context.

We mine customer complaint documents about ignoring their subscriptions of a telecommunication firm in Turkey. TM and natural language processing techniques are used to extract useful knowledge from customer feedback documents.

III. Data Preprocessing

A total of 841 documents are used dating between the dates January 1 2013 - 31 December 2013. We use the TF-IDF (Term Frequency - Inverse Document Frequency) numerical static which reflects how important a word is to a document in a collection. It can be seen in Equation (1) and (2).

$$t_f(t,d) = f(t,d) / \max\{ f(t,d) : W \in d \} \quad (1)$$

$$idf(t,D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \quad (2)$$

After tokenization process, we remove 220 unnecessary Turkish stopwords. Stemming, which is a technique for the reduction of words into their roots, is applied. The outputs from the preprocessing techniques consist of word lists and document vectors. TF-IDF scores are achieved, with attribute name, total occurrences and document occurrences.

TABLE I: Example Sets of the Firms

Word	*In English*	*Total occurrences*	*Number of documents*
mesaj	message	32682	838
detay	detail	31692	841
şikayet	complaint	54020	841
müşteri	customer	7033	841
beğenmek	like	4954	838

One of the outputs of the preprocessing process is the example set. The example set of the firm has 841 examples

with 10 special attributes and 4,626 regular attributes. The term "message" is the most frequently used word in the documents, and "detail, complaint, customer, like" are the other attributes that are most frequently used. The extracted document vectors are used in similarity analysis. Table I presents the most frequently used words in the documents.

IV. Modeling

Although similarity between documents is an essential ingredient in organizing unlabeled documents into distinct groups, measuring the similarity of documents is an end in itself. Measuring the similarity between documents is fundamental to most forms of document analysis, especially information retrieval [26]. Given two documents $\vec{t}_a$ and $\vec{t}_b$, their cosine similarity is as follows:

$$SIM_c(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a . \vec{t}_b}{|\vec{t}_a| \times |\vec{t}_b|} \quad (3)$$

where $\vec{t}_a$ and $\vec{t}_b$ are *m*-dimensional vectors over the term set $T = \{t_1, ..., t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0, 1] [27].

Cosine similarity is used. The histograms of similarities are given in Fig. 1. The similarity graphs are presented in Fig. 2.

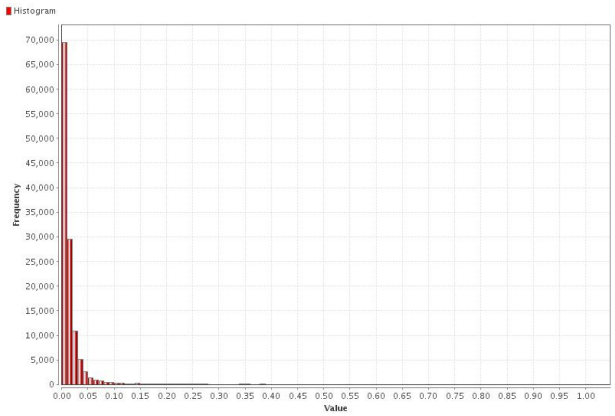
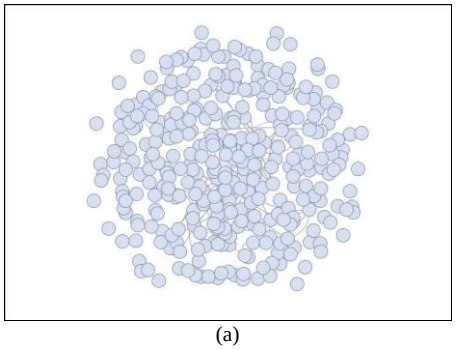
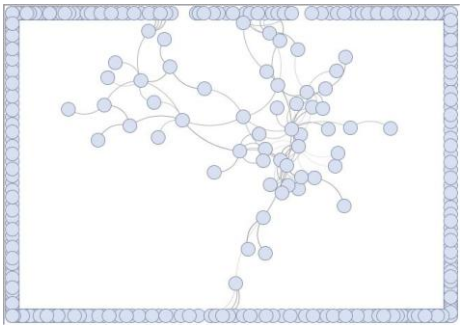


Fig. 1. Histograms of cosine similarity.

K-medoids clustering is applied to vectors. Four groups are selected (*k*=4). Eaclidean distance is used. After clustering, 191, 236, 119, 295 items were obtained. The centroid plot views of clusters can be seen from Fig. 3.



(a)



(b)

Fig. 2. Similarity graphs obtained with RapidMiner.

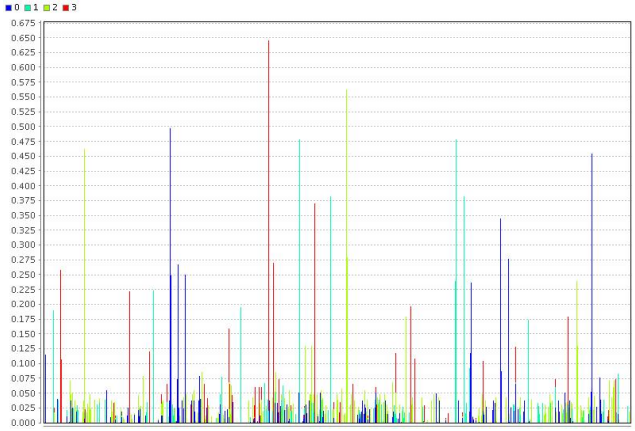


Fig. 3. The centroid plot views of clusters.

Performance operators, which can be used to derive a performance measure (in the form of a performance vector) from the dataset, are used. The performance vector of the model's cluster number index is 0.984.

V. Conclusion

TM has gained increasing attention in recent years. In addition, firms need to analyze unstructured feedback data to make good decisions because there is important hidden knowledge in textual databases. Customer complaints must be systematically mined to discern patterns and make better decisions.

Managers save time by understanding the huge amount of unstructured data with TM, instead of reading it through page by page. In addition, there is no danger of overlooking data. Consequently, analyzing complaint documents will result in a competitive advantage for a firm.

In this research, DM and TM techniques, and natural language processing are applied to extract knowledge. Documents were clustered. Similarity analysis was used to determine similar documents. The similarities of firms about the subject were determined. Graphs and tables were obtained.

In time, this model will lose its validity. Proposed method must be integrated in the complaint management system of the firm to prevent this.

In the future we plan the further evaluation and implementation of this framework. Alternative methods could be applied in this framework. Further work could be done. For example, 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.

ACKNOWLEDGMENT

This work was supported by the strategic grant POSDRU/159/1.5/S/133255, Project ID 133255 (2014), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007–2013. This work was supported by Erciyes University Research Fund, Project Number FBA-2014-5364. This project is also supported by Bournemouth University, Faculty of Science and Technology, Software Systems Engineering.

REFERENCES

[1] F. V. Ordenes, B. Theodoulidis, J. Burton, T. Gruber, and M. Zaki, "Analyzing customer experience feedback using TM: A linguistics-based approach," *Journal of Service Research*, pp. 1-18, 2014.

[2] E. Kahya-Özyirmidokuz and M. H. Özyirmidokuz, "Analyzing customer complaints: A web TM application," in *Proc. International Conference on Education and Social Sciences*, Ferit Uslu, pp. 734-743, İstanbul: Ocerint, February 2014.

[3] 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.

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

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

[6] E. K. Ozyirmidokuz, "Analyzing unstructured facebook social network data through web TM: A study of online shopping firms in Turkey," *Information Development*, pp. 1–12, 2014.

[7] 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.

[8] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis," in *Proc. the 20th international conference on Computational Linguistics*, pp. 841-847, PA, USA: Association for Computational Linguistics Stroudsburg, 2004.

[9] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," *LNCS*, pp. 121-132, Heidelberg, Berlin: Springer-Verlag, 2005.

[10] A. R. Ittoo, Y. R. Zhang, and J. Jiao, "A TM based recommendation system for customer decision making in online product customization," in *Proc. International Conference on Management of innovation and technology*, vol. 1, pp. 473-477, Singapore, China: IEEE, 2006.

[11] K. Coussement and D. V. den Poel, "Improving customer complaint management by automatic email classification using linguistic style features as predictors," *Decision Support Systems*, vol. 44, pp. 870–882, 2008.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] D. Thorleuchter and D. V. 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.

[22] D. Thorleuchter, D. V. 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.

[23] 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.

[24] Z.-G. Hao, "A new text clustering method based on KSEP," *Journal of Software*, vol. 7, no. 6, pp. 1421-1425, 2012.

[25] 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.

[26] S. M. Weiss, N. Indurkhya, T. Coussement *et al*., *TM: Predictive Methods for Analyzing Unstructured Information*, USA: Springer, 2005.

[27] A. Huang, "Similarity measures for text document clustering," in *Proc. New Zealand Computer Science Research Student Conference (NZCSRSC)*, J. Holland, A. Nicholas, D. Brignoli, eds., pp. 49-56. 2008.

Eduard Alexandru Stoica was born in Sinaia, Romania. He is a lecturer, at University Lucian Blaga of Sibiu, Faculty of Economic Sciences. He has a master degree in economics and a Ph.D. in the field of economic cybernetics and statistics at Bucharest University of Economic Studies, with the thesis entitled "Generative mechanisms of economic processes". Now he is a postdoctoral research on business use of technology and how companies can increase Digital IQ. His research study is on the border between "cybernetics" (by creating an adaptive feedback model), e-business (through the model's connection with e-marketing, e-commerce and e-government) and "software engineering" (by projecting and implementing software instruments). His research activity includes grants CNCSIS and European projects H2020, POS-DRU, POS-CCE, as well as a Socrates mobility program, Erasmus+ and Leonardo program.

Esra Kahya Özyirmidokuz was born in Kayseri, Turkey. She is an assistant professor at Erciyes University Computer Technologies Department, Turkey. She received the B.S. degree in control and computer engineering, Faculty of Engineering, Erciyes University; M.S. and Ph.D. degrees in production and marketing, Erciyes University Faculty of Economics. In her master thesis, she designed a personnel selection expert system which uses fuzzy logic. The title of her Ph.D. thesis is "Analyzing and modeling manufacturing data by using data mining techniques". She is the supervisor of two mining projects and researcher at a text mining project at Erciyes University. Now she is a visiting researcher at Bournemouth University.