International Journal of Business Quantitative Economics and Applied Management Research
Volume-3, Issue-7, December-2016 ISSN No: 2349-5677

CLUSTERING ONLINE CUSTOMER COMPLAINTS

Ozyirmidokuz Kahya Esra
Computer Technologies Department
Erciyes University
Kayseri, Turkey
esrakahya@erciyes.edu.tr

Stoica Eduard Alexandru
Faculty of Economic Sciences
Lucian Blaga University of Sibiu
Sibiu, Romania
eduard.stoica@ulbsibiu.ro

Abstract

Nowadays one of the biggest needs of a firm is to extract knowledge by analysing unstructured big data in the strategic decision making process in order to improve customer satisfaction. The aim of this research is automatically clustering the online complaints which are about the customers ignoring their subscriptions in order to understand a specific group of the complaints. 809 customer complaints which are about ignoring subscriptions are collected from huge amount of online complaints with web mining from a telecommunication firm and those of its biggest competitor in Turkey. Text mining and natural language processing techniques are used to analyse the data. The positive feedback of the customers are re-analysed to make an adaptation.

We present an adaptive feedback model to achieve knowledge which helps in making business decisions. We clustered online complaints while determining similarities between the groups. New, interesting and hidden knowledge about customer complaints are found.

Index Terms— Web text mining, Customer feedback, Natural language processing, Clustering

I. INTRODUCTION

Analysing online customer complaints is important for a number of reasons to garner many types of information from customers. Customer complaints contain the direct voice of the customer and unfiltered information about what has happened to quality of both a product and service [1]. Customer complaints represent an opportunity for a firm to extract valuable information and gain insightful knowledge. They help firms determine what's important to their customers. Analysing customer complaints also allows firms to better understand how

customers rate their products versus competitives. They can help managers to make better business decisions at run-time. For this, nowadays, firms need to analyse online data including social media [2] while giving business decisions.

We need to expand the scope of business intelligence to include textual information for a number of reasons [3]:

We have tools at our disposal for analysing text and extracting key information to create a document warehouse with distilled, useful business intelligence (BI) information. Steady advances in computational linguistics since the 1960s have left us with a wide range of tools for extracting key features, categorizing documents, indexing by topic as well as by keywords, automatically summarizing texts, and grouping similar documents. These tools are the keys to successful integration of text into the BI infrastructure.

The Internet and the WWW are making vast amounts of information easily accessible. With the right tools we can find information about the financial, marketing and technology plans of competitors.

Organizations have depended upon writing systems of some form to record information. A significant portion of this information is text-based, and organizations are beginning to realize the need to deal with this text from a decision support perspective. Our current means of dealing with, or ignoring, text are no longer sufficient to meet the needs of decision makers.

Successful organizations are not just driven by managing core operations such as selling products, tracking changes in quality control measures, or analysing trends in cash flow. More and more, the intangible aspects of organizations, such as knowledge about process management, patented technologies, and methodologies, are fundamental factors influencing the course of a business. Managers and executives need to understand the competitive advantage created by their intellectual property as well as how the market responds to innovations by competitors. This kind of information is not available by looking at data extracts from transaction processing systems. It is, however, available to those who know where to look and how to extract the key information.

Decision makers think strategically. This means that they need information about what is going on outside the organization as well as inside. They need to understand industry structure and dynamics, which shed light on the competitive environment in which companies operate. Macroenvironmental analysis, another aspect of strategic management, examines the economic, political, social, and technological events that influence an industry. Monitoring the macroenvironment is no small task. With techniques, this type of analysis can at least be less ponderous and information far more accessible.

The drawbacks of structured studies are the expense associated with the design and administration of the survey, the limit that is necessarily imposed on the free expression of opinions by customers, and the corresponding risk of missing trends and opinions that are not expressed in a controlled situation. Additionally, there is the risk of missing whole segments of the customer population that do not like responding to a guided and structured set of questions.

Another potential source of information for business intelligence, which is becoming more and more pervasive and voluminous, is spontaneous customer feedback. This feedback can be gathered from blogs, newsgroups, feedback email from customers, and websites that collect free-

form product reviews. These can be rich sources of information, but these sources are much less structured than traditional surveys.

The information is contained in free text, not in a set of answers elicited for a specific set of questions [4]. Marketing departments have become increasingly aware of the importance of textual feedback and use manual or automatic approaches to analyse information. Companies that run analysis on a manual basis can gain a deeper understanding of customer feedback, but if their analysis lacks procedural models, they tend to be inconsistent when reviewing large quantities of data. At the same time, companies that have adopted automated analysis of textual feedback have failed to realize their expectations in using this method. Specifically, a lack of accuracy in predicting customer sentiments and the inflexibility of methods in adapting to different business domains represent the main causes of this disillusionment. The deployment of text mining (TM) models has clear managerial implications, including the availability of accurate and timely information, for better informed decision-making [5]. In conclusion, firms need to apply TM, which is a rapidly growing field, to such huge amounts of online customer feedback data, to summarize them, do trends analyses, and visualize the results. Web TM can analyse large volumes of online customer complaints in order to identify the reasons behind consumer behavior.

The paper is organized as follows: Next section provides the literature. A brief introduction is given about the methodology which is used to cluster the online complaints. After experimental analysis of the study the results are given. A brief conclusion ends the paper.

II. LITERATURE

TM has become an increasingly important tool in order to extract useful and unknown knowledge from business text-based online data. Hu and Liu [6] proposed a number of techniques for mining opinion features from product reviews based on DM (data mining) and NLP (natural language processing) methods to summarize all the customer reviews of a product sold online. Weng and Liu [7] proposed a template for e-mails with multiple questions. They used text classification techniques applied to e-mail reply template suggestions in order to lower the burden of customer service personnel in responding to e-mails. Tsai and Kwee [8] explored the feasibility and performance of novelty mining and database optimization of business blogs. Novelty mining could help to single out novel information from a massive set of text documents. They also used TM, NLP and categorization techniques. The results showed that the novelty mining system detected novelty in their dataset of business blogs with very high accuracy. Onishi and Manchanda [9] assembled a unique data set from Japan that contained market outcomes (sales) for new products, new media (blogs) and traditional media (TV advertising) in the movie category. The authors adopted a primarily empirical approach in the analysis. They specified a simultaneous equation log-linear system for market outcomes and the volume of blogs, and they used novel text-mining analysis. Armentano, Godoy and Amandi [10] aimed to determine the impact of different profiling strategies based on the text analysis of micro-blogs as well as several factors that allowed the identification of users acting as good information sources. Experimental evaluation using a dataset containing a sample of a Twitter social graph and the tweets of each user in this graph was carried out to validate the approach and compare the

performance of the proposed profiling strategies. Thorleuchter and Van den Poel [11] analysed the impact of textual information from e-commerce companies' websites on their commercial success. They extracted information from the web content of e-commerce companies divided into the Top 100 most successful companies worldwide and into the Top 101 to 500 most successful companies worldwide. It was shown that latent semantic concepts extracted from the analysis of textual information could be adopted as success factors for a Top 100 e-commerce company classification. Kahya-Ozyirmidokuz and Ozyirmidokuz [12] mined 200 popular Turkish shopping firms' Facebook websites in order to extract patterns about the firms.

Sentiment analysis is an important tool to cluster online business data according to determine whether the online data is positive, negative or neutral. There are several studies which detect sentiment groups of the customers. Breen [13] used sentiment analysis to estimate an airline's consumer tweet sentiments by counting the number of occurrences of positive and negative words with R software. He compared the score distributions for five international airlines. Barbosa et al. [14] analysed more than one million reviews and numerical ratings of hotels which were extracted from TripAdvisor web site. Reviews were classified as positive or negative using three sentiment analysis tools. Chong et al. [15] presented a model for big data architecture, in combination with sentimental and neural network analysis that can facilitate future business research for predicting product sales in an online environment.

Researchers used TM [16] to understand, to cluster, or to classify the online customer feedback. Gamon [4] demonstrated that it was possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. The data set consisted of 11399 feedback items from a Global Support Services survey, and 29485 feedback items from a Knowledge Base survey for a total of 40884 items. Gamon used NLP 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. [4] presented a prototype system, code-named Pulse, for mining topics and sentiment orientation jointly from free text customer feedback. They described the application of the prototype system to a database of car reviews; it is a simple but effective technique for clustering sentences, the application of a bootstrapping approach to sentiment classification, and as a novel user-interface. The database contained 406,818 customer car reviews written over a four year period, with no editing beyond simple filtering for profanity. Ittoo, Zhang and Jiao [17] 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. The system employed TM techniques to learn product features, and accordingly recommends products that match the customers' preferences. The effectiveness of the suggested recommendation methodology was validated by experimental evaluations. Coussement, and Van den Poel [18] 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 Term Frequency Inverse Document Frequency (TF-IDF)

were used. Boosting was applied as the main classification technique for discriminating complaints from non-complaints throughout the study.

Zhan, Loh, and Liu [19] discovered and extracted salient topics from a set of online textual reviews and further ranked these topics via mining techniques. The results of the study demonstrated that the proposed approach achieved better summarization performance and users' satisfaction when compared to the approaches of opinion mining and clustering summarization. Ozyurt and Kose [20] analysed chat conversations, which were very different from conventional text. They determined the characteristics of conversations via machine learning and DM methods. They used Naive Bayes, k-Nearest Neighbor and Support Vector Machines. Thorleuchter, Van den Poel and Prinzie [21] introduced idea mining as a process for 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. They used methods from TM and text classification (tokenization, term filtering methods, Euclidean distance measure etc.) and combined them with a new heuristic measure for mining ideas. In addition, Sitko-Lutek et al. [1] analysed complaint document reviews, discussion groups and interviews with social network analysis formulation to examine existing information flows formally and informally within a customer complaint handling process, and to identify possible improvement areas to strengthen the effectiveness of this process in the workplace. Key features such as connection strength, point connectivity, and degree centrality were examined. They used UCInet software for the social network analysis.

Zhang, Narayanan and Choudhary [22] presented a feature-based product ranking technique that mined thousands of customer digital camera and television reviews of products from Amazon.com. Sentiment orientations for the sentences were used. The authors focused on the feature-specific ranking obtained by mining the individual product graphs generated for each product feature. In the same year, Somprasertsri and Lalitrojwong [23] proposed an approach for mining product features and opinions based on the consideration of syntactic information and semantic information. They used customer reviews on digital cameras. The sentences in the dataset manually generated tags indicating product features and opinions. They conducted 5-fold cross validation on that dataset, and employed the OpenNLP maximum entropy package for classification. In addition, they used precision, recall, and F-score to measure the effectiveness of their approach.

Thorleuchter, Van den Poel and Prinzie [24] used Web TM techniques. They showed how latent semantic concepts from the textual information of existing customers' websites could be used to uncover the characteristics of companies' websites that could be used to produce profitable customers. They used a regression model based on these concepts to predict new customers.

Ur-Rahman and Harding [25] focused on the use of hybrid applications of TM or textual DM techniques to classify textual data into two different classes. They applied clustering techniques at the first stage and Apriori Association Rule Mining at the second stage. Additionally, studies were made to improve the classification accuracies of the classifiers i.e. C4.5, K-NN, Naïve Bayes

and Support Vector Machines. The classification accuracies were measured and the results compared with those of a single term based classification model. The methodology proposed could be used to analyse any free formatted textual data, and in the experimental study it was demonstrated on an industrial dataset consisting of Post Project Reviews collected from the construction industry. He, Zha, and Li [26] increased competitive advantage and effectively assessed the competitive environment of businesses. Companies need to monitor and analyse 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 analyse 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 and Ozyirmidokuz [12] analysed the top seven heating systems firms' customer complaint documents. They collected data over the period between December 2012 and October 2013. Not only did they extract knowledge about the customers but also about the firms in the sector. Ordenes et al. [5] 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 co-creation processes, interactions, and context. The proposed TM model showed high accuracy levels and provides flexibility through training. They used the Cross Industry Standard Process for DM (CRISP-DM) and the IBM Statistical Package for the Social Sciences Modular as a TM tool to implement two iterations of the TM model. Stoica and Kahya-Ozyirmidokuz [27] mined customer feedback documents to extract knowledge from unstructured customer feedback documents. They used NLP techniques. They determined the similarities of the customers. Jiang et al. [28] proposed a two-stage approach that employs latent class analysis (LCA): the feature-mention matrix construction stage and the LCA-based customer segmentation stage. The approach considered reviewers' mention on product features, and the probability-based LCA method is adopted upon the characteristics of online reviews, to effectively cluster reviewers into specified segmentations. Lee and Suh [29] analysed ideas and comments from MyStarbucksIdea.com community. They extracted features from idea, comment and user information. They applied sentiment analysis and developed classification models to identify potential idea launchers, using DM techniques such as artificial neural network, decision tree and Bayesian network.

Analysing customer complaints must start with an effective segmentation strategy, analysing vast mountains of past data to create clusters of customers and prospects which demonstrate similar purchasing behaviors that are meaningfully predictive to derive statistically significant results. An adaptive cluster model based on TM is built that is successful in grouping the documents. The aim of this research is automatically clustering the online complaints which are about the customers ignoring their subscriptions in order to understand a specific group of the complaints. In this research the customer complaints of a telecommunication firm and those of its biggest competitor are analysed in order to understand the feedback. The positive secondary feedback of the customers are also analysed to make an adaptation in the process. In this manner, the two biggest telecommunication firms' customer complaint feedback which are about ignoring the subscriptions are mined to automate the analysis of customer feedback through a

TM model. In the experimental analysis, a total of 809 documents are used dating between the dates January 1 2013- 31 December 2013. In this research, a customer adaptive feedback methodology is presented. Online Turkish customer complaint web pages which are under the specific complaint topic are automatically mined. The web pages are transformed to a collection of documents by generating a document for each record. TM preprocessing process which represents each document as a feature vector. After the finalization of the NLP preprocessing phase [30] which includes tokenization, transforming, filtering and stemming algorithms, the clustering models begin the knowledge extraction stage. The clustering phase employs the document Cosine similarity process by clustering algorithms.

III. MINING UNSTRUCTURED DATA

DM extracts information from a structured database and it has limited ability in handling huge amounts of unstructured textual documents. TM is the set of techniques and methods used for the automatic processing of natural language text data available in reasonably large quantities in the form of computer files, with the aim of extracting and structuring their contents, and themes, for the purposes of rapid (non-literary) analysis, the discovery of hidden data, or automatic decision making [31]. TM, also referred to as text DM, involves the application of techniques from areas such as information extraction, NLP and information retrieval. After TM extracts meaningful numeric indices from unstructured data, it makes the information contained in the text accessible to the various DM algorithms.

There is a CRISP-DM process model for industrial DM applications which provides a structured approach to planning a knowledge discovery process project. It moves away from this focus on technology by addressing the needs of all levels of users in developing DM technology to solve business problems. Starting from the embryonic knowledge discovery processes used in industry today and responding directly to user requirements, this project defined and validated a DM process that is generally applicable in diverse industry sectors. This will make large DM projects faster, more efficient, more reliable, more manageable, and less costly [32]. CRISP-DM was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting DM into the general problem solving strategy of a business or research unit. According to CRISP-DM, a given DM project has a life cycle consisting of six phases, as illustrated in Fig. 1 [33].

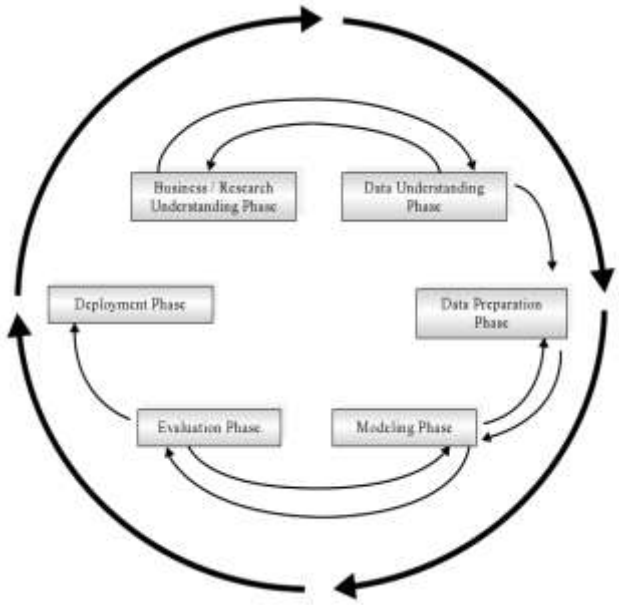


Figure 1. CRISP-DM process
Source: Larose, 2005, p.6

3.1 Data preparation

Text generally requires more preparation than the structured data used for other types of DM. After identifying the source of the text, the biggest challenge is to parse it into words and phrases [34]. One of the challenges of TM is converting unstructured and semi structured text into the structured vector-space model. This must be done prior to doing any advanced TM or analytics. The possible steps of text preprocessing are the same for all TM tasks, though which processing steps are chosen depends on the task. The basic steps are as follows [35]:

- Choose the scope of the text to be processed (documents, paragraphs, etc.).
- Tokenize: Break text into discrete words called tokens.
- Remove stopwords ("stopping"): Remove common words such as the.
- Stem: Remove prefixes and suffixes to normalize words—for example, run, running, and runs would all be steamed to run.
- Normalize spelling: unify misspellings and other spelling variations into a single token.
- Detect sentence boundaries: Mark the ends of sentences.
- Normalize case: Convert the text to either all lower or all upper case.

After text preprocessing has been completed, the individual word tokens must be transformed into a vector representation suitable for input into TM algorithms. This vector representation can take one of three different forms: a binary representation, an integer count, or a float-valued weighted vector. Storing text as weighted vectors first requires choosing a weighting scheme. The most popular scheme is the TF-IDF weighting approach. The term frequency for a term is the number of times the term appears in a document. Document frequency for a term is the

number of documents that contain a given term. The Equations for these values are given in Equation 1, 2 and 3. The assumption behind TF-IDF is that words with high term frequency should receive high weight unless they also have high document frequency. The word the is one of the most commonly occurring words in the English language. The often occurs many times within a single document, but it also occurs in nearly every document. These two competing effects cancel out to give a low weight [35].

$$tf \cdot idft, = tft, d \cdot idf(t) \quad (1)$$

$$tft, = i \in dd1di = t \quad (2)$$

$$id(t) = logDd \in DD \mid t \in d \quad (3)$$

A word vector is just a fancy name for a table, where each row is a document, and each column is a unique Word in the corpus (all of the words in all of your documents). The values inside the table depend on the type of word vector you are creating. In this case we are using Term Occurrences, meaning that a value in a cell represents the number of times that Word appeared in that document. You could also use Binary Term Occurrences, meaning the value in the cell will be zero if the Word did not appear in that document, an done if the Word appeared one or more times in that document. It is always a good idea to examine your data, in order to "get a feel" for it, and to look for strange anomalies [36].

3.2 Clustering

Clustering is related to many disciplines and plays an important role in a broad range of applications. The applications of clustering usually deal with large datasets and data with many attributes. Exploration of such data is a subject of DM [37]. Clustering is an automated process that groups all input documents into clusters, based on similarities. It is an unsupervised process, where no prior information is available about the documents. Early cluster analysis was focused on providing an adaptive method of browsing a document collection when a query could not be created. Subsequently, clustering was applied to query-based clustering on document collections using a hierarchical clustering method. The basic idea in clustering of document collections is to form some sort of similarity or distance measure and then group documents together so the similarities or distances meet some objective function. The next phase of NLP was concerned primarily with understanding the meaning and context of the information, rather than focusing just on the words themselves. Subsequent developments in NLP moved into bibliometrics to consider the context of documents [35].

Since clustering basically involves grouping objects based on their interrelationships of the original feature space. The key insight is that if one can find a similarity measure that is appropriate for the problem domain, then a single number can capture the essential "closeness" of a given pair of objects, and any further analysis can be based only on these numbers. Once this is done, the original high-dimensional space is not dealt with at all; we only work in the transformed similarity space, and subsequent processing is independent of the dimensionality of the data [38].

Before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems [39]. 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 [40]. Given two documents $\vec{t}_a$ and $\vec{t}_b$, their cosine similarity is as follows:

$$SIM_c(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a . \vec{t}_b}{\left|\vec{t}_a\right| \times \left|\vec{t}_b\right|} \quad (4)$$

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] [39].

k-medoids which is a clustering algorithm In k-medoids methods a cluster is represented by one of its points. We have already mentioned that this is an easy solution because it covers any attribute type and medoids are insensitive to outliers because peripheral cluster points do not affect them. When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and the corresponding medoid [37]. In contrast to k-means algorithm, k-medoids is more robust to noise and outliers.

IV. EXPERIMENTAL ANALYSIS

Customers generally use Internet when they have a problem about a product. Therefore, the ever increasing volumes of complaint data shared on social media and web sites have become an important source of knowledge. It is also hard to mine unstructured Turkish data to achieve clusters. To achieve this, we used CRISP-DM process which is presented in Figure 1. In the preparation phase, NLP techniques are used and then clustering is applied.

Figure 2 presents the adaptive feedback model of the study. Turkish customer complaints are collected via a website (WWW) which is a third party marketing research company about the two biggest telecommunication firms in Turkey. We download the data with web mining techniques and we store them in a database. The data which we used are already grouped online. We used a specific topic of the complaints which are about ignoring subscriptions of the customers. After applying web mining techniques, the complaints are converted to free-formed

documents, which then become ready for clustering by applying TM techniques. After clustering, managers see the results and they are effective in deciding the rules. The rules are extracted to send emails to the customers. An example of a rule is given as follows, "if the document is in Cluster 1 then send email1". Customers receive the emails automatically. According to Fig. 2, the customers can re-write a positive or negative response in the process. Then there could be a new clustering for their responses.

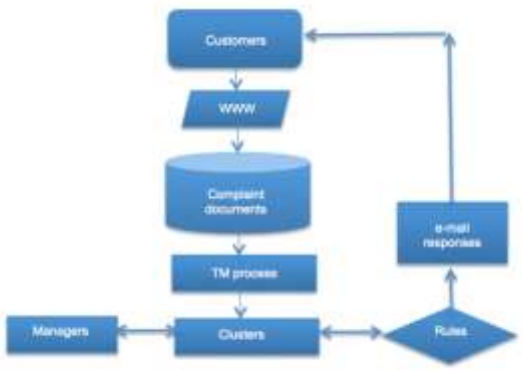


Figure 2. Customer feedback process
Source: Own processing

The complaints were automatically collected in 2013; 355 and 454 Turkish text-based documents for the Firms 1 and 2 were used covering a one year time period, respectively. The customer responds of the firms were also collected. The web pages are converted to documents.

In the TM process, TF-IDF numerical static [35] is applied which reflects how important a word is to a document in a collection. Data are transformed to a collection of documents by generating a document for each record. Every complaint is transformed to a document. In the TM process, we firstly applied tokenization which is the process of breaking up a stream of text into tokens which are meaningful elements. Then we used the Snowball Turkish stemming algorithm which finds the stems of the words. We removed all the words composed of less than 2 characters and more than 25 characters from the documents by filtering tokens by length. Also, we deleted all the Turkish stop-words (and, the, because, although, etc.) because they are redundant. We also removed all the html codes, the other unnecessary codes and repeats from the documents. Finally, we applied the n-grams algorithm to include the word groups and count them together, and used TM techniques. After the TM process, documents are grouped into 4 clusters. According to the 4 clusters, similar complaints are replied to with the same e-mail. Then we collect the secondary customer feedback and measure the ratios and compare the feedback per firm.

We used the RapidMiner [41] tool. RapidMiner, which is used by business analysts and scientists around the world, looks beyond what has happened, and helps predict what might happen -

customer churn, factory breakdowns, better results from changes in advertising and marketing campaigns, as examples [42].

V. RESULTS

One of the outputs of the text preprocessing process is the example set. The example set of Firm 1 has 355 examples with 10 special attributes and 6,181 regular attributes. The term "complaint" is the most frequently used word in the documents, and "detail, message, firm" are the other attributes that are most frequently used. The example set of Firm 2 has 454 examples with 15 special attributes and 3,824 regular attributes. The term "message" is the most frequently used word in the documents, and "detail, complaint, like, ignore" are the other attributes that are most frequently used. Table 1 presents the most frequently used words in the complaint documents for the firms. According to the Firm 1, "complaint" is the most commonly used word as compared to the Firm 2.

TABLE I. THE FREQUENCIES OF THE MOST USED WORDS IN DOCUMENTS PER FIRM

Firm id	Words in English meaning	Total occurrences	Number of documents
1	complaint	22932	355
1	message	15601	354
1	detail	13422	355
1	category	3985	355
1	firm	3894	355
2	message	17472	448
2	detail	16653	453
2	complaint	29907	453
2	like	10788	448
2	ignore	6371	451

Source: Own computation

The extracted document vectors are used in similarity anlaysis. Cosine similarity is used to visualize the relationships between documents. Fig. 3 presents the similarity graphs. One can see from the graphs that the documents about customers ignoring their subscriptions are similar to each other. In addition, these similarities are not connected to the other nodes.

Once a data matrix has been computed from the documents and words found in those documents, various well-known methods can be used for further processing those data including methods for clustering, factoring, or predictive DM [43]. We group together

documents that look similar to each other by using k-medoids clustering [37]. In k-medoids, the centroid is a record of the cluster that minimizes the sum of the distances with respect to all the other points of the cluster. Four groups are obtained (k=4). We decide the appropriate number of clusters (k value) for the documents according to a trial-and-error process. In contrast to the k-means algorithm, the k-medoids is more robust to noise and outliers. The distance measure is defined as a parameter in order to enable evaluation of the models with the same distance measure, as was used for in the creation of models [44]. Euclidean distance is used. After execution of the process, 59,49,153 and 94 items for Firm 1 were obtained. For Firm 2, 29,19,189 and 217 items were obtained. In general, when the content of documents are investigated, one can easily understand that Cluster 1 belongs to "withdrawal fee". Cluster 2 has documents about the "cancellation problems". The "unnecessary or too many bill problems" is clustered in the documents which belong to Cluster 3. Cluster 4 documents are about the "other reasons" for ignoring the sim card.

According to the similarity distances, and Fig. 3 (a), document 1 is similar to document 15 with a distance "1". Document 1 is similar to document 2 with a distance "1.41421". In Fig. 3 (b), document 319 is similar to document 338 with a distance "0.99999". All the distance measures and all the word lists per document can be seen from the extracted tables. In general, the complaint documents per firm show similar results.

Firm 1's customers generally complain about the high price of the ignoring process although 4% of their customers are happy because their problems are solved when compared with Firm 2. Both firms can use the feedback model to respond the customers automatically. The customers of Firm 2 talk about "liking something".

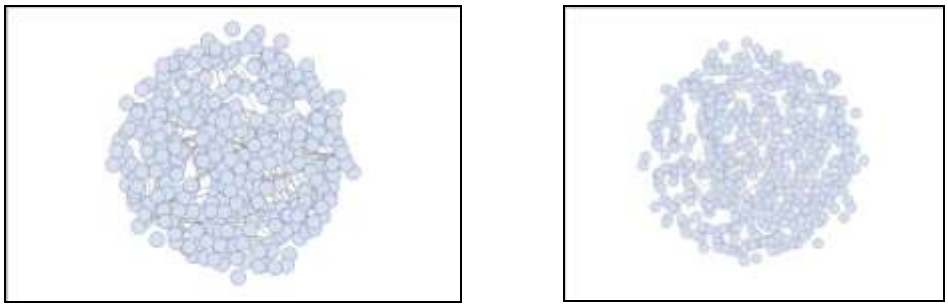
Performance operators, which can be used to derive a performance measure (in the form of a performance vector) from the dataset, are used. The performance vectors of the models' cluster number indexes are both 0.996.

We also measure the secondary feedback of the customers about their complaints after the firms' responses to the customers to make an adaptation. The secondary feedback data are automatically collected from the website. Two hundred documents for Firm 1 and 240 documents for Firm 2 are examined. Forty documents for Firm 1 and 60 for Firm 2 are positive feedback. Comparing negative complaints with positive complaints, one can see that firms have very low positive response ratios; 4% of positive feedback from Firm 1, and 2% of positive feedback from Firm 2. Most importantly, firms could not reply to all of the complaints. If they could response to the customers automatically, the positive feedback would have been a better ratio. Thus, it is necessary to cluster and automatically reply to customers.

NLP methods are applied to the secondary positive feedback documents. Cosine similarity is used to visualize the relationships between documents. Fig. 4 shows the similarity graphs of the customers' secondary positive feedback per firm. In addition, k-means clustering with Bregman Divergences with Squared Euclidean Distance is also used. Two clusters are decided (k=2).

According to the Firm 1, Cluster 1 has only 6 elements. The other documents are all about thanking the firm for providing solutions. The term "problem" is the most frequently used word in the documents, and "solution, thank, service" are the other attributes that are most frequently used. For the Firm 2, "interest" is the most commonly used word. The other attributes are "thank", "short time", "problem" and "authorized person". Cluster 1, in which the documents are related with short time problem solving, has 27 elements.

The performance vector of the model's cluster number index is 0.967 for Firm 1 and for Firm 2, the cluster number index is 0.950. After examining both firms' positive feedback, one can easily see that the Firm 2's documents have text about solving problems in a short time in comparison to the Firm 1.



(a) (b)

Figure 3. Customer feedback process
Source: Own processing

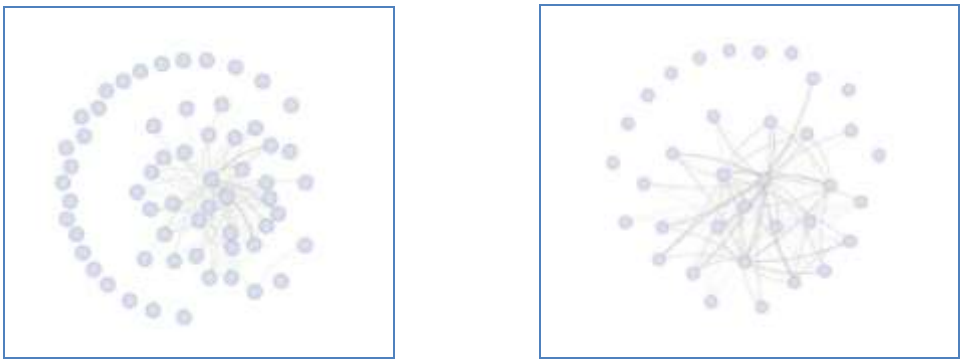


Figure 4. Similarity graphs for secondary feedback per firm
Source: Own processing

Similarity analysis is used to determine similar documents. The similarities of firms about the subject are determined. Graphs and tables are obtained. Every cluster has similar solution documents. We also compare the firms in terms of complaint groups. Similar queries about complaint return documents are grouped into categories such as similar customers' response

mails. These findings are also valuable for understanding customer behaviors. In conclusion, the customer feedback system's adaptive results are achieved.

VI. CONCLUSIONS

Analyzing customer complaints represents an opportunity for a firm to understand their customers. Feedback data must be clustered without loss of information for an effective feedback management system. Additionally, managers can understand feedback without being overwhelmed if the documents are clustered automatically. Managers can use these extracted knowledge to understand customers and then to give business decisions.

The cost of ignoring online customer complaints is very important for a firm. Analysing customer complaints represents an opportunity for a firm to understand their customers. Customer feedback data must be clustered without loss of information for an effective feedback management system. By clustering the customers, managers could be able to make better decisions and could identify significant business opportunities. Consequently, analysing complaint documents will result in a competitive advantage for a firm. We present an adaptive feedback model to cluster the customer complaints. Managers will monitor their customers' opinions and they can give decisions by using this model. In addition, they save time by understanding the online huge amount of unstructured data with TM, instead of reading it through page by page. In addition, there is no danger of overlooking data. New, interesting and hidden knowledge which can be used in managerial decision making are achieved in this research.

Cluster analysis is an important starting point for other purposes. Therefore, in further work we could apply qualitative research to the clustered documents to extract the behaviors of customers and then we could consider extending our analysis to include a predictive approach. In further research, we should collect customers' online big data to integrate customers and the power of the crowd at runtime.

ACKNOWLEDGMENT

This work was supported by Erciyes University Research Fund, Project Number FBA-2014-5364.

REFERENCES

[1] Sitko-Lutek, A., Chuancharoen, S., Sukpitikul, A., Phusavat, K.: Applying social network analysis on customer complaint handling, Industrial Management & Data Systems 110 (9) 1402-1419 (2010)
[2] Stoica, E.A., Pitic, A.G., Calin, B.: New Media E-marketing Campaign. Case Study for a Romanian Press Trust, Procedia Economics and Finance, 16, 635 – 640 (2014)I. S. J
[3] Sullivan, D.: Document warehousing and TM: Techniques for improving business operations, marketing, and sales, John Wiley&Sons, Inc., Canada (2001)

[4] Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: Mining customer opinions from free text. In:Famili, A.F., et al. (eds.) IDA 2005, LNCS 3646. pp. 121-132. Springer-Verlag Berlin Heidelberg (2005)
[5] . Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M.: Analysing Customer Experience Feedback Using TM: A Linguistics-Based Approach, Journal of Service Research 1-18 (2014)
[6] . Hu, M., Liu, B.: Mining Opinion Features in Customer Reviews. In: Proceedings of Nineteenth National Conference on Artificial Intellgience (AAAI, American Association for Artificial Intelligence). pp.755-760. http://www.aaai.org/Papers/AAAI/2004/AAAI04-119.pdf (2004)
[7] . Weng, S.-S., Liu, C.-K.: Using text classification and multiple concepts to answer e-mails, Expert Systems with Applications, 26, 529–543 (2004)
[8] . Tsai, S., Kwee, A.T.: Database optimization for novelty mining of business blogs, Expert Systems with Applications, 38, 11040–11047 (2011)
[9] . Onishi, H., Manchanda, P.: Marketing activity, blogging and sales, Intern. J. of Research in Marketing 29, 221–234 (2012)
[10] Armentano, M.G., Godoy, D., Amandi, A.A.: "Followee recommendation based on text analysis of micro-blogging activity", Information Systems, 38, 1116-1127 (2013)
[11] Thorleuchter, D., Van den Poel, D.: Predicting e-commerce company success by mining the text of its publicly-accessible Website, Expert Systems with Applications, 39, 13026–13034 (2012)
[12] Kahya-Ozyirmidokuz, E., Kahya-Ozyirmidokuz, M.H.: Analysing customer complaints : A Web TM application. In: F. Uslu (Ed.) International Conference on Education and Social Sciences (INTCESS14) pp.734-743. Ocerint, İstanbul-Turkey (2014)
[13] Breen, J.O: Mining Twitter for airline consumer sentiment, Delen, D., Fast, A., Hill, T., Elder, J. Miner, G., Nisbet, B. (Eds.), Practical TM and Statistical Analysis for Non-structured Text Data Applications, Elsevier, pp.133-149 (2012)
[14] Barbosa, R.R.L., Sánchez-Alonso, S., Sicilia-Urban, M.A.: Evaluating hotels rating prediction based on sentiment analysis services, Aslib Journal of Information Management, 67 (4) 392-407 (2015)
[15] Chong, A.L.Y., Li, B., Ngai, E.W.T., Ch'ng, E., Lee, F.: Predicting online product sales via online reviews, sentiments, and promotion strategies, International Journal of Operations & Production Management, 36 (4) 358-383 (2012)
[16] Kahya-Ozyirmidokuz, E.: Analysing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey, Information Development, SAGE, 1-12 (2014)
[17] Ittoo, A.R., Zhang, Y., R. Jiao, J.: A TM based recommendation system for customer decision making in online product customization, International Conference on Management of innovation and technology, IEEE, 1, 473-477 (2006)
[18] Coussement, K., Van den Poel, D.: Improving customer complaint management by automatic email classification using linguistic style features as predictors, Decision Support Systems, 44, 870–882 (2008)
[19] Zhan, J., Loh, H. T., Liu, Y.: Gather customer concerns from online product reviews – A text summarization approach, Expert Systems with Applications, 36, 2107–2115 (2009)
[20] Ozyurt, O., Kose, C.: Chat mining: Automatically determination of chat conversations' topic in Turkish text based chat mediums, Expert Systems with Applications, 37, 8705–8710 (2010)
[21] Thorleuchter, D., Van den Poel, D., Prinzie, A.: Mining ideas from textual information, Expert Systems with Applications, 37, 7182–7188 (2010)

[22]Zhang, K., Narayanan, R., Choudhary, A.: Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking. In: 3rd Conference on Online social networks (WOSN'10) USENIX Association Berkeley. pp. 11. CA, Berkeley, USA, https://www.usenix.org/legacy/event/wosn10/tech/full_papers/Zhang.pdf (2010)
[23]Somprasertsri, G., Lalitrojwong, P.: Mining feature-opinion in online customer reviews for opinion summarization, Journal of Universal Computer Science 16 (6) 938-955 (2010)
[24]Thorleuchter, D., Van den Poel, D., Prinzie, A.: Analysing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing, Expert Systems with Applications, 39, 2597–2605 (2012)
[25]Ur-Rahman, N., Harding, J.A.: Textual DM for industrial knowledge management and text classification: A business oriented approach, Expert Systems with Applications, 39, 4729–4739 (2012)
[26]He, W., Zha, S., Li, L.: Social media competitive analysis and TM: A case study in the pizza industry, International Journal of Information Management, 33 (3) 464-472 (2013)
[27]Stoica, E.A., Kahya Ozyirmidokuz, E.: Mining customer feedback documents, International Journal of Knowledge Engineering, 1 (1) 68-71 (2015)
[28]Jiang, S., Cai, Shuqin, Olle, G.O., Qin, Z.: Durable product review mining for customer segmentation, Kybernetes, 44 (1) 124 – 138 (2015)
[29]Lee, H., Suh, Y.: Who creates value in a user innovation community? A case study of MyStarbucksIdea.com, Online Information Review 40 (2) 170 – 186 (2016)
[30]Kao, A., Poteet, S.R.: NLP and TM, Springer (2007)
[31]Tuffery, S.: DM and Statistics for Decision Making. Wiley. (2011).
[32]Sumathi, S., Sivanandam, S.N.: Introduction to DM and its applications, Springer (2006)
[33]Larose, D.T.: Discovering Knowledge in Data: An Introduction to DM. Wiley. (2005).
[34]Berry, M.J.A., Linoff, G.S.: Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Third Ed., Wiley, New York (2011)
[35]Delen, D., Fast, A., Hill, T., Elder, J. Miner, G., Nisbet, B.: Practical TM and Statistical Analysis for Non-structured Text Data Applications, Elsevier (2012)
[36]McGuigan, N.: Detecting text message spam. In: Hofmann, M., Klinkenberg, R. (eds.) RapidMiner: DM use cases and business analytics applications. pp. 199-212. CRC Press, USA (2014)
[37]Berkhin, P.: A survey of clustering DM techniques. In: Kogan, J, Nicholas, C., Teboulle, M. (eds.) Grouping multidimensional data: Recent advances in clustering. pp. 25-71. Springer, USA (2006)
[38]Ghosh, J, Strehl, A.: Similarity-based text clustering: A comparative study. In: Kogan, J, Nicholas, C., Teboulle, M. (eds.) Grouping multidimensional data: Recent advances in clustering. pp. 73-97. Springer, USA (2006)
[39]Huang, A.: Similarity Measures for Text Document Clustering. In: J. Holland, A. Nicholas, D. Brignoli, (eds.) New Zealand Computer Science Research Student Conference NZCSRSC. pp. 49-56. http://www.milanmirkovic.com/wp-content/uploads/2012/10/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf (2008)
[40]Weiss, S.M., Indurkhya, N., Zhang, T., et al.: TM: Predictive Methods for Analysing Unstructured Information Springer, USA (2005)
[41]Ertek, G., Tapucu, D., Arin, I.: Text mining with RapidMiner. In: Hofmann, M., Klinkenberg, R. (eds.) RapidMiner: DM use cases and business analytics applications. pp. 241-261. CRC Press, USA (2014)

[42]Fayyad, U.: Why Rapidminer? By Usama Fayyad, a Top Data Scientist and Entrepreneur, http://www.kdnuggets.com/2013/12/why-rapidminer-by-usama-fayyad-top-data-scientist-entrepreneur.html (2015)
[43]Manning, C.D., Schütze, H.: Foundations of Statistical NLP, MIT Press (1999)
[44]Vukicevic, M., Jovanovic, MM., Delibasic, B., Suknovic, M.: Grouping higher education students with RapidMiner. In: Hofmann, M., Klinkenberg, R. (eds.) RapidMiner: DM use cases and business analytics application. pp. 185-195. CRC Press, USA, (2014)