Ontology Based Aspect Oriented Opinion Summary Methodology

Dervis Kanbur
Kuveyt Turk P. Bank R.D. Center, Kocaeli, Turkey
Department of Computer Engineering
Yildiz Technical University
Istanbul, Turkey
E-mail: dervis.kanbur@std.yildiz.edu.tr

Mehmet S. Aktas
Department of Computer Engineering
Yildiz Technical University
Istanbul, Turkey
E-mail: mehmet@ce.yildiz.edu.tr

Abstract— **Enterprises selling over the web generally request feedbacks from customers on their purchased products/services. Thanks to e-commerce in an increasingly developing structure, the number of customer feedbacks grows rapidly. Due to the great increase in the number of e-commerce enterprises and customers, it becomes difficult for a potential customer to read these feedbacks while making decisions. It becomes almost impossible for the producer to monitor these feedbacks as well. Product feature extraction from customer reviews is an important operation in opinion mining. The extracted features help to assess the opinions written by customers who have purchased specific products and they provide opinions of customers regarding their positive/negative experiences. Because most of the customer reviews are asyntactic plain texts, methods should be developed for extraction of implicit and explicit product features expressed in customer reviews and comments. In this research, we aim to develop a methodology that examines the opinions, thoughts and comments, written in Turkish, about the products and summarizes the features of the products and the emotional processes related to these properties. Our study differs from others in that product characteristics expressed using synonyms or word groups can be identified, and features can be detected with greater accuracy using ontologies that include product specifications. Our study offers successful results in product feature extraction. Experimental work on opinion/idea/comment for some products on e-commerce sites shows that our methodology yields positive results.**

Keywords-text mining; sentiment classification; review summarizing; synonym word group; ontology;

I. INTRODUCTION

In the developing and globalizing world, measuring and evaluating customer satisfaction for organizations operating in the trade field and using this information to improve customer experience has become a key to gain advantage in today's competitive market. Thus, organizations that develop new products / services or market improved products / services will be able to effectively increase service qualities, capture customer trust and reach a stable customer portfolio in product usage.

Because the number of skilled workers is likely disproportional to the workload, the number of customers and products in today's commercial enterprises, monitoring customer satisfaction becomes quite difficult/impossible. The features such as obtaining written opinions/suggestions of customers on products over the Internet, rapid analysis of these data and assessments on products with minimum human resources in a fast way, shall free the way of doing business actively and developing products according to customers' requests for enterprises and enable maximum customer satisfaction.

With the development of technology from the other side, the rapid development of the virtual store network increases the number of online shopping and the variety of products sold. Through the virtual shopping world, the need for comments of customers who have purchased the same product before has become prevalent. Nevertheless, customers become used to writing feedbacks for products they purchase. As a result, thereof, products with more comments indicate an increased popularity. Thus, some products collect more comments. Hundreds of comments on these products from different sources (in other words, different e-commerce sites) make it difficult to assess the general satisfaction degree of customers browsing through these comments. It is not possible to make a general deduction for customers who read a part of these comments.

At the early undeveloped stages of data handling, we observe that data analysis reveals the competitive aspects of firms. As time passed, the competitors acquiring these capabilities have removed this advantage. So, for today's enterprises, development of more complex data analysis approaches become obligatory, in order to get a step ahead. To help marketing people get ahead safely in this challenging competition environment, an independently sponsored research unit, Harvard Business Review Analytic Services [7] within the Harvard Business Review Group has researched how leading marketing enterprises integrated customer and marketing data and how they used it to increase their companies' performance. This research has revealed that the data source used for customers and marketing activities were the most important element in competition.

As the Internet world grows up and social media is used more frequently, it becomes possible to take customer reviews through various channels. It is essential to respond positively to customer needs in sales & marketing. Monitoring and responding to these needs shall help companies in reaching maximum customer satisfaction and provide an improvement of marketing data at a significant rate. In this case, the need for an application capable of being implemented rapidly and easily by companies, analyzing

 ISBN: 978-1-61208-557-9

customer reviews, performing certain deductions on the product according to the analysis is among the reason motivating our research.

The aim of this research is to increase the quality of products/services in line with feedbacks of customers and ensure customer satisfaction within the product sales and service sector and to develop a methodology for creating mutual satisfaction through win-win strategy. The method we propose in this research shall collect customers' feedbacks dynamically, make deductions on product features by analyzing the reviews, and enable figuring out the positive/negative aspects of products.

A. Research Questions

In order to achieve the motivation mentioned above, the research questions we have determined are as follows:

1) R.Q-1: What should be a method that can process customer comments for products sold on the Internet and automatically extract the properties of the products and then summarize the positive or negative opinions of the customers based on their product characteristics? How should the architecture of the software that implements such a method be and how can it be developed?

2) R.Q-2: How should a methodology be developed that will reveal whether the product features revealed by the analysis of the data obtained from the customer comments are the actual product characteristics?

B. Contribution of Research

The increase in online resources where different ideas/opinions/comments are shared, and the fact that these shares are in different formats, reveals the need to develop methods to collect data. Within the scope of this research, a web crawler application has been developed to find, download, parse, and collect customer comments (texts containing comments/ideas/ thoughts) in such sources. In order to extract product features from collected customer comments and to perform feature based sentiment analysis, the following functions have been implemented in order: (1) Extracting product features from Turkish comments, (2) To determine whether each feature is used with the word group that contains positive or negative expression according to Turkish grammar rules (3) Identifying Product features, identified with different words but with the same meaning, and combining under one feature title (4) taking advantage of ontology, comparing actual product properties with determined properties (5) summarizing sentiment analysis results. In the scope of this research, a methodology which can determine the properties of products from the interpretations entered by the customers and can analyze the emotions related to these properties is proposed. A prototype implementation of the developed methodology has been developed and the success of the methodology has been demonstrated through experimental work on the prototype.

C. Organizational Structure

In the Introduction, we talked about our motivation to trigger this research. The main issues related to our similar researches and researches are summarized in literature review. In the part "Proposed Software Architecture", we talked about our methodology for analyzing customer comments and determining product characteristics. In the part "Development", how the methodology used in the system developed in the research is realized is explained. In Section IV, the success of the proposed methodology is examined. In Section V, the results we have obtained after conducting the research are discussed.

II. LITERATURE REVIEW AND BASIS TOPICS

In this part, we will discuss the studies performed on extracting product feature from customer reviews. Generally, these studies are sorted into different classes as frequency based approach [1], statistical approach [11], and relational based approach [2]. Since the methodology we provide in this research is in the frequency-based approach category, we concentrate on studies in this area especially if we do not review the literature. The most prominent work in this category is the method proposed by Hu and Liu, which extracts the most frequent terms from the reviews as product features [1]. In this research, It was assumed that the product features shall be expressed clearly by nouns or noun phrases in the sentence. Association Rule Mining was implemented to find frequently repeated words. When the frequently repeating features' list is generated, all words expressing an opinion around these features are figured out. The main disadvantage of this method is that some words that are not actually product features may be extracted as a feature because they are used frequently. After identifying the features in the texts, it is necessary to reveal the opinions associated with these features. The works done in this area is categorized as opinion mining. In opinion mining studies focuses on the following processes:

A. Subjective Classification

In this category of studies, a given text/sentence is searched if it contains any opinion [1] [12]. These studies consider whether a sentence contains a positive or negative opinion rather than the general document.

B. Sentiment Classification

It is the process of trying to find out whether a given sentence is positive/negative or neutral. In this category, sentiment based classification studies are inspired by cognitive linguistics [13]. In some studies, the texts were classified with certain specific sentimental words [14]. In these researches, each word was added to a dictionary and it was marked to detect if it has a positive/negative expression. The statistics found using a search engine in a study, which was developed for common languages such as English and an effective study, which uses WordNet, were matched with words and documents and a learning technique was generated [8]. In some studies, sentiment classifications were performed using machine learning techniques [9]. The other studies in this area have not been designed to give knowledge of which opinions indicate which feature.

 ISBN: 978-1-61208-557-9

C. *Opinion Summarizing*

It is a function that is worked on to express multiple opinions or a long opinion text in a short way. Several summarizing techniques basically work on two categories. Creating templates and extracting passages. Most studies in this area were developed in a single document. Whereas some studies were performed on multiple documents containing the same information in order to search for similarities and differences [10].

Our study was performed on Turkish reviews and different words, abbreviations or loan word groups are checked through an available list and the recall value is being increased by combining words which were extracted as different features. It is distinguished from previous studies in its being performed on Turkish reviews, assessment of loan words, and detecting synonyms.

The study we have proposed in this paper is based on Turkish interpretations and allows us to extract more accurate features by combining different word, abbreviation or foreign word groups which express the same feature (in other words, the words that are expressed as different features but which express the same feature). By using field ontologies, it is checked whether the determined properties are met in ontology. Our developed feature extraction method is separated from existing studies by being able to determine on the basis of Turkish interpretations, use of ontologies and synonyms and to collect them under the same heading.

Product features are extracted from multiple customer reviews in our work. In this respect, it differs from the traditional text summarization. A method that works on multiple comments / text clusters instead of a single text has been developed. Our goal here is to extract product features on multiple texts, rather than finding similarities or differences in texts. Again, within the scope of our study, a list of words and phrases for adjectives with positive or negative meaning in the Turkish language was extracted and their sentiment ratings were determined. In our study, sentiment analysis was done at the sentence level, not at the document level. By using the sentimental adjectives in the sentences, positive or negative opinions related to the product features in the sentence were determined.

III. THE PROPOSED METHODLOGY

The details of our methodology given in Figure 1 are given below. The Turkish Feature Identification System works on data (eg: e-commerce site selling X Television Brand, the source of X product) obtained by finding, gathering, decomposing comment texts of a product with a specific source. These data may be obtained from a single source where the product is sold, or it can occur in different environments if the product is sold in more than one source (e-commerce site). These environments may be sampled as e-commerce websites, social media groups, opinion expressing websites. As shown in Figure 1 the study was performed on data acquired from selected e-commerce and opinion expressing web sites. The system is designed to be used for analysis and reporting of customer reviews by integrating it to these web sites. Through the integration of this system with companies' e-commerce websites, it would be possible to observe how the product generates opinions from performed reviews. This would increase the customer satisfaction and contribute positively to the product marketing.

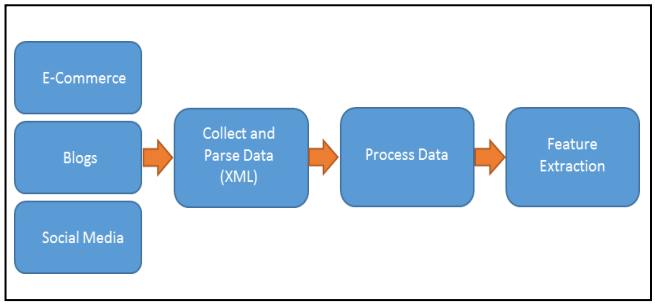

Figure 1. The Proposed Methodoloy Workflow

A. *Data Collection*

In this phase, texts containing comments/ideas/opinions are obtained in the data source. Within the scope of this research, because of the difficulty in finding a ready dataset prepared in Turkish, reviews from two selected websites (arabainceleme.com, hepsiburada.com) were extracted and prepared for processing. However, since the method we propose is extensible, it can be integrated into the system from other data sources so that it can provide output in the format the system expects.

B. *Data Pre-Processing*

In our study, some preprocessing techniques were used. Primarily, stop words likely to recur in reviews, which are required by the language, were cleared. Then, the roots of words were obtained by stemming on the words which made the texts containing comments/opinions/ideas. Within the scope of word rooting function, because some words from the Turkish language had the same roots, there was a problem of not to be able to pick the right structure in stemming. In such cases, the next words were examined and formation of a combined structure was checked.

For example, a sentence including the expression "kullanım kılavuzu(user manual)" may be given as an example of this problem. When we observe the stem of the word "kullanım", the verb "kullan" appears. But its suffix "ım" indicates that this word may be used as a noun. In such words, the following word was examined to recognize its type. Because the word "kılavuzu" has a noun root and it has a noun generating suffix "u", the previous word "kullanım" was assumed as a noun. In following parts of the contribution, the algorithm which discovers Turkish compound words shall be explained in detail.

C. *Extracting Feature From Data*

In the process of extracting features from data, the Zemberek Natural Language Processing library was used [15]. An ontology related to the product features was created by making use of the data dictionary of the Turkish Language Association. Classification Based on Associations

 ISBN: 978-1-61208-557-9

(CBA) algorithm was used to find the frequency of the words in the obtained sentences. Recall and Precision metrics were used to measure the obtained values. The process of extracting properties is made up of the following sub-steps: word tagging, compound word detection, extraction of frequently repetitive product features, extraction of opinion words, extraction of opinion sentence tendencies, and combination of related words. We explain these sub-steps in detail below.

1) Word Tagging

After the removal of stop words, we have tagged the remaining words. Figure 2 shows a tagged sentence after the process.

The example sentence in Figure 2 is as follows: “Tam ihtiyaçlarıma göre bir araç, düşük viteslerde gaza biraz yüklenmek gerekiyor ama hızını aldıktan sonra gerçekten etkileyici (the car design as my needs, in low gears the needs to load to gas a bit, but it's really impressive)”. For each sentence in the comment contained in each comment text, an identifier is assigned to each word, and the tagged data are recorded in XML as shown in Figure 2.

```
<Tag>ADJECTIVE</Tag>Tam</Text>
<Tag>NOUN</Tag><Text>ihtiyaçlarıma</Text>
<Tag>PREPOSITION</Tag><Text>göre</Text>
<Tag>NUMERAL</Tag><Text>bir</Text>
<Tag>NOUN</Tag><Text>araç</Text>
<Tag>NOUN</Tag><Text>düşük</Text>
<Tag>NOUN</Tag><Text>viteslerde</Text>
<Tag>NOUN</Tag><Text>gaza</Text>
<Tag>NOUN</Tag><Text>biraz</Text>
<Tag>VERB</Tag><Text>yüklenmek</Text>
<Tag>VERB</Tag><Text>gerekiyor</Text>
<Tag>NOUN</Tag><Text>ama</Text>
<Tag>NOUN</Tag><Text>hızını</Text>
<Tag>VERB</Tag><Text>aldıktan</Text>
<Tag>TIME</Tag><Text>sonra</Text>
<Tag>ADJECTIVE</Tag><Text>gerçekten</Text>
<Tag>ADJECTIVE</Tag><Text>etkileyici</Text>
```

Figure 2. Example of Tagging Sentence

2) Turkish Compound Word Detection

In our implemented word tagging structure, the word’s structure is extracted in relation to its stem. However, some features of the products may be expressed phrases constructed from noun phrases, adjective clauses, loan words and different compound words. Detection of these structures was determined within the scope of this study. A compound word detection algorithm specifically developed for Turkish language is shown in Figure 3. Within the context of this study, it is assumed that the words that express opinions are adjectives. However, considering that a product feature can also be expressed by adjective phrases, the adjectives used in such a case are considered to be part of the word groups expressing the product characteristic, and no sentiment analysis is performed on the adjectives in this case.

a) Detection of Various Structures from the Same Stem

In the tagging process, data root detection is performed by stemming on the word. However, detection of real tags of words in two different structures with roots having the same letters has appeared as a problem in front of us. We have tried to tag these words with multiple structures by combining them with following words.

For example; "gaz pedalı çok hafif" and "araç gaz ile çalışırken performansı düşüyor", the word "gaz" in the sentences has a different meaning in each term. While the "gas pedal" word group is a name proposition in the first sentence, the other sentence does not have such a formation. For this reason, it is necessary to correctly determine in what sense the word is used in order analyze the sentences correctly. In order to distinguish the words in this situation, we proposed a compound word detection algorithm as shown in Figure 3.

In the given algorithm detail, the sequential operations are as follows:

The block between lines 4-45 processes words in the given sentence.

The block between lines 5-25 handles the cases in which the word root is a noun. In this case:

- If the next word has got a noun-phrase suffix
- If the next word has got one of the nouns –den and –in suffixes
- If the next word is a verb and has got one of the –me (passive voice) -im -in suffixes
- If the word has got noun-phrase -ın suffix
- If the word has got a possession suffix –lı or absence suffix –sız
- If the word has got a noun –de suffix and if the previous word is verb

if the word has one of these conditions, it is deduced that "this word may be part of the combined word".

The block between lines 26-44, handles the cases when the word root is a verb. In this case:

- If the word has got only one suffix and the next word is an adjective
- If one the suffixes of the word is a negation suffix, and the next word is a noun and has got a noun-phrase –ı suffix.
- If the word has got two suffixes and one of them is necessarily an –erek suffix and the previous word is a noun.
- If the word has got two suffixes and one of them is –dik –en or –im, and the next word is a noun and has one of the noun-phrase suffixes –ı –ın –de.

if the word has one of these conditions, it is deduced that "this word may be part of the combined word".

In line 45, if the word is marked as a compound word, it is added to the compound word list.

Lines 50-70 deal with the words on the compound word list. If a word / word group has the following conditions, this word / word group is combined with the next word itself.

- If it is not compound and is a noun and is nominative case

 ISBN: 978-1-61208-557-9

- If it is not compound and has got a noun-phase suffix –ın

```
1 Procedure PhraseDetector
2 List phrasedList=new List;
3 for each word in sentence{
4   isPhrased=false;
5   if word_root is NOUN{
6     if count(word_eco) = 1 and exists(next_word){
7       if (next_word_eco contains ("ISIM_TAMLAMA_I" or  "ISIM_BELIRTME"
8      or "ISIM_CIKMA_DEN" or "ISIM_SAHIPLIK_SEN_IN" or
9      "ISIM_TAMLAMA_IN"))
10       or (next_word_root is VERB and next_word_eco contains
11     ("FIIL_DONUSUM_ME" or "FIIL_EDILGENSESLI_N" or
12     "FIIL_DONUSUM_IM"))
13     or (next_word_root is NOUN and word is Lean and next_word is Lean)
14         isPhrased=true;
15     }
16     else if (word_eco contains("ISIM_BULUNMA_LI" or
17     "ISIM_YOKLUK_SIZ")){
18       if(phrasedList.add(word_root,next_word_root); continue next;
19     }
20     else if (word_eco contains("ISIM_KALMA_DE") and exists(before_word)){
21       if(before_word_root is VERB and !(before_word_eco
22       contains("ZAMAN")))
23         isPhrased=true;
24     }
25   }//end if word_root is NOUN
26   else if (word_root is VERB){
27     if (word_eco.size = 1 and next_word  is not ADJECTIVE) isPhrased=true;
28     else if (word_eco contains("OLUMSUZLUK_ME") and word_eco.size = 2){
29       if(next_word_root is NOUN and next_word_eco
30       contains("ISIM_TAMLAMA_I")) isPhrased=true;
31     }
32     else if (word_eco contains("FIIL_SUREKLILIK_EREK") and
33     word_eco.size = 2){
34       if(exists(before_word) and before_word is NOUN)
35       {    phrasedList.add(before_word_root,word_root); continue next;}
36     }
37     else if (word_eco contains("FIIL_BELIRTME_DIK" or
38     "FIIL_DONUSUM_EN" or "FIIL_DONUSUM_IM") and
39      word_eco.size >= 2){
40         if (next_word_root is NOUN and next_word_eco
41          contains("ISIM_TAMLAMA_DE" or "ISIM_TAMLAMA_IN" or
42          "ISIM_TAMLAMA_I")) isPhrased=true;
43     }
44   }//end if word_root is VERB
45   if (isPhrased)phrasedList.add(word_root,next_word_root);
46 } end for
47
48 List newPhrasedList = new List;
49 for each phrased_word in setBirlesikKelime{
50   boolean isAdd = true;
51   if((phrased_word[0] is FOREIGN_WORD) or phrased_word.size = 1 and
52    phrased_word[0] is NOUN and phrased_word[0] is Lean or
53    phrased_word.size = 1 and phrased_word[0].eco
54    contains("ISIM_TAMLAMA_IN") or phrased_word.size = 2 and
55   phrased_word[1].eco contains("ISIM_TAMLAMA_IN" or
56   "ISIM_TAMLAMA_I")){
57     if (exists next_phrased_word){
58       if(next_phrased_word.size =1 and next_phrased_word[0].eco
59        contains("ISIM_TAMLAMA_I") or next_phrased_word.size =2 and
60       next_phrased_word[1].eco contains("ISIM_TAMLAMA_I")){
61         List<WordSet> kList = new ArrayList(phrased_word[0]);
62         if (phrased_word.size = 2) kList.add(phrased_word[1]);
63         kList.add(next_phrased_word[0]);
64         if (next_phrased_word.size = 2) kList.add(next_phrased_word[1]);
65         newPhrasedList.add(kList);
66         isAdd = false;
67       }
68     }
69   }
70   if (isAdd) { newPhrasedList.add(phrased_word);  }
71 }
72 return newPhrasedList;
--
```

Figure 3. Turkish Compound Word Detection Algorithm

- If it is compound and the second word has got –ın – ın suffixes.

The algorithm finally extracts the combined word list in the sentence.

This algorithm was developed to extract Turkish conjugated words and does not contain all grammar rules. To be able to contain all the rules of the Turkish language, it needs to be developed further. Since the algorithm is used on the sentences established in the daily speech language, it has been seen that not all grammatical rules are sufficient for this process in such studies. Some exclusion rules have been added to control situations such as foreign word usage.

b) Detection of Foreign Words and Abbreviations

Foreign words and abbreviations appeared as another issue to work on at tagging. Since the words in the sentence which are commonly used or which originated from traditional words used for technologies and product features cannot be solved with a Turkish natural language processor, a controlled dictionary containing these words has been created.

For these cases, the words / word groups encountered in the sentences are labeled using the abbreviations and foreign word controlled-dictionary created for these situations. Since the foreign words have been overused in the domains of the comment texts we have worked on, this controlled dictionary has produced very effective results in extracting the product features. Since the foreign words have been overused in the domains of the comment texts we have worked on, this controlled dictionary has produced very effective results in extracting the product features.

3) Detection of Frequently Repeated Product Feature

After tagging studies, an a priori algorithm was used to calculate repetition frequencies of words and word groups [16]. Repetition frequency of tagged items consisting of word roots in reviews, and the details about in which review and which sentence they appear, was kept in a file based storage environment for analysis. The reviews generally include sentences within the scope of the same topic. When opinions on the same product are combined in this way, noun and noun groups in sentences indicate product features. Frequently used words and word groups are assumed to be product features.

4) Extraction of Opinion Expressing Words

We have handled adjective and adjective expressing words in sentences to extract sentiment expressing words. Sentimental words express personal thoughts of reviewers about products. In case of multiple features in a sentence, the sentimental adjectives are evaluated relative to the most frequently repeated word or word group. Positive or negative expression forms in the Turkish language are not limited to adjectives. In order not to miss adjectives from words with the same letter clusters but different structures, adjective building statuses of suffixes rather than word stems are handled.

When we look to the stem of the word “etkileyici” in the sentence given in III.C as an example, the word “etki” does not express the meaning of an adjective. However, while the suffix “le” gives a verb meaning, the suffix “ici” makes the

word an adjective and indicates a sentiment. In the method we propose, accepting the word groups that express such opinions as adjectives, sentiment sentences are revealed.

5) Extraction of Opinion Sentence Tendency

Studies on sentiment analysis aim at classification of the given text as positive, negative, or neutral. For example, it is possible that all of the reviews/comments about a soccer team or a TV program on social media are classified as positive or negative using natural language processing and text mining techniques. Here, the right situation or the wrong idea is not sought and the current situation is detected. Various methods for opinion mining may be developed. For example, distinguishing words as positive/negative and classification of reviews as positive/negative according to the number of words is one of the essential methods.

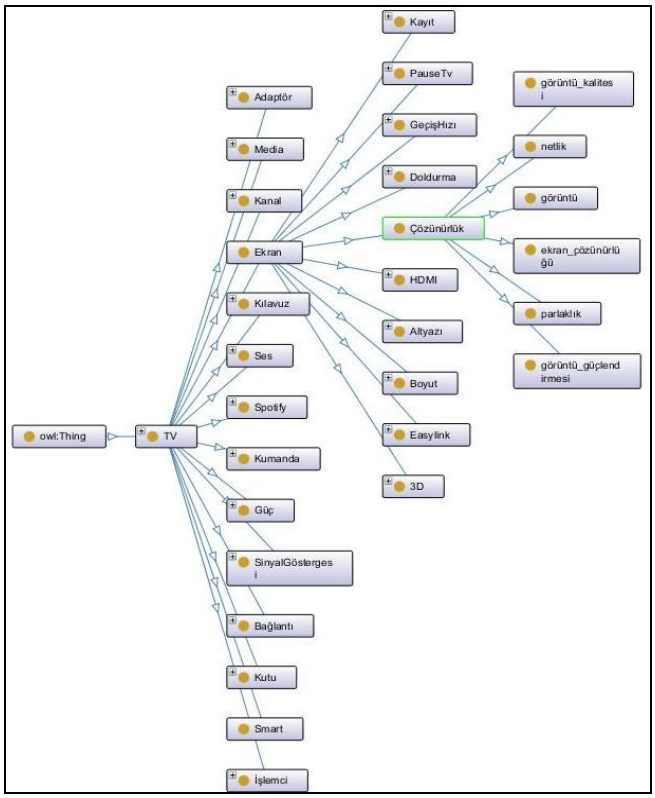

Figure 4. Ontology Diagram for Product Features – TV

While performing object-oriented opinion extraction through reviews, opinion mining through number of words, frequencies of words such as nouns, adjectives, adverbs, or verbs was a technique that we have used. In this approach, primarily noun word groups are detected, and by adding adjective word groups defining these noun word groups the infrequently repeating sentiments are eliminated. Scoring is done through adjective or adjective groups. Utilizing the adjective dictionary that exists at our disposal, the adjectives in the sentence are found to have negative or positive levels. The adjectives are subjected to these levels of aggregation and a score is calculated for the generic. All scores in the comment are summed up and a score is calculated for the comment. According to the calculation made, it is concluded that the score is positive if it is +, negative if it is - or neutral if it is 0. In this calculation, the WordOrientation and SentenceOrietation algorithms shown in the study of Hu and Liu [1] are used. The calculated scores for some features are shown in Figure 8 and Figure 13 by the OrientationScore tag.

6) Combination of Associated Words

In the processes up to this point, the identification of words which frequently repeat and express positive / negative opinions has been emphasized. Because of

```
1 Procedure SynonymController()
2 begin
3  for each word sentence si
4  begin
5       var synword = synonymList.find(word.stemmed)
6       if synword is not null
7            word.synonymId=synword
8            word.text=synword.Text
9  endfor;
10 end
```

Figure 5. Algorithm for Turkish Associated Word Extraction

variety in the grammar of the Turkish language, and the use of foreign words, we have seen that words expressing the same features have qualified as different features, and some very frequently expressed features remained under a threshold value and were not qualified as features. In order to reduce this situation as much as possible, we have focused on the integration of the words which are synonymous and express the same situation.

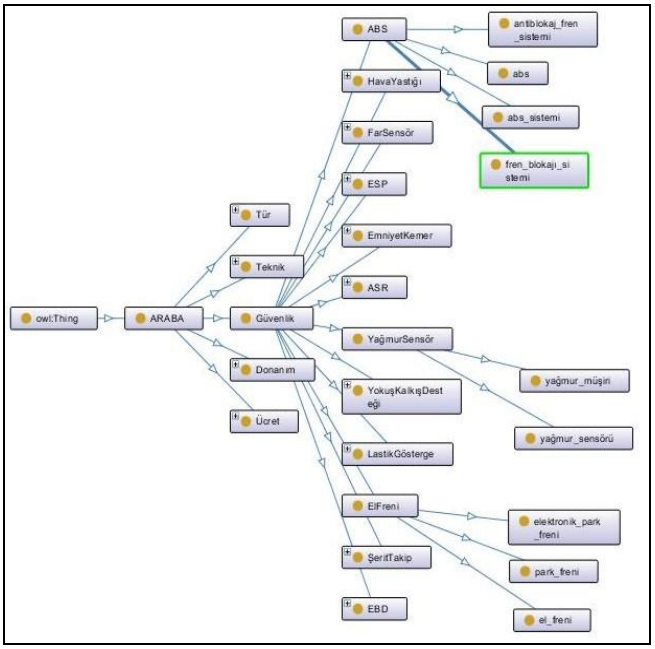

Figure 6. Ontology Diagram for Product Features - Car

 ISBN: 978-1-61208-557-9

Since it is not enough to check the comments one by one and remove synonyms, the abbreviations and foreign words are also grouped. For this purpose, ontologies have been created that cover the product features related to the products. These ontologies are given in Figure 4 and Figure 6.

In this way, controlled dictionaries created by experts for product specifications are created. This allows us to further increase the success of our feature extraction method. The algorithm we use is presented in Figure 5. Synonyms are detected on line 5 and values are assigned to synonyms on lines 7-8.

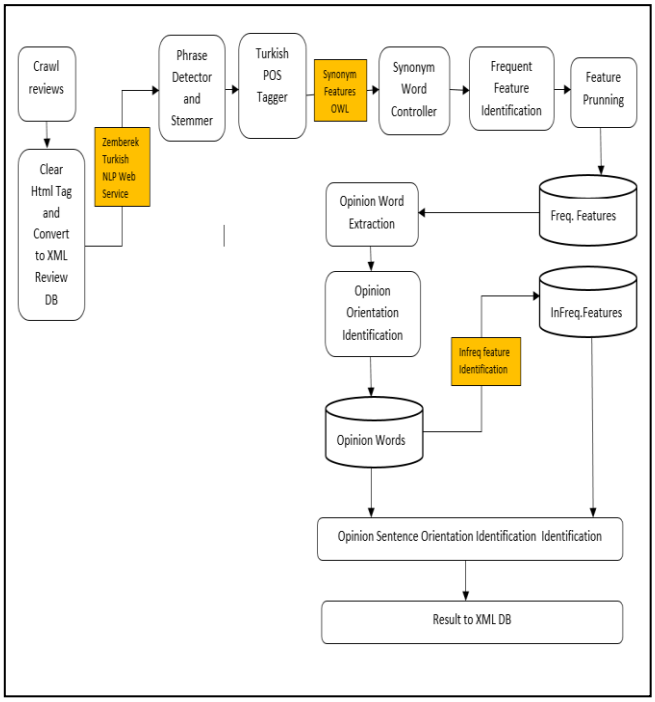

Figure 7. System Architecture Model

The flow architecture of the system is shown in Figure 7 above.

IV. IMPLEMENTATION AND EVALUATION

Our proposed methodology was implemented using C# and Java technologies. Java version 1.8.91 and C# version 4.6.1 were selected. The open source Zemberek Natural Language Processor was used as a Turkish word processor [15]; Protege 5.0.1 (protege.stanford.edu) was used for ontology development; the Jena library (jena.apache.org) has been used to process ontology data. These projects, which we use to improve our methodology, have been chosen because they are open source projects with specific developer support. The general architecture of our developed prototype software is shown in Figure 7.

The developed prototype software was designed in order to parse the product reviews performed instantly or asynchronously on the e-commerce website or product review website, to detect product features, and to perform extractions from product reviews.

In our proposed system, each idea/opinion/comment is considered to be a resource. In this structure, customers create their own comments and evaluations on e-commerce or evaluation sites.

```
<ReviewXmlDataContract>
  <ReviewID>3</ReviewID>
  <Date>2016-11-24T00:00:00</Date>
  <Author>ufuk tıkız</Author>
  <Sentences>
    <ReviewSentence>
      <SentenceId>4</SentenceId>
      <Text>ekran gayet uygun boyutlarda</Text>
      <Words>
        <ReviewWord>
          <WordId>1</WordId>
          <Tag>NOUN</Tag>
          <Text>ekran</Text>
          <StemmedText>ekran</StemmedText>
        </ReviewWord>
        <ReviewWord>
          <WordId>2</WordId>
          <Tag>ADJECTIVE</Tag>
          <Text>gayet</Text>
          <StemmedText>gayet</StemmedText>
        </ReviewWord>
        <ReviewWord>
          <WordId>3</WordId>
          <Tag>ADJECTIVE</Tag>
          <Text>uygun</Text>
          <StemmedText>uygun</StemmedText>
        </ReviewWord>
        <ReviewWord>
          <WordId>4</WordId>
          <Tag>NOUN</Tag>
          <Text>boyutlarda</Text>
          <StemmedText>boyut</StemmedText>
        </ReviewWord>
      </Words>
      <OrientationScore>3</OrientationScore>
    </ReviewSentence>
  </Sentences>
</ReviewXmlDataContract>
```

Figure 8. Parsed Review Content

Figure 8 contains this sentence as an example “ekran gayet uygun boyutlarda (the screen is of a very good size)”. Figure 8 depicts our data model of a TV model, based on the comment text entered by a user, in XML format. As can be seen from Figure 8, the data structure has upper data fields related to comment text such as ReviewId, Date, and Author. In addition, it shows the content of the text, such as "SentenceId", "WordId", "Text", "Tag", "StemmedText", "FrequencyId", "SynonymId", "Date", "Author" There are data items in the text that indicate the frequency of their passage in the text.

We have carried out our experiments on texts that express comments/opinions/thoughts for two different product types: Automobile and Television. We used data sets consisting of customer reviews of 5 car brands and 5 television models. Review texts were collected via "arabainceleme.com" and "hepsiburada.com" sites. For each comment text, the title of the text, the content of the text, the date and time the text was entered, the name of the author, the location of the author, and the rating value of the author are exhibited.

For each product, 100 comments were captured and downloaded by scanning the text and converted to XML documents as the example in Figure 8 shows. The texts were

 ISBN: 978-1-61208-557-9

separated using the Zemberek Natural Language Processing library, and the words and phrases in the texts were tagged.

For each sentence in the product interpretation, these terms are tagged if they contain the user's opinions. The features of these sentences are also determined. For feature identification, the method of feature identification proposed by Hu and Liu [1] is expanded. Here, unlike the work of Hu and Liu, the synonyms and feature words are united. It also provides the tagging of the different words that have the same meaning. The terms entered for the product features entered by the users are compared with the terms of the ontologies we create to ensure that the correct product attributes are tagged. For each product, a list of features included in the comment text is produced.

TABLE I. PRODUCT FEATURES PRECISION AND RECALL RESULTS

Product Feature Extraction	**Hu and Liu Approach**			**Proposed Approach**		
Product	***Precision***	***Recall***	***F-Score***	***Precision***	***Recall***	***F-Score***
VW Golf	0.8667	0.1585	0.2680	0.9444	0.4146	0.5763
Seat Leon	0.9474	0.2195	0.3564	0.9706	0.4024	0.5690
Renault Megane	0.7667	0.5610	0.6479	0.8133	0.7439	0.7771
Opel Mokka	0.8082	0.7195	0.7613	0.8108	0.7317	0.7692
Nissan Qashqai	0.8036	0.5488	0.6522	0.8406	0.7073	0.7682
LG 32LF580N	0.6667	0.1167	0.1986	0.6885	0.9767	0.8077
Vestel 40FA5050	0.7143	0.2326	0.3509	0.8571	0.5581	0.6761
Vestel 40FB7100	0.6667	0.0667	0.1212	0.6531	0.7442	0.6957
Vestel 48FA8200	0.6429	0.2093	0.3158	0.7727	0.3953	0.5231
Vestel 48UA9300	0.7500	0.2791	0.4068	0.8667	0.6047	0.7123

For each product, the number of features extracted using the proposed method is summarized in Table 1. Here, all features in the comment text are manually tagged to determine whether the features detected by the method are correct features. Values for the recall and precision metrics were determined by comparing the properties found by our method with the properties of the products manually identified. Here, the detection of product properties (Hu and Liu's method) based on only the frequency values was compared to the frequency-based product feature detection (recommended method) attachment and discovery metrics after combining the synonymous features under a single feature. In extracting the feature count, the threshold value we use for the frequency rate is taken as 10%.

When we examine the comparison given in Table 1, we see the effect of combining the synonyms as the reason for the differences in the results. The results show that we can achieve higher finding and fixation values because of a large number of frequently used features and the presence of different feature names that mean the same. We also did the evaluation based on our product. The results are shown in

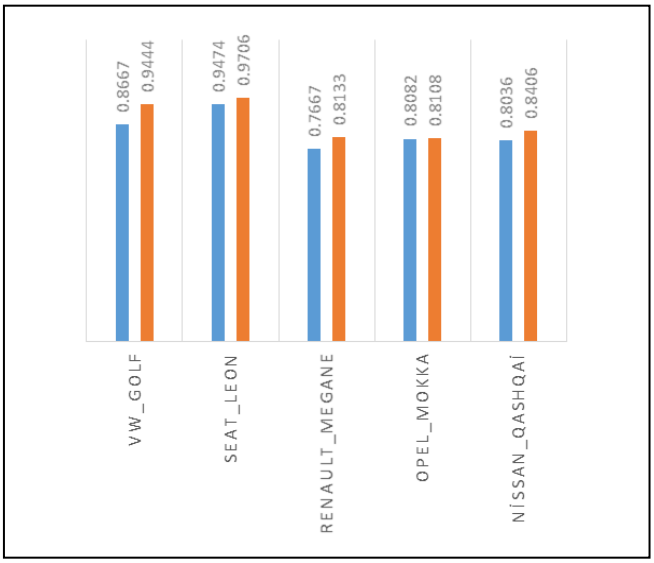

Figure 9. Precision metric values in different product feature detection approaches for automotive products

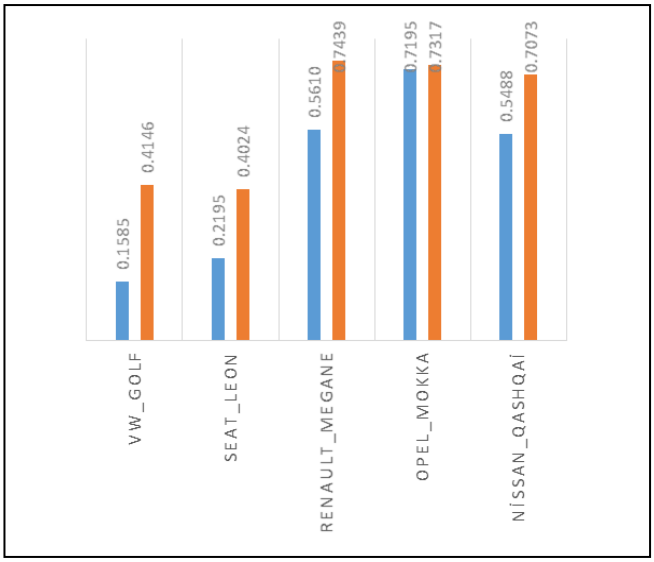

Figure 10. Recall metric values for different product feature detection approaches for automotive products

Figures 9-12. Based on the product-based evaluation, our approach suggests that it allows for more accurate traits. However, we see that the features that rarely appear in comment texts are features that are not of great interest to the general public.

Outputs from the sentiment analysis are shown in Figure 14 and Figure 15. Sentimental analysis values of the sentences we operate are calculated by taking advantage of the grades of the adjectives they possess. In this research, adjectives we have identified in the sentence are assigned values from a controlled list composed of rated adjective

words and word groups. The values corresponding to the adjectives that existed in the properties we extracted were given value assignments at the frequency of repetition of the properties and sentiment analysis values were obtained. Each of the rated adjectives in our hand is rated as positive (1), negative (-1), or neutral (0). These sentimental analysis values given to the extracted features are calculated in proportion to the frequency of repetition of the feature, the frequency of the adjectives it possesses, and the grades of these adjectives.

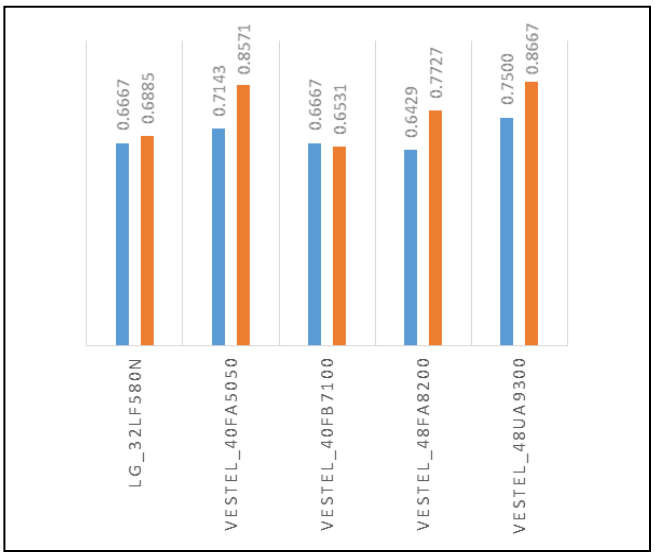

Figure 11. Precision metric values for different product feature detection approaches for TV model

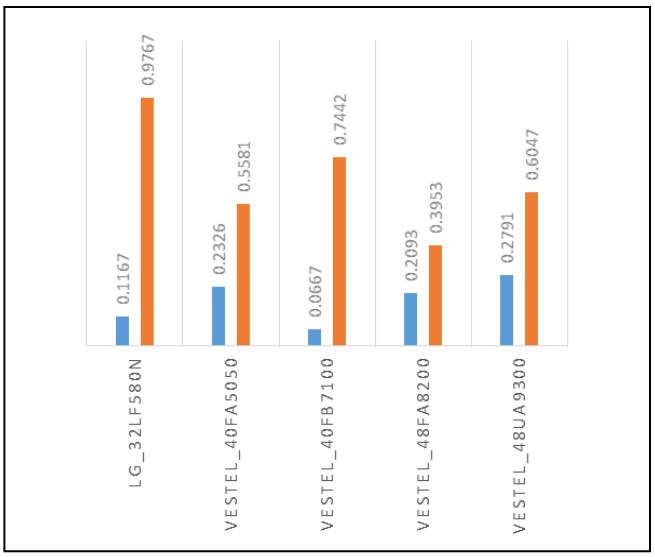

Figure 12. Recall metric values for different product feature detection approaches for TV model

Figure 13 deals with the statement "Ancak(CON) donanım (NOUN) konusunda (NOUN) eksik (ADJ)", (however, the hardware is a bit lacking). The approach we have proposed is that the "donanım konusu" sentence is a compound word. "eksik" is perceived as a negative adjective.

```
<ReviewSentence
xsi:type="ReviewSentenceWithOrientation">
  <SentenceId>7</SentenceId>
  <Text>Ancak donanım konusunda eksik.</Text>
  <Words>
    <ReviewWord>
      <WordId>1</WordId>
      <Tag>CONNECTIVE</Tag>
      <Text>ancak</Text>
      <StemmedText>ancak</StemmedText>
    </ReviewWord>
    <ReviewWord xsi:type="ReviewPhraseWord">
      <WordId>2</WordId>
      <Tag>NOUN</Tag>
      <Text>donanım konusunda</Text>
      <StemmedText>donanım konu</StemmedText>
    </ReviewWord>
    <ReviewWord>
      <WordId>3</WordId>
      <Tag>ADJECTIVE</Tag>
      <Text>eksik</Text>
      <StemmedText>eksik</StemmedText>
    </ReviewWord>
  </Words>
  <OrientationScore>-1</OrientationScore>
</ReviewSentence>
```

Figure 13. Sentence Example for Output of Sentiment Analysis

When viewed in this way, the word "donanım konusu" is marked as a property by being associated with the "donanım" attribute from the ontology diagram, and has a score of -1 in the sentiment analysis calculation because it is passed along with a negative adjective.

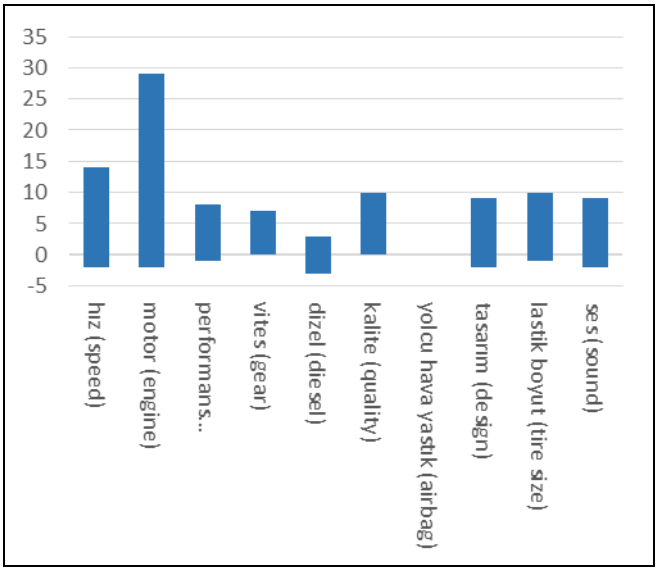

Figure 14. Sentiment Analysis values for the top 10 features in the VW Golf car

The sentiment analysis results of the features that we have obtained and shown with OrientationScore tag, appear in Figure 14 and Figure 15. While positive comments gain

 ISBN: 978-1-61208-557-9

weight in some features, some seem to gain negative comments, others are neutral. These figures give us information about the good and bad features of the products.

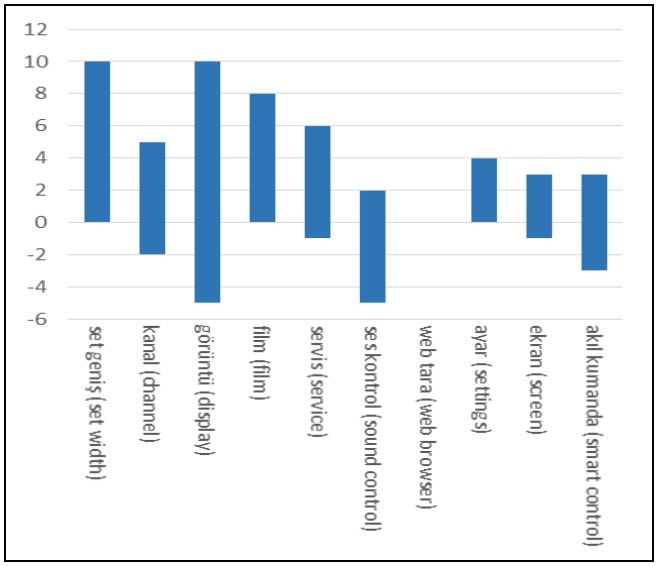

Figure 15. Sentiment Analysis values for the top 10 features of Vestel 48FA200 television

V. RESULTS AND FUTURE WORK

Within the scope of this research, a method is proposed that can process the Turkish comments made by customers for the products sold online, automatically extract the features of the products, and then summarize the positive or negative opinions of the customers based on the product features.

The methodology we propose and the software we have developed for this methodology yields very successful results in deriving the actual features used in the sentence. The experiments we have conducted to test the success of our proposed method have shown the effectiveness of our proposed approach, including the fact that the recall and precision metric values we obtained in the results are high.

Within the scope of this study, the performance ratios for product feature inferences were tested from texts that expressed comments / opinions / ideas. In future studies, we will develop our ontologies to extract product features more effectively. In addition to this, we will expand the scope of our experiments by increasing the number of texts that specify the comments / opinions / thoughts we have selected.

ACKNOWLEDGMENT

This study is being supported by the TUBITAK-3501-Career Development Program (CAREER) with the Project ID: 114E781.

REFERENCES

[1] M. Hu and B. Liu. “Mining and summarizing customer reviews”. *In Proceedings of ACM-KDD 2004*, pp.168-177.

[2] R. Kumar V. and K. Raghuveer. “Web User Opinion Analysis for Product Features Extraction and Opinion Summarization”. *International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.4, October 2012*, pp.69.

[3] V. Brindha and M. Kathiravan “Text Mining For Infrequent Noun Feature Extraction And Sentiment Classification”. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), ISSN: 0976-1353 Volume 13 Issue 4 – March 2015*, pp.323-326.

[4] S. H. Ghorashi, R. Ibrahim, S. Noekhah, and N. S. Dastjerdi “A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews”. *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012*, pp.29-35.

[5] L. Ferreira, N. Jakob, and I. Gurevych “A Comparative Study of Feature Extraction Algorithms in Customer Reviews”. *2008 IEEE International Conference on Semantic Computing, August 2008*, pp.144-151.

[6] L. Zhuang, F. Jing, and X. Zhu “Movie review mining and summarization”, *Proceedings of the 15th ACM international conference on Information and knowledge management, Virginia, November 2006*, pp.43-50.

[7] Harvard Business Review Analytic Services Report, *Marketing in the driver’s seat: Using Analytics to create customer value February 2016*

[8] P. D. Turney, “Thumbs Up or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews”. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, July 2002, Philadelphia*, pp.417-424.

[9] B. Pang, L. Lee, and S. Vaithyanathan “Thumbs up? Sentiment Classification Using Machine Learning Techniques.” *Procedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002*, pp.78-96.

[10] I. Mani and E. Bloedorn “Multi-document Summarization by Graph Search and Matching”. *Procedings of the 14th national conference on artificial intelligence and 9th conference on Innovative applications of artificial intelligence, Rhode Island, July 1997*, pp.622-628.

[11] J. Yi and W. Niblack “Sentiment Mining in WebFountain”. *Proceedings of the 21st International Conference on Data Engineering, April 2005, Volume 0*, pp.1073-1083.

[12] V. Hatzivassiloglou and J. Wiebe “Effects of Adjective Orientation and Gradability on Sentence Subjectivity”. *Proceedings of the 18th conference on Computational linguistics - Volume 1*, pp.299-305.

[13] M. Hearst ”Direction-based Text Interpretation as an Information Access Refinement”. *In Paul Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, 1992*, pp.257-274..

[14] R. Tong “An Operational System for Detecting and Tracking Opinions in on-line discussion” *SIGIR 2001 Workshop on Operational Text Classification, Volume 1*, p 6.

[15] A.A. Akın and M.D. Akın “Zemberek, an open source NLP framework for Turkic Languages” https://github.com/ahmetaa/zemberek-nlp [retrieved: 04, 2017]

[16] F. Coenen “LUCS KDD implementation of CBA”, Department of Computer Science, The University of Liverpool http://www.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cba.html [retrieved: 04, 2017]

 ISBN: 978-1-61208-557-9