International Journal of Computer Engineering and Information Technology

VOL. 9, NO. 1, January 2017, 15–23
Available online at: www.ijceit.org
E-ISSN 2412-8856 (Online)

Multi-word Aspect Term Extraction Using Turkish User Reviews

Ekin Ekinci[1], Hazal Türkmen[2] and Sevinç İlhan Omurca[3]

[1, 2, 3] Department of Computer Engineering, Kocaeli University, Kocaeli, İzmit, 41380 Turkey

[1]ekin.ekinci@kocaeli.edu.tr, [2]hazalturkmen91@gmail.com, [3]silhan@kocaeli.edu.tr

ABSTRACT

Nowadays, when an individual wants to buy any product or a company wants to take the pulse of public opinion about its product, user reviews of this product have become a valuable source of information. As a consequence of that, aspect based sentiment analysis has become popular research field which has also attracted the attention of researchers. In this study, we devised a method which extracts multi-word aspects from the Turkish user reviews. To investigate the reliability and the performance of the system, the frequency basis method based on N-gram by unifying finite state automata which are set for the recognition of the Turkish grammar rules were preferred. The success of the system was measured by using cell phones and by using hotel reviews. As a result, the success obtained is averagely 82% for cell phone domain and averagely 79% for hotel domain.

Keywords: *Aspect Bases Sentiment Analysis, Aspect Extraction, Multi-word Aspect Extraction, Finite State Automata.*

1. INTRODUCTION

Automatic sentiment analysis of online customer reviews has become a needed research area due to the rapid growth of user-generated reviews on the internet. The institutions who manage voice of the customer and social media insight programs particularly require the sentiment analysis of online reviews. There is wide range of products and services being reviewed on the Web, thus Web has become an excellent source for extracting customer reviews. Due to the huge number of online reviews regarding products and services, the supervised machine learning methods are not practical to implement. As the number of reviews expands, it is essential to develop an unsupervised sentiment analysis model which is capable of extracting the product aspects and determining the sentiments for these. Aspect extraction is a critical stage in sentiment analysis, therefore, for an overall sentiment analysis of the reviews, a word or multi-word aspects (MWA) of the products must be detected.

For the recent years, sentiment analysis for online reviews has attracted a great deal of attentions from researchers and product manufacturers [1-4]. Product manufacturers need online reviews to understand the general responses of customers to their products for finding weaknesses of the products and improving their products accordingly. An analysis has been being developed to define the tendency of the customers however the need for a new method was born when we figure out that this kind of analysis is too inadequate to precise weak and strong sides of the product specifications.

Along with getting generally a negative or positive result of a product in a document leveled analysis, it is not correct to assume that if the product result is negative; all the product specifications are weak. Accordingly it doesn't mean that all the product specifications are fine if the product is resulted positively. Sentiment classification based on aspect is a much more correct approach seeing that it gives sentiments separately for the product specifications instead of reflecting general sentiments about the product. What we mean by the word 'aspect' is anything that defines, completes a product. In aspect based sentiment analysis, as the sentences not including the word aspect doesn't mean anything in terms of sentiment analysis, it is also possible to use the sentiments in reviews effectively with this method.

In this paper, we concentrate on MWA extraction from customer reviews and propose a novel unsupervised and domain-independent hybrid model for detecting Turkish MWAs. There is no known study realized with this objective until today. The unsupervised and domain-independent product abolishes the need for labeled data during the process of excluding MWAs. Seeing that this method based on N-gram and heuristic rules particular to a language is easily compatible with any language, it makes possible to adjust the system for other languages. A successful system has been achieved thanks to PMI method which is mostly preferred for natural language processing (NLP). In order to see the performance of the proposed system, a human-generated MWA corpus is used.

The precision and recall values which determine the coverage between the human-generated and automatically generated MWAs are used as performance evaluator.

The rest of the paper is organized as follows: Section 2 reviews the literature; Section 3 explains the proposed MWA extraction model. The experimental results presented in Section 4. Finally, discussion and conclusions for the future work are summarized in Section 5.

2. RELATED WORKS

When the literature is evaluated it is shown that aspect extraction is one of the cornerstones of the sentiment analysis studies. To design a powerful sentiment analysis system, aspect extraction process should be carried out successfully. To better understand the subject literature is summarized in this section.

In both studies, Hu and Liu [5, 6], utilized a data mining algorithm, which is called association rule mining based on Apriori, to find all frequent itemsets. In this context, the itemset was a set of noun or a noun phrase that occurs together. Among these itemsets to remove unlikely aspects they applied filtering techniques which were compactness pruning and redundancy pruning. Popsecu and Etzioni [7] devised OPINE, an unsupervised information extraction system to extract noun phrases from reviews and these noun phrases were eliminated according to threshold. To remove non-aspects their system evaluated each noun phrase by computing the PMI score between the phrase and some meronymy discriminators associated with the product class of interest. Yi and Niblack [8], developed heuristics and selection algorithms to extract feature terms from online reviews. A feature term can be a part-of relationship with the given topic, an attribute-of relationship with the given topic, an attribute-of relationship with a known feature of the given topic. Candidate feature term extraction, which was noun phrases, was achieved with bBNP heuristics and to select the feature terms among these extracted terms, they used a likelihood ratio based algorithm. Wei et al. [9], proposed a semantic-based product feature extraction (SPE) technique. In this technique, subjective adjectives obtained from General Inquirer were used to eliminate non-product features and opinion-irrelevant product features and explore infrequent product features. Zhu et al. [10], constructed an aspect based opinion polling system. In their system, to extract candidate MWA related terms C-value method was preferred. C-value is often utilized as multi-word extraction method. Bagheri et al. [2], performed an unsupervised domain-independent aspect detection model for online reviews. After applying POS tagging and stemming to extract candidate aspect, a generalized statistical measure was performed for MWAs. In this study, extracted MWAs were ranked with word scoring method which is called FLR. Yan et al. [11], developed EXPRSS method an extended pagerank algorithm to extract product features from Chinese online consumer reviews. In this study, noun and noun phrases and dependency relations were identified with Chinese lexical analysis tool (ICTCLAS). Candidate extraction was performed with NodeRank algorithm an extended pagerank algorithm. Li et al. [12], enacted a method to extract candidate aspects based on frequent noun and noun phrases and PMI-IR score were used to prune among these candidates. While PMI score is computed between aspect and discriminator, PMI-IR score is computed between aspect and target entity. RCut algorithm was also implemented for determination of threshold to select candidate aspects.

3. THE PROPOSED MULTI-WORD ASPECT EXTRACTION MODEL

The system architecture designed within the scope of study is shown in Fig. 1. The first step of the system is to precise the web sites including the reviews of domain which will be worked on and is to crawl the reviews from the related pages with a web crawler. When the reviews are examined, a lot of typos are detected. Since these typos affect the aspect determination negatively, the step of Syntax error correction is realised with the help of Turkish NLP Library Zemberek. After preprocessing, word N-grams are excluded along with their frequencies from the document. Over precised N-grams, candidate MWA have been identified by realising stop word, digit and punctuation based elimination. Candidate MWA set is simplified via frequencies basis Compactness pruning method and after that by applying heuristic rules. The final MWA set has been achieved with a PMI basis method of elimination to define the aspects related to chosen application domain.

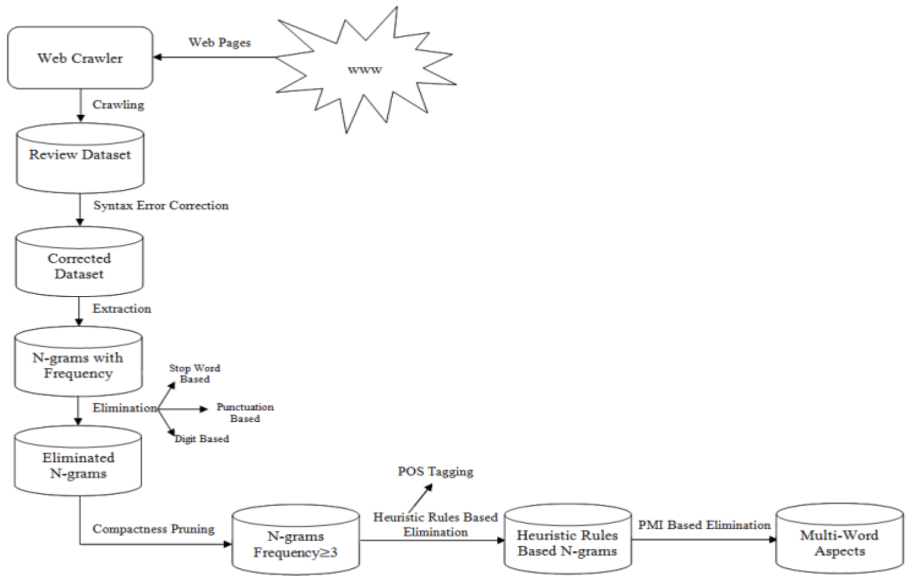

Fig. 1. Overall process of multi-word aspect extraction system.

3.1 N-grams

N-gram basis method is often preferred in NLP applications for being powerful and easy to use. While its independence from languages and low grammar knowledge necessity provides an easy realisation, its successful results show that this method is powerful.

N-grams are named as a series of groups composed of N members in document. These members are identified as characters composing the document (letters, punctuation, space etc.), words, POS labels or specifications that can be found successively. The selection of a member can be change according to objective of the study. Character N-grams are usually preferred for author, title and language identification applications [13-15]. Word N-grams are preferred for finding and using multi-words [16, 17]. Pos-tag is among the methods called N-grams [18].

In this study, word N-grams are used. The objective of the study is to achieve MWAs in a certain review cluster. When MWAs in Turkish are examined, it is clear that they are usually composed of two words. This is why N=2 is chosen in this study (battery charger, memory card, customer satisfaction etc). The binary combinations in all the reviews are excluded with their frequencies.

3.2 Compactness Pruning

It is aimed to realise a filtration and compactness pruning [5, 6] through the combinations and their frequencies got by N-gram. According to compactness pruning method, the combinations that are precised as MWA by using N-gram are eliminated according to a threshold value.

3.3 Heuristic Rules

It might be impossible for any notion to be explained in details, to be talked about its specifications. To be able to overcome this situation, it might be necessary that a lot of words come together under some rules. As for an explication of this situation over a computer notion, we can show as examples computer cabinets, computer prices, computer engineering etc. However; more than one words might not always come together to give details about a notion. Some notions correspond to more than one words. For example; holiday camp, video card, dishwashing machine etc. In some cases, certain words come together to modificate a notion and compose a phrase. In the examples of boiled corn, kissable hands, familiar faces, cold weather, we see that first words compose a phrase by modification of second ones. In brief, the words gathered under certain rules for expressing notions in details, for notions expressed with more than one words and for modificating notions compose phrases [19]. When these rules are examined, it is seen that the word that is aimed to emphasized is found at the very end and the emphasizing word or words are found at the beginning. In noun phrases, while the word aimed to emphasized which is at the end is named determinated, the word(s) emphasizing is named determining. As compound nouns are categorized in two (adjective or noun) according to type of the determining, possessive construction are divided in three according to the suffix that determining takes [20].

- Defined compound nouns: In defined compound nouns, both determinated and determining are nouns. Both of them takes compound suffixes; determinated takes its suffixes (ın,in,un,ün) and determining takes its (ı,i,u,ü). Example: "Okulun bahçesi (school's garden)", "Bilgisayarın kasası (computer's cabinet)".
- Undefined compound nouns: In undefined compound nouns, both determinated and determining are nouns. Only determining takes suffixes which are (ı,i,u,ü). Example: Ekran çözünürlüğü (screen resolution), domates salçası (tomato paste), çilek reçeli (strawberry jam), biber turşusu (pepper pickle).
- Compound nouns with no suffixes: In compound nouns with no suffixes, both determinated and determining are nouns. Neither of them takes suffixes. Example: Ölü deniz (Dead Sea).

A Finite State Automata (FSA) represents the relational patterns of noun phrases and adjective clauses. These patterns rely on the simple pos-tagger outputs i.e. they consist of pos-tags. A Finite State Automata (FSA) begins from one of the states (called the start state), goes through transitions depending on inputs to different states and end in one certain set of states marking a successful flow of operation (called final states). An example of a simple noun phrase automata that can recognize noun phrases in Turkish is illustrated in Fig. 2.

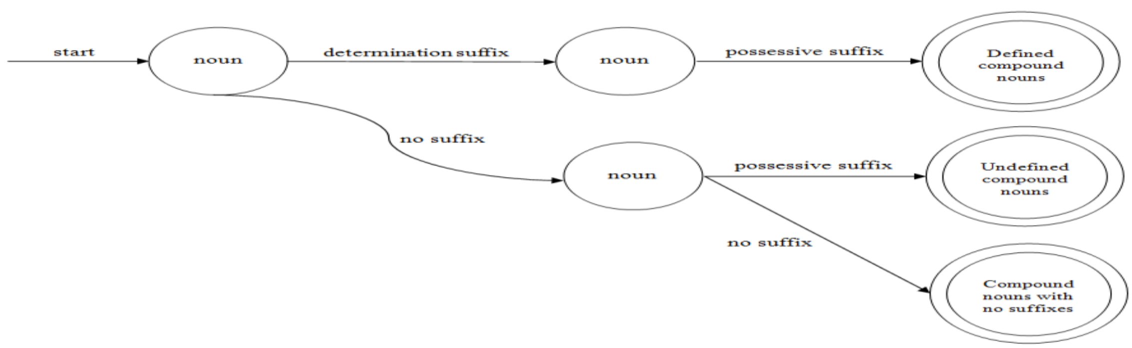

Fig. 2. A FSA of noun phrases.

- Adjectival determinatives: The participles are composed as getting suffixes at the end (-en,-esi,-mez,-ar,-dik,-ecek,-miş). The participles, just like adjectives, come before nouns and compose an adjective clause. Example: Haşlanmış mısır (boiled corn), tanıdık yüzler (familiar faces). In this study, rules for adjectival determinatives are taken into consideration and an automata for this is illustrated in Fig. 3.

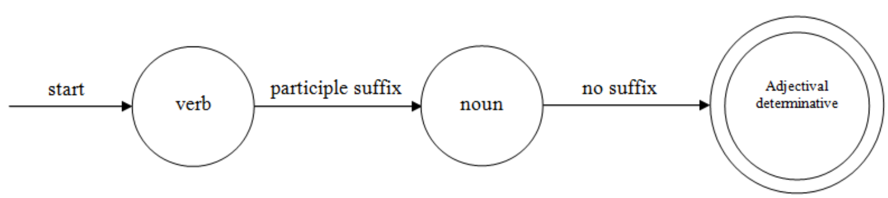

Fig. 3. A FSA of adjectival determinatives.

3.4 PMI

It is necessary to determine if each member of candidate MWA set is associated with the main entity (hotel and cell phone). For this purpose, PMI which is a criterion often used to measure the relation between words and the main entity in natural language processing [3]. For main entity let's say d and for any specification in candidate aspect set let's say A_i, in that case the here, relation between A_i and d is defined with PMI method as in Equation 1.

$$PMI(A_i, d) = hits(A_i, d) / (hits(A_i) \times hits(d)) \qquad (1)$$

Here, $hits(A_i, d)$ is co-occurrence number of A_i and d and $hits(A_i) \times hits(d)$ is the co-occurrence number of the

A_i and d if they are statistically independent. The value which is calculated for every single specification is compared with the predefined threshold value. MWA that cannot reach to the threshold value are excluded. Let's suppose that candidate aspect set is presented with $A = \{A_1, A_2, ..., A_l\}$. Then the PMI threshold value that is used to determine if any aspect is related enough to the chosen domain is calculated with this formula:

$$PMI_{threshold}(d) = (1/l) \times \sum_{i=1}^{l} PMI(A_i, d) \quad (2)$$

This kind of calculation of PMI threshold value provides achievement of all the MWA lists from the most frequent to the most rare.

4. EXPERIMENTAL RESULTS

In this section we evaluate the experimental result of the proposed MWA extraction model. There are not any benchmarking methods for aspect extraction process in Turkish. Therefore we only compare our results by the results obtained by human experts.

4.1 Dataset Collection and Description

Previous studies on aspect extraction have mostly evaluated English reviews, because there are conventional dataset of online reviews in English for several domains. However there is no study realising aspect extraction by using Turkish reviews. In this study we conduct the experiment with Turkish reviews to extract MWAs. Therefore we crawled user reviews from frequently visited web pages, such as www.otelpuan.com for hotel reviews and www.hepsiburada.com for cell phone reviews. Each datasets contain textual reviews which are randomly selected from web pages collected by our crawler. Table 1 presents the descriptive information about these data collections.

Table 1: Dataset description

Dataset	***# of Reviews***	***# of Sentences***	***# of MWAs***
Hotel	1518	6852	46
Cell Phone	1592	6314	45

Three human annotators were asked to extract MWAs from user reviews independently for each domain. Only the aspects that all annotators had agreed were included in the final MWA set. The agreed aspect numbers of two domains are shown in the last column of Table 1. These data sets can be used in similar studies.

4.2 Evaluation Metrics

In the previous researches the precision, recall and F-measure used as the metrics to assess the effectiveness of the proposed approaches. Alternatively, in this study the aspect extraction model is considered as an Information Retrieval (IR) system. In IR usually no decision is made on whether a document is relevant or irrelevant to another document. Instead, a ranking of the documents is produced [21]. Aspect extraction system is also an IR system where the most important aspects are extracted from a given domain. Based on this admission, our study examines not only the MWAs of the user reviews, but also the ranked list of the aspects according to their popularity and frequency in user reviews. The ranked list consideration of MWAs provides a better analysis about their importance with regard to user opinions.

Given a user review D, the evaluation method first computes relevance scores for all MWAs in D and then produces a ranking $R_{mwa} = \{mwa_1, mwa_2, ..., mwa_n\}$ of these aspects based on their relevance scores. The mwa_1 is the most relevant aspect to the review text and mwa_n is the most irrelevant aspect to the review text. The precision and recall values at each mwa_i in the ranking are computed. A general representation of ranked MWA list is shown in Table 2.

Table 2: General rank list representation

Rank Order	***Aspect Name***	***Agreed/Not Agreed***	***p(i)***	***r(i)***
1	.	+	1/1	1/n
2	.	-	1/2	1/n
.	.	+	.	.
i	.	-	.	.
.	.	-	.	.
n	.	+	.	.

Recall at position i denoted by $r(i)$ is the fraction of relevant multi-word aspects from mwa_1 to mwa_i in R_{mwa}. The recall value is computed as in Equation 3.

$$r(i) = relevant_i / |R_{mwa}| \quad (3)$$

Where, $relevant_i$ represents the number of relevant multi-word aspects in the related range i. Precision at position i, denoted by $p(i)$, is the fraction of multi-word aspects from mwa_1 to mwa_i in R_{mwa} and computed as in Equation 4.

$$p(i) = relevant_i / i \quad (4)$$

The computed precision and recall values enable the evaluation of the coverage among the manually and automatically generated multi word aspects. Further, an average precision can be computed based on the precision at each level in the ranking R_{mwa} as in Equation 5.

$$p_{avg} = \sum_{mwa_i \in Relevant_{mwa}} p(i) / R_{mwa} \quad (5)$$

Here, $Relevant_{mwa}$ represents the set of relevant multi word aspects which are defined by human experts. The performance evaluation is made based on judgments of human experts. The quality of a computer generated multi-word aspect list is tested due to precision and recall values of a manually generated MWA list which is represented by $Relevant_{mwa}$ in Equation 5.

4.3 Empirical Results

When suggested system model is applied to the cell phone reviews, for the beginning 14007 candidate MWA are found via N-gram method. After that, the compactness pruning is realised according to the threshold value which is determined as 27 by a human expert. As a result of this step, the number of candidate aspects is withdrawn to 863. When each of these aspects are presented to FSA which is defined for noun and adjective phrases, MWAs that can make it to the final state are accepted as valid noun or adjective according to Turkish grammar. 97 MWA are recognized by FSA. Finally, A PMI list containing 45 MWA is achieved from heuristic list by operating PMI method. The flowchart for the cell phone domain is displayed in Fig. 4.

The precision and recall numbers of PMI list containing final MWA's achieved for cell phones are shown in Table 3 as ranked list.

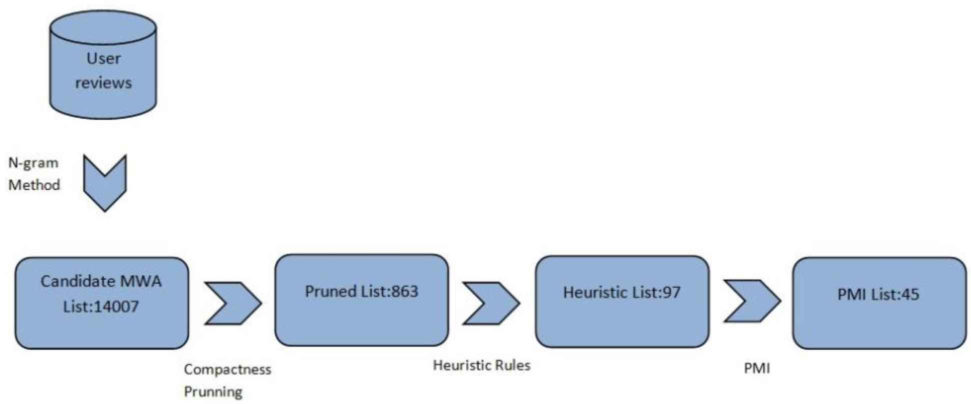

Fig. 4. The flowchart for the cell phone domain.

Table 3: MWA list for cell phone domain

Rank Order	***Aspect Name (English/Turkish)***	***Agreed/Not Agreed***	***p(i)***	***r(i)***
1	cell phone (cep telefonu)	+	100	2
2	smart phone (akıllı telefon)	+	100	4
3	battery charger (şarj aleti)	+	100	7
4	operating system (işletim sistemi)	+	100	9
5	charging adapter (şarj cihazı)	+	100	11
6	user manual (kullanım klavuzu)	+	100	13
…	…	…	…	…
10	guarantee certificate (garanti belgesi)	+	80	18
...	…	…	…	…
21	telephone game (telefon oyun)	-	81	38
...	…	…	…	…

39	screen quality (ekran kalitesi)	+	79	69
40	picture video (resim video)	-	78	69
41	ideal phone (ideal telefon)	-	76	69
42	super screen (ekran süper)	-	74	69
43	speech qualification (konuşma özelliği)	+	74	71
44	phone suggestion (telefon tavsiye)	-	73	71
45	video quality (video kalitesi)	+	73	73

As listed in Table 3, the "cell phone" domain contains 45 MWAs. Our system ranked the "cell phone" as the most critical aspect and the "video kalitesi" as the less critical aspect among them. As a result of our application, the aspects are ordered from most voted and to least voted as ranked list. Moreover A system that will be useful to make a more flexible evaluation for experts is developed. If there is an obligation for choosing a few MWAs, it is possible to make a choice from top of the list as many as wanted. For example, when the first 6 aspects are chosen as a final set, both precision value and average precision value of the system at 6th level will be 100%. Alternatively, for the first 10 aspects, while p(10) value 80%, it will be 75,8%. When all of the list is chosen as final set, average precision value will be 82% as p(45) value is 73%. The precision and recall values of cell phone are shown in Table 3.

For hotel domain, we got 46 MWAs from totally 10924 candidate MWAs. Among the specified MWAs, while the most voted feature is "food and drink", the least voted is " employee attention" with a rank number 43. The features number 44, 45 and 46 are not taken into consideration as human expert don't evaluate them as aspects. When we consider the first 43 members are chosen as related MWAs, for final set the precision value is 88% and recall value is 83%, as for average precision value for hotel domain, it is calculated as 79,2%. The obtained results for "hotel" domain are given in Table 4.

Table 4: MWA list for hotel domain

Rank Order	***Aspect Name (English/Turkish)***	***Agreed/Not Agreed***	***p(i)***	***r(i)***
1	food and drink (yiyecek içecek)	+	100	2
2	holiday camp (tatil köyü)	+	100	4
...	…	…	…	…
18	quality of service(servis kalitesi)	+	100	39
19	human relations (insan ilişkileri)	+	100	41
...	…	…	…	…
43	employee concern (personel ilgisi)	+	88	83
44	over the age of …(yaş üstü)	-	86	83
45	food beverage (yemek içecek)	-	84	83
46	hotel full (otel dolu)	-	83	83

5. CONCLUSIONS

In these days, understanding the features of a product from its customer reviews become an important application and necessity domain. Which features are the most eye catching for both potential customers and producing company is a very fundamental point. In this study, a system which provides automatic exploration of binary phrased aspects by using Turkish reviews of products from different domains is exposed. Automatic detection of MW As in hotel and cell phone domains is realised with the use of Turkish customer reviews. It is observed that the product features found in customer reviews are either single or binary phrases. In this study, instead of single phrases we mostly focused on getting binary phrased aspects which are dependent to natural language. The existing heuristic rules for making Turkish binary phrases are recognized as noun and adjective FSAs in this study. Moreover the recognized phrases are seen valid in Turkish and are accepted as MWAs in our study. To form input MWAs for noun and adjective FSA, firstly all the candidate MWAs in customer reviews are detected. Candidate MWA set is pruned with the Compactness

pruning method. The domain suitability of MWAs recognized by FSAs is tested with PMI method.
Ranked list method, which is a kind of information retrieval evaluation method and which isn't used before in sentiment analysis, is used for evaluating the success of the system. Achieving MWAs from customer reviews in hotel domain is realised with precision value of 79% and achieving MWAs from customer reviews in cell phone domain is realised with precision value of 82%.

ACKNOWLEDGMENTS

This study is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under project number 114E422.

REFERENCES

[1] Brody, S., Elhadad, N., Unsupervised Aspect-Sentiment Model for Online Reviews, In: Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, USA, 2010, pp 804-812.

[2] Bagheri, A., Saraee, M., de Jong, F., Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews, Knowledge-Based Systems, 2013, 52, pp 201-213.

[3] Zhang, W., Xu, H., Wan, W., Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis, Expert Systems with Applications, 2012, 39 (11), pp 10282-10291.

[4] Shi, H., Zhan, W., Li, X., A Supervised Fine-Grained Sentiment Analysis System for Online Reviews, Intelligent Automation & Soft Computing, 2015, 21 (4), pp 589-605.

[5] Hu, M., Liu, B., A Mining Opinion Features in Customer Reviews, In: Proceedings of 19th National Conference on Artificial Intelligence, California, USA, 2004, pp 755-760.

[6] Hu, M., Liu, B., Mining and Summarizing Customer Reviews, In: Proceedings of International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 2004, pp 168-177.

[7] Popescu, A. M., Etzioni, O., Extracting product features and opinions from reviews, In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2005, pp 339-346.

[8] Yi, J., Niblack, W., Sentiment Mining in WebFountain, In: Proceeding ICDE '05 Proceedings of the 21st International Conference on Data Engineering, Washington, USA, 2005, pp 1073-1083.

[9] Wei, C., Chen, Y., Yang, C., Yang, C. C., Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews, Information Systems and e-Business Management, 2010, 8 (2), pp 149-167.

[10] Zhu, J., Wang, H., Zhu, M., Tsou, B. K., Ma, M., Aspect-Based Opinion Polling from Customer Reviews, Affective Computing, IEEE Transactions on, 2011, 2 (1), pp 37-49.

[11] Yan, Z., Xing, M., Zhang, D., Ma, B., EXPRS: An extended pagerank method for product feature extraction from online consumer reviews, Information & Management, 2015, 52 (7), pp 850-858.

[12] Li, S., Zhou, L., Li, Y., Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures, Information Processing & Management, 2015, 51 (1), pp 58-67.

[13] Stamatatos, E., On the Robustness of Authorship Attribution Based on Character n-gram Features, Journal of Law and Policy, 2013, 21, pp 421-439.

[14] Gencosman, B., Ozmutlu, H., Ozmutlu, S., Character n-gram application for automatic new topic identification, Information Processing & Management, 2014, 50 (6), pp 821-856.

[15] Shi Hayta, Ş., Takçı, H., Eminli, M., Language Identification Based on N-gram Feature Extraction Method by Using Classifiers, IU-Journal of Electrical & Electronics Engineering, 2013, 13 (2), pp 1629-1639.

[16] Antonia, A., Craig, H., Elliott, J., Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution, Literary & Linguistic Computing, 2014, 29 (2), pp 147-163.

[17] Baayen, H., Hendrix, P., Ramscar, M., Sidestepping the Combinatorial Explosion: An Explanation of N-gram Frequency Effects Based on Naive Discriminative Learning, Language and Speech, 2013, 56 (3), pp 329-347.

[18] Hasan, F. M., UzZaman, N., Khan, M., Elleithy, K. (ed), Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla, In: Advances and Innovations in Systems, Computing Sciences and Software Engineering (Springer), 2007, pp 121-126.

[19] Özkan, M., Sevinçli, V., Türkiye Türkçesi Söz Dizimi, İstanbul, 2008.

[20] Gökdayı, H., Türkiye Türkçesinde Öbekler, International Periodical For the Languages, Literature and History of Turkish or Turkic, 2010, 5 (3), pp 1297-1319.

[21] Liu, B., Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, In: Data-Centric Systems with Applications, 2011.

AUTHOR PROFILES:

Ekin Ekinci is a Research Assistant of Computer Engineering Department at Kocaeli University in Turkey. She has received her BEng. in Computer Engineering from Çanakkale Onsekiz Mart University in 2009 and ME. in Computer Engineering from Gebze Technical University in 2012. She is currently working towards Ph.D. degree in Computer Engineering from Kocaeli University. Her main research interests include text mining, sentiment analysis, natural language processing and machine learning.

Hazal Türkmen received her BEng. in 2013 and ME. in 2012 in Computer Engineering from Kocaeli University. She is currently working towards Ph.D. degree in Computer Engineering from Kocaeli University, Turkey. Her main research interests include text mining, sentiment analysis, natural language processing and machine learning.

Sevinç İlhan Omurca is an associate professor at the Kocaeli University Computer Engineering Department in Turkey. She has a Ph.D. at the Kocaeli University Electronics and Communications Engineering. Her main research interests include text mining, sentiment analysis, natural language processing, machine learning and data mining.