Text Summarization of Multi-Aspect Comments in Social Networks in Persian Language

Hossein Shahverdian*
Department of computer engineering, Faculty of Technical & Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Hassan Saneifar
Department of Computer, Raja University of Qazvin, Iran

***Abstract*—Now-a-days, there are increasingly huge amount of user generated comments on the web. The user generated comments usually contains useful and essential information reflecting public's or customers' opinions. Since the information in the comments could be used for decision making, production or service improvement, and achieving user satisfaction, the systematic analysis of these comments is an essential need in so many domains including e-commerce, production, and social network analysis. However, the analysis of large volume of comments is a difficult and time-consuming task. Therefore, the need for a system which can convert this massive volume of comments to a useful and efficient summary is felt more and more. Text summarization leads to using more resources at higher speeds and getting richer information. According to numerous studies conducted in the field of multi-document summarization, few studies can be found that have been focused on the user generated comments in Persian language. In this paper, we propose a novel approach to summarize huge amount of comments in Persian, which is enough close to a human summarization. Our approach is based on semantic and lexical similarities and uses a graph-based summarization. We also propose a clustering to deal with multiple aspects (subjects) in a corpus of comments. According to the experiments, the summaries extracted by the proposed approach reached an average score of 8.75 out of 10, which improves the state-of-the-art summarizer's score about 14 percent.**

Keywords*—Text mining; comments analysis; summarization; graph summarization; Persian language

I. Introduction

With increasing number of user on Web 2 platforms like social networks, weblogs, and online review sites, the user generated comments is dramatically increasing. The user generated comments contain primordial and useful information about public's opinion, social interactions, cultural events, customer's satisfaction, market analysis, etc. These online comments affect also the customers' behavior and could be useful in decision making as well as improving the services or products. Also, the user generated comments contain short and useful information which is beneficial for manufacturers or service providers. Also, most of people have accepted to read online comments as one of the steps before making a purchase [1]. A huge volume of user comments are generated through social networks. A social network is defined as a set of social institutions including people and organizations that are linked together by a set of meaningful social relationships, while sharing some values [2]. The social networks are defined in different ways. According to [3], the society is not beyond the individuals and their social relations. Social networks express a common association for the representatives of anthropology, sociology, history, social psychology, political science, human geography, biology, economics, communication sciences and other disciplines that are interested in studying the empirical structure of social relations [4]. Social networks also create a significant target area for the marketers to interact with the users. Social networks are the websites that provide the opportunity for people in the form of online communities to share the content created by the user [5]. The studies show that many people who connect to the social networks' sites check their profiles or do another online activity at least once a day [6]. The use of comments on the social media is increasing. The individuals and organizations use comments on the social networks to affect the buyer's decisions, decision making in the election, marketing and product design. The number of these comments is increasing day by day. Although this rapid growth has many benefits and provides more information, it raise also some considerable analysis challenges according to the huge volume of comments. The main challenges is accessing to the useful and required information in the shortest time when there are considerable amount of comments. To this problem, one of the possible solution is summarizing all comments and providing a comprehensive summary which contains all essential information. Automatic text summarization consists of producing a shorter version of the original document by a computer program so that the main features and main points of the primary document to be maintained [7], [8]. According to the definition presented in the standard ISO 215 in 1986, summary is "a brief retelling the document" [9]. Since human is able to understand the concepts in the text and their relationship using his own knowledge and intelligence, the human summarization is much better than machine summarization. But, human summarization is a tedious and time-consuming task. The ultimate goal of summarization systems is to make a summary with a quality close to the human summaries [7] in a short time. A good summary has a high continuity and readability in addition to the proper coverage of the contents. According to Hovy & Lin [10], the automatic summarization system can be categorized based on source, target, and output. In general, there are four types of text summarization:

1) Extraction summarization: In the extraction summarization, a selection of the original text is returned unchanged as the summary, and often the sentence is considered as the selection unit [11]. The structure of

*Corresponding author

sentences does not change in this method. In the extraction summarization, the sentences must be selected in such a way that there is no redundancy and repetitive sentences while fully covering the content of the text as well as have a high legibility and accuracy.

2) Abstract summarization: In the abstract summarization, the structure of sentences are generally changed, this type of summary is an interpretation of the original text. In this method, first, the system analyzes the text and then expresses its perception of the text in the form of an understandable language for the user [12].

3) Single text summarization: In the single text summarization, the input of the summarization system is just one document. Because in this model of summarization, we are faced with only one document, it is more likely that it talks continuously about a topic and there is no sub-topics [13].

4) Multi-document summarization: In the multi-document summarization, we have multiple documents as input of summarization system. In fact, the multi-document summarization is done on the documents which are related to a topic, but their view angle (aspect) are different from each other. Thus we are facing with multiple sub-topics. In the multi-document summarization we are faced with more complexity than single text summarization.

In this work, we aim at proposing a novel approach for summarizing a huge volume of online comments in Persian language. Our approach is mainly based on a clustering and graph scoring techniques. Since, in a set of comments, different aspects are usually addressed, we are dealing with a multi-document summarization case. In other words, we should deal with multiple sub-topics in the summarization process. Using clustering helps to better identify different subjects in the set of comments and thus provide a more relevant summary. The final phrase extraction to produce the summary is performed using a ranked graph. In our approach, we also different similarity measures to calculate the distance between comments and phrases. In the rest of the paper, we first present a study of the current related works in Section 2. The proposed approach to comment summarization is detailed in Section 3. Then, in Section 4, we present the results of our experiments as well as comparing our approach to a state-of-the-art Persian text summarization method. Finally, we conclude this paper with a conclusion in Section 5.

II. Work Study

In this Section, we present few works in the text summarization domain. Term frequency-based summarization as one of the first summarization method was used [14] in 1958. The title-based method is also one of the first methods of text summarization [15] and its main idea is that the subject and title of the text always represent the text's content. The importance of referring expressions including specific phrases and the importance of their subsequent sentences are discussed in [16]. Also, the Swesum system that is a multilingual summarization system operates on the same basis [17]. In [16], a single text summary is made using sentences getting the highest scores. Then, the sentences are clustered using syntactic and semantic similarities in order to specify the parts of the text that should be included in the summary. Finally, the summary is generated by extracting a sentence from each cluster [16].

Graph-based method provides a way for identifying the topics raised in the document. After the usual preprocessing steps, the sentences of the documents are displayed in the form of nodes in a graph without the direction. The nature of nodes and edges will be defined due to the type of text. Each node contains a sentence of the text. The weight of the edges displays the semantic and lexical relationship or the common points between the two nodes [18]. The method based on Latent Semantic Analysis in the text is used to extract and present the contextual meaning of the word and the similarity of sentences based on the observation of the words co-occurrence. The method based on the neural networks is also a machine learning approach that provides a summary with a desired length using the artificial neural networks. In [19], author also uses a fuzzy logic based method considers each features of a text such as the sentence length, sentence resemblance to the title, resemblance to the keywords, and so on as the fuzzy system input. Conroy used Markov chains in the summarization of the text for the first time in 2001 [20]. Markov chains are sequences of random variables that all of these random variables have the same sample space, but their distribution of probabilities can be different. Text summarization in Persian language raises some specific challenges. Actually, according to special characteristics of Persian, the preprocessing methods and similarity measures need to be adjusted. As an example, since there are so many unique words containing space character, to tokenize sentences in word level, using space character is not relevant in Persian. Also, most of user generated comments in Persian are written in spoken language which dramatically and lexically different from standard written language. We note most famous summarization systems in the Persian language as follow. FarsiSum: this system is a web-based summarization tool for the Persian language which has been created based on SweSum. This system is able to summarize the Persian newspaper texts with HTML format and encoded text with Unicode format [21].

Ijaz is a summarization system for single text and multi-document summarization of the Persian news. This system was created by the Information Technology Organization of Iran and the Web Technology Lab of Ferdowsi University of Mashhad [22]. We compare our approach with Ijaz in experiments in Section 5.

III. Comments Summarization

In this section, we detail our approach to summarize a corpus of comments. The proposed approach is a graph-based summarization method. To give a brief description, first, all input sentences are preprocessed. After preprocessing, all sentences are semantically and lexically clustered. Therefore, a few clusters of sentences are generated that each cluster contains a number of similar sentences. The sentences of each cluster are scored according to their specific characteristics and relation with the other sentences in the same cluster. The sentences will be in the final summary which have the highest

score. Actually, in each cluster, the sentence that has the most relevance to other sentences is the pivotal sentence, and is more suitable for expressing the information in that cluster than other sentences.

In general, this method consists of three phases: preprocessing, clustering and constructing graphs. Fig. 1 illustrates three steps of the proposed approach.

A. Preprocessing

Preprocessing is the first step to prepare and put the text documents in a suitable format. It has been proven that only 33% of the words of a text are useful and can be used to extract the information [23]. Here, the preprocessing consists of five steps which are described in the following.

Tokenization: In this step, all comments in the corpus are divided into the meaningful units. This is done by the tokenizer function. The tokenized helps better identification of stop words and stemming in next steps.

Normalization: We replace all text characters with its standard equivalents. Actually, in Persian most comments are written in spoken language which has no predefined structure. Also, some words are written in brief. Non-structured texts have no default structure, and we consider them as an arranged set of the sentences [24]. Taking all these problems, the normalization aims at converting the comments to a standard format.

Stop-words: Stop words are the words with little importance in terms of meaning despite the frequent repetition in the text. Several lists of stop words have been also created for the Persian language which has an average of 1000 words [11].

Stemming: In this step, every single fragmented word is given to the Stemmer function. By means of stemming, we transforms different forms of a word into a similar standard one.

POS Tagging: After stemming, recognition of parts of speech (POS) in comments is performed using [15]. In fact, POS tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

B. Clustering

As mentioned, in summarization of comments, we are facing with multiple sub-topics. That is, although all comments are about a single topic, but different aspects are addressed in comments. Taking this fact into account, we are dealing with a multi-document summarization. To better identify all aspects (sub-topics) and reflecting all of them into the final summary, we first perform a clustering. By means of a clustering based on semantic and lexical similarity between comments, we obtain cluster of comments that in each one, a unique topic is addressed. Then, for each cluster of comments, we produce a summary. At final phase, these middle summaries are gathered and used to produce the final summary. In the following, we describe the clustering process. Clustering is a no supervised machine learning method. It consists of categorizing a set of elements into several clusters when each clusters contains the most similar elements [5]. Document clustering has many applications in text analysis such as fast data recovery, document organizing with no supervision, and so on [22].

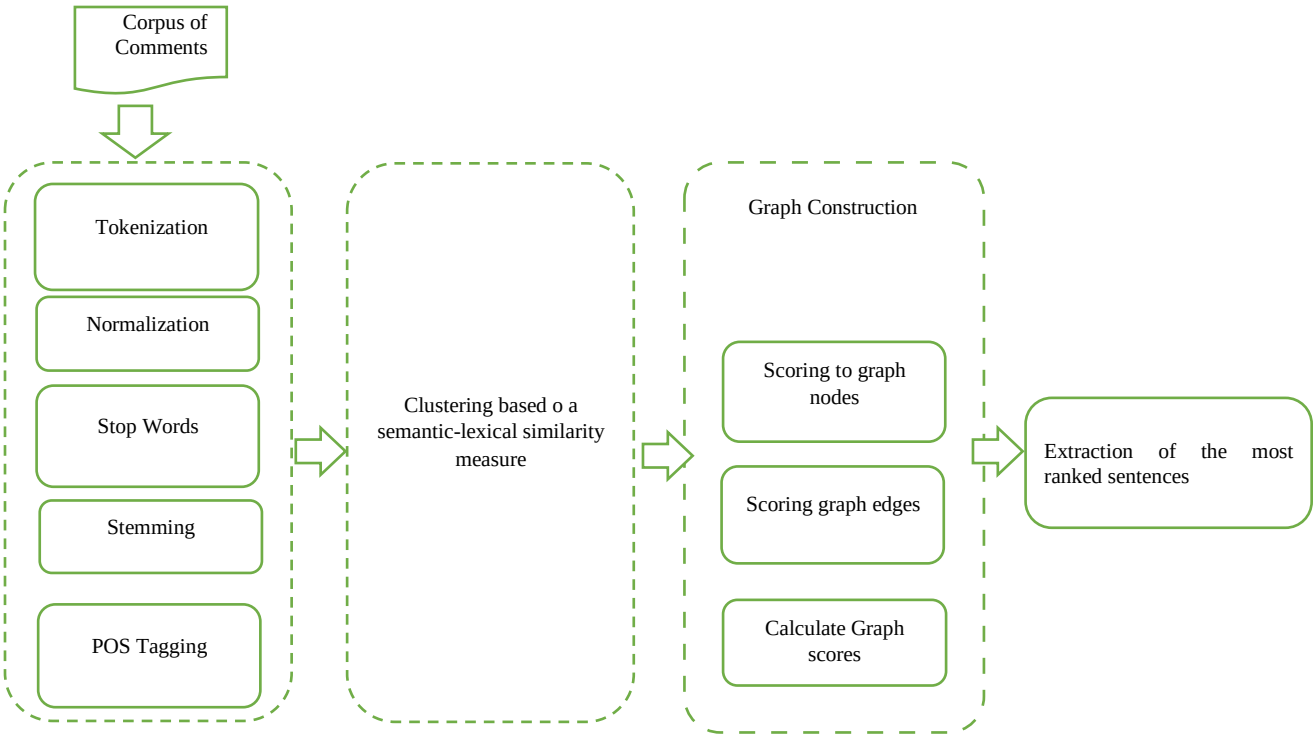

Fig. 1. Flowchart diagram of the proposed approach.

In our approach, we use K-Mean clustering algorithm. K-Means starts with K random points (cluster centers) and repeatedly assigns data to the nearest cluster centers and then updates the cluster centers taking the new points into account [23]. Along with all the advantages, this algorithm has some disadvantages such as the dependence of the results on its initial conditions, such as the number of clusters [23]. To overcome this disadvantage, we use Elbow method which determine the best number of clusters (optimal K). This method was proposed by Robert et al. in 1953 [25]. In this method, two-dimensional graph is used to determine the optimal cluster number. In this graph, the X axis represents the number of clusters and the Y axis represents the value of clustering variance for different clustering. The idea is that the final clusters have to be the best and most optimal [26]. In Elbow method, the elbow part of the graph show the optimal number of clusters. In order to better identify similar comments, in K-Means algorithm, we use a semantic and lexical similarity measure which is different from the default K-Means similarity measure. Calculating the similarity between the sentences is a very difficult and complex process. The popular methods of measuring the similarity are divided into the three main categories based on the used criteria: common words, TF-IDF, and use of linguistic criteria [27]. The similarity measure used in our approach is based on cosine similarity and semantic and lexical distance between sentences [28]. In this similarity measure, to measure the similarity of a word pair, first a string similarity is determined, and then the semantic one. This technique is then enhanced by using a term frequency ponderation. Thus, words appearing in both texts are treated as a word pair which consists of two identical words. After calculating their similarity scores, they are added up into a similarity sum S_{same}, and then these words are discarded from further consideration [28]. The goal is to match words across the two texts according to their mutual similarity score. Hence, we search for the highest value within the final similarity matrix, and add it to a similarity sum $S_{different}$. We then remove the row and the column of the matrix to which the selected cell belonged, thereby discarding all other word pairs in which words from the chosen pair appeared. We repeat this procedure until there are no more rows and/or columns left in the matrix [28]. Equation (1) show how the final similarity between two comments is calculated [28].

$$S(P.R) = \frac{\left(S_{same} + S_{diffrence}\right) * (m+n)}{2mn} \quad (1)$$

In other words, the final similarity score S(P,R) is gained by summing up the similarity scores of words that appear in both texts (S_{same}) and the scores of word pairs formed from words unique to one of the texts ($S_{different}$). Lastly, this sum is multiplied by a reciprocal harmonic mean function of the lengths of both texts, so as to achieve a final text similarity score between 0 and 1.

C. Graph Construction

In this step, we take comments in every cluster and represent them in a form of scored graph. Using this graph, we determine the sentences with highest score as the representative sentence of the cluster. The sentence score is calculated according to node and edge scores in the graph. In the following we detail the graph construction process and how nodes and edges scores are calculated.

5) Scoring the Graph Nodes

Each graph node is representing one of the sentences in the corpus, and each node receives its own special score based on the importance of the corresponding sentence. The score of each node is calculated based on the following measures:

- *Term Frequency*: The term frequency is one of the oldest techniques for measuring the relevance and importance of a sentence for text summarization [29] and was introduced by Luhn for the first time in 1958 [30]. According to this assumption, the words that have a high frequency in the text are more important than other words. Of course, stop words are removed and all other words are stemmed before calculating the term frequency [31]. The importance of a sentence based term frequency is calculated according to (2):

$$(S_i) = \sum_{j=1}^{N} freq(W_j) \quad (2)$$

Where N is the number of all the words in the sentence and Wj is the frequency of each word.

- *Sentence Resemblance to the Title*: The title usually indicates the main topics discussed in a document. This is an efficient method to calculate the value of a sentence in a document. This scoring technique assumed that the sentence that has a higher relevance to the title is the main sentence in the document [29] [31]. This score is calculated based on (3):

$$SentRST(S_i) = \frac{W_{si} \bigcap W_t}{|W|} \quad (3)$$

Where Wsi is the existing words in the sentence Si, Wt is the existing words in the title, and |Wt| is the total sum of the existing words in the title.

Sentences containing Cue-Phrases: Here, the sentences that contain the Cue-Phrases such as "In short", "as a result", "as a summary", in this paper, and so on are considered very important. This method relies on a predefined dictionary of sign expressions. This technique is calculated according to (4) [31]:

$$\text{CuePhr (Si)} = \frac{\text{Number of CuePhrases in Si}}{\text{Total of CuePhrases in the Document}} \quad (4)$$

- *Sentences containing numerical data*: Numerical information refers to the important information such as date, percentage, cost, feature, and so on [29], [31]. The score of the sentence Si is calculated using this feature according to (5):

$$\text{Num Data (Si)} = \frac{\text{Number of Numerical data in Si}}{\text{Total of Word in Si}} \quad (5)$$

- *Sentences containing Noun Phrases*: A Noun Phrases is a group of nouns and their transformation. In a sentence, a Noun Phrases can play the role of a subject, an object or its

complement. The score of the sentence Si is calculated using this feature according to (6):

$$NP(S_i) = \frac{\text{Number of Noun Phrases in Si}}{\text{Total of Word in Si}} \quad (6)$$

These five measures apply to each single sentence of each cluster. The final score of a node is sum of these five measure normalized by the number of words in the corresponding sentence. The normalization is performed since all sentences do not have the same length.

6) Scoring the Graph Edges

Once all graph nodes are scored, we assign a score to each edge according to relationship between adjacent nodes. Actually, an edge represent a relationship between two sentences (represented by nodes). To score the graph edges, we use three measures:

- The number of common words between two nodes: in this step, the node which has acquired the highest score is considered as the main node in the cluster, and the number of their common words with all the other nodes is calculated. Then, the number of common words between the node which has acquired the second score and the other nodes is calculated. This process continues until the last node of each cluster. The score for this step is calculated according to (7):

$$\text{Score (i.j)} = \frac{\text{The Number of Common Words Between (i,j)}}{\text{Total of Words in i.j}} \quad (7)$$

- Calculate the semantic similarity between the two nodes: the semantic similarity between two nodes is calculated based on the semantic measured in (1).
- The number of common keywords between two nodes: extracting keywords is an important step to retrieve the document, retrieve the web page, clustering the document, summarization, text mining and so on [32]. Keywords can be extracted by different methods [33]. In next Sub-section we describe how we extract keywords in a corpus of comments. This score is also normalized by the number of words in two sentences. The score for common keywords is calculated according to (8):

$$\text{Score (i.j)} = \frac{\text{The Number of Word keys Between (i.j)}}{\text{Total of Words in (i.j)}} \quad (8)$$

7) Keyword Extraction

Fig. 2 illustrates how we extract the keywords in the comments. First, all sentences are preprocessed. Then, using the term frequency, we find the M most occurred terms in the corpus. According to our experiments, we consider M=15 in this work. At next step, a matrix is created where the columns are the most occurred terms and rows correspond to all other words. Each element in the matrix show the number of times that the two words co-occurred. At this step, we find the P words that co-occurred the most. According to experiments, we consider P=10 in this work. Then, we calculate the TF-IDF value for all selected words. Finally, the words having the highest TF-IDF value are selected as keywords.

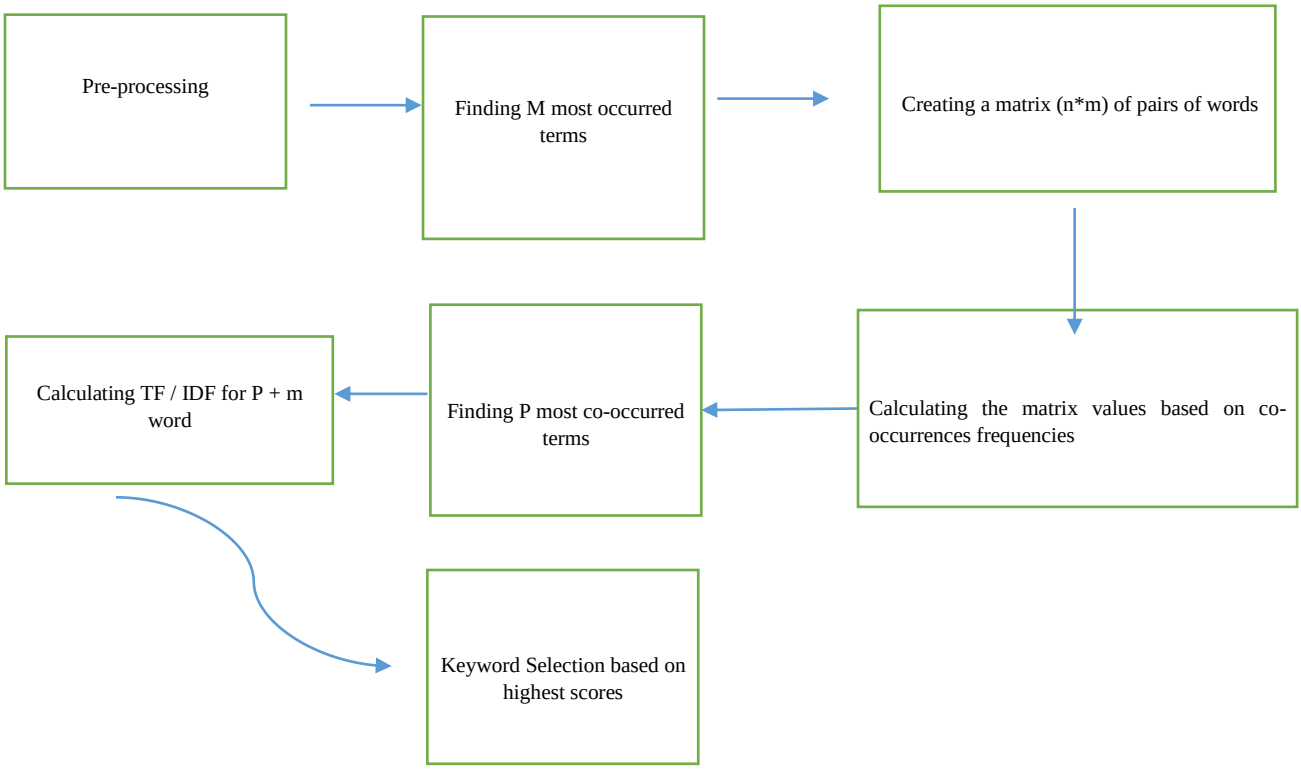

Fig. 2. The proposed algorithm to extract the keywords in a corpus of comment.

IV. Experiments

There are generally two ways to evaluate the summaries generated by automatic summarization methods: intrinsic evaluation and extrinsic evaluation. In intrinsic evaluation, the quality of summaries is directly evaluated. This is done by comparing the summary with reference summaries or with the direct opinion of a few human experts. In extrinsic evaluation, the quality of summaries is evaluated based on how it is useful in performing a specific task (such as categorization). This is also called task-based evaluation. In this work, we first evaluate the quality of summaries extracted by our approach with an intrinsic evaluation. Then, we compare the results of our approach with those of Ijaz summarization system [9].

A. Results

To evaluate the quality of the extracted summaries by each approach, four linguistic experts investigate all final summaries. Each expert gives a score from 0 to 10 to each summary considering how accurate and comprehensive it is. The final score of each summary is sum of scores assigned by each expert.

Table I shows the expert's scores for the first three summaries extracted by the proposed approach. As shown in Table I, the average score given to the top summary extracted by the proposed approach is about 9.25 out of 10.

To evaluate the impact of clustering on summarization performance in our approach, we also extracted summaries without performing the clustering. Then, the summaries extracted without clustering are scored by the experts. Table II shows scores given to top three summaries extracted without clustering.

According to Tables I and II, the clustering improves considerably the summarization performance. Without clustering, the average score for extracted summaries is about 7.66 out of 10 when using the clustering the average score for extracted summaries is increased to 8.75 out of 10.

TABLE I. Expert's Scores to the Top Three Summaries Extracted by the Proposed Approach

Sentence	Expert 1	Expert 2	Expert 3	Expert 4	Total average
First summary	10	9	9	9	9.25
Second summary	9	9	8	9	8.75
Third summary	9	8	8	8	8.25
Total average	9.33	8.66	8.33	8.66	8.75

TABLE II. Expert's Score to the Top Three Summaries Extracted without Clustering

Sentence	Expert 1	Expert 2	Expert 3	Expert 4	Total average
First summary	9	9	8	9	8.75
Second summary	8	7	7	7	7.25
Third summary	7	7	7	7	7
Total average	8	7.66	7.33	7.66	7.66

TABLE III. The Expert's Score to the Summaries by the Ijaz System

Sentence	Expert 1	Expert 2	Expert 3	Expert 4	Total average
First summary	9	9	9	9	9
Second summary	7	7	7	7	7
Third summary	6	6	6	6	6.25
Total average	7.66	7.33	7.33	7.33	7.4

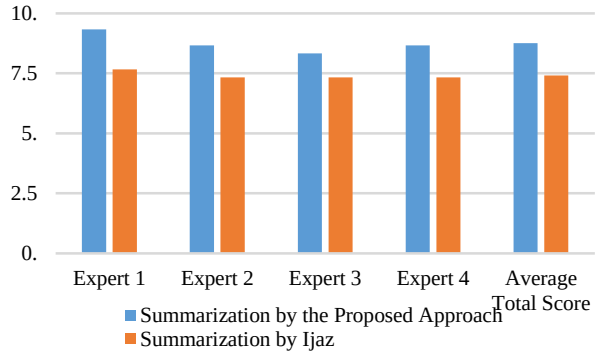

Fig. 3. Comparison of the proposed approach with Ijaz system.

B. Comparing with Ijaz

At this step, we compare the summaries obtained by our approach with summaries extracted by Ijaz summarization system presented in Section 2. As mentioned before, Ijaz is one the of the most famous summarization method for Persian texts. In this step, we also asked the same four linguistic experts to give a score to summaries extracted by Ijaz on our corpus of comments. Table III shows the scores given to top three summaries extracted by Ijaz summarization system.

In Fig. 3, we demonstrate the score given to each summary extracted by the proposed approach and summaries extracted by Ijaz system. According to results, the average score given to summaries by Ijaz is about 7.4 out of 10 when the average score given to summaries by the proposed method is about 8.75 out of 10.

V. Conclusion

In this work, we proposed an approach to summarizing a huge corpus of user generated comments in Persian language. Our approach uses a clustering technique to deal with multiple aspects within comments. It is also based on a scored graph summarization method. We use several measures to calculated the semantic and lexical distance between sentences in the corpus.

According to experiments and by an intrinsic evaluation, the summaries extracted by our approach obtained an average score of 8.75 out of 10. We also evaluated the impact of clustering on final summaries. According to results, the clustering help to obtain more accurate summaries. We also compared our approach with the state-of-the-art summarizer in Persian language called Ijaz [22]. The average score given to summaries extracted by the proposed method is considerably higher than the average score given to Ijaz summaries.

As future work, we aims at investigating other similarity measures to better calculate the similarity between short texts like comment and tweets in terms of lexical, semantic, and structural criteria. Also, it is needed to study the use of other summarization methods other than graph-based ones.

REFERENCES

[1] R. Keshvarian, A. Taei Sadeh, A. Jami, "The role of social networks in online advertising and marketing". Second National Conference on Applied Research in Computer Science and Information Technology, 2010.

[2] S. Memar, S. Adlipour. "Virtual social networks and identity crisis". Scientific-research quarterly of the Social Studies and Research in Iran / Vol. 1, No. 4, 2012

[3] F. Akhavan, M. Noghani, "Virtual social networks and happiness", Treatise on Culture, Research Institute for Humanities and Cultural Studies, Year 4, No. 2, 2014

[4] M. Everett, T.Valente, "Social Networks" An International Journal of Structural Analysis, ISBN: 0378-8733.

[5] C. H. Shah, A. Jivani. "Comparison Of Data Mining Clustering Algorithms". Nirma University International Conference On Engineering. 2013

[6] F. Christian. "Social Networking Sites and Thd Sutveiilance Society Austria Vienna". Forschungs ruppe Unified Theory of Information Cresearch Group Unitifed Theoty of Information, 2009

[7] I. Mani, M. T. Maybury, (Eds.). "Advances in Automated Text Summarization". Cambridge, the Mit Press, 1999.

[8] I. Mani, M. Maybury: Advances in Automatic Text Summarization.The Mit Press, 1999

[9] E. Delgado López, Iso 215 Documentation – "Presentation of Contributions to Periodicals and Other Serials. Iso 215": Technical Report, International Organisation for Standardisation, 1986.

[10] E. Hovy, C. Y. Lin, "Automatic Text Summarization in Summarist". In Proceedings of the Acl Workshop Onintelligent Scalable Text Summarization (Pp. 18–24), Madrid, Spain, 1997.

[11] Y. Kumar Meena, P. Deolia, D. Gopalani. "Optimal Features Set for Extractive Automatic Text Summarization", IEEE, 2015.

[12] T. Akhavan, M. Shams Fard, M. Erfani Jorabchi, "single text summarization and multi-document summarization of the Persian texts", Fourteenth Annual National Conference on the Iranian Computer Society, Amirkabir University of Technology, 2008.

[13] M. Moshki, M. Analoye, "Multi-document summarization of the Persian texts using a clustering-based method", First National Conference on Software Engineering in Iran, Technical and Vocational School of Sama in Roudehen, 2009

[14] T. Zhu, Q. Liu, Q, "Sentence Descending Algorithm for Automatic Text Summarization". In International Conference on Computational and Information Sciences, Pp. 301-304, 2011.

[15] Tashakori. "Bon: The Persian Stemmer" In Eurasia-Ict 2002: Information and Communication Technology, Ed: Springer, 2002.

[16] E. Hovy, C. Y. Lin, "Automatic Text Summarization in Summarist". In Proceedings of the Acl‟97/Eacl‟97 Workshop Onintelligent Scalable Text Summarization, Pp. 18–24, 1997

[17] H. Dalianis, "Swesum A Text Summarizer for Swedish", Technical Report, Tritana-P9915, Iplab-174, 2000

[18] B. Samei, M. Eshtiagh, F. Keshtkar, S. Hashemi, "Multi-Document Summarization Using Graph-Based Iterative Ranking Algorithms and Information Theoretical Distortion Measures," Proceedings of The Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, 2014.

[19] L. Suanmali, M. S. Binwahlan, N. Salim, "Sentence Features Fusion for Text Summarization Using Fuzzy Logic". In Hybrid Intelligent Systems, His'09. Ninth International Conference on, Vol. 1, Pp. 142-146, 2009.

[20] J.M. Conroy, D. P. O'leary, "Text Summarization via Hidden Markov Models". Proceedings Of The 24th Annual International Acmsigir Conference On Research And Developmentin Information Retrieval, Sept. 9-12, New Orleans, Louisiana, United States, Pp: 406-407, 2001.

[21] H. Mazdak, "Farisum-A Persiantext Summarizer", Master Thesis Department of Linguisticsm, Stockholm University, 2004

[22] A. Masoumi, M. Kahani, S. A. Tusi, A. Steari, "IJaz: An operational system for single text summarization for the Persian news texts", Ferdowsi University of Mashhad, 2014

[23] J. Stephen, R. Conor Heneghan, "A Method for Initialising the K-Means Clustering Algorithm Using Kd-Trees". Pattern Recognition Letters،Vol.28. Pages: 965-973, 2007.

[24] M. H. Moattar, M. M. Homayounpour, N. Farzinfar, "Normalization of the Persian texts using matching to the phrases pattern", The 10th Annual Conference of the Iranian Computer Society, 2004

[25] L. Robert, Thorndike, "Who Belongs In The Family", Psychometrika-Vol, 18, No. 4, No. 4, 1953.

[26] P. Bholowalia, A. Kumar, "Ebk-Means: A Clustering Technique Based on Elbow Method and K-Means", International Journal of Computer Applications, November 2014

[27] D. Metzler, Y. Bernstein, W. Croft, A. Moffat, J. Zobel, "Similarity measures for tracking information flow". Proceedings of CIKM, pp. 517–524, 2005.

[28] B. Furlan, V.Batanović, B. Nikolić, "Semantic Similarity of Short Texts in Languages with A Deficient Natural Language Processing Support", Elsevier, 2013.

[29] R. Ferreira, L. De Souza Cabral , R. Dueire Lins , G. De Frana Silva , F. Freitas ,G.D.C. Cavalcanti , R. Lima , J. Steven, S. Luciano Favaro. "Assessing Shallow Sentence Scoring Techniques And Combinations For Single And Multi-Document Summarization", Elsevier, 2013.

[30] H. P. Luhn, "The Automatic Creation of Literature Abstracts". Ibm Journal of Re-Search and Development, pp.159–165, 1958.

[31] H. Oliveira , R. Ferreira , R. Lima , R. Dueire Lins , F. Freitas , M. Riss , J. Steven Simske, "Assessing Shallow Sentence Scoring Techniques And Combinations For Single And Multi-Document Summarization" , Elsevier Ltd, 2016.

[32] C. H. W. Rogier Brussee, W. Slakhorst, "Keyword Extraction Using Word Co-Occurrence". Workshops on Database and Expert Systems Applications, 2010.

[33] M. Islami Nasab, R. Javidan, "providing a method based on cosine similarity and the vocabulary network to find the amount of semantic similarity between the texts", 7th International Conference on Information and Knowledge Technology, Urmia University 5-7 June 2014.