JACET Journal of Advances in Computer Engineering and Technology, 2(4) 2016

Analysis of users' comments about the divorce factors

Fatemeh Eghrari Solout[1], Mehdi Hosseinzadeh[2]

Received (2016-08-16)
Accepted (2016-08-11)

Abstract — **One of the most important issues related to knowledge discovery is the field of comment mining. Opinion mining is a tool through which the opinions of people who comment about a specific issue can be evaluated in order to achieve some interesting results. This is a subset of data mining. Opinion mining can be improved using the data mining algorithms. One of the important parts of opinion mining is the sentiment analysis in social networks. Today, the social networks contain billions of users' comments about different issues. In previous researches in this area, various methods have been used for Persian comments analysis. In these studies, preprocessing is one of the most important parts. It arranges the data set for analysis in a standard form. The number of hashtags selected for analysis is limited. To detect the positive and negative comments, knowledge extraction or neural network techniques have been used. The current research presents a method of analysis which can analyze any hashtag for each group of users and has no limitations in this regard. Type of hashtag, the number of likes, type of user and type of positive and negative sentences can be analyzed by this method. The results of simulation and comparison of divorce data set show that the proposed method has an acceptable performance.**

Keywords - **social networks, content analysis, comment mining, divorce, users' comments**

1,2- Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran. (Eghrari.1361@yahoo.com)

I. INTRODUCTION

The use of social media creates many opportunities for expressing the people opinions, but it makes some problems when they have some beliefs. Opinion mining is a type of natural language processing which can follow people moods about a particular product through reviewing. In this article, a method is provided to investigate the users' comments about the divorce and to determine the reasons of divorce from the users' perspective and the effects of each factor.

Opinion mining is a range of natural language processing and text analysis which aims to discover and exploit the intellectual quantities from the textual resources. Generally, opinion mining tasks can be referred to as sentiment analysis. Its purpose is to prove the polarity of the source text (for example, to distinguish between negative, neutral and positive beliefs). Its second mission is to identify the degree of objectivity and subjectivity of a text (the identification of reality data in contrast with comments). This task is sometimes called as opinion mining. The third purpose is to discover or summarize the outspoken opinions about the selected features of the evaluated products. Some of the authors describe this task as sentiment analysis. All three classes of opinion mining missions can greatly benefit from the additional data provided by the social networks.

In this research, a method is provided which investigates users' comments about divorce. The proposed method consists of several steps and uses some tools such as hashtag, tagging and likes. In this method, users are divided into different categories. Based on users' comments, the causes of divorce will be determined. The

rest of this article includes the main concepts of opinion mining, review of literature, the proposed method and conclusion.

II. OPINION MINING

Sentiment analysis or opinion mining is a process which extracts and classifies the comments and emotions about an issue. The issue can be anything including a product, a person, etc. [1].

The comment about an issue can be expressed directly or in a comparative form. The direct comment about an issue expresses the opinion, but the comparative comment provides a comparison. The expression of emotions can be classified according to the sentence level, positive and negative characteristics, etc. [2].

Emotion is the result of a person's relationship with the environment. It includes sophisticated sets and coordinated components to respond. These responses may include physical adjustments and emotional or practical expressions. In person interaction with his environment, the person experiences a series of problems in his life. Emotions have different intensities caused by personal experiences regarding the environment. These emotions can be positive or negative [3].

1. The main phases of the process of opinion mining in social networks

Opinion mining is a process which involves many technological fields. Data recovery, data mining, artificial intelligence and computational linguistics are some fields which play the important roles in this area. In general, there are two main phases in comment mining. The first phase is the preprocessing of documents. The output of this phase can have two different forms [4]:

• Document-based output

• Concept-based output

In document-based display, the method of displaying the documentation is important. For example, the documents conversion into an intermediate semi-structured format or the use of an index on them or any other display which makes working with documents more efficiently. Any entity in this display will be a document eventually. In the second type, the documents display is improved. Concepts and meanings contained in the document, the relationship between them and any other extractable conceptual information are extracted [5].

In this type of display, the documents are not referred as the entity, but the concepts extracted from these documents are considered. The next step is to extract the knowledge from the documents display intermediate forms. Based on the type of displaying a document, the method of extracting the knowledge is different. The document-based display is used for grouping, classification, visualization and so on [6].

2. Comments ranking

Comments should be categorized in different classes such as sport, art, emotional, political groups. When a new comment is recorded, the relevant class must be determined. There are some methods to rank the comments. After ranking, the desired class must be tested in terms of efficiency. The created comment is tagged. These tags are compared with the real tags of the desired class. The proportion of correctly classified documents to the total number of documents is called accuracy. Two criteria are used to compare the classifiers [7].

• Precision: A part of the retrieved documents which are relevant

• Recall: A part of relevant retrieved documents

3. Comments classification

Comments classification is done based on tagging a concept to each comment and placing the similar comments in a similar category. In other words, each comment belongs to a certain category in accordance with the embedded similar semantic units. One of the major categories is the category of emotional comments in social networks [8].

The criteria used in opinion mining are as follows:

• Attitude: The author's view about a specific issue which can be negative, positive, or neutral

• Aggregation: The relationship between two mentioned elements and the extent of their similarity

• Mentality: It specifies the subjectivity or objectivity of the comment

• Length: The length of a comment (short, long, medium, etc.)

4. Polarity-based classification

It specifies that the sentence which is tagged as intellectual is positive, negative or neutral. Classification at the sentence level differs from the classification at the document level since the number of words is less than the number of words in the document. Recent researches indicate that the efficiency of polarity-based classification at the sentence level can be increased by using the learning algorithms with two features; polarity information and linguistic features.

Linguistic features include the speech, the depth of the words combination, the type of expression and the field. Not only the supervised learning methods can be implemented for symmetric classification, but also the semi-supervised methods can be used in this regard [9].

III. REVIEW OF LITERATURE

Opinion mining is divided into two main categories; level-based opinion mining (document level and sentence level) and feature-based comment mining. In the first category (document or sentence), a comment is examined completely as a comment and its positive or negative emotion is specified. In the final report of this category, the users' final satisfaction with the entity will be expressed. In the second category, the specific features of an entity are regarded. For example, in the case of commenting about a cellphone, the battery, the screen and the body are the features of this cellphone. In the final report of this category, the users' satisfaction rate with each feature is expressed. In this research, a method for opinion mining at sentence level is provided [10].

1. Lexicon-based

The first research describing the lexicon-based method for opinion mining has been conducted by Turney. In the first step, the comments are tagged by the POS (part of speech) tagger. Four specific patterns are considered as the candidate patterns to express the opinion. In the second step, using PMI (point-wise mutual information) measure, the ratio of the relationship between the words which have the same pattern is investigated. This measure is expressed as follows:

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right) \quad (1)$$

In this formula, Pr (term1 ^ term2) indicates the simultaneous occurrence of two terms. Pr (term2) Pr (term1) shows the ratio of occurrence of these two terms when they are not dependent on each other statistically. Therefore, the tendency of this word is extracted based on the following formula:

SO (phrase) = PMI (phrase, "excellent")-PMI(phrase, "poor") (2)

In the third step, the SO average calculated for all words specifies the negative or positive orientation of the document.

In [11], it was tried to discover the syntactic properties of the comment which are independent of range by tagging the words. In the meantime, the syntactic tags of noun, adjective and verb are considered. The main idea of this method is based on the fact that, among the features used by the syntactic tagger of the words, some of them are dependent of the range and some of them are independent of the range. In lexicon-based methods, the syntactic pattern or the words are focused. In the methods mentioned so far, the main focus was on the selection of syntax [11]. A set of words and phrases with their tendency was used. In this method, the phrases were used along with a set of semantic resonators such as "very" and negator such as "not" to determine the final orientation of the words. The method used in this research enjoyed the tagging process.

[12] can be mentioned as another research which used syntactic tagger as the basis for its work. In all these researches, there is a fixed patterns to explore the comment. Table I shows the syntactic patterns of the words which represent the commenting in the text. For example, The first row of the table states that if the first word is an adjective and the next word is a noun, there is likely an opinion in these two words. Table II also defines the signs used by the syntactic tagger [13].

TABLE I
Words pattern

S.NO	First Word	Second Word	Third Word
1	JJ	NN or NNS	Anything
2	RB,RBR, or RBS	JJ	not NN n or NNS
3	JJ	JJ	not NN n or NNS
4	NN or NNS	JJ	not NN n or NNS
5	RB,RBR, or RBS	VB,VBD,VBN or VBG	Anything

TABLE II
Used signs

S.NO	Tag	Description
1	NNP	Noun, proper, singular
2	NNPS	Noun, common , plural
3	RB	Adverb
4	JJ	Adjective or numeral, ordinal
5	JJR	Adjective ,superlative
6	NN	Noun, common, singular
7	RBR	Adverb, comparative
8	VB	Verb, base, form
9	VBD	Verb, present participle, or Gerund
10	WDT	WH-determiner
11	CC	Conjunction, coordinating
12	CD	Numeral ,cardinal
13	DT	Determiner

In the proposed method, a large data set containing 10 million words and tags was used.

2. Researches done in Persian language

Among the large number of articles published over the past decade in the field of comment mining, few researches are related to the non-English languages. In the meantime, limited studies are about Persian language. Chi-square method was used for proper features selection in classification algorithm. The classification algorithm used in this study is a decision tree.

In the final report, it was stated that the use of chi-square method for features selection does not improve the opinion mining in the Persian language. Then, the LDA-based non-supervisory approach called LDASA was introduced. To create linguistic resources and sensitive phrases database, the translation of English vocabulary database was used. Eventually, the research was implemented on the three data sets (cellphones, hotels and digital cameras) [14].

IV. THE PROPOSED METHOD

The proposed method for analysis of comments about divorce has several stages as follows:

1. Preprocessing

At this stage, the data set of the users' comments is filtered and some items are removed.

- Non-Persian posts
- Posts using inappropriate and immoral words
- Posts containing no comment
- Posts unrelated to divorce

In the following, the main hashtags for content analysis are divided according to the following table III. The hashtags are classified and each group has similar hashtags.

TABLE III
Hashtags classification

Hashtag	Phrases
Divorce	I got divorced, I divorced, separation, divorce
Court	#Justice #The Revolutionary Court #Family Court #Court of Appeal #Court #Court #Assets #Legal
Consultation	#Consultation #lawyer #Consultation #Right #Legal Consultation #Uncontested Divorce #Divorce #Attorney #Free legal advice #Family Legal Consultancy #Institution #Legal Institute #Lawyer #Family lawyer #Criminal lawyer #Lawyer #Tehran lawyer #Case
Lawyer	#Lawyer, Attorney, legitimacy, civil rights, family lawyer, lawyer, free lawyer
Treason	Treason, traitor, disloyal, nerveless, stranger, corruption
Separation	Separation, being left alone, living alone,

In analysis of divorce comments, the comments which use the mentioned hashtags are separated.

2. Users classification

The users who used the hashtags of the above table are separated. These users are divided into the categories of lawyers, consultants and individuals involved in divorce. Each section has a criterion according to which the users are divided.

3. Posts classification

The posts related to the hashtags and the posts related to each group of users are separated. The information can be seen in Figure 1.

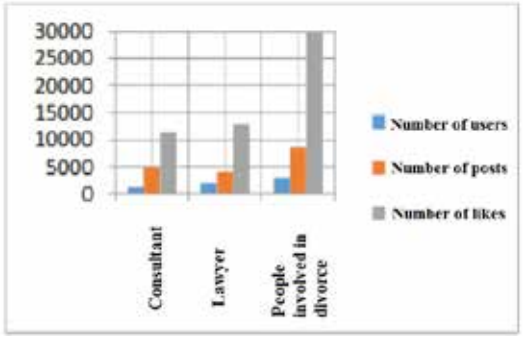

Fig. 1. Content analysis based on user type

Figure 2 shows the users' followers.

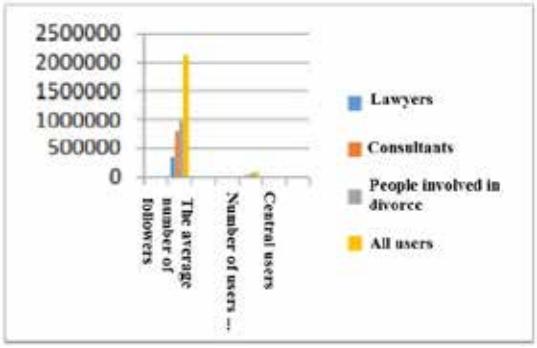

Fig. 2. The number of filtered users' followers

Figure 3 shows the hashtags content analysis based on the users.

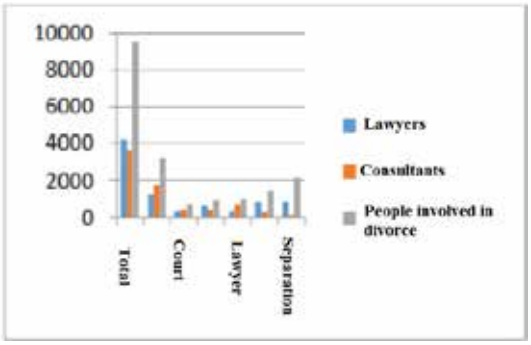

Fig. 3. Hashtags content analysis based on the users

4. Tagging stage

In this research, the reasons of divorce from the users' perspective were analyzed and tagging was done accordingly. The tags in Table IV were used for users' comments.

TABLE IV
Divorce tags

The name of divorce Tag
Financial problem
Immorality
Unemployment
Sexual problem
Treason
Capricious
Addiction
Apathetic

After tagging, the comments were classified based on the comments' tags. It means that each tag contains a set of comments. Each comment can be a reason of divorce. It should be determined that the comments are positive or negative. The positive sentence is a sentence which is confirmed by the user and approves something. The negative sentence denies a manner. The result of tagged comments classification can be seen in Table V.

Table V
The number of comments of each tag

The name of tag	The number of comments
Financial problem	3246
Immorality	2578
Unemployment	2045
Sexual problem	2162
Treason	3854
Capricious	2908
Addiction	2875
Apathetic	1366

1. Results analysis

In this section, to detect positive and negative sentences, the comments were tagged again. A set of tags from intelligent processing center was used for tagging. It contains 10 million dominant words with different tags. These words are tagged according to the positive or negative tags [15].

After tagging the positive and negative comments, the percentage of the negative and positive sentences for each reason of divorce is surveyed. The result shows the reasons of divorce from the users' perspective. It can be seen in

Table VI.

TABLE VI
The number of likes for positive and negative sentences of each tag

The name of tag	The number of comments	The number of positive comments	The number of positive likes	The number of negative comments	The number of negative likes
Financial problem	3246	2041	9532	1753	7245
Immorality	2578	1057	7965	982	6421
Unemployment	2045	1146	8695	851	6308
Sexual problem	2162	1425	9337	635	7561
Treason	3854	2745	10267	843	8435
Capricious	2908	2078	9397	742	6718
Addiction	2875	1578	5463	953	6512
Apathetic	1366	523	6452	451	5178

The results of divorce reasons can be seen in Figure 4.

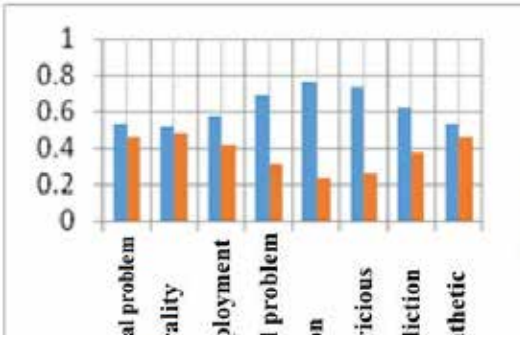

Fig. 4. The reasons of divorce from the users' perspective

V. SIMULATION AND COMPARISON

Excel and MATLAB software were used to simulate the proposed method. Two methods Banitalebi and Peikari were used for comparison [16] [17].

The first criterion for comparison is precision.

This criterion is one of the most important accuracy criteria in the sentiment analysis. It can be calculated by the following equation.

$$precision = \frac{TP}{TP + FP} \qquad (3)$$

In this equation, TP is the number of data which are correctly diagnosed as positive and FP is the number of data which are mistakenly diagnosed as positive. The simulation result based on the above criteria for all positive and negative comments is shown in Figure 5.

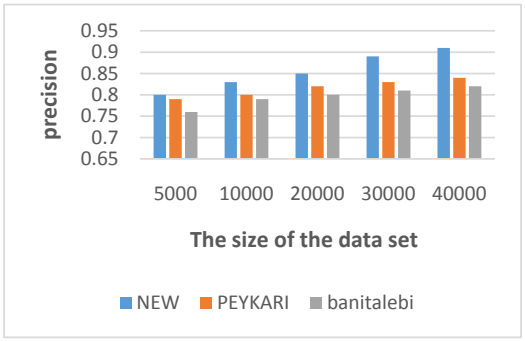

Fig. 5. The comparison of precision criterion

The proposed method for precision caused a better result.

The next criterion for accuracy is recall which can be calculated by the following equation.

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

In this equation, TP is the number of data which are correctly diagnosed as positive and FN is the number of data which are mistakenly diagnosed as negative. The simulation result based on the above criteria for all positive and negative comments is shown in Figure 6.

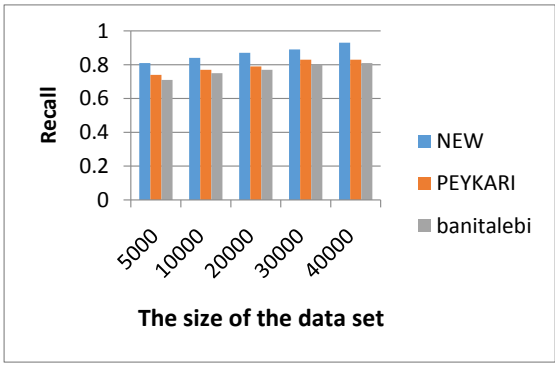

Fig. 7. The comparison of F-measure

According to the obtained result, the proposed method performed better (18%) compared to previous methods.

The next criterion is the F-measure. This measure is between the measures of recall and precision. It can be calculated by the following equation.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \qquad (5)$$

The simulation result for this criterion is shown in Figure 7.

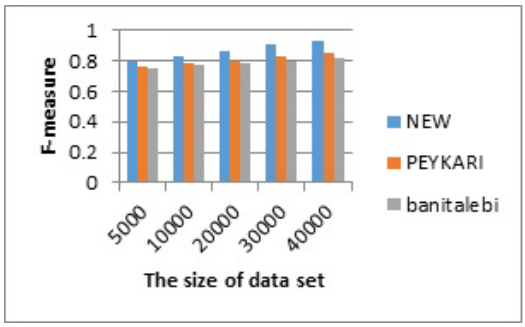

Fig. 7. The comparison of F-measure

According to the above figure, the proposed method performed better (14%) compared to previous methods.

2. Conclusion

In this article, the users' comments about the reasons of divorce have been investigated. A new multi-stage method was used for content analysis. This method is based on hashtags, users, the positive and negative sentences and the number of likes. The result of implementing this method on the total comments shows the reasons of divorce from the users' perspective. The number of likes presented by other users shows an interesting result about a certain reason. The proposed method was compared with other methods according to three important criteria. Simulation results show that the proposed method has a better performance.

REFERENCE

[1] Vijay.B.Raut et al, 2014. “Survey on Opinion Mining and Summarization of User Reviews on Web”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol 5(2). 1026-1030.

[2] G.Angulakshmi, 2014. An Analysis on Opinion Mining: Techniques and Tools, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7.

[3] Raisa Varghese, Jayasree, 2013. “A Survey on Sentiment Analysis and Opinion Mining”, International Journal of Research in Engineering and Technology (IJRET), Vol 2 Issue 11 Nov.

[4] Martin Mikula and KristínaMachová, 2015.” Classification of opinions in conversational content”, IEEE 13th International Symposium on Applied Machine Intelligence and Informatics • January 22-24,

[5] Sagar Bhuta and Uchit Doshi, 2014. “A review of techniques for sentiment analysis of twitter data”, Issues and Challenges in Intelligent Computing Techniques (ICICT), International Conference on, pp. 583-591

[6] Xiuzhen Zhang, Zhixin Zhou, Mingfang Wu, 2014. “Positive, Negative, or Mixed? Mining Blogs for Opinions”, Melbourne Australia.

[7] Jianxin Li, 2014. Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-Aware Approach, Utility and Cloud Computing (UCC), IEEE/ ACM 7th International Conference on

[8] Peiman Barnaghi, 2016. Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment, Big Data Computing Service and Applications (Big Data Service), IEEE Second International Conference on

[9] R. Feldman, 2013. “Techniques and applications for sentiment analysis”, Communications of the ACM, vol. 56, pp. 82-89.

[10] R. Piryani, (2015) Analytical mapping of opinion mining and sentiment analysis research during 20, Information Processing and Management, Elsevier Ltd. All rights reserved

[11] Wang, H., & Wang, W. (2014). Product weakness finder: An opinion-aware system through sentiment analysis. Industrial Management & Data Systems, 114 (8), 1301–1320

[12] Weichselbraun, A., Gindl, S., & Scharl, A. (2014). Enriching semantic knowledge bases for opinion mining in big data applications. Knowledge-Based Systems, 69 , 78–85

[13] P. D. Turney, (2002) “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, , pp. 417- 424.

[14] Khatibi, T., Sepehri, V. & Hamidpour, B. (2008). Efficacy of chi square method to select the parameters in the Persian text comment mining, Presented at the National Conference on Electrical Engineering, Computer and Information Technology, Hamedan, Sama educational and cultural center.

[15] www.rcisp.com

[16] Peikari, N. & Yaghoubi, S. (2015). The sentiment analysis in Twitter social network using the text mining techniques, International Conference on Web Researching.

[17] Banitalebi, A. (2016). Opinion mining of social networks with the use of machine learning algorithms, International Conference on Computer Engineering and Information Technology.