Journal of AI and Data Mining
Vol 3, No 2, 2015, 169-179. 10.5829/idosi.JAIDM.2015.03.02.06

Feature extraction in opinion mining through Persian reviews

E. Golpar-Rabooki[1], S. Zarghamifar[2*] and J. Rezaeenour[3]

1. Department of Mathematics, University of Qom, Qom, Iran
2. Department of Computer Engineering, University of Qom, Qom, Iran
3. Department of Industrial Engineering, University of Qom, Qom, Iran

Received 12 May 2015; Accepted 8 August 2015
**Corresponding author: saghi_zarghami@yahoo.com (S. Zarghamifar).*

Abstract
Opinion mining deals with an analysis of user reviews for extracting their opinions, sentiments and demands in a specific area, which plays an important role in making major decisions in such areas. In general, opinion mining extracts user reviews at three levels of document, sentence and feature. Opinion mining at the feature level is taken into consideration more than the other two levels due to orientation analysis of different aspects of an area. In this paper, two methods are introduced for a feature extraction. The recommended methods consist of four main stages. First, opinion-mining lexicon for Persian is created. This lexicon is used to determine the orientation of users' reviews. Second, the preprocessing stage includes unification of writing, tokenization, creating parts-of-speech tagging and syntactic dependency parsing for documents. Third, the extraction of features uses two methods including frequency-based feature extraction and dependency grammar based feature extraction. Fourth, the features and polarities of the word reviews extracted in the previous stage are modified and the final features' polarity is determined. To assess the suggested techniques, a set of user reviews in both scopes of university and cell phone areas were collected and the results of the two methods were compared.

Keywords: *Opinion Mining, Feature Extraction, Opinion-mining Lexicon, Corpus, Parts-of-speech Tagging, Syntactic Dependency Parsing.*

1. Introduction

As the Web 2.0 and the social networks evolve, many data were published on the Internet. Such data have newly potential applications, different groups of which are sporadically detected. Generally, data contained text documents published on the web can be classified in two groups: Objective (realistic) and Subjective. Realities are real and observable commands about independent identities and the events happened around the world. However, subjective commands reflect on human emotions and observations and the people have about the outside world and its events [1].Search engines can retrieve data from realistic documents based on keywords referring to realities. Yet, to retrieve and analyze subjective documents, it seems inefficient to use them [2].

Opinion mining and sentiment analysis have drawn much attention since the last decade, while they extract users' reviews and detect their polarity inside subjective texts. Among the applications of opinion mining, we suggest to the followings:

- *Analysis of Online Customers' reviews*

Increasing number of websites, which attempt to collect visitors' reviews about a particular product or service, reveals the significance of opinion mining. It can be utilized as an offer to buy or not to buy a particular product or use special services, as well as a consultant for manufacturers, to extract customers' desirable features and provide high quality products and services [3].

- *Representation of Proper Advertisement*

Investigating the issues and reviews discussed in a blog or a forum, we can display an advertisement with higher probability to be seen. For example, if the reviews brought up in a forum about a specific product is positive, advertisements of that product would be very likely to be seen by users of the related forum. However, were the reviews negative, it might be better to display competing products in advertisements [4].

- *Investigation Of Public Opinions*

To investigate public reviews on a particular issue several sources on the Internet (including specific forums, Twitter, etc.) can be examined and collect and evaluate users' reviews about the issue in question.

As it is defined in [5], opinion mining is only to identify positive, negative and/or neutral reviews; however, any opinion word is given a weight based on the subject of text and its polarity, in sentiment analysis. The weight means the probability or number considered for positivity or negativity of a word. As an example in [6], a weight of 0.01 is assigned to word "dirty" provided the subject is hotel and the polarity is negative; while the same word with the same subject and positive polarity gains a weight of 0.00001.

The opinion words are used to express positive and/or negative sentiments. For example, the words such as "good", "beautiful" and "wonderful" induce positive feelings in human and the words like "bad", "ugly" and "terrible" are some words with negative polarity. Polarity of any means feelings and estimation brought into the mind by such a word. It should be noted that most of the opinion words are adjectives and adverbs; however, some nouns including "junk" and "hell" and verbs such as "hate" and "love" also carry sentiment information and thus need to be considered.

In opinion mining lexicon, any opinion word is mentioned along with its polarity. It might be weighted or non-weighted.

Though a variety of methods have been introduced to establish opinion mining lexicon and several opinion mining lexicons have been created which are available to the public, it seems very unlikely to develop an opinion mining lexicon that contain all opinion words and include all areas and languages. A word can have a positive polarity in an area and a negative or neutral polarity in another area. For example, the word "unpredictable" has a negative polarity in the field of electronic instruments, but it has a positive polarity in the field of movie.

In this paper, two methods are introduced for extracting the features. The first one extracts nouns with the highest frequency as features only by using parts-of-speech tagging. Then, it will extract all other features, making use of the extracted features and the opinion words that described them. The second method deals with extracting the features and expanding opinion mining lexicon, using parts-of-speech tagging, syntactic dependency parsing and a number of Persian grammar rules.

The proposed method consists of four main steps. First, two lexicons are established for two suggested methods in order to extract the features. Second, the preprocessing stage includes unification of writing, tokenization, creating parts-of-speech tagging and syntactic dependency parsing. Third, extracted features use two proposed methods and fourth, the features and opinion words gained in the previous step are modified. Finally, the polarity of the features are determines.

2. Review of literature

The first opinion mining lexicon was established in 1997, using syntactic structure [7].

In 2002, Pang and Lee classified the texts into two neutral and polarized groups, making use of machine learning algorithms. They used three algorithms including Support Vector Machines (SVM), Naïve Bayes and Maximum Entropy Model [2].

In 2003, an opinion mining lexicon was established based on dependency criteria, which include two main stages. At the first stage, syntactic phrases including adjectives or adverbs are extracted from different sentences according to syntactic category label of phrases. At the second stage, the polarity of each extracted phrase is determined [8].

In the same year, Riloff et al devised a method to extract subjective sentences using Bootstrapping method in which the sentences are firstly categorized into two classes (sentences related to user's opinion and all other sentences) from a lexicon and an unlabeled set of data by using two classifiers. Then, some patterns are extracted from such sentences that will be returned to the classifier in the form of an iterative algorithm [9].

Yi et al extracted the features of users' reviews, using hybrid model presented in [10-11]. Their method was based on parts-of-speech and feature tagging using training set. They merely considered accuracy evaluation criteria.

Liu and Hu (2004) extracted the features by identifying and frequency nouns in the collection of documents [12]. They used parts-of-speech tagging to identify nouns.

In 2005, OPINE method was introduced including four steps of features identification, identifying the reviews related to each feature, determination of reviews' polarity and the final ranking [13]. In this method, Pointwise Mutual Information (PMI) calculation was used to identify the words.

Mei et al (2007) proceeded to extract features by creating a pattern in a specific area using Hidden Markov Model (HMM) [14].
Many other methods were presented based on pattern creation in order to extract the features in a specific area [15-16-17].
In 2008, Titove and McDonald extracted the features using Dirichlet Allocation Method and finally ranked each feature considering the user's opinion on the feature in question [18]. In this study, features are divided into two groups including fine-grained and coarse-grained.
Liu et al extracted the features and extended the lexicon by making use of syntactic dependency parsing and Persian rules. Their method was using only a basic lexicon containing a limited number of opinion words [19].
In 2012, Shams introduced an unsupervised method to determine polarity of Persian documents in which each word is weighted using two PLSASA and LDASA algorithms based on the subject in which it lays [6].

The review of literature shows that many methods suggested to extract features on a specific area require training data specific to such an area. Since there are now no training sets for this purpose in different areas of Persian, we apply two methods not dependent on a specific area, which use merely parts-of-speech tagging and syntactic dependency parsing to extract features [12-19].

3. The Proposed method

The method suggested in this study is at the feature level and includes four main stages including creation of lexicon, pre-processing, feature extraction and post-processing. Each of these stages will be explained in details later in this paper.
Overview of the proposed method is shown in the figure 1 below:

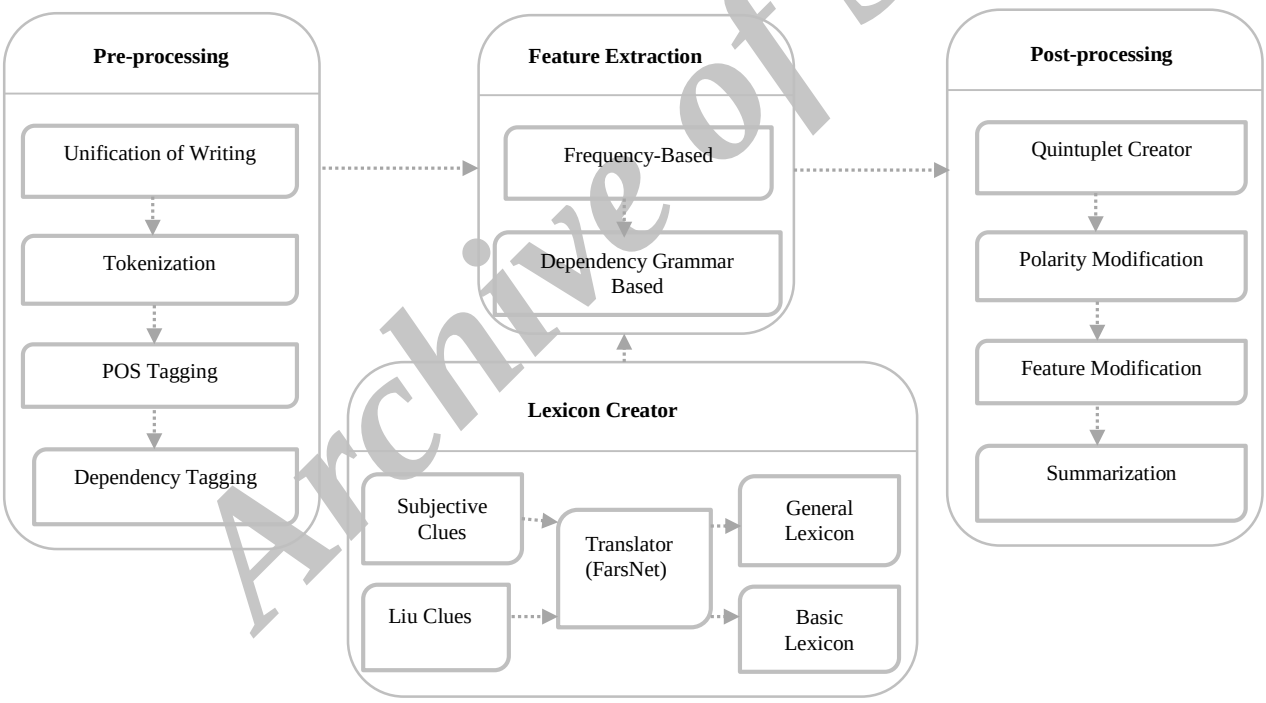

Figure 1. Overview of the suggested method.

3.1. Creation of opinion mining lexicon

The first step in opinion mining is creation of lexicon. Considering the works done on all languages (other than English) indicate that the method used in most languages for creation of lexicon is to translate an existing lexicon into the target language and then make any modification to it. The same method is used in this investigation to create two lexicons. The first lexicon is a comprehensive one for using in frequency-based feature extraction. The other one including much smaller number of positive and negative words than the first lexicon is created for using in a dependency grammar based feature extraction. The words contained in this lexicon such as "good", "beautiful", "bad" and "ugly" could be approximately seen in all areas to express emotions. The presented algorithm develops this lexicon. Subjective Clues [20] lexicon as one of the best-known and the most

important opinion-mining lexicons in English were translated into Persian. For this translation, FarsNet tools as a free lexicon was used [21-22]. In this method, Persian equivalent of the word in question is found and then is added to Persian lexicons along with its synonyms. To determine the polarity of the words translated in the lexicon, all words inherit their polarity from their English equivalents. It means that if a word has a positive label in English, all its equivalents will also gain positive label after translation. Inheritance of polarity is made, because concepts are usually independent from languages [6] and a word suggesting a positive concept in a language has almost its positive concept after translation into another language. The lexicon created in [12-23] was used to create a basic lexicon. We have used FarsNet for translation at this stage. To create a basic lexicon, Persian equivalents of the word are used and synonyms are ignored. This is because we aimed to create a primary lexicon which will be expanded in later stages by analyzing the documents. Due to the lack of polarity in English lexicon, polarity of the words after translation is determined manually. Furthermore, a group of words, which are not seemingly common in Persian texts were deleted manually.

Simple translation of a lexicon has some problems. In order to modify the Persian lexicon, all the words were checked and the ones labeled incorrectly were modified manually.

Finally, 6746 words were created in the comprehensively generated lexicon where 3866 words had negative polarity, 273 words had neutral polarity and the remaining with positive polarity.

In the basic lexicon production, the total number of words is 575 in which 288 ones had negative polarity, 36 ones had neutral words, and the remaining were with positive polarity.

3.2. Pre-processing

This stage includes several steps to create data required by feature extraction algorithm at the next stage. These steps consist of unification of writing, tokenization, parts-of speech tagging and syntactic dependency parsing.

3.2.1. Unification of writing

There are some letters in Persian, which are written by different methods in a variety of character encoding standards. As an example, each of the letters "ی" and "ک" are found in different forms in Persian texts. In the unification of writing stage, such letters are uniformed.

3.2.2. Tokenization

Each document is segmented into its constituent words. In order to determine the words, each review is firstly segmented based on punctuation marks («!» «.» «،»«.»،«:»«؟»), and then, the resulting sentences are divided into their constituent words.

3.2.3. Parts-of speech tagging

A Part-Of-Speech Tagging (POS) assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. There are two main steps to create parts-of-speech system based on data. At the first step, labeling pattern is resulted by making use of a training set. At the next step, an appropriate label would be provided for any input word based on the pattern resulted in the previous step.

Parts-of-speech tagging on the reviews segmented into words in the previous step is labeled by TNT tagger software and Bijankhan corpus at this stage **[**24**]**.

3.2.4. Syntactic dependency parsing

A syntactic dependency will be semantically defined as a binary operation that takes as arguments the denotations of the two related words (both the head and the dependent), and gives as a result for a more elaborate arrangement of their denotations [25]. Generally, for any input sentence in dependency parsing, one graph is constructed and there are two general approaches including data and grammar.

In supervised learning method, there are two main steps for constructing a dependency parsing system. At the first step, dependency grammar is gained by using a training set. As the dependency grammar is achieved, parsing pattern will be gained. At the next step, a dependency graph will be constructed for each input sentence based on the pattern resulted in the previous step.

In this step, syntactic dependency tagging is done on the reviews using MST Parser software and Dadegan Persian dependency framework [25-26].

3.3. Feature extraction

Creating lexicon and preparing documents, we will introduce the third stage. Positive polarity of any document does not mean user's positive opinion about all features of such a document. This status is also true about the negative documents. The comments expressed by the users are a collection of positive and negative reviews on different aspects of an issue. At this stage as the most important suggested method, features (aspects) of an object commented by the users are

extracted. Two methods are suggested to extract such features including frequency-based feature extraction and dependency grammar based feature extraction, which will be introduced later in the next section.

3.3.1. Frequency-based feature extraction

In this method, a set of nouns and noun phrases is gained per document. For this purpose, the words with part-of-speech tag of "N" are known as noun and the set of nouns with part-of-speech tag of "N N" are considered as noun phrases and will be added to set of nouns in such a document. As an example, in the sentence "university environment was extremely good", the phrase, "university environment" as a noun phrase and "environment" and "university" each as a noun are selected and added to set-of-words. At the next step, we determine the number of each of the nouns (bag-of-word) gained at the previous step among total current lists. To do this, a new set including all words extracted at the previous stage is constructed and then, the frequency of each word is specified. At the next step, nouns with a frequency higher than a threshold are extracted as important features. Frequency threshold can be any number, which is usually determined by experience.

At the final step, we will use the following idea to extract features with a frequency lower than defined frequency threshold. The opinion words can be utilized to describe different features. For example, noun phrase "university environment" in the previous example is selected as a feature and tagged in the documents; considering the sentence in the previous example, a commenter has used the word "good" to describe this feature. Now, we can search the word "good" in entire documents and then extract the noun found before it as a feature. As a result, in a sentence like "university staff were very good". "University staff" is extracted as a feature. Opinion words are found using the general lexicon constructed at the first stage.

3.3.2. Dependency grammar based feature extraction

This is a bootstrapping method which starts to work merely by a basic lexicon. However, the extracted features are used in the next round to extract other features and expend the lexicon.

This method is based on rules naturally existing in language dependency relationships. As an example, in sentence "this phone has a good appearance", if we know the word "good" as an opinion word, we could extract the word "appearance" as a feature through dependency grammar.

In table 1, the rules applied in this method for extracting features and expanding opinion words are shown:

Table 1. Rules of dependency grammar.

Rule	Relations and constraint	Output
1	(*OW* Dep *POS(ADJ)*) or (*OW* Dep *POS(ADV)POS(ADJ)*) Depϵ {*CONJ*}	*ADJ* is new *Opinion word* *If(CONJ* ϵ *Contrary words)* Polarity(*ADJ*) = - Polarity(*OW)* *Else* Polarity(*ADJ*) = Polarity(*OW)*
2	(*SBJ(POS(N))* Dep *OW*)	*SBJ(N)* is new *Feature*
3	(*F* Dep *MOS(POS(ADJ))*)	*MOS(ADJ)* is new *Opinion word*
4	(*F* Dep *POS(N)*) or *(F MOZ(POS(N)))* Depϵ {*CONJ*}	*N* is new Feature
5	(*F NPOSTMOD(POS(ADJ))* Dep *MOS(POS(ADJ)))*	*F+ NPOSTMOD* is new Feature

Description:

OW = opinion word	*F= Feature*	*MOS =Mosnad*	*SJB=Subject*
ADJ =Adjective	*CONJ= Conjunction*	*NPOSTMOD=Subsequent Adjective*	*MOZ= Mozaf*

Rule1: if a word is an opinion word and is followed by a conjunction and an adjective and/ or the conjunction is followed by an adverb and an adjective, respectively, the word tagged as an adjective will be selected as an opinion word. This rule is considered as the inverse. It means that if a word contains an opinion word with a conjunction and an adjective before it, the word tagged as an adjective will be an opinion word. To identify the polarity of a new word, if conjunction is in

contrary word group, polarity of the new word is opposite to that by which we extracted the new word. If the conjunction is not in this group, both words have the same polarity.

For example, suppose that the word "beautiful" is tagged as OW with positive polarity. In this case, taking this rule into consideration, we can extract the word "attractive" in a sentence like "it has a beautiful and attractive appearance" as a new opinion word with the positive polarity. Furthermore, the word "fragile" in the sentence "it has a beautiful, but fragile frame" is an opinion word with negative polarity.

3.3.3. Contrary word

The contrary words include former and anterior sentences/words with different polarities. For example, the polarity of the sentence after "but" is opposite to the polarity of the sentence followed by "but". We translated and modified a list of contrary words shown in [19] and used it to identify such words in an opinion-mining system.

Rule 2: If a word is an opinion word (OW) with a POS tag of MOS (Mosnad is a property of a noun, an adjective or a pronoun ascribed to the subject of a sentence whose main verb is a linking verb. The relation between the verb and Mosnad is MOS), we will consider a word tagged subject (SJB) in the sentence as a new feature. As an example, in the sentence "my university is very beautiful", the word "university" is a new feature.

Rule 3: this rule is exactly the opposite of the previous rule; that is if a word in the sentence is a feature, then the word with MOS role in the sentence which is also an adjective, will be extracted as an opinion word.

Rule 4: if a word is a feature (F) followed by a conjunction and a noun and/or a noun with a MOZ (Ezafe dependents in Persian are nouns or pronouns which follow a head noun and signify a possessed-possessor, first name-last name, etc. relation with the head noun. The relation between a noun and its Ezafe dependent is MOZ) role, the noun is selected as the feature. This rule is considered as the inverse. It means that if a word is a feature with a conjunction and a noun before it, the related noun will be a new feature. For example, in the sentence "it has a library and a small buffet". If the word "library" is tagged as a feature, the word "buffet" is also extracted as a new feature.

Rule5: according to observations, if the feature and opinion word are specified in a sentence and the feature is followed by a word with adjective (POS tagging) and NPOSTMOD (Adjectives in their positive and comparative forms together with post-noun numerals are considered post-modifiers of noun.

The relation between a noun and its post-modifiers is NPOSTMOD) roles which have not been separated from an opinion word through conjunction, the feature and the adjective after it can be considered as a noun phrase and a new feature. The structure of the noun phrase is shown in [27]. According to studies, most of noun phrases used in comments include only noun and a subsequent adjective. As an instant, in the sentence "امکانات رفاهی آنجا افتضاح بود", the words "امکانات رفاهی" can be considered as a new feature, provided that "امکانات" is a feature, and "افتضاح" is an opinion word.

The algorithm suggested is shown in figure 2. Inputs of this algorithm are the basic lexicon and a set of users' reviews. At the first step, words of lexicon will be searched in all documents and any words found in any document are tagged newly as an opinion word.

At the next step, the rules are applied sequentially and the words extracted in the document are tagged as a feature and/or opinion word, considering the type of such words and are also added to the related set-of-words. In the next round, we will search the newly extracted words (features and opinion words) until the time when no new word is found.

3.4. Post-processing

At this stage, we proceed to modify the features and the polarity of opinion words extracted at the previous stage. The step of opinion words' polarity correction is used only for dependency grammar based feature extraction, which has also expanded the lexicon while implementing the algorithm.

3.4.1. Establishment of quintuples and set of opinion words

In this section, a record is created per any feature in each review, which includes five characteristics including feature, polarity, date, writer and type. Furthermore, a set of opinion words describing the feature is created for each record. To clarify this issue, the record created for the sentence "speaker on the camera is too weak and unqualified to play video sounds" is shown in figure 3.

The polarity of the feature is determined by adding the polarities of opinion words describing the feature and considering the negative-makers' roles in the sentence.

```
Input: Opinion word Dictionary {O}, Review Data R
Output: All possible Features {F}, The Expanded Opinion Lexicon {O-Expanded}
Function:
1. {O-Expanded} = {O}
2. {F} =∅, {O} =∅, {TempF} =∅
3.  for each {O}
4.          For each parsed sentence in R
5.                  Label Opinion words based on Opinion words in {O}
6.          Endfor
7.          Remove {O}
8. Endfor
9. for each {TempF}
10.         For each parsed sentence in R
11.                 Label Features based on Features in {TempF}
12.         Endfor
13.         Remove {TempF}
14. Endfor
15. For each parsed sentence in R
16.         Extract Opinion word {O'} using rule1 and add to {O} and {O-Expanded}
17. Endfor
18. For each parsed sentence in R
19.         Extract Feature {F'} using rule2 and add to {F} and {TempF}
20. Endfor
21. For each parsed sentence in R
22.         Extract Opinion word {O'} using rule3 and add to add to {O} and {O-Expanded}
23. Endfor
24. For each parsed sentence in R
25.         Extract Feature {F'} using rule4 and add to {F} and {TempF}
26. Endfor
27. For each parsed sentence in R
28.         Extract Feature {F'} using rule5 and add to {F} and {TempF}
29. Endfor
30. Repeat 3 till size ({TempF}) =0, size ({O}) =0
```

Figure 2. Dependency grammar based feature extraction.

Type	Author	Date	Polarity	Feature
S	93	2012	-1	speaker

Polarity	Opinion Word
-1	Weak
-1	Unqualified

Figure 3. A Sample of created quintuples.

Negative-makers in Persian. The negative-maker means a word or words, which reverse the polarity of a sentence. Since the opinion mining aims to determine positivity or negativity of an opinion, it will be highly important to study negative-makers' roles in this field. In Persian, most negative-makers appeared in the verb of a sentence. To identify negative-makers of Persian verbs, we will follow the method explained in **[**6**]**. In this method, we used Bijankhan corpus to identify verbs. In this method, all negative verbs were first tagged manually. Then, for further expansion and coverage, all verbs with negative-maker suffixes such as "ن", "نمی" and all forms of "نخواه" were expanded. Some of the resulted words will be definitely meaningless and inapplicable. For example, a word like "ناست" (ن+است) created by this method is not a correct word, but no problem will arise because such words are not included in data sets. Negative verbs are used to reverse polarity and examine the role of negative-makers.

3.4.2. Modification of extracted opinion words' polarity

Some opinion words extracted by the second method do not assigned any polarity. The reason is that such words were extracted using the features. The features have themselves no polarity and the opinion words describe them and determine their polarity. To solve this problem, we will work according to the following method:

We will firstly determine quintuplet of opinion word. Then, we observe polarity of features before and after such a record. If the polarity of

both records is positive and/or negative, we will allocate it to an opinion word and modify also the polarity of features in quintuplet. Observation shows that if two sentences in comments are positive, the sentence between them is often positive and vice versa. For example, in the following text: "Scientific knowledge of the university was so low. Class times were organized carelessly. Also, class features were so low" if we do not set the polarity for the second sentence, because of the polarity similarity between first and third sentence, second sentence polarity will be set to negative; equal to the first and third sentence. If this rule is not applied, we will determine polarity of the entire opinion in the review and will regard the polarity of the intended word as a document and accordingly, we will modify its feature polarity.

3.4.3. Modification of feature

In [28], a list of common unintelligible words in Persian documents is provided. For example, the word "viewpoint"("لحاظ") is tagged as noun and subsequently as a feature in many reviews. To resolve this problem, prepared list is used to modify the feature. However, a group of these words do not have noun roles and are not applicable in this section. Then, using [29], synonym features are specified and the features with the highest frequency in documents are considered as the main feature and the main feature replaced their synonyms. Also, we provided a list of names of universities, models and cell phone manufacturers and used them to modify the extracted features, because such names are tagged as features according to observations. For this purpose, each feature containing one of these words is corrected and the related word is deleted from this feature.

3.4.4. Feature-based summarization

After constructing quintuplets and correcting them, a summarization of reviews can be provided and represented for each feature as resulted by opining mining. In **[**23**]**, visual tools like a bar graph are used to show summarizations. This study has such a capability and is possible to prepare reports using quintuplets.

4. Analysis of results

In this section, we will evaluate the method proposed for the opinion mining. Then, we compare the results of the dependency grammar based method with the results published in Liu's article [19] using the above method in English sentences. The first step in evaluation of each system is to select a data set on which the system's performance is evaluated. This is the reason we start this section by introducing the established data set and will explain the results obtained. To evaluate this method, we will make use of the most important basic approaches in the opinion mining.

4.1. Opinion mining data set

Lack of adequate data in natural language processing areas and its subset is one of the current problems in this group of activities. For this reason, two data sets were prepared from users' reviews in the fields of university and cell phone. Users' reviews on the cell phone collected and classified form the site http://www.digikala.com. The reviews about the university were obtained by a group of academic people filling in the form designed for this purpose. In the field of university, 90 reviews were totally selected, 45 ones out of which were negative and the other 45 reviews were positive. In the field of cell phone, 250 reviews were selected from the mentioned site including 125 positive reviews and 125 negative ones. All collected documents were reviewed and spell-checked by Virastyar software. The most important reason why we used this software was to delete the spaces and putting virtual space between the words. In any document, features of each document, opinion words with their polarity, sentence polarity and document polarity were tagged manually for final evaluation of the suggested methods.

Table 2 shows the information related to data sets.

Table 2. Data sets.

Data Set	Number of reviews	Number of sentences
University area	90	598
Cellphone area	250	1409

4.2. Evaluation of the proposed method

To evaluate the proposed method for extraction of features and opinion words in this study, three measures including Precision, Recall and F-measure were used. Accuracy measures were used to evaluate the polarity assigned to dependency grammar based feature extraction method in this study.

4.2.1. Evaluation of extracted features

The results show that though the frequency-based extraction algorithm has a higher precision, but recall and f-measure evaluation in the dependency

grammar-based extraction method has been considerably improved.

Table 3. Precision for extracted features.

	Precision	
Data set	**Frequency based**	**dependency grammar**
University area	0.92	0.88
Cellphone area	0.94	0.92

Table 4. Recall for extracted features.

	Recall	
Data set	**Frequency based**	**dependency grammar**
University area	0.64	0.83
Cellphone area	0.72	0.86

Table 5. F-measure for extracted features.

	F-Measure	
Data set	**Frequency based**	**dependency grammar**
University area	0.75	0.85
Cellphone area	0.81	0.89

4.2.2. Evaluation of extracted opinion words

The results of opinion words extraction evaluation using dependency grammar based feature extraction in two areas including university and cell phone is shown below. Due to the use of general lexicon in frequency-based extraction algorithm which was not expended in this method, we are going to evaluate merely the dependency grammar based feature extraction.

Table 6. Evaluation for extracted opinion words.

Data set	**dependency grammar**		
	Precision	**Recall**	**F-Measure**
University area	0.83	0.79	0.81
Cellphone area	0.88	0.82	0.85

4.2.3. Evaluation of extracted opinion words' polarity

As mentioned earlier, the polarity of a word is sometimes dependent on the area where it is used. The Accuracy related to the opinion words' polarity extracted by dependency grammar based feature extraction is evaluated in accordance with the table 7.

Table 7. Accuracy for extracted opinion words' polarity.

	Accuracy
Data Set	**dependency grammar**
University area	0.73
Cellphone area	0.66

4.3. Comparison of dependency grammar based method in Persian and English data set

To evaluate the dependency grammar based method in English text, Liu [19] has used five different data sets.

Table 8. Liu data set.

Data Set	Number of reviews	Number of sentences
D1	45	597
D2	34	346
D3	41	546
D4	95	1716
D5	99	740

Results of "D1" dataset are compared with the results of the university dataset and the results of the "D4" dataset are compared with the result of cell phone dataset, since there is the same number of sentences in each dataset.

Table 9. Comparison D1 and university data set.

Data Set	Precision	Recall	F-Measure
D1	0.87	0.81	0.84
University area	0.88	0.83	0.85

Table 10. Comparison D4 and cellphone data set.

Data Set	Precision	Recall	F-Measure
D4	0.81	0.84	0.82
Cellphone area	0.92	0.86	0.89

5. Conclusions and future work

In this investigation, two methods were applied to extract features in Persian reviews which are not limited to a specific area and do not require training data set for such an extraction. In frequency-based feature extraction method, the parts-of-speech tagging and noun frequency were only used in entire documents for feature extraction. In the dependency grammar based feature extraction method, syntactic dependency parsing was merely used to extract features and expand the opinion words.

The results indicate that the dependency-grammar -based method has a better performance compared to frequency-based in extracting features. Furthermore, using this method, there will be no problem in creating a comprehensive lexicon which will cover all areas, because this method starts its performance by making use of a basic lexicon with limited number of words and will expand it later using users' reviews.

The results suggest that the dependency grammar based method does not work properly in determining polarity of newly extracted opinion words.

As the last discussion in this paper, the works which could be conducted in the future to improve and expand the proposed method are suggested as follows:

- Identification of co-reference resolution in Persian texts
- Not ignoring the sentences which contain opinion words implicitly
- Identification of a feature indicator: many opinion words can be used for any features. For example, the words "good", "bad", etc., but, some of these words are indicators of specific features. As an example, the word "large" in the sentence "this phone is very large" is an indicator of size feature. Identifying such words in the system, further and more precise features can be extracted.

- Considering the subject in an opinion mining
- Conversion of colloquial writing to formal writing
- Analysis of sentiments expressed using weighting algorithms

References

[1] Stavrianou, A. & Chauchat, H. (2008). Opinion Mining Issues and Agreement identification in Forum Texts. Atelier FOuille des Données d'OPinions (FODOP 08), France, 2008.

[2] Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. The ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86.

[3] Sepehri, M. (2009). Chi-Square for features selection in Opinion mining in Persian text. 2nd National Conference on Computer/Electrical and IT Engineering (CEIC'09), Hamedan, Iran, 2009.

[4] Nicholls, C. & Song, F. (2010). Comparison of Feature Selection Methods for Sentiment Analysis. Advances in Artificial Intelligence, vol. 6085, pp. 286-289.

[5] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan and Claypool. Atlanta, USA.

[6] Shams, M. (2012). Opinion mining and Sentiment Analysis in Persian Documents. University of Tehran, Iran.

[7] Hatzivassiloglou, V. & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. The eighth conference on European chapter of the Association for Computational Linguistics, Madrid, Spain, pp. 174-181.

[8] Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from associaCon. ACM TransacCons on InformaCon Systems (TOIS), vol. 21, no. 4, pp. 315-346.

[9] Riloff, E., Wiebe, J. & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. The seventh conference on Natural language (CONLL), Edmonton, Canada, vol. 4, pp. 25–32.

[10] Zhai, C. & Lafferty, J. (2001). Model-based Feedback in the Language Modeling Approach to Information Retrieval. The tenth International Conference on Information and Knowledge Management. Berlin, Heidelberg, Springer-Verlag. pp. 403-410.

[11] Yi, J., Nasukawa, T., Bunescu, R. & Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. The Third IEEE International Conference on data mining, pp. 427 – 434.

[12] Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA, pp. 168-177.

[13] Popescu, A. & Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. The conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, pp. 339-346.

[14] Mei, Q., Ling, X., Wondra, M., Su, H. & Zhai, C. (2007). Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. The 16th international conference on World Wide Web, pp. 171-180.

[15] Liu, Y., Huang, X., An, A. & Yu, X. (2007). ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. The 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 607-614.

[16] McDonald, R., Hannan, K., Neylon, T., Wells, M. & Reynar, J. (2007). Structured Models for Fine-to-Coarse Sentiment Analysis. The 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic. pp. 432–439.

[17] Su, Q., et al. (2008). Hidden Sentiment Association in Chinese Web Opinion Mining. The 17th international conference on World Wide Web. Beijing, China. pp. 959-968.

[18] Titov, I. & McDonald, R. (2008). A Joint Model of Text and Aspect Ratings for Sentiment Summarization. The Association for Computational Linguistics, pp. 308-316.

[19] Qiu, G., Liu, B., Bu, J. & Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics. vol.37, no.1, pp. 9-27.

[20] Wiebe, J. M. (2000). Learning subjective adjectives from corpora. The Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on InnovaCve ApplicaCons of ArCficial Intelligence, 2000, pp. 735–740.

[21] Shamsfard, M., et al. (2010). Semi-Automatic Development of FarsNet; the Persian WordNet. 5th Global WordNet Conference, Mumbai.

[22] The FarsNet website (2013), Available: http://nlp.sbu.ac.ir/site/farsnet/.

[23] Liu, B., Hu, M. & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. The 14th International World Wide Web conference. Chiba, Japan. pp. 342-351.

[24] Raja, F., et al. (2007). Evaluation of part of speech tagging on Persian text. The Second Workshop on Computational approaches to Arabic Script-based Languages, Linguistic Institute Stanford University, Stanford, California, USA, pp. 120-127.

[25] Dadegan Research Group. (2012). Persian Dependency Treebank Version 1.0, Annotation Manual and User Guide. Supreme Council of Information and Communication Technology (SCICT).

[26] Rasooli, M., Kouhestani, M. & Moloodi, M. (2013). Development of a Persian Syntactic Dependency Treebank. The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT). Atlanta, USA.

[27] Kavoosi Nejad, S. (2000). Delete the noun group in Persian language. Academy of Letters. vol.16, pp. 109-127.

[28] Sanji, M. & Davar-Panah, M. R. (2010). Identification of non-sense words (common) Automatic indexing of documents in Persian. Journal of Library and Information Science, vol.12, no.48, pp. 23-35.

[29] Khoda-Parasti, F. (1997). Comprehensive dictionary of synonyms and antonyms in Farsi. Fars Encyclopedia.

[30] Tasharofi, S., Raja, F. & Oroumchian, F. (2007). Evaluation of Statistical Part of Speech Tagging of Persian Text. International Symposium on Signal Processing and its Applications. Sharjah, United Arab Emirates.

[31] Elahi-Manesh, M. H. & Minaee, B. (2011). The hidden Markov model part of speech labeling in Persian texts. Journal of information science, computer science education and Islamic Studies, vol.34, pp.102-106.

نشریه هوش مصنوعی و داده کاوی

دانشگاه شاهرود

استخراج ویژگی‌ها در اندیشه کاوی مورد استفاده در متون فارسی

عفت گلپر رابوکی[1]، ساقی‌السادات ضرغامی‌فر[2,*] و جلال رضایی نور[3]

[1] دانشکده ریاضی، دانشگاه قم، قم، ایران.

[2] دنشکده مهندسی کامپیوتر، دانشگاه قم، قم، ایران.

[3] دانشکده مهندسی صنایع، دانشگاه قم، قم، ایران.

ارسال ۲۰۱۵/۰۵/۱۲؛ پذیرش ۲۰۱۵/۰۸/۰۸

چکیده:

اندیشه کاوی به تحلیل اظهار نظرات کاربران جهت استخراج نظرات، احساسات و خواسته‌های کاربران در یک حوزه خاص می‌پردازد. دانستن نظرات افراد در یک حوزه خاص می‌تواند نقش مهمی در تصمیم‌گیری‌های کلان آن حوزه ایفا کند. به طور کلی اندیشه کاوی در سه سطح سند، جمله و ویژگی به استخراج نظرات کاربران می‌پردازد. اندیشه کاوی در سطح ویژگی به دلیل تحلیل جهت‌گیری جنبه‌های مختلف یک حوزه از دو سطح دیگر بیشتر مورد توجه قرار دارد. در این مقاله، دو روش به منظور استخراج ویژگی‌ها ارائه شده است. روش پیشنهادی شامل چهار گام اصلی است. در گام نخست لغت‌نامه اندیشه‌کاوی برای زبان فارسی ایجاد می‌شود. این لغت‌نامه به منظور تعیین جهت‌گیری نظرات کاربران مورد استفاده قرار می‌گیرد. گام دوم مرحله پیش‌پردازش شامل یکسان سازی نگارشی، تقطیع، ایجاد برچسب‌های ادات سخن و برچسب وابستگی نحوی اسناد است. گام سوم استخراج ویژگی‌ها با استفاده از دو روش استخراج ویژگی بر مبنای تکرار و استخراج ویژگی بر اساس قوانین وابستگی است و در گام چهارم ویژگی‌ها و قطبیت کلمات حاوی نظر استخراج شده در مرحله قبلی اصلاح شده و در نهایت قطبیت ویژگی‌ها تعیین می‌گردد. برای ارزیابی روش‌های پیشنهادی، مجموعه عقاید کاربران در دو حوزه دانشگاه و تلفن همراه جمع‌آوری شده و نتایج حاصل از دو روش با یکدیگر مقایسه می‌شوند.

کلمات کلیدی: اندیشه کاوی، استخراج ویژگی، لغت‌نامه اندیشه کاوی، برچسب ادات سخن، برچسب وابستگی نحوی.