A Supervised Method for Constructing Sentiment Lexicon in Persian Language

Faranak Ebrahimi Rashed [a], Neda Abdolvand [b,*]

[a] *Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
[b] *Department of Social Sciences and Economics, Alzahra University, Tehran, Iran*

Received 1 December 2015; revised 20 August 2016; accepted 10 November 2016; available online 16 March 2017

Abstract

Due to the increasing growth of digital content on the internet and social media, sentiment analysis problem is one of the emerging fields. This problem deals with information extraction and knowledge discovery from textual data using natural language processing has attracted the attention of many researchers. Construction of sentiment lexicon as a valuable language resource is a one of the important fields of study in this domain. The main researches in the area of sentiment analysis have focused on English language and few works considered the sentiment analysis in Persian language due to the lack of resources. This paper aims to introduce a supervised method for creating a sentiment dictionary in Persian language with extracting linguistic features in reviews and statistical mutual information to determine the sentiment orientation and sentistrength of words. To evaluate the proposed method, a set of existing reviews in the online retail site is used in various domains and the present dictionary is compared with Sentiwordnet. The results show the proposed method achieves an accuracy of 80% in determining the orientation of sentiment word.

Keywords: *Sentiment Analysis; Semantic Orientation; Point Wise Mutual Information; Sentiment Dictionary.*

1. Introduction

With the rapid development of information technology, user generated content can be easily produced and posted online. High volume and exponential growth of this information provides the potential value for governments, businesses and even users themselves. These reviews, on the one hand, are used to adjust business strategies of the e-commerce websites, and on the other hand it works as a guide for customers who want to buy products, so the development of such tools that automatically extract the opinions from online reviews has been taken much attention in recent years [1-3]. Opinion mining is a process of extracting opinions in textual data, that also in some studies is known as sentiment analysis, sentiment mining, subjectivity analysis and review mining, mainly aims to determine the emotional attitudes (positive or negative whether subjective or objective) as expressed in the reviews by combining text mining and natural language processing techniques[3-5].

Sentiment words and phrases are the basic linguistic units to express feelings so constructing sentiment lexicon plays an important role in

* Corresponding author. Email: abdolvand@gmail.com

recognizing the sentiment polarity of reviews [6]. However the construction of such terms manually is time-consuming and difficult. In recent years, researchers have suggested two approaches to deal with this issue: dictionary based approach and corpus based approach [6-9]. Dictionary approach [10-13] utilizes semantic relations, synonyms, concepts and glosses in dictionary to determine the sentiment orientation of words. Corpus approach uses the statistical co-occurrence information in large collection of documents and is based on the assumption that the sentiment words that have same polarity occur together in corpus [8, 14-17]. Dictionary method determines the orientation of words independently of the textual content and it has some limitations: it does not have good performance to find the domain and context dependent orientation of sentiment words. It cannot be applied to languages which lack thesaurus knowledge [7, 15] also multi-word expressions and slangs that indicate the author’s opinion but, do not exist in dictionary [18, 19]. The corpus based approach can solve these problems and it has a better performance in determining the semantic orientation of sentiment words in the domain; however it is not able to create the rich and full list of sentiment words [7, 20].

The only available study that creates Persian sentiwordnet presents in [21], which is based on dictionary approach.

This paper presents a supervised and automated method for creating a sentiment dictionary in Persian language using a corpus-based approach. The proposed approach consists of five steps: preprocessing Persian reviews and extracting linguistic features, part of speech tagging (determining the role of words in sentences), extracting subjectivity words, determining the semantic orientation of words (word’s polarity) and calculating score for sentiment words.

The contribution of this work is creating a sentiment lexicon from Persian reviews that can cover the morphological characteristics of Persian language and considered score for each word in sentiment dictionary.

The rest of this paper is organized as follows: Section 2 surveys the related works on building the sentiment dictionaries, Section 3 describes each step in our proposed approach in details, Section 4 presents and discuses the empirical experiments, and finally, Section 5 concludes this paper and recommend for future works.

2. Literature Review

So far, the conducted studies have presented several methods to build the sentiment lexicon. Most of them use a list of seed words and word similarities to construct the sentiment dictionary. According to the manner of obtaining similarities, these studies can be mainly classified into two categories: dictionary based approach and corpus based approach. Table 1 shows basic properties, advantages and disadvantages of each method.

2.1. Dictionary Based Approach

Dictionary based approach utilizes synonyms and semantic relations in order to determine the positive and negative polarity of words. These methods produce sentiment lexicon using a dictionary like wordnet [9]. In these methods, at first a set of seed words with known positive and negative orientation is collected manually and then using bootstrapping algorithms to find their synonyms and antonyms in dictionary, the newly found words in each iteration are added to the positive and negative lists until no more new words can be found[3].

Hu and Liu [10] utilized the adjective synonyms and antonyms set in wordnet to predict the semantic orientation of adjectives. The general idea is that synonyms of adjective have the same similar polarity to it and its antonyms hold its opposite polarity. Kamps and et al. [11] presented a new strategy by using the lexical network construction and determining the value semantic relations between words. For interpretating the semantic orientation according to their study, two words have the same semantic orientation. If they have the strong relation with each other so they focus on the lexical relations existing in wordnet for calculating the distance between adjectives and the sentiment polarity of each word accordingly is defined by its distance from the initial two adjectives “good” and “bad”. Kim and Hovy [22] determined the sentiment polarity of words using probabilistic methods, synonyms, and antonyms in wordnet. Andreevskaia and Bergler [13] proposed a method to find the polarity of words that relies on a fuzzy algorithm, synonyms, antonyms and glosses in wordnet. As well, the sentiment strength of each word was determined. Esuli [23] proposed a semi-supervised approach to words classification in two categories: positive and negative. In this approach, he used glosses and expressions, which are found in the dictionary or wordnet to recognize the orientation of the terms. Its basic assumption is that if a word is semantically oriented in one direction, then the words in its glosses tend to be oriented in the same direction. He also built sentiwordnet dictionary using wordnet and a supervised method based on the relationship between words [12]. At first three classes of positive, negative and neutral are defined, then the sentiment polarity of each word is found with calculating its placement in three classes.

2.2. Corpus Based Approach

These approaches are based on this assumption that polar terms conveying the same polarities, co-occurred with each other in domain corpus and use the statistical measurements to calculate the sentiment orientation of words.

The first study in this area is Hatzivassiloglou and Mckeown’s study [24]. They used English features to detect adjective polarity and constructed a lexical graph using conjoined words “and” and “but” then they finally used the clustering method to create a list of positive and negative adjectives. Turney [17] determined the orientation of sentiment words and phrases with pointwise mutual information. In this study adjective and adverb phrases were first extracted as candidate sentiment terms using part of speech tagging rules from sentences then polarity of each terms were determined based on co-occurrence with two paradigm words (“excellent with positive polarity” and “poor with negative polarity”). Turney and Litman [14] proposed weakly supervised method to detect semantic orientation of words using pointwise mutual information and latent semantic association. They used 14 seed words with known polarity in their study. Gamon [25] expanded the Turney method and added a new assumption, namely that sentiment terms of opposite orientation tend not to co-occur at the same sentence. Ding and Liu [15] investigated the multi-sense words and features. They used several rules to recognize the polarity of words. First, they separated sentences and specified the product features, then they marked the sentiment words around features and give score to words, and then used conjoined words like “but” and “and” to determine the polarity of ambiguous words. Du and et al. [26] created sentiment dictionary for different domain using association measures of words in the same domain and in different domains. Tan and et al. [8] described an extraction to Du’s method. He considered the association of words with documents and determined the orientation of words in different domains.

Due to the lack of required resources such as a comprehensive dictionary of sentiment words and datasets with positive and negative reviews in Persian language, very few studies have been done in opinion mining and sentiment analysis in Persian.

Some studies have used the translation existence resource in English to create sentiment dictionary. Shams [27] used English subjectivity lexicon and Basiri [28] used sentistrength and translated words list in Persian. Then they used them as a Persian sentiment dictionary for sentiment analysis. The only study which has created the sentiment dictionary in Persian conducted by Alimardani [21] that has used the Persian wordnet which consists 17000 lexical entries in the category of noun, adjective, adverb and verb. Each synset contains some information such as part of speech, glossary and example of word usage. Besides, for each word an equivalent in English wordnet has been considered, So each word is mapped as equivalent to sentiwordnet to determine semantic orientation of it.

Compared with the method proposed in this paper, this method is generally dependent on the existence of strong dictionary, it also determines the context free orientation of words. On the other hand it is faced with the scalability problem and has no availability to detect the orientation of words that do not exist in dictionary [8, 15, 26, 29]. Since both positive and negative words can have different polarities in different domains [17], this paper has used the corpus based approach for calculating the semantic similarity of words and sentiment lexicon construction. Also this study determined the score for each word in sentiment dictionary using Min-max normalization.

Table. 1. Comparison method

Method	General features	Advantages and Disadvantages
Dictionary based approach	•need a small set of sentiment words with positive and negative polarity •extended set using semantic relations like synonyms, antonyms and glosses in wordnet or dictionary	**Advantages:** •Easy and quickly find huge number of sentiment words **Disadvantages:** •Unable to identify domain dependent sentiment words •Rely on a thesaurus or on lexical database such as wordnet
Corpus based approach	•Consider syntactic patterns, linguistic clues and structural clues of the words in the document •use of statistical co-occurrence information and distributional context similarity in large collection of documents	**Advantages:** •domain and context dependent •create the rich and full list of sentiment words **Disadvantages:** •Not cover all of sentiment words • need large corpus

3. Proposed Model

In this section, the proposed method for sentiment dictionary construction in Persian language is described. Fig 1 presents a general framework of approach.

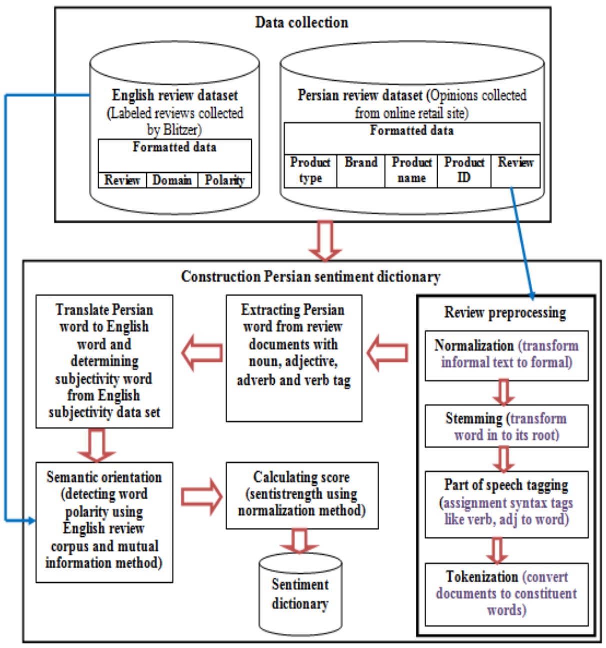

Fig. 1. Proposed framework

3.1. Text Pre-Processing

Unlike other languages including English, text mining in Persian faced with many problems due to the much complexity. These problems are due to the lack of tools, wide variety of declensional suffixes, word spacing and many informal or colloquial words. Since the creation of sentiment lexicon needs pre-processing, the words are derived from pre-processed data [30].

Text pre-processing is the process of cleaning and preparing the text for classification. In this paper, this process consists of the steps of normalization, stemming, part of speech tagging and tokenization [31, 32].

3.2. Feature Selection

Feature selection is one of the essential steps in sentiment classification. Adjectives, adverbs, verbs and nouns are features that express opinions [15]. At this step for sentiment lexicon construction, words with these label derived from existing terms.

3.3. Extraction of Subjectivity Features

This process involves the translation of Persian words in English and searches the words in the common subjectivity words list in English language [33]. The words in this list are used in the next step.

3.4. Semantic Orientation of Words

To determine the semantic orientation of words and phrases, the study uses the similarity measure theory based on point wise mutual information. PMI is a very simple method in the field of information theory [34]:

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

In practice, p(x) can be approximated as the number of times that x appears in the corpus, p(y) as the number of times y appears in the corpus, and p(x,y) as the number of times the two words co-occur in a context or document. The above equation is changed to the following form to determine the word's polarity:

$$A(word, pos) = \log_2 \frac{p(word, pos)}{p(word)p(pos)} \tag{2}$$

$$A(word, neg) = \log_2 \frac{p(word, neg)}{p(word)p(neg)} \tag{3}$$

Here, p(word, pos) is the co-occurrence probability of word in positive document, p (word, neg) is the co-occurrence probability of word in negative document and p (word), p (pos) and p (neg) respectively are the occurrence probability of word, positive document and negative document. Finally the semantic orientation of word is calculated by following equation:

$$SO(word) = A(word, pos) - A(word, neg) \tag{4}$$

Word is classified as having a positive semantic orientation when So (word) is positive and negative orientation when So (word) is negative and if the value is zero, word has no positive and negative polarity and it is neutral.

$$wordpolarity \begin{Bmatrix} positive \to so(word) \succ 0 \\ negative \to so(word) \prec 0 \\ neutral \to so(word) = 0 \end{Bmatrix} \tag{5}$$

To get the score to the words, the study uses Min-max normalization. This method performs a linear transformation on the original data. Min-max normalization maps a value d of p to in the range [new_min (p), new_max (p)]. It is calculated by the following formula [35]:

$$d' = \frac{[d - \min(p)] \times [new\_max(p) - new\_min(p)]}{\max(p) - \min(p)} + new\_min(p) \tag{6}$$

Where min(p) is minimum value of attribute, max(p) is maximum value of attribute. In our case Min-max normalization maps a value d of p to in the range [-1, 1]. So put new_min (p)= -1 and new_max (p)= 1 in the above equation. Now the formula changes to the following form:

$$d' = \frac{[2 \times d] - \min(p) - \max(p)}{\max(p) - \min(p)} \tag{7}$$

After normalization, the values obtained from the above formula, multiplied in the amount of 5 and score range put in [-5, 5] and finally discretization is done. 5 score means that the word or phrase is quite positive and -5 means that the word or phrase is quite negative:

$$mainscore = 5 \times d' \tag{8}$$

$$discretization = \begin{Bmatrix} \lceil mainscore \rceil \xrightarrow{if} mainscore \succ 0 \\ \lfloor mainscore \rfloor \xrightarrow{if} mainscore \prec 0 \\ 0 \xrightarrow{if} mainscore = 0 \end{Bmatrix} \tag{9}$$

In this section, we have discussed the implementation of the proposed model on real data and the results of each steps of model are defined. Also evaluation of sentiment dictionary will be investigated here.

4. Experimental Results

4.1. Collecting Dataset

For the experiments in this study, the following dataset were used:

1) Opinions Collected from Online Retail Site: the corpus includes 7500 reviews in area of digital camera, laptop, television, tablet and mobile phones that were collected manually. This corpus is used for feature selection to create list of sentiment words.

2) Labelled English Reviews: we used the multi-domain review corpus collected by Blitzer [36]. The collected dataset consists of both positive and negative reviews in clothing, car, digital camera, mobile phone, computer, video, DVD and electronics domain. 31335 reviews were selected from his dataset and used for semantic orientation phase.

4.2. Preparation and Preprocessing Reviews

For normalization, stemming and part of speech tagging of text reviews have been used natural language processing tool created in Mashhad's web technology laboratory [37]. To extract the words and terms from document corpus, open source software named Rapid miner was used. This software converts documents in to a set of words using vector creation tool. In general created linguistic features were 12608. Among them only features with adjective, adverb, verb and noun label were selected. The numbers of words with any of these tags were respectively 1543, 106, 7621 and 3904. Only 3005 of them were used as an input to the next step. Table 2 shows the preprocessing of the sampled reviews and Table 3 presents subjectivity words in Persian language.

4.3. Construction Dictionary

To calculate the semantic orientation of words, we measure the relevance of them with positive and negative English reviews. Some of these words are presented in Table 4. Table 5 shows the number of positive and negative word and their percentage in dictionary.

Table. 2. Sample review

Review	Pre-processing
حرفی واسه گفتن ندارم فقط میتونم بگم عالیه!!!	Normalization : حرفی برای گفتن ندارم فقط می‌توانم بگویم عالی هست!!! Stemming : حرف برای گفت نداشت فقط توانست گفت عالی هست!!! Part of speech tagging : حرف<N> برای<PREP> گفت<N> نداشت<V> فقط<ADV> توانست<V> گفت<N> عالی<ADJ>هست<V>!!! ا<PUNC>ا
اگه تخصصی بهش نگاه کنید میشه گفت عالیه..	Normalization : اگر تخصصی به آن نگاه کنید می‌شود گفت عالی هست... Stemming : اگر تخصصی به آن نگاه کرد کرد گفت عالی هست... Part of speech tagging : اگر<SUBR> تخصصی<ADJ> به<PREP> آن<PR> نگاه<N> کرد<V> کرد<V> گفت<V> عالی<ADJ> هست<V>... ا<PUNC>ا
این دوربین عالیه تو رو خدا توی فروش ویژه بذارین !!!!!	Normalization : این دوربین عالی هست تو را خدا در فروش ویژه بگذارید!!!!! Stemming : این دوربین عالی هست تو را خدا در فروش ویژه گذارد!!!!! Part of speech tagging : این<PREM> دوربین<N> عالی<ADJ> هست<V> تو<PR> را<POSTP> خدا<N> در<PREP> فروش<N> ویژه<ADJ> گذارد<V>!!!!! ا<PUNC>ا

Table. 3. Persian subjectivity words sample

word	Equivalent to English	Part of speech
آبروریزی	Scandal	N
آرام	Calm	ADJ
آزار	Abuse	N
آزاردهنده	Annoying	ADJ
آزرد	Hurt	V
آسان	Easy	ADJ
آسوده	Ease	ADJ
آسیب	Damage	N
آسیب‌دیده	Hurt	ADJ
آسیب‌دیدگی	Injury	N
خوب	Good	ADJ
عالی	Excellent	ADJ
وحشتناک	Terrible	ADJ

Table. 4. Polarity of sentiment words

word	Association with positive	Association with positive	Semantic orientation	polarity
آبروریزی	-0.5373	0.3894	-0.9267	Negative
آرام	0.5197	-0.8164	1.3361	Positive
آزار	-0.0805	0.0759	-0.1564	Negative
آزاردهنده	-0.7456	0.4877	-1.2333	Negative
آزرد	-0.8304	0.5222	-1.3526	Negative
آسان	0.6186	-1.1003	1.7189	Positive
آسوده	0.4095	-0.5715	0.981	Positive
آسیب	-0.2369	0.2028	-0.4397	Negative
آسیب‌دیده	-0.1799	0.1594	-0.3393	Negative
آسیب‌دیدگی	-0.7955	0.5084	-1.3039	Negative

Table. 5. Score of sentiment words

Word	polarity	normalization	main score	discretization
آبروریزی	Negative	-0.4215	-2.107	-3
آرام	Positive	0.3216	1.608	2
آزار	Negative	-0.4202	-2.101	-3
آزاردهنده	Negative	-0.6618	-3.309	-4
آزرد	Negative	-0.4774	-2.378	-3
آسان	Positive	0.3565	1.782	2
آسوده	Positive	0.2892	1.446	2
آسیب	Negative	-0.6575	-3.287	-3
آسیب‌دیده	Negative	-0.7443	-3.721	-4
آسیب‌دیدگی	Negative	-0.7710	-3.855	-4

Table. 6. Number of positive and negative word in dictionary

Class	number	percent
Positive word	2007	55%
Negative word	1698	45%

4.4. Performance Evaluation

To evaluate the sentiment dictionary in a more qualitative manner, the polarity of terms were compared with sentiwordnet lexicon. This analysis was made in two steps. At first, common words in both dictionaries were extracted, which includes 2022 word. Then the polarity of words in sentiwordnet that have several senses obtained by averaging the values. at the second step, 97 words in sentiwordnet that have neutral polarity are extracted from two common lists and they are excluded, so only common words that have positive and negative polarity are remained. According to the following confusion matrix, “precision”, “accuracy” and recall are calculated:

Table. 7. Confusion matrix

class	predicted	
	Positive word	Negative word
Actual Positive word	# True positive samples (TP)	# False negative samples (FN)
Actual Negative word	#False positive samples (FP)	# True negative samples (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$precision^{+} = \frac{TP}{TP+FP}, precision^{-} = \frac{TN}{TN+FN} \quad (11)$$

$$recall^{+} = \frac{TP}{TP+FN}, recall^{-} = \frac{TN}{TN+FP} \quad (12)$$

The polarity of words in each category in both dictionaries is presented in table 6. From 1925 common words in both dictionaries, the number of words with positive and negative polarity is as follows:

Table. 8. Comparison method

Class	Number of words in sentiwordnet	Corrected word using proposed method	Percentage corrected
Positive word	1087	881	81%
Negative word	838	672	79%

Table. 9. Evaluation proposed model

Class	Precision	Recall	Accuracy
Positive word	84%	81%	80%
Negative word	76%	79%	

As you can see in table 7, the proposed method has acceptable accuracy in determining the sentiment polarity of words and the overall accuracy is 80%.

5. Conclusion

In this paper, we presented a supervised method for creating sentiment dictionary in Persian language, which used part of speech tagging and mutual information to detect the sentiment polarity of words. Since the proposed method is based on corpus approach, can be covered morphological and linguistic features existing in reviews. Also the proposed method considered the sentistrength of each polarity word in range 5 to -5. The evaluation results on created sentiment lexicon indicate that created sentiment dictionary is appropriate for sentiment classification. The proposed method has some limitations: it has limitation of corpus approach this means that it can not cover all words. Also mutual information is easy in term of implementation and it is not limited to adjectives but it needs a large scale corpus for better performance.

In future we are going to expand and improve sentiment lexicon by combining dictionary based approach and with corpus based approach and also apply mutual information method with other techniques for association measurement of words with positive and negative category.

References

[1] Wang, G.; et al., Sentiment classification: The contribution of ensemble learning. Decision Support Systems. 57(0): pp: 77-93 (2014).

[2] Zhang, C.; et al., Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology. 60(12): pp: 2474-2487 (2009).

[3] Pang, B.; Lee, L., Opinion mining and sentiment analysis. Foundations and trends in information retrieval. 2(1-2): pp: 1-135 (2008).

[4] Khan, A.; Baharudin, B.; Khan, K., Sentiment classification from online customer reviews using lexical contextual sentence structure, in Software Engineering and Computer Systems. Springer. pp: 317-331 (2011).

[5] Martín-Valdivia, M.T.; et al., Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications. 40(10): pp: 3934-3942 (2013).

[6] Huang, S.; Niu, Z.; Shi, C., Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems. 56: pp: 191-200 (2014).

[7] Liu, B., Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 5(1): pp. 1-167 (2012).

[8] Tan, S.; Wu, Q., A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Systems with Applications. 38(10): pp: 12094-12100 (2011).

[9] Kaji, N.; Kitsuregawa, M., Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. in EMNLP-CoNLL. pp: 1075-1083 (2007).

[10] Hu, M.; Liu, B., Mining and summarizing customer reviews. in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. pp:168-177 (2004).

[11] Kamps, J.; et al. Using WordNet to Measure Semantic Orientations of Adjectives. In LREC. Citeseer. pp: 1115-1118 (2004).

[12] Baccianella, S.; Esuli, A., Sebastiani, F.; SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. in LREC. pp: 2200-2204 (2010).

[13] Andreevskaia, A.; Bergler, S., Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. in EACL. pp: 209-216 (2006).

[14] Turney, P.D.; Littman, M.L., Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS). 21(4): pp: 315-346 (2003).

[15] Ding, X.; Liu, B.; Yu, P.S., A holistic lexicon-based approach to opinion mining. in Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM (2008).

[16] Kanayama, H.; Nasukawa, T., Fully automatic lexicon expansion for domain-oriented sentiment analysis. in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2006).

[17] Turney, P.D., 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. Association for Computational Linguistics: Philadelphia, Pennsylvania. pp: 417-424 (2002).

[18] Souza, M.; et al., Construction of a portuguese opinion lexicon from multiple resources. STIL, (2011).

[19] Montejo-Ráez, A.; et al., Ranked wordnet graph for sentiment polarity classification in twitter. Computer Speech & Language. 28(1): pp: 93-107 (2014).

[20] Medhat, W.; Hassan, A; Korashy, H.; Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 5(4): pp: 1093-1113 (2014).

[21] Alimardani, S.; Aghaei, A.; Opinion Mining in Persian Language Using Supervised Algorithms.

[22] Kim, S.-M.; Hovy, E.; Determining the sentiment of opinions, in Proceedings of the 20th international conference on Computational Linguistics. 2004, Association for Computational Linguistics: Geneva, Switzerland. p. 1367.

[23] Esuli, A.; Sebastiani, F.; Determining the semantic orientation of terms through gloss classification. in Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. ACM.

[24] Hatzivassiloglou, V.; McKeown, K.R.; Predicting the semantic orientation of adjectives. in Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics (1997).

[25] Gamon, M.; Aue, A.; Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. in Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. Association for Computational Linguistics (2005).

[26] Du, W.; et al. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. in Proceedings of the third ACM international conference on Web search and data mining. ACM. pp:111-120 (2010).

[27] Shams, M.; Shakery, A.; Faili, H., A non-parametric LDA-based induction method for sentiment analysis. in Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium on. IEEE (2012).

[28] Basiri, M.E.; Naghsh-Nilchi, A.R.; Ghassem-Aghaee, N., A Framework for Sentiment Analysis in Persian. open transactions on information processing, (2014).

[29] Lu, Y.; et al., Automatic construction of a context-aware sentiment lexicon: an optimization approach. in Proceedings of the 20th international conference on World wide web. ACM (2011).

[30] Saraee, M.; Bagheri, A., Feature Selection Methods in Persian Sentiment Analysis, in Natural Language Processing and Information Systems, E. Métais, et al., Editors. Springer Berlin Heidelberg. pp: 303-308 (2013).

[31] Ravi, K; Ravi, V., A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems. 89: pp: 14-46 (2015).

[32] Haddi, E.; Liu, X.; Shi, Y., The role of text pre-processing in sentiment analysis. Procedia Computer Science. 17: pp: 26-32 (2013).

[33] Wiebe, J.; Riloff, E., Creating subjective and objective sentence classifiers from unannotated texts, in Computational Linguistics and Intelligent Text Processing. Springer. pp: 486-497 (2005).

[34] Recchia, G.; Jones, M.N., More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. Behavior research methods. 41(3): pp: 647-656 (2009).

[35] Jain, Y.K.; Bhandare, S.K., Min max normalization based data perturbation method for privacy protection. International Journal of Computer & Communication Technology (IJCCT). 2(8): pp: 45-50 (2011).

[36] Blitzer, J.; Dredze, M.; Pereira. F., Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. in ACL (2007).

[37] Kahani, M.; et al., Designing Persian Language Parser Tool. International Conference on Asian Language Processing, (2012).