Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining

MOHAMMAD EHSAN BASIRI and ARMAN KABIRI, Shahrekord University

Lexicon-based sentiment analysis (SA) aims to address the problem of extracting people's opinions from their comments on the Web using a predefined lexicon of opinionated words. In contrast to the machine learning (ML) approach, lexicon-based methods are domain-independent methods that do not need a large annotated training corpus and hence are faster. This makes the lexicon-based approach prevalent in the SA community. However, the story is different for the Persian language. In contrast to English, using the lexicon-based method in Persian is a new discipline. There are rather limited resources available for SA in Persian, making the accuracy of the existing lexicon-based methods lower than other languages. In the current study, first an exhaustive investigation of the lexicon-based method is performed. Then two new resources are introduced to address the problem of resource scarcity for SA in Persian: a carefully labeled lexicon of sentiment words, PerLex, and a new handmade dataset of about 16,000 rated documents, PerView. Moreover, a new hybrid method using both ML and the lexicon-based approach is presented in which PerLex words are used to train the ML algorithm. Experiments are carried out on our new PerView dataset. Results indicate that the accuracy of PerLex is higher than the existing CNRC, Adjectives, SentiStrength, PerSent, and LexiPers lexicons. In addition, the results show that using PerLex significantly decreases the execution time of the proposed system in comparison to the above-mentioned lexicons. Moreover, the results demonstrate the excellence of using opinionated lexicon terms followed by bigrams as the features employed in the ML method.

CCS Concepts: • **Information systems → Content analysis and feature selection**; *Data mining*; *Web searching and information discovery*;

Additional Key Words and Phrases: Sentiment analysis, Persian language, lexicon-based approach, opinion mining, machine learning, PerView dataset

ACM Reference format:
Mohammad Ehsan Basiri and Arman Kabiri. 2018. Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 17, 4, Article 26 (May 2018), 18 pages.
https://doi.org/10.1145/3195633

1 INTRODUCTION

Sentiment analysis (SA) is a subfield of natural language processing (NLP) and data mining (DM) that concentrates on the process of computationally identifying and extracting people's opinions and attitudes expressed in their comments on the Web [8].

This work was financially supported by the research deputy of Shahrekord University (grant 95GRN1M1874).
Authors' addresses: M. E. Basiri and A. Kabiri, Department of Computer Engineering, Shahrekord University, Rahbar Boulevard Shahrekord, 105, Iran; emails: basiri@eng.sku.ac.ir and Arman.Kabiri94@gmail.com.

https://doi.org/10.1145/3195633

Research on SA started in the early 2000s, and since then it has become an active research topic in DM and NLP communities. There are plenty of academic and industrial applications for SA, including extracting customers' attitudes toward a product or service [34], social media monitoring [15], analysis of political tweets [14], and predicting sales performance [33].

Existing approaches are classified into two main categories: the corpus-based machine learning (ML) approach and the lexicon-based method [30, 31]. Although ML approaches offer some advantages such as the ability to identify implied sentiment [32], they suffer from several drawbacks such as needing a corpus of human-annotated reviews for training and depending on the domain on which they were trained [22, 30]. Lexicon-based approaches are robust, domain-independent methods that can be easily improved using different sources of knowledge [30].

Most researchers in the SA field have investigated widespread languages such as English, Chinese, or Arabic, and few studies have targeted the Persian language [7, 24]. Persian is spoken by more than 100 million speakers in Iran, Afghanistan, and many states of the former Soviet Union [7]. However, the Persian language has not received the attention it deserves, and hence there are limited available linguistic resources for it.

As pointed out earlier, SA applications use either lexicon-based or ML methods. The resource exploited in the former is a lexicon of labeled sentiment words, whereas a human-annotated dataset is the resource used in the latter. The more precise the resources are, the more accurate results will be obtained. This article introduces two new resources: a carefully labeled lexicon of sentiment words, PerLex, and a new dataset, PerView.

PerLex is an accurate lexicon of common sentiment-bearing words augmented with a list of emoticons. In the process of creating this lexicon, we selected two well-known existing lexicons—NRC and SentiStrength—as the base lexicons. Having conducted different experiments on these lexicons to demonstrate their shortcomings, we found that the main drawback of these lexicons is that they are directly translated from English. To overcome their shortcomings, we remove all of the words that do not convey sentiment in Persian from NRC and SentiStrength. Then we remove those words corresponding to long phrases in Persian that are never matched with phrases in a real comment. In the next step, all tokens are carefully reviewed and those with incorrect labels are corrected. Finally, new missing words and phrases are added to PerLex. More detailed information explaining how PerLex is developed is provided in Section 3.

Almost all previous studies on SA in the Persian language suffer from the unavailability of a large dataset [7, 24]. PerView is introduced in this article to fill this gap. This dataset contains about 16,000 user reviews and was labeled at the document level. More details about these resources will be presented in Section 3.

To the best of our knowledge, existing corpus-based methods for SA in the Persian language use either *n*-gram features or semantic features [24]. To enhance the accuracy of the corpus-based approach, a new hybrid method for SA in Persian is presented in this article. This method is an ML-based method exploiting sentiment words listed in PerLex as the training features.

The remainder of the article is organized as follows. Section 2 reviews the background and related work. Section 3 illustrates the methodology and the proposed system. Section 4 reports the experimental results and presents a discussion of the examined methods. Section 5 presents our conclusion and future work.

2 RELATED WORK

SA has attracted a lot of attention in recent years, especially for widespread languages such as English [20], and numerous studies for SA have been published so far [12, 18, 19]. However, we do not intend to review SA studies on the English language in this section. Instead, we will present a comprehensive literature review of SA studies focusing on the Persian language.

The first published study on SA investigating the Persian language was reported by Shams et al. [27]. They suggested an unsupervised LDA-based method and evaluated their method on three manually created datasets about hotels, cell phones, and digital cameras. Although they reported a 9% improvement in comparison to a baseline algorithm, their study had some limitations. First, their method is applicable only for polarity detection. Second, they did not deal with language-specific problems of SA in the Persian language. Finally, the datasets on which they reported their results were relatively small.

Bagheri et al. [5] proposed a model for SA in the Persian language employing a naive Bayes algorithm for classification. They also presented a feature selection method based on the mutual information and evaluated their model on a manually gathered collection of cell phone reviews. This study has the same limitations as that of Shams et al. [27].

Later on, Hajmohammadi and Ibrahim [17] compared the performance of two standard ML techniques—SVM and Naive Bayes—on a dataset of online Persian movie reviews. According to the previous studies, this method was restricted to the polarity detection problem. Moreover, only *n*-gram features were used for training the classifier.

Basiri et al. [7] proposed a framework for SA in the Persian language in which some of the Persian text processing difficulties were considered. Their proposed system could be considered as the first lexicon-based method for SA on the Persian language. Three ML algorithms, namely naive Bayes, SMO, and J48, were compared to the lexicon-based approach, and the authors stated that their "proposed approach outperforms machine learning methods in terms of MAE and F-score" [7]. This study was also limited to polarity detection problem.

In a similar study, Bagheri and Saraee [6] addressed some of the Persian text processing difficulties and investigated different feature selection methods for polarity detection. This study had two limitations: first, it only utilized the naive Bayes learning algorithm, and second, the dataset on which they evaluated their method was too small and domain specific.

Golpar-Rabooki et al. [16] proposed a feature extraction method for SA on Persian reviews. Specifically, they first created a lexicon and performed some preprocessing steps on the reviews. Then, they applied two feature extraction methods: a frequency-based method and an association rule-based method. Finally, they assessed the performance of their methods on a dataset of user reviews. Similar to the previous reported research on Persian SA, this study focused on polarity detection. Another limitation of this study was the size of the dataset used for evaluation that contained only 340 reviews.

Recently, Alimardani and Aghaei [2] proposed a method for polarity detection applying the combination of Persian SentiWordNet and three ML algorithms. Specifically, they first created a Persian SentiWordNet using the existing English SentiWordNet and Persian WordNet. Finally, they used the Persian SentiWordNet to weight the features.

More recently, Dashtipour et al. [13] published a freely available lexicon, PerSent, containing 1,500 phrases and their part-of-speech (POS) tags. They evaluated their lexicon with two ML methods and reported an average overall accuracy of about 62%. One of the advantages of this study is the POS tags associated with sentiment-bearing words. However, the main drawback of their lexicon is that it contains many unconventional Persian phrases that are barely seen in informal Web texts. Moreover, the accuracy of sentiment labels could be higher. For example, in PerSent, sentiment words such as "beautiful," "correct," and "detrimental" are all considered neutral words.

In summary, all of the preceding studies have some similar limitations. They all have addressed the polarity detection problem. This could be considered a limiting factor for SA methods since recent applications of SA need more detailed analysis, such as rating predition. For example, to utilize the history of reviewers' comments, sentiment polarity is not sufficient for the method proposed by Basiri et al. [8]. Moreover, the dataset used for evaluation in almost all of the studies

Table 1. Review of SA Studies in Persian

Author	Title	Year	Limitations
Shams et al. [27]	A non-parametric LDA-based induction method for sentiment analysis	2012	Limited to polarity detection. Ignores language-specific problems. Employs a relatively small dataset.
Bagheri et al. [5]	Sentiment classification in Persian: Introducing a mutual information-based method for feature selection	2013	Limited to polarity detection. Ignores language-specific problems. Employs a relatively small dataset.
Hajmohammadi and Ibrahim [17]	A SVM-based method for sentiment analysis in Persian language	2013	Limited to polarity detection. Only *n*-gram features are used.
Basiri et al. [7]	A framework for sentiment analysis in Persian	2014	Limited to polarity detection.
Bagheri and Saraee [24]	Feature selection methods in Persian sentiment analysis	2013	Limited to polarity detection. Only utilized naive Bayes. Employs a small and domain-specific dataset.
Golpar-Rabooki et al. [16]	Feature extraction in opinion mining through Persian reviews	2015	Limited to polarity detection. Employs a very small dataset.
Alimardani and Aghaei [2]	Opinion mining in Persian language using supervised algorithms	2015	Limited to polarity detection.
Dashtipour et al. [13]	PerSent: A freely available Persian sentiment lexicon	2016	Limited to polarity detection.
Basiri and Kabiri [9]	Sentence-level sentiment analysis in Persian	2017	Limited to sentence level.

have had less than 1,000 records. This increases the randomness of results, which in turn makes the reported results unreliable. In addition, almost all of them are either pure lexicon-based or ML-based approaches. A review of the preceding studies is depicted in Table 1.

3 PROPOSED SYSTEM

As mentioned earlier, the study of SA in Persian language has just started since 2012 and the first lexicon-based approach for SA in Persian is reported in 2014 [7]. In the lexicon-based approach, a dictionary of words and their corresponding sentiment label is used to specify the overall sentiment of a sentence or a document [18]. This approach, compared to the ML method, has several advantages, such as robustness, domain independency, ease of implementation, and the ability to be improved using different sources of knowledge. Therefore, we focus on this approach. The overall view of the proposed lexicon-based approach is depicted in Figure 1.

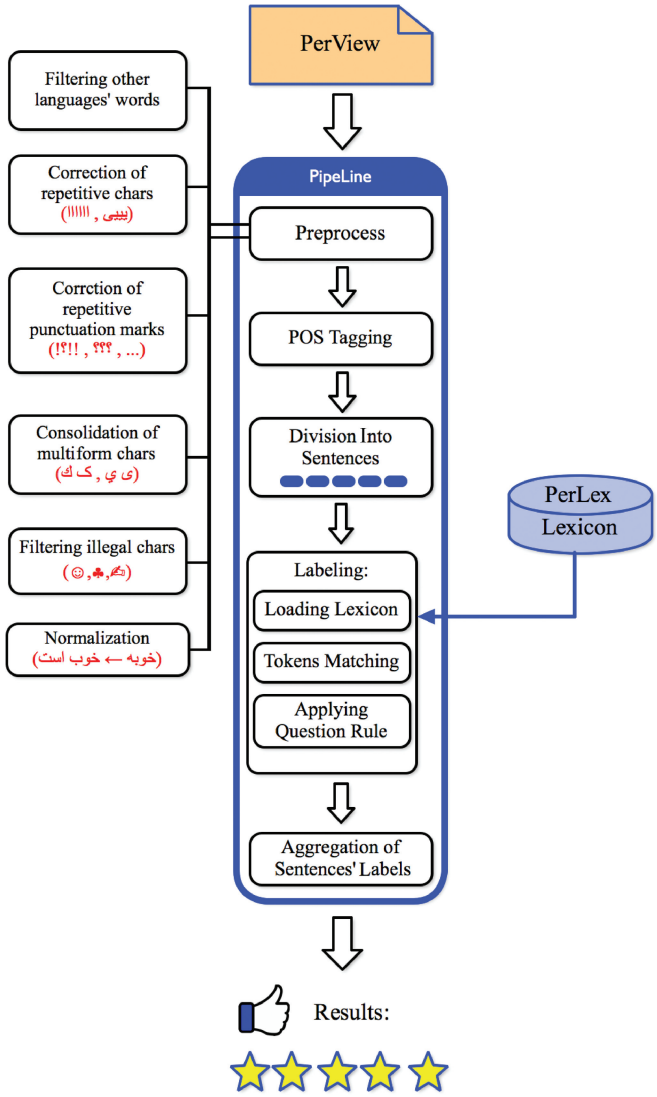

Fig. 1. Proposed lexicon-based approach.

The input to the system in Figure 1 is a review containing at least one sentence. The output of the system, however, is a five-star score for every test review. In fact, in contrast to the previous studies on SA in Persian, we focus on the rating prediction problem instead of the polarity detection problem. Different steps of each part of Figure 1 will be described in detail in the following sections.

3.1 Preprocessing

In the preprocessing phase of Figure 1, six preprocessing steps are applied: filtering, correction of repetitive chars, correction of punctuation, consolidation of multiform chars, filtering illegal chars, and normalization:

Table 2. Typical Examples of Common Simplification Rules in Informal Writing

Replacement pattern	Standard form	Informal form	English translation
ون by ان	ارزان	ارزون	Cheap
	گران	گرون	Expensive
	جوان	جوون	Young
د by ه	می پرد	می پره	Jumps
	می خرد	می خره	Buys
	می برد	می بره	Wins
است by ه	مهمتر است	مهمتره	It is more important
	کمتر است	کمتره	It is less than
	جالبتر است	جالبتره	It is more interesting
هستم by م	خوشحال هستم	خوشحالم	I am happy
	مطمئن هستم	مطمئنم	I am sure
	ناامید هستم	ناامیدم	I am hopeless

- *Filtering other languages' words*: Those words not belonging to Persian are removed because the lexicon does not contain any non-Persian words. For example, many words such as models' and brands' names, dates, and place names are written in English, and omitting them has no effect on the performance of the system.
- *Correction of repetitive chars*: Repetitive characters that are used for emphasis are removed to enhance the matching process.
- *Correction of punctuation*: Similar to the previous case, repetitive punctuation is also removed.
- *Consolidation multiforms chars*: Some letters in Persian have different Unicode. In particular, this problem occurs for those words containing letters such as ی and ک(for "i" and "k," respectively).
- *Filtering illegal chars*: Some illegal characters are used as abbreviations and are not necessary in lexicon-based methods.
- *Normalization*: This step is used to convert informal-style to formal-style writing. In the informal style, grammatical rules are usually ignored and some simplifying rules are applied to the words (Table 2). Although this informal style is not common in news, books, and newspapers, it has become widespread in recent years in social media. We used *NLP-Tools*, a freely available toolbox developed by Web Technology Lab at Ferdowsi University of Mashad [4].

3.2 Score Detection

Although in recent years some methods have been proposed to detect the sentiment score in English, all of the reported studies for Persian SA have targeted the polarity detection problem [3, 9, 27]. Score detection methods are used to detect the degree to which a review is positive or negative. As mentioned earlier, there are three common approaches for SA: lexicon-based, ML-based, and hybrid approaches.

In the lexicon-based approach, having performed the preprocessing steps on the input review, we use POS tags to specify verbs in order to separate sentences. We do so because usually each

verb has an independent meaning, and thereby it can be used to specify sentences' boundaries. Then, all words of each sentence are looked up in a lexicon of sentiment words, and the average of the score of the matched words is considered as the score of the sentence. Additionally, due to the nature of the Persian language, numerous suffixes associate with Persian words, most of which are pronouns. This issue decreases the chance of full matching of words with lexicon words. For example, in the sentence من ماشین زیبایش را دیدم ("I saw her beautiful car"), the character ش is a pronoun suffix accompanying the adjective زیبا (beautiful). To address this difficulty, we use partial matching instead of full matching in the labeling phase in Figure 1. The partial-matching approach first tries to divide the preknown suffixes from the main word. Then, it looks up the separated word in the lexicon. This process is somehow similar to what usually is done in stemming phases.

Finally, just those sentences with the indicative mood are passed to the next phase. The rationale behind this policy is that usually interrogative sentences do not carry reliable facts to which we can base our prediction. For instance, although the sentence "Do you think Toyota is a good car?" contains a positive adjective (good), its writer does not express a positive idea about that car; instead, it just means to ask to see whether it is a good car or not. That is why we employ this policy on our prediction. Moreover, we consider punctuation as a reliable source to determine the mood of a sentence.

3.3 Score Aggregation

The aim of the aggregation mechanism is to calculate the overall sentiment score of a review based on the scores calculated for its sentences. In the document-level SA, score aggregation is a data fusion step to combine sentence scores into a single review score. Despite its importance, score aggregation in SA has not received the attention it deserves so far [8], and in most studies simple methods such as maximum of scores, majority voting, and simple averaging [11] are used. Recently, a formally defined method for score aggregation based on the Dempster-Shafer (DS) theory of evidence [26] was proposed by Basiri et al. [11]. It has been shown that the DS-based aggregation method clearly outperforms other aggregation methods used in SA [11]. There may be two reasons for this: first, the DS-based method takes all pieces of evidence into account, and second, it preserves maximal agreements among the evidence [25].

To use the DS-based method for score aggregation, we first define the sentence scores computed by the score detection module of the previous section as the evidence. Then, we define the mass function (a basic probability assignment (BPA)) as follows:

$$m\_S(A) = \frac{score - min}{max - min}, \tag{1}$$

where S is a sentence; *score* is the output of the score detection module for this sentence; and *max* and *min* are the maximum and minimum scores of all sentences, respectively. This mass function reflects the degree to which a review is positive. As can be seen, this function is a BPA and has the following necessary properties:

$$m(\phi) = 0 \quad \text{and} \sum_{m \in 2^{\theta}} m(A) = 1, \tag{2}$$

where θ is a finite set of mutually exclusive hypotheses, called *frame of discernment*. These properties must be held for any mass function in the DS theory.

The next step for using DS in our method is to use Dempster's rule of combination for aggregating n sentence scores as follows:

$$m_{1,\ldots,n}(A) = \frac{\sum_{\cap_{i=1}^{n} X_i = A} (\prod_{j=i}^{n} m_j(X_i))}{1 - K}, \tag{3}$$

where the denominator, K, is a normalization factor to ensure that $m_{1,\ldots,n}(A)$ remains a BPA and is computed as follows:

$$K = 1 - \sum_{\cap_{i=1}^{n} X_i = \phi} \left(\prod_{j=1}^{n} m_j(X_i) \right). \tag{4}$$

Since the DS rule of combination is both commutative and associative, we can iteratively apply the following equation to avoid the computational complexity of Equation (3):

$$m(A) = \frac{\sum_{X \cap Y = A} m_n(X) m_o(Y)}{1 - \sum_{X \cap Y = \phi} m_n(X) m_o(Y)}, \tag{5}$$

where m_n and m_o correspond to the new and old existing evidence. In other words, m_o is the aggregated value from the previous iteration of Dempster's rule of combination and m_n is calculated for the current sentence. Eventually, the final aggregated $m(A)$ is scaled to a five-star score as follows:

$$FiveStarScore = round((m(A \times 4) + 1). \tag{6}$$

3.4 Proposed Hybrid Method

To utilize the benefits of lexicon-based and ML-based approaches, we have proposed a hybrid method as depicted in Figure 2.

As shown in Figure 2, the proposed hybrid method is a feature-level combination of lexicon-based and ML-based methods. Specifically, having preprocessed the input review, we use lexicon terms and bigrams in the feature extraction step. The reason for combining lexicon-based and bigram features is that previous studies have shown that the best performance can be obtained through unigrams and bigrams [28]. However, in the current study, the unigrams are replaced by lexicon terms. The rationale behind using lexicon features is that in contrast to the nonsentiment-bearing unigrams, lexicon terms are determinants of the overall sentiment of the review. Hence, using lexicon terms should improve the performance of the system.

As in ML-based methods, the proposed hybrid method would suffer from the large size of feature space if feature selection was not used. The following feature selection steps are used in the proposed systems:

- *Occurrence filter*: In this step, those features occurring fewer than 10 times are considered rare features. To simultaneously reduce the size of the feature vector and increase the precision of the proposed hybrid system, we prune the feature vector by removing the rare features.
- *Stop word filter*: Although they are very frequent features, stop words not only do not play a significant role in sentiment prediction process but also decrease the performance of the system.

The final step in Figure 2 is the training and validation phase, which is necessary in ML-based methods [23].

3.5 Creating the PerLex[1]

Although some previous studies followed the lexicon-based approach, they addressed the problem simplistically [3, 7]. For example, almost all of the existing lexicon-based approaches have used automatically translated lexicons from English. This rises different problems, such as the following:

[1] All resources introduced in this article are available at the file menu of the homepage of the first author.

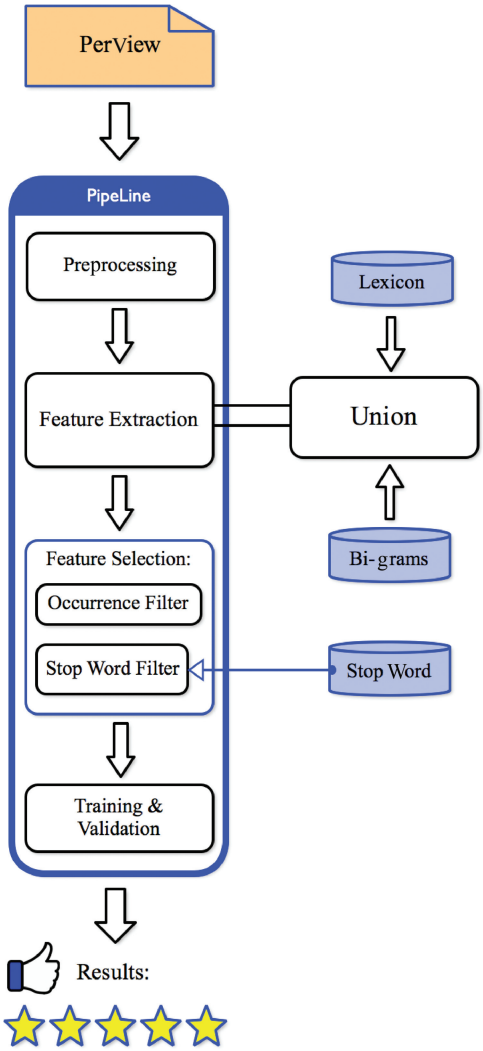

Fig. 2. Proposed hybrid methods for SA.

- There is not always a one-to-one relationship between sentiment words of the source and destination languages. For example, a word such as *ball* in English is not a sentiment word, but in informal Persian, this word (i.e., its translation) is used instead of *perfect*.
- Some sentiment words in the source language have different translation in the destination language, each with a different sentiment score.
- Some words in the source language correspond to a phrase or sentence in the destination language, making the translated entry pointless or useless.
- Automatic translation errors change the meaning and hence the polarity of some words. For example, the word *abba* in the NRC lexicon is a positive word mistakenly translated to *avoid* in Persian, which is obviously a negative word.

Since the core of every lexicon-based approach is the lexicon that it exploits, the preceding problems can significantly decrease their performance. To overcome such problems, having analyzed

the performance of NRC and SentiStrength lexicons for SA in Persian, we designed a new lexicon, PerLex, which can be used for both polarity detection and rating the prediction problems. The overall process of creating PerLex is shown in Figure 3.

As can be seen in Figure 3, PerLex can be seen as the intersection of three lexicons—CNRC, Adjectives, and Persian SentiStrength—in which a postprocessing step is performed. Introduced in our previous study, CNRC is the corrected version of the NRC lexicon [9]. To create Persian SentiStrength, we first converted its score to a five-star scale and then used machine translation to translate its sentiment words into Persian. The process of creating the Adjectives lexicon is as follows.

According to the previous studies on SA, adjectives are one of the most important signs of sentiment [11, 29]. Keeping this fact in mind, we use the PerView dataset to extract the adjectives. First, a preprocessing step is applied to the dataset in which the following four tasks are performed as described in Section 3.1:

- Filtering non-Persian words
- Correction of the repetitive characters
- Consolidation of multiform words
- Normalization.

Having preprocessed the dataset as described earlier, we use a POS tagger to specify POS labels of the words. Based on the POS tags, we filter the dataset by ADJ tags to keep just adjectives. Finally, the resulted adjectives are labeled manually in a five-star scale.

After the preceding steps, three Persian lexicons are intersected and the following postprocessing steps are applied on the intersection result to form PerLex:

- *Pointless words removal*: In this step, all words that do not convey sentiment in the Persian language are removed.
- *Long phrases removal*: This step is considered to remove those words corresponding to long phrases in Persian that are never matched with phrases in a real comment.
- *Semantic filtering*: In this step, all tokens are carefully reviewed and those with incorrect labels are corrected.

Finally, the labels of each word in PerLex is calculated by employing the DS theory of evidence on CNRC, Adjectives, SentiStrength, PerSent, and LexiPers lexicons [8, 21].

4 RESULTS AND DISCUSSION

To show the effectiveness of PerLex, two series of experiments are conducted on thew PerView dataset. In the first experiment, we aim to answer the following research questions:

(1) With respect to their constituent words, what is the difference between PerLex and the existing lexicons?
(2) Does PerLex produce more accurate results when it is used in a pure lexicon-based approach?

In the second experiment, our goal is to answer the following research question:

(1) Is the proposed hybrid method superior to lexicon-based and ML-based methods for SA in Persian?

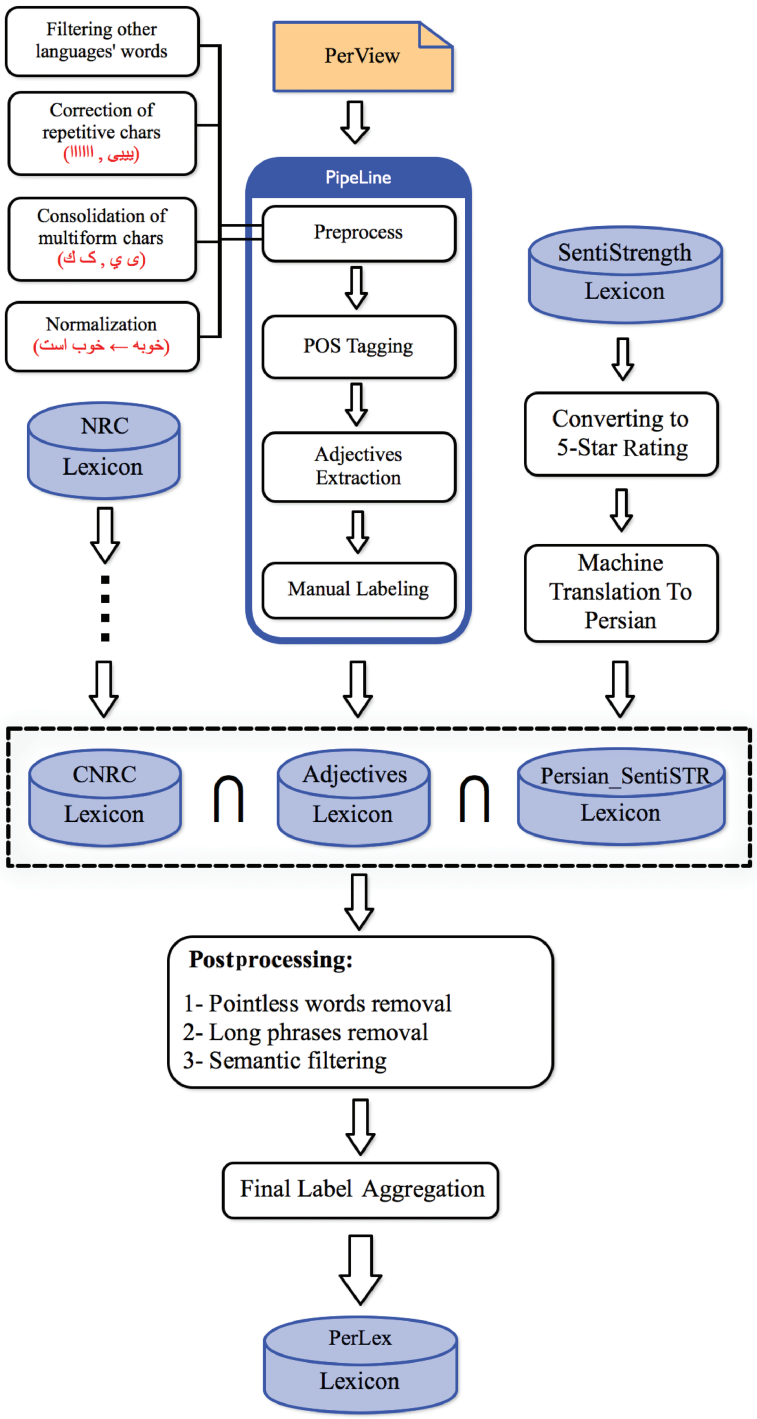

Fig. 3. Process of creating PerLex.

Table 3. Specifications of Seven Lexicons

Lexicon	Number of Words	Words' Occurrence	Max Occurrence	Unique Words
NRC	9,449	33%	9,003	63%
CNRC	2,697	38%	2,229	0%
SentiStrength	2,765	35%	2,618	45%
Adjectives	1,677	91%	3,527	54%
LexiPers	6,500	12%	4,315	83%
PerSent	1,470	17%	2,709	77%
PerLex	174	98%	2,229	0%

4.1 Dataset and Evaluation Metrics

As mentioned earlier, one of the greatest shortcomings of the previous studies on SA in Persian is the small size of the dataset they used. In this study, we introduced PerView as a large manually labeled dataset that can be used for document-level SA in Persian. This dataset contains 16,000 user comments collected from Digikala.com, the biggest e-commerce startup in Iran and the Middle East [1]. The PerView comments were collected from July 2016 to February 2017. It contains customers' comments about digital equipment including cell phones, cameras, and computer peripherals.

In our experiments, we used five evaluation criteria: precision (π), recall (ρ), F-measure, accuracy, and MAE. These criteria are common in the previous studies [7, 11, 24] and are defined as follows:

$$\pi = \frac{TP}{TP + FP},$$
$$\rho = \frac{TP}{TP + FN},$$
$$\text{F-Measure} = \frac{2 \times \pi \times \rho}{\pi + \rho},$$
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$
$$MAE = \frac{\sum_1^n |p_i - r_i|}{n},$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively [7]. For MAE, n is the number of test comments, and p_i and r_i are the predicted and real rate of the ith test comment, respectively.

4.2 Differences Between Lexicons

As mentioned earlier, seven lexicons are tested in this study: NRC; CNRC; SentiStrength; Adjectives; LexiPers; PerSent; and our proposed lexicon, PerLex. To answer the first research question, we analyze the lexicons. Specifications of these lexicons are presented in Table 3.

The second column in Table 3 shows the percentage of words occurring at least one time in the PerView dataset. The third column shows the frequency of the most frequent words of each lexicon, and the fourth column shows the percentage of unique words. As can be seen in Table 3, NRC contains more words compared to other lexicons, but as shown in the experiments, most of its words are not prevalent sentiment-bearing words.

To clarify the differences between the lexicons, their word clouds are depicted in Figure 4.

As can be seen in Figure 4, there are many frequent nonopinionated words in NRC. Such words can severely decrease the quality of this lexicon. However, the other six lexicons seem similar in

(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 4. Word cloud of seven lexicons: Adjectives (a), CNRC (b), PerLex (c), SentiStrength (d), NRC (e), LexiPers (f), and PerSent (g).

Table 4. Specifications of Seven Lexicons

Lexicon	Top 10 Frequent Words
NRC	and, hello, it, very, until, excellent, work, good, opinion, then
CNRC	excellent, good, difficulty, friend, no, dear, quality, little, model, slow
SentiStrength	better, excellent, good, difficulty, friend, dear, ever, price, little, goodness
Adjectives	better, excellent, good, difficulty, thankful, dear, quality, open, little
LexiPers	slow, good, later, problem, two, low, no, thankful, friend, open
PerSent	all, good, later, have been, low, to make, soft, done, right, to be
PerLex	excellent, good, difficulty, friend, dear, little, important, hard, satisfied, comfortable

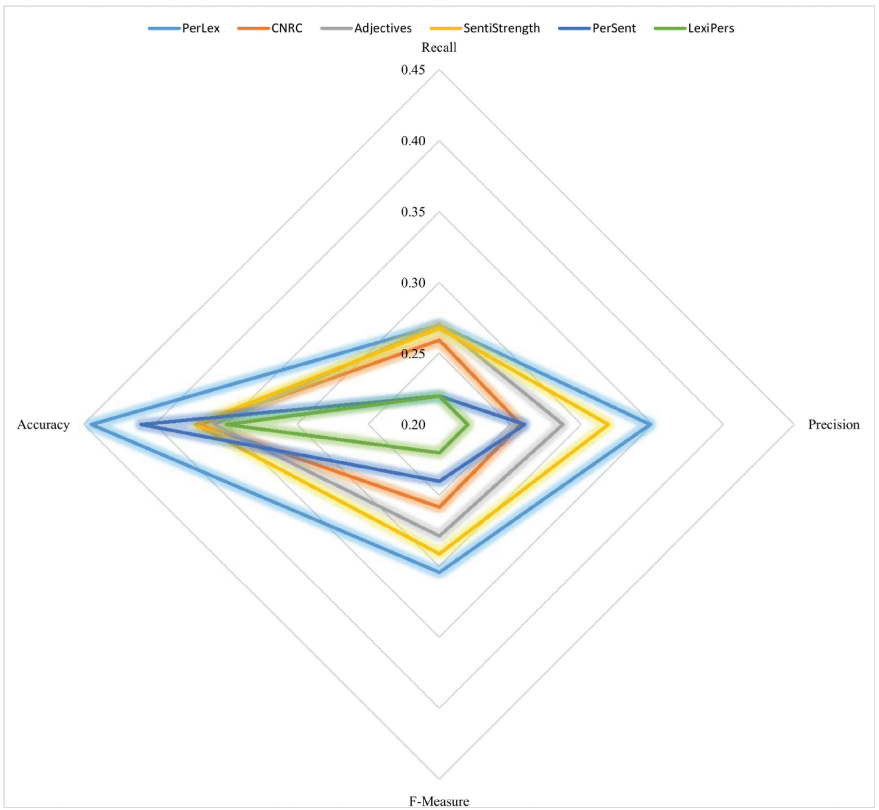

Fig. 5. Comparison of the performance using PerLex with other lexicons.

that the most frequent words in all of them are positive words. To show the differences between lexicons more clearly, the top 10 frequent words of lexicons are listed in Table 4. Each row in this table shows the 10 most frequent words sorted according to their frequency from left to right.

To answer the second research question, we tested each lexicon in a pure lexicon-based system described in Figure 1. The comparison of the results obtained using different lexicons are presented in Figures 5 and 6.

Results obtained using the NRC lexicon are omitted because of its poor performance. A significant point in Figure 5 is that the recall of all lexicons are nearly identical but their precisions are different. This shows that the false positive is different for different lexicons. Specifically,

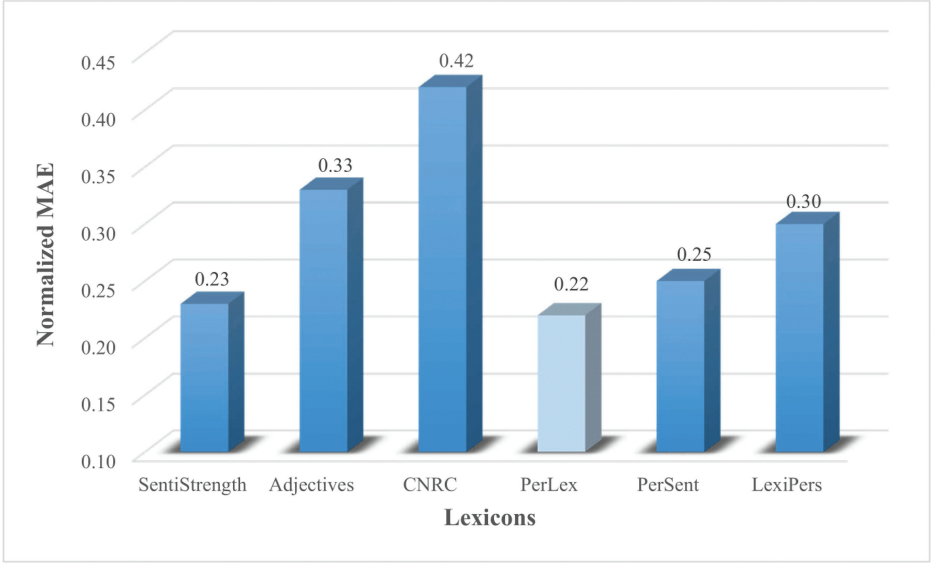

Fig. 6. Comparison of the normalized MAE of using PerLex and other lexicons in the proposed system.

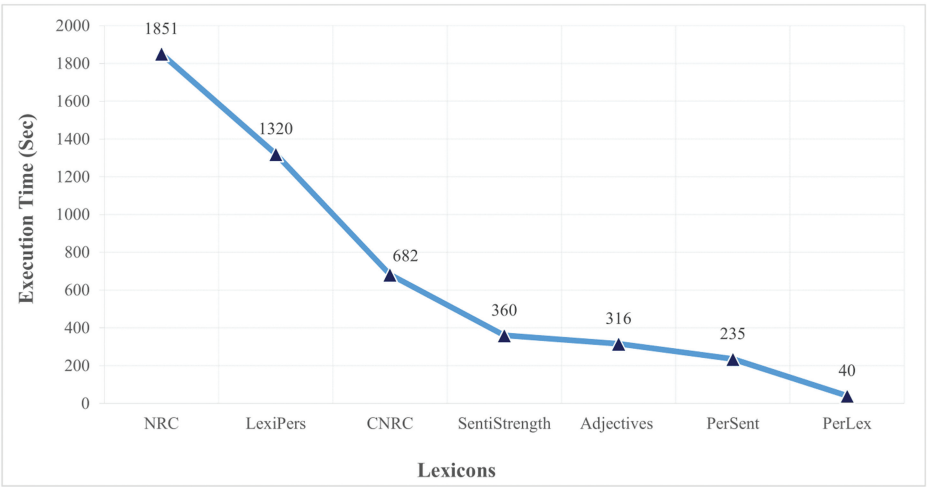

Fig. 7. Comparison of the time performance of the lexicon-based method using different lexicons.

according to Figure 5, PerLex has the lowest false positive, and hence its precision is higher than the other lexicons. Another point in Figure 5 is that the performance of LexiPers is lower than other lexicons. These results answer the second research question successfully.

An important factor for preferring one lexicon over other lexicons is its size, because it directly affects the overall time complexity of system. To show this, we compare the execution time of the proposed system using different lexicons in Figure 7. All steps were implemented in Java 8 on a 3740QM-i7 machine with a 3.7Ghz CPU, 6MB cache, and 16GB RAM.

As can be seen in Figure 7, the execution time of the system using LexiPers and NRC is more than three times that of using CNRC, whereas, as pointed out earlier, the performance of CNRC

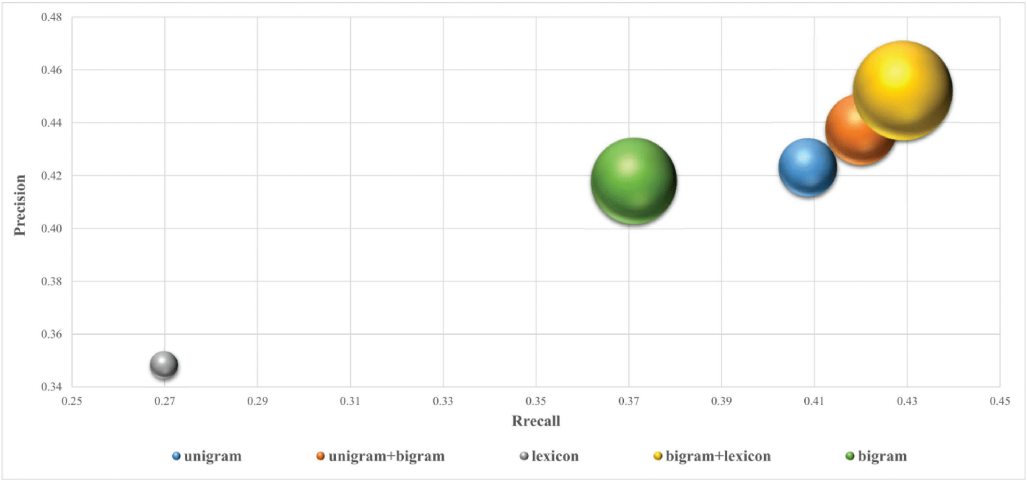

Fig. 8. Comparison of the performance of the lexicon-based, hybrid, and ML-based methods. Sphere size for each method has a direct relation with (1 – MAE) of using corresponding classifier.

is significantly higher than that of those two lexicons. Moreover, PerLex has the lowest execution time, which, besides its higher performance, makes it the best choice among the tested lexicons. This significantly lower execution time also makes PerLex a suitable choice for online applications.

The third research question can be answered by comparing the performance of the proposed hybrid method described in Figure 2 with lexicon-based and ML-based methods. Both ML-based and hybrid methods use a naive Bayes classifier was previously shown to be a successful ML-based classifier for SA [8, 11].

As can be seen in Figure 8, with respect to all three performance measures, the proposed hybrid method (the yellow sphere) outperforms both the lexicon-based (the silver sphere) and ML-based methods (the green, blue, and orange spheres). This justifies the fact that although ML-based methods outperform the lexicon-based method, the ML-based method can be enhanced when unigram features are replaced by PerLex terms. Hence, the third research question is successfully addressed.

5 CONCLUSIONS

The Persian language is the official language of Iran, and more than 100 million people around the world speak Persian. However, SA in the Persian language is a young research field. Although early studies preferred ML methods to the lexicon-based approach, lexicon-based SA methods have attracted increasing attention in recent years. Compared to their counterparts in English, existing lexicon-based methods for SA in Persian have lower performance. To address this problem and to improve the performance of lexicon-based methods, an exhaustive investigation of the lexicon-based method is performed in the current study. The investigation results showed that the main reason for the low performance of SA in the Persian language is the resource scarcity problem. To address this problem, two new resources are introduced: a carefully labeled lexicon of sentiment words, PerLex, and a new handmade dataset of about 16,000 rated documents, PerView.

In the construction of PerLex, three lexicons are used and several preprocessing and postprocessing steps are applied on the resulting lexicon. To show the performance of PerLex, several experiments are carried out on the PerView dataset. Results indicate that the accuracy of PerLex is higher than those of existing lexicons. Moreover, a new hybrid method using both ML and the

lexicon-based approach is presented in which PerLex words are used to train the ML algorithm. This hybrid method is shown to be more effective when PerLex terms and bigrams are employed as the features. This shows the higher quality of the PerLex in comparison to unigram features.

Several directions are suggested for future research. For example, improving the proposed lexicon using ML methods may be a promising suggestion. Another line of research may be employing more contextual features in the proposed hybrid method. Finally, enhancing the PerLex with contextual heuristic rules may also be considered for future work.

ACKNOWLEDGMENTS

The authors would like to thank the M.Sc. students of Safahan institute of higher education who voluntary participate in the process of gathering PerView comments.

REFERENCES

[1] Digikala. 2017. Home Page. Retrieved March 23, 2018, from http://www.digikala.com.

[2] Saeedeh Alimardani and Abdollah Aghaei. 2015. Opinion mining in Persian language using supervised algorithms. *Journal of Information Systems and Telecommunication* 3, 3, 1–7.

[3] Fatemeh Amiri, Simon Scerri, Mohammad H. Khodashahi, Fraunhofer Iais, and Sankt Augustin. 2015. Lexicon-based sentiment analysis for Persian text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. 9–16.

[4] Ehsan Asgarian, Reza Saeedi, Ahmad Stiri, Behdad Bahmadi, and Hadi Ghaemi. [n. d.]. NLPTools. Available at https://wtlab.um.ac.ir.

[5] Ayoub Bagheri, Mohamad Saraee, and Franciska de Jong. 2013. Sentiment classification in Persian: Introducing a mutual information-based method for feature selection. In *Proceedings of the 2013 21st Iranian Conference on Electrical Engineering (ICEE'13)*. IEEE, Los Alamitos, CA, 1–6. DOI: https://doi.org/10.1109/IranianCEE.2013.6599671

[6] Ayoub Bagheri and Mohamad Saraee.2014. Persian sentiment analyzer: A framework based on a novel feature selection method. *International Journal of Artificial Intelligence* 12, 2, 115., http://www.scopus.com/inward/record.url?eid=2-s2.0-84926213301&partnerID=40&md5=69f8a916da14f0362bc2cbded411a2f3 (2014), 115–129

[7] Mohammad Basiri, Ahmad Nilchi, and Nasser Ghassem-Aghaee. 2014. A framework for sentiment analysis in persian. *Open Transactions on Information Processing* 1, 3, 1–14. DOI: https://doi.org/10.15764/OTIP.2014.03001

[8] Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee, and Ahmad-Mohamad SaraeeReza Naghsh-Nilchi. 2014. Exploiting reviewers' comment histories for sentiment analysis. *Journal of Information Science* 40, 3, 313–328. DOI: https://doi.org/10.1177/0165551514522734

[9] Mohammad Ehsan Basiri and Arman Kabiri. 2017. Sentence-level sentiment analysis in Persian. In *Proceedings of the 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA'17)*. IEEE, Los Alamitos, CA, 84–89. DOI: https://doi.org/10.1109/PRIA.2017.7983023

[10] Mohammad Ehsan Basiri, Ahmad Reza Naghsh-Nilchi, and Nasser Ghasem-Aghaee. 2014. Sentiment prediction based on Dempster-Shafer theory of evidence. *Mathematical Problems in Engineering* 2014, 1–13. http://www.hindawi.com/journals/mpe/2014/361201/abs/.

[11] Farah Benamara, Sabatier Irit, Carmine Cesarano, Napoli Federico, and Diego Reforgiato. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*. 1–4. DOI: https://doi.org/citeulike-article-id:9387439

[12] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems* 28, 2, 15–21. DOI: https://doi.org/10.1109/MIS.2013.30

[13] Kia Dashtipour, Amir Hussain, Qiang Zhou, Alexander Gelbukh, Ahmad YAHawalah, and Erik Cambria. 2016. PerSent: A freely available persian sentiment lexicon. In *Advances in Brain Inspired Cognitive Systems: 8th International Conference (BICS'16)*. Springer, 310–320. DOI: https://doi.org/10.1007/978-3-319-49685-6_28

[14] Andrea Ceron, Luigi Curini, and Stefano M. Iacus. 2015. Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United Sates and Italy. *Social Science Computer Review* 33, 1, 3–20. DOI: https://doi.org/10.1177/0894439314521983

[15] Andrea Ceron, Luigi Curini, Stefano M. Iacus, and Giuseppe Porro. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society* 16, 2, 340–358. DOI: https://doi.org/10.1177/1461444813480466

[16] Effat Golpar-Rabooki, Saghi-Al-Sadat Zarghamifar, and Jalal Rezaeenour. 2015. Feature extraction in opinion mining through Persian reviews. *Journal of Artificial Intelligence and Data Mining* 3, 2, 169–179. DOI: https://doi.org/10.5829/idosi.JAIDM.2015.03.02.06

[17] Mohammad Sadegh Hajmohammadi and Roliana Ibrahim. 2013. A SVM-based method for sentiment analysis in Persian language. In *Proceedings of SPIE 8768: International Conference on Graphic and Image Processing (ICGIP'12)*. 876838. DOI: https://doi.org/10.1117/12.2010940

[18] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5, 1, 1–167. DOI: https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[19] Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*.Cambridge University Press.

[20] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4, 1093–1113. DOI: https://doi.org/10.1016/j.asej.2014.04.011

[21] Shahla Nemati and Ahmad Reza Naghsh-Nilchi. 2016. Incorporating social media comments in affective video retrieval. *Journal of Information Science* 42, 4, 524–538. DOI: https://doi.org/10.1177/1045389X14554132

[22] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2, 1–135. DOI: https://doi.org/10.1561/1500000011

[23] Bo Pang, Lillian Lee, Harry Rd, and San Jose. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*. 79–86.

[24] Mohamad Saraee and Ayoub Bagheri. 2013. Feature selection methods in Persian sentiment analysis. In *Natural Language Processing and Information Systems*. Lecture Notes in Computer Science, Vol. 7934. Springer, 303–308. DOI: https://doi.org/10.1007/978-3-642-38824-8_29

[25] Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 3, 813–830. DOI: https://doi.org/10.1109/TKDE.2015.2485209

[26] Glenn Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

[27] Mohammadreza Shams, Azadeh Shakery, and Heshaam Faili. 2012. A non-parametric LDA-based induction method for sentiment analysis. In *Proceedings of the 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP'12)*. IEEE, Los Alamitos, CA, 216–221. DOI: https://doi.org/10.1109/AISP.2012.6313747

[28] Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 90–94.

[29] Venkatramana S. Subrahmanian and Diego Reforgiato. 2008. AVA: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems* 23, 4, 43–50.

[30] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37, 2, 267–307. DOI: https://doi.org/10.1162/COLI_a_00049

[31] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology* 63, 1, 163–173. DOI: https://doi.org/10.1002/asi.21662

[32] Mike Thelwall, Kevan Buckley, George Paltoglou, Marcin Skowron, David Garcia, Stephane Gobron, Junghyun Ahn, Arvid Kappas, Dennis Küster, and Janusz A. Holyst. 2013. Damping sentiment analysis in online communication: Discussions, monologs and dialogs. In *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, Vol. 7817. Springer, 1–12. DOI: https://doi.org/10.1007/978-3-642-37256-8_1

[33] Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering* 24, 4, 720–734. DOI: https://doi.org/10.1109/TKDE.2010.269

[34] Wenhao Zhang, Hua Xu, and Wei Wan. 2012. Weakness finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems With Applications* 39, 11, 10283–10291. DOI: https://doi.org/10.1016/j.eswa.2012.02.166

Received October 2017; revised January 2018; accepted March 2018