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

MOHAMMAD EHSAN BASIRI, Department of Computer Engineering, Shahrekord University, Iran
ARMAN KABIRI, Department of Computer Engineering, Shahrekord University, Iran

Lexicon-based sentiment analysis aims to address the problem of extracting people's opinions from their comments on the Web using a pre-defined lexicon of opinionated words. In contrast to machine learning approach, lexicon-based methods are domain-independent methods which do not need a large annotated training corpus and hence are faster. This makes the lexicon-based approach to be prevalent in the sentiment analysis community. However, the story is different for Persian language. In contrast to English, using lexicon-based method in Persian is a new discipline. There are rather limited resources available for sentiment analysis in Persian making the accuracy of the existing lexicon-based methods lower than that of other languages. In the current study, first an exhaustive investigation of lexicon-based method is performed. Then, two new resources are introduced in order to addresses the problem of resource scarcity for sentiment analysis in Persian; a carefully labeled lexicon of sentiment words, PerLex, and a new hand-made dataset of about 16000 rated documents, PerView. Moreover, a new hybrid method using both machine learning and lexicon-based approach is presented in which PerLex words are used to train the machine learning algorithm. Experiments are carried out on our new PerView dataset. Results indicate that the accuracy of PerLex is higher than that of the existing lexicons. Also, the results show that using PerLex significantly decreases the execution time of the proposed system in comparison to using existing lexicons. Moreover, the results demonstrate the excellence of using opinionated lexicon terms followed by bigrams as the features employed in machine learning 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 Persian Language by Lexicon Refining. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 1, 1, Article 1 (January 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].

Research on SA started in 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

Authors' addresses: Mohammad Ehsan Basiri, Department of Computer Engineering, Shahrekord University, Rahbar Blvd. Shahrekord, 105, Iran, basiri@eng.sku.ac.ir; Arman Kabiri, Department of Computer Engineering, Shahrekord University, Shahrekord, Iran, Arman.Kabiri94@gmail.com.

2375-4699/2018/1-ART1 $15.00
https://doi.org/10.1145/3195633

including extracting customers' attitudes toward a product or service [34], social media monitoring [14], analysis of political tweets [12], and predicting sales performance [33].

Existing approaches are classified into two main categories; corpus-based machine learning (ML) approach and lexicon-based method [30, 31]. Although machine learning 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 they were trained on [21, 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, 23]. Persian is spoken by more than a hundred million speakers in Iran, Afghanistan, and many states of the former Soviet Union [7]. However, 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 labelled sentiment words, while 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 paper introduces two new resources; a carefully labelled 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. In order to overcome their shortcomings, we remove all the words which 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 label are corrected. Finally, new missing words and phrases are added to the PerLex. More detailed information explaining how PerLex is developed is provided in Section 3.

Almost all previous studies on SA in Persian language suffer from the unavailability of a large dataset [7, 23]. PerView is introduced in this paper to fill this gap. This dataset contains about 16000 users' reviews and was labelled 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 Persian language use either n-grams features or semantic features [23]. In order to enhance the accuracy of corpus-based approach, a new hybrid method for SA in Persian is presented in this paper. This method is an ML-based method exploiting sentiment words listed in PerLex as the training features.

The remainder of the paper 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; Finally, Section 5 sets out the conclusion and future work.

2 RELATED WORK

Sentiment analysis has attracted a lot of attention in recent years, especially for widespread languages such as English [19] and numerous studies for SA has been published so far [11, 17, 18]. However, we do not intend to review SA studies on English language in this section. Instead, we will present a comprehensive literature review of SA studies focusing on Persian language.

The first published study on SA investigating Persian language has been reported by Shams et al. [27]. They suggested an unsupervised LDA-based method and evaluated their method on three

Table 1. Review of sentiment analysis 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 and Saraee [5]	Sentiment classification in Persian: Introducing a mutual information-based method for feature selection	2012	Limited to polarity detection. Ignores language-specific problems. Employs a relatively small dataset.
Hajmohammadi and Ibrahim [16]]	A SVM-based method for sentiment analysis in Persian language	2013	Limited to polarity detection. Only n-grams features are used.
Basiri et al. [7]	A Framework for Sentiment Analysis in Persian	2014	Limited to polarity detection.
Bagheri and Saraee [19]	Feature Selection Methods in Persian Sentiment Analysis	2013	Limited to polarity detection. Only utilized Naive Bayes. Employs a small and domain-specific dataset.
Rabooki et al. [15]	Feature extraction in opinion mining through Persian reviews	2015	Limited to polarity detection. Employs a very small dataset.
Alimardani and Aghaei [3]	Opinion Mining in Persian Language Using Supervised Algorithms	2015	Limited to polarity detection.
Dashtipour et al. [13]	PerSent: A Freely Available Persian Sentiment Lexicons	2016	Limited to polarity detection.
Basiri and Kabiri [9]	Sentence-level sentiment analysis in Persian	2017	Limited to sentence-level.

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 Persian language. Finally, the datasets on which they reported their results were relatively small.

Bagheri and Saraee [5] proposed a model for SA in Persian language employing 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 [16] 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-grams features were used for training the classifier.

Basiri et al. [7], proposed a framework for SA in 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 Persian language. Three ML algorithms, namely Naive Bayes, SMO, and J48 were compared with 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 [24] 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 Naive Bayes learning algorithm. Second, the dataset on which they evaluated their method was too small and domain-specific.

Rabooki et al. [15] proposed a feature extraction method for SA on Persian reviews. Specifically, they first created a lexicon and performed some pre-processing 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 the 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 [3] 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, Dastipour et al., published a freely available lexicon, PerSent, containing 1500 phrases and their 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 which 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 as neutral words.

In summary, all the above-mentioned studies have some similar limitations. They all have addressed the polarity detection problem. This could be considered as a limiting factor for SA methods since recent applications of SA need more detailed analysis such as rating predition. For example, in order 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 the studies have had less than 1000 records. This increases the randomness of results which in turn makes the reported results unreliable. Also, almost all of them are either pure lexicon-based or ML-based approaches. A review of the above-mentioned 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 [17]. This approach, compared to the machine learning 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.

The input to the system in Figure 1 is a review containing at least one sentence. The output of the system, on the other hand, is a 5-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 details in the following sub-sections.

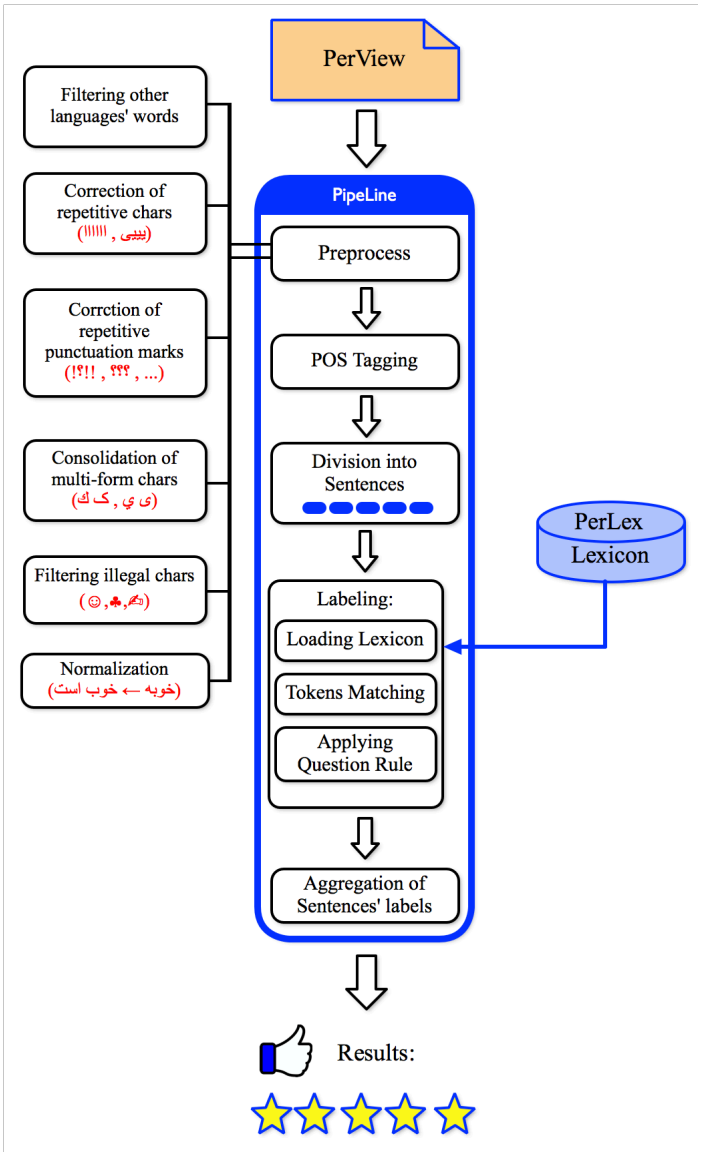

Fig. 1. The proposed lexicon-based approach.

3.1 Pre-processing

In the preprocessing phase of Figure 1, six preprocessing steps are applied; filtering, correction of repetitive chars, correction of punctuation, consolidation of multi-forms chars, filtering illegal chars, and normalization.

- *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.

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
م by هستم	خوشحال هستم	خوشحالم	I am happy
	مطمئن هستم	مطمئنم	I am sure
	ناامید هستم	ناامیدم	I am hopeless

- *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 are also removed.
- Consolidation multi-forms chars: Some letters in Persian have different Unicode. In particular, this problem occurs for those words containing letters like ی and ک(for 'i' and 'k', respectively).
- *Filtering illegal chars*: Some illegal characters 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 too 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 [2].

3.2 Score Detection

Although in recent years some methods have been proposed to detect sentiment score in English, all the reported studies for Persian SA have targeted the polarity detection problem [4, 9, 13, 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 pre-processing steps on the input review, we use part-of-speech (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 are considered as the score of the sentence. Besides, due

to the nature of 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). In order to address this difficulty, we use partial-matching instead of full-matching in labeling phase in Figure 1. Partial-matching approach first tries to divide the pre-known 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 much attention it deserves so far [8] and in most studies simple methods such as maximum of scores, majority voting, and simple averaging [6] are used. Recently, a formally defined method for score aggregation based on the Dempster-Shafer (DS) Theory of Evidence [26] has been proposed by Basiri, et al. [6]. It has been shown that the DS-based aggregation method clearly outperforms other aggregation methods used in SA [6]. There may be two reasons for this; firstly, the DS-based method takes all pieces of evidence into account, secondly, it preserves maximal agreements among the evidence [25].

In order 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) 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, 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 could be seen, this function is a *basic probability assignment* (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:

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

Since the DS rule of combination is both commutative and associative, we can iteratively apply the following equation in order 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 The Proposed hybrid method

In order 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 pre-processed 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 could 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 non-sentiment bearing unigrams, lexicon terms are determinants of the overall sentiment of the review. Hence, using lexicon terms should improve the performance of the system.

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

- *Occurrence filter*: In this step, those features occurring less than 10 times are considered as rare features. In order 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 training and validation phase which is necessary in ML-based methods [22].

3.5 Creating the PerLex[1]

Although some previous studies followed the lexicon-based approach, they addressed the problem simplistically [4, 7]. For example, almost all the existing lexicon-based approaches have used automatically translated lexicons from English. This rises different problems such as follows.

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

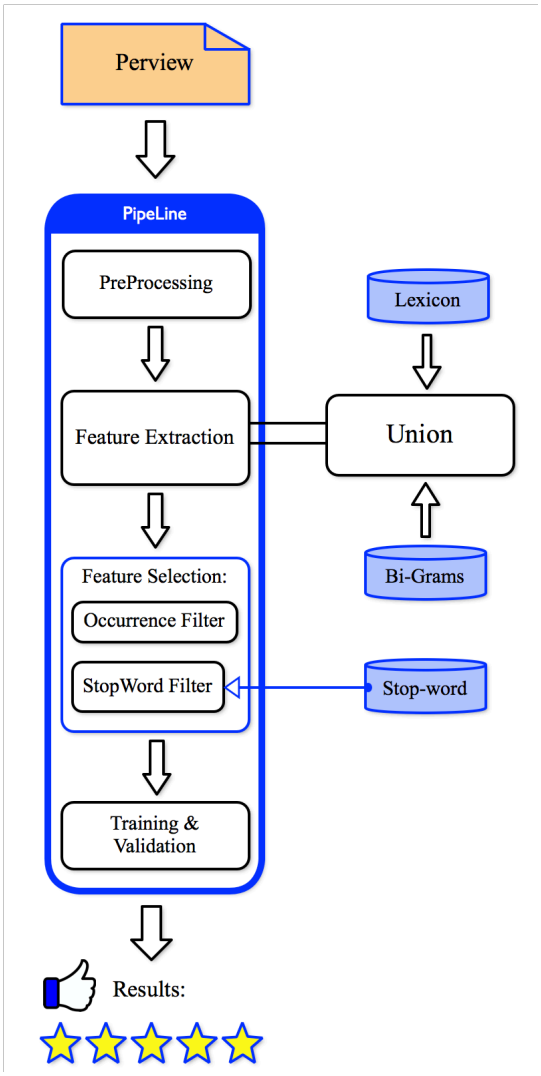

Fig. 2. The Proposed hybrid methods for SA.

- There is not always a one-to-one relationship between sentiment words of the source and destination languages. For example, a word like *ball* in English is not a sentiment word while in the 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 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 it exploits, the above-mentioned problems can significantly decrease their performance. In order to overcome such problems, having analyzed the performance of NRC and SentiStrength lexicons for SA in Persian, we design 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 post-processing step is performed. Introduced in our previous study, CNRC is the corrected version of NRC lexicon [9]. In order to create Persian SentiStrength, we first converted its score to 5-star scale and then used machine translation to translate its sentiment words into Persian. The process of creating the Adjective lexicon is as follows.

According to the previous studies on sentiment analysis, adjectives are one of the most important signs of sentiment [10, 29]. Keeping this fact in mind, we use the PerView dataset to extract the adjectives. First, a pre-processing 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 multi-form words.
- Normalization.

Having pre-processed the dataset as described above, 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 labelled manually in a 5-star scale.

After the above-mentioned steps, three Persian lexicons are intersected and the following post-processing steps are applied on the intersection result to form PerLex.

- *Pointless words removal*: in this step, all words that do not convey sentiment in 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 label are corrected.

Finally, the labels of each word in PerLex is calculated by employing Dempster-Shafer (DS) Theory of Evidence on all available three lexicons [8, 20].

4 RESULTS AND DISCUSSION

In order to show the effectiveness of PerLex, two series of experiments are conducted on 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:

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

4.1 Dataset and Evaluation Metrics

As mentioned earlier, one of the most 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

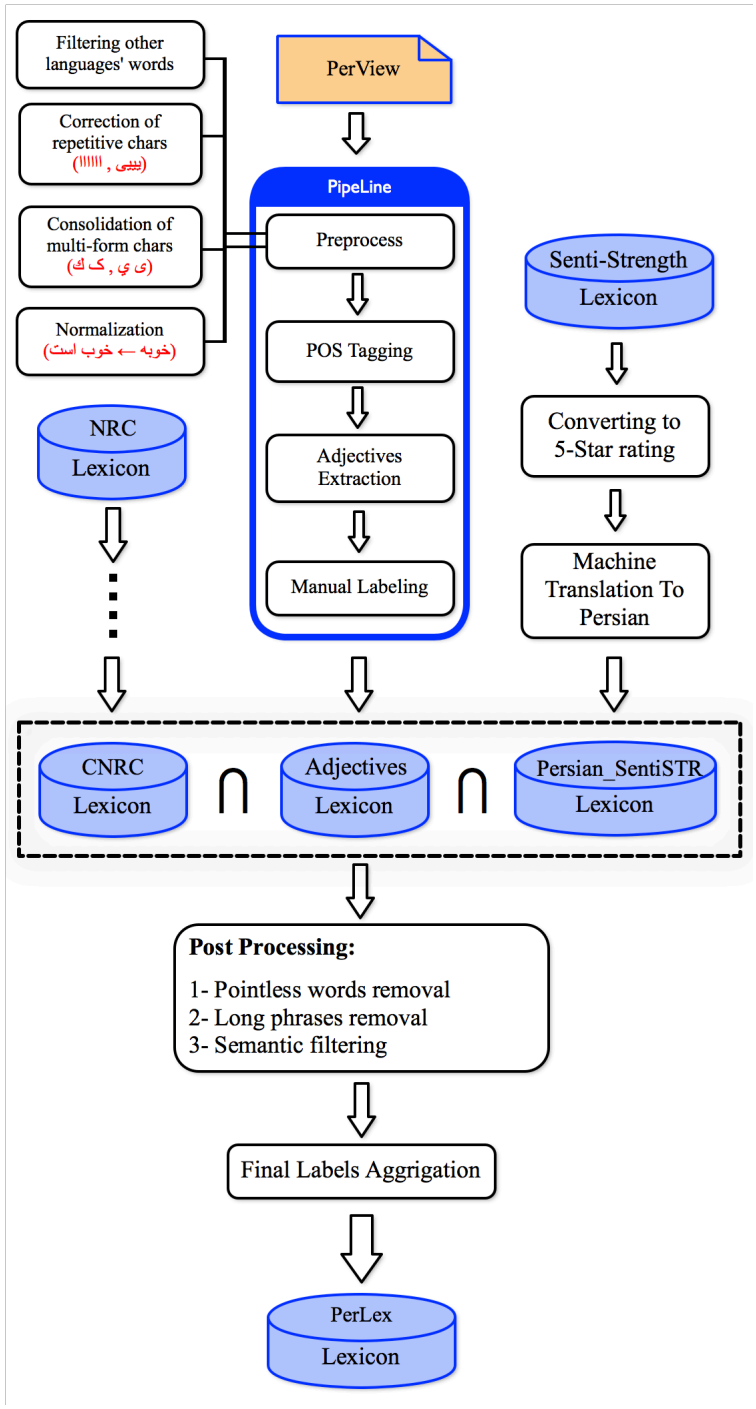

Fig. 3. The process of creating PerLex.

Table 3. Specifications of seven lexicons.

Lexicon	Number of words	Words' occurrence	Max occurrence	Unique words
NRC	9449	33%	9003	63%
CNRC	2697	38%	2229	0%
SentiStrength	2765	35%	2618	45%
Adjectives	1677	91%	3527	54%
LexiPers	6500	12%	4315	83%
PerSent	1470	17%	2709	77%
PerLex	174	98%	2229	0%

labelled dataset which can be used for document-level SA in Persian. This dataset contains 16000 user comments collected from *Digikala.com*, the biggest e-commerce start-up in Iran and the Middle East [1]. The PerView comments have been collected since July 2016 to February 2017. It contains customers' comments about digital equipment including cell-phones, cameras, and computer peripheral.

In our experiments, we have used five evaluation criteria; Precision (π), recall (ρ), F-Measure, accuracy, and MAE. These criteria are common in the previous studies [6, 7, 23] and are defined as follows:

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

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 predicted and real rate of the i^{th} test comment, respectively.

4.2 Differences of lexicons

As mentioned earlier, in this study, seven lexicons are tested, NRC, CNRC, SentiStrength, Adjectives, LexiPers, PerSent, and our proposed lexicon, PerLex. In order 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 percent of words occurred at least one time in the PerView dataset. The third column shows the frequency of the most frequent words of each lexicon and the forth column shows the percent 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.

In order to clarify the differences between the lexicons, their word clouds are depicted in Figure 4.

As could be seen in Figure 4, there are many frequent non-opinionated words in NRC. Such words can severely decrease the quality of this lexicon. On the other hand, the other six lexicons seem similar in that the most frequent words in all of them are positive words. In order to show the

(a)

(b)

(c)

(d)

(e)

(f)

(g)

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

Table 4. Specifications of seven lexicons.

Lexicon	Top ten 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

differences between lexicons more clearly, top ten frequent words of lexicons are listed in Table 4. Each row in this table shows 10 most frequent words sorted according to their frequency from left to right.

In order to answer the second research question, we have 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 Figure 5 and Figure 6.

Results obtained using 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 while their precisions are different. This shows that the false positive is different for different lexicons. Specifically, according to Figure 5, the 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. In order 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 3.7 Ghz CPU, 6 MB cache, and 16 GB RAM.

As could be seen in Figure 7, the execution time of the system using LexiPers and NRC is more than three times of using CNRC while as pointed out earlier, the performance of CNRC is significantly higher than that of those two lexicons. Moreover, PerLex has the lowest execution time which, beside 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 Naive Bayes classifier that has been previously shown to be a successful ML-based classifier for sentiment analysis [6, 8].

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

Persian language is the official language of Iran and more than a hundred million people around the world speak in Persian. However, sentiment analysis in Persian language is a young research field.

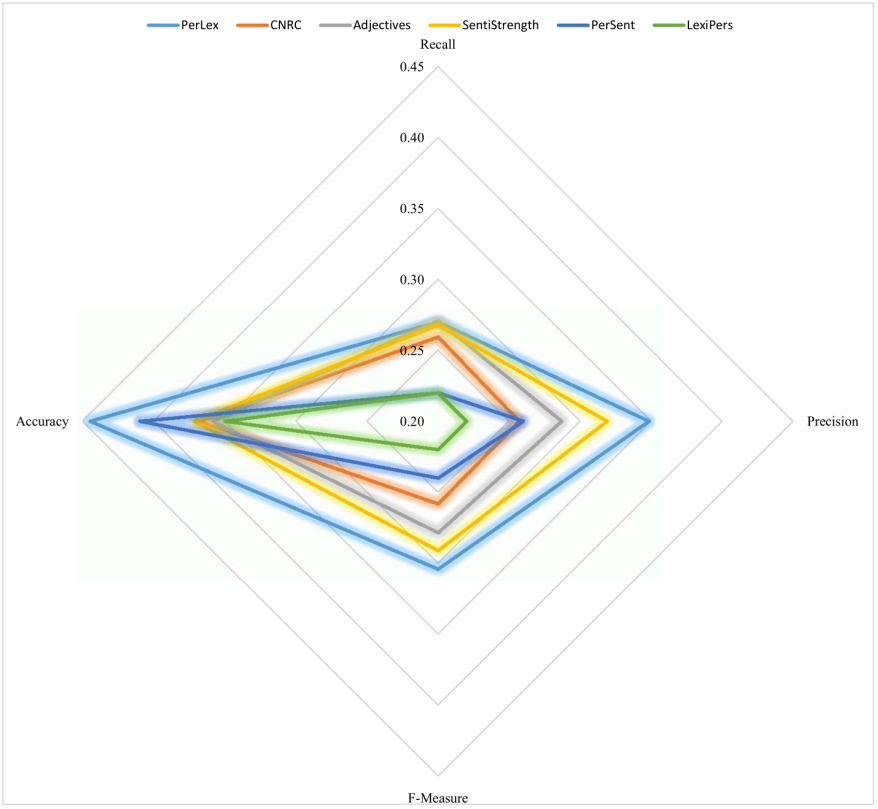

Fig. 5. The Comparison of the performance of using PerLex with other lexicons.

Although early studies preferred machine learning methods to lexicon-based approach, lexicon-based sentiment analysis methods have attracted an increasing attention in recent years. Compared to their counterparts in English, existing lexicon-based methods for sentiment analysis in Persian have a lower performance. In order to address this problem and to improve the performance of lexicon-based methods, an exhaustive investigation of lexicon-based method is performed in the current study. The investigation results showed that the main reason for the low performance of sentiment analysis in Persian language is the resource scarcity problem. In order to address this problem, two new resources are introduced; a carefully labelled lexicon of sentiment words, PerLex, and a new hand-made dataset of about 16000 rated documents, PerView.

In construction of PerLex, three lexicons are used and several pre-processing and post-processing steps are applied on the resulted lexicon. In order to show the performance of the PerLex, several experiments are carried out on PerView dataset. Results indicate that the accuracy of PerLex is higher than the existing lexicons. Moreover, a new hybrid method using both machine learning and lexicon-based approach is presented in which PerLex words are used to train the machine learning 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 may be suggested for future research. For example, improving the proposed lexicon using machine learning methods may be a promising suggestion. Another line of research

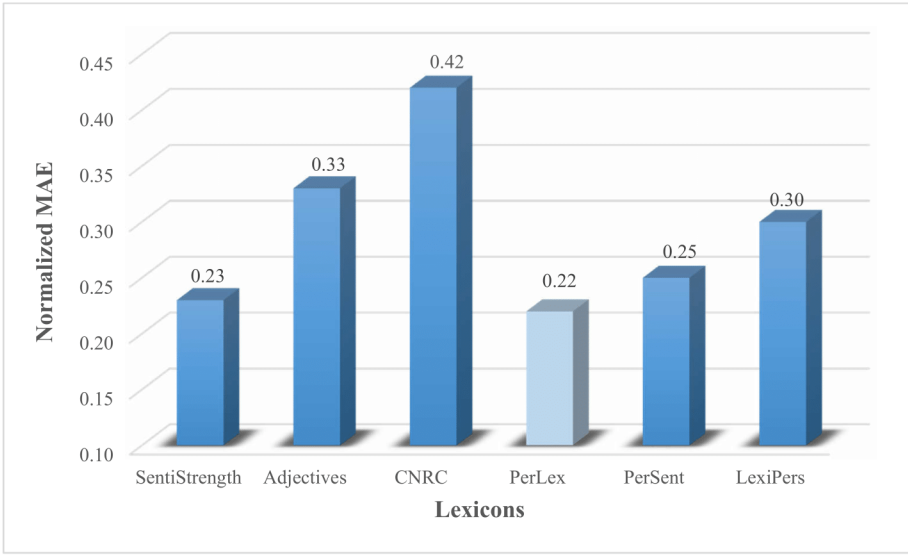

Fig. 6. The Comparison of the normalized MAE of using PerLex and other lexicons in the proposed system.

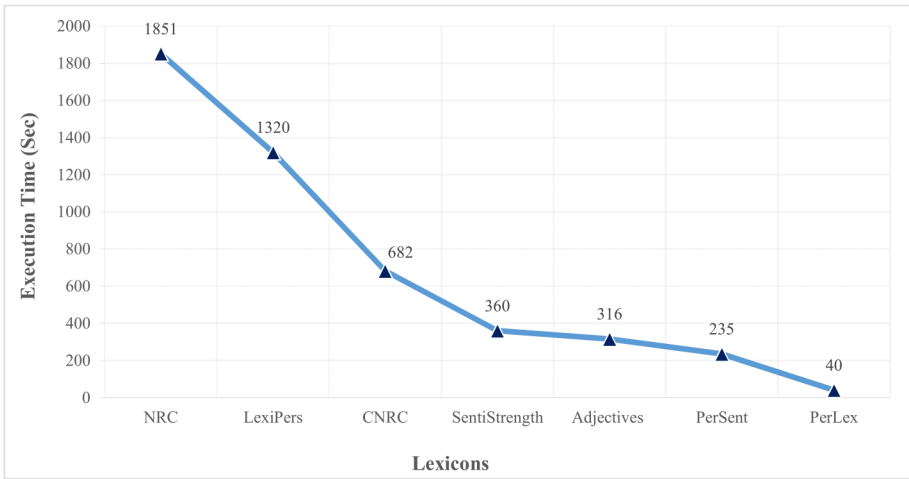

Fig. 7. The Comparison of the time performance of lexicon-based method using different lexicons.

may be employing more contextual features in the proposed hybrid method. Finally, enhancing the PerLex with contextual heuristic rules may be also 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.

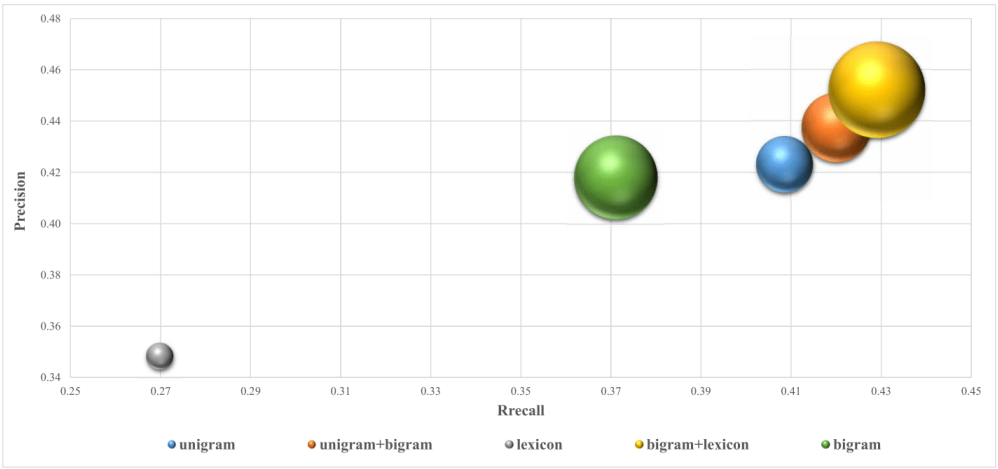

Fig. 8. The Comparison of the performance of lexicon-based, hybrid, and ML-based method. Ball size shows (1 - MAE) of using corresponding classifier.

This work has been financially supported by the research deputy of Shahrekord University. The grant number is 95GRN1M1874.

REFERENCES

[1] Digikala. (2017). http://www.digikala.com
[2] NLPTools. (2012). https://wtlab.um.ac.ir
[3] Saeedeh Alimardani and Abdollah Aghaei. 2015. Opinion Mining in Persian Language Using Supervised Algorithms. *Journal of Information Systems and Telecommunication* 3 (2015).
[4] Fatemeh Amiri, Simon Scerri, Mohammad H Khodashahi, Fraunhofer Iais, and Sankt Augustin. 2015. Lexicon-based Sentiment Analysis for Persian Text. (2015), 9–16.
[5] Ayoub Bagheri, Mohamad Saraee, and Franciska de Jong. 2013. Sentiment classification in Persian: Introducing a mutual information-based method for feature selection. In *2013 21st Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 1–6. https://doi.org/10.1109/IranianCEE.2013.6599671
[6] 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 (2014). https://doi.org/10.1155/2014/361201
[7] Mohammad Ehsan Basiri, Ahmad-Reza Nilchi, and Nasser Ghassem-Aghaee. 2014. A Framework for Sentiment Analysis in Persian. *Open Transactions on Information Processing* 1, 3 (2014), 1–14. 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 (2014), 313–328. https://doi.org/10.1177/0165551514522734
[9] Mohammad Ehsan Basiri and Arman Kabiri. 2017. Sentence-level sentiment analysis in Persian. In *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, IEEE (Ed.). IEEE, Shahrekord, 84–89. https://doi.org/10.1109/PRIA.2017.7983023
[10] Farah Benamara, Sabatier Irit, Carmine Cesarano, Napoli Federico, and Diego Reforgiato. 2007. Sentiment Analysis : Adjectives and Adverbs are better than Adjectives Alone. In *In Proc of Int Conf on Weblogs and Social Media*. CO, USA, 1–4. https://doi.org/citeulike-article-id:9387439
[11] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* 28, 2 (mar 2013), 15–21. https://doi.org/10.1109/MIS.2013.30
[12] Andrea Ceron, Luigi Curini, and Stefano M. Iacus. 2015. Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters–Evidence From the United States and Italy. *Social Science Computer Review* 33, 1 (feb 2015), 3–20. https://doi.org/10.1177/0894439314521983

[13] Kia dashtipour, Amir Hussain, Qiang Zhou, Alexander Gelbukh, Ahmad YA Hawalah, and Erik Cambria. 2016. PerSent: A Freely Available Persian Sentiment Lexicon. In *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, November 28-30, 2016*. Springer, 310–320. https://doi.org/10.1007/978-3-319-49685-6_28

[14] 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 & Society* 16, 2 (mar 2014), 340–358. https://doi.org/10.1177/1461444813480466

[15] 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 (2015). https://doi.org/10.5829/idosi.JAIDM.2015.03.02.06

[16] Mohammad Sadegh Hajmohammadi and Roliana Ibrahim. 2013. A SVM-based method for sentiment analysis in Persian language, Zeng Zhu (Ed.). 876838. https://doi.org/10.1117/12.2010940

[17] Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (may 2012), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[18] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions* (first ed.). Cambridge University Press.

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

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

[21] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* 2, 12 (2008), 1–135. https://doi.org/10.1561/1500000011

[22] Bo Pang, Lillian Lee, Harry Rd, and San Jose. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. July (2002), 79–86.

[23] Mohamad Saraee and Ayoub Bagheri. 2013. Feature Selection Methods in Persian Sentiment Analysis. 303–308. https://doi.org/10.1007/978-3-642-38824-8_29

[24] Mohamad Saraee and Ayoub Bagheri. 2014. Persian sentiment analyzer: A framework based on a novel feature selection method. *International Journal of Artificial Intelligence* 12,2,115, (2014), 115–129.

[25] Kim Schouten and Flavius Frasincar. 2016. Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2016), 813–830. https://doi.org/10.1109/TKDE.2015.2485209

[26] Glenn Shafer. 1976. *A mathematical theory of evidence*. Princeton: Princeton university press.

[27] Mohammadreza Shams, Azadeh Shakery, and Heshaam Faili. 2012. A non-parametric LDA-based induction method for sentiment analysis. In *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*. IEEE, 216–221. 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*. Association for Computational Linguistics, Jeju, Republic of Korea, 90–94.

[29] Venkatramana S. Subrahmanian and Diego Reforgiato. 2008. AVA: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems* 23, 4 (2008), 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 (jun 2011), 267–307. 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 (jan 2012), 163–173. 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. 1–12. 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 (apr 2012), 720–734. 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 (2012), 10283–10291. https://doi.org/10.1016/j.eswa.2012.02.166