Research Paper

JIS

Journal of Information Science
1–17
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0165551519827886
journals.sagepub.com/home/jis
SAGE

HOMPer: A new hybrid system for opinion mining in the Persian language

Mohammad Ehsan Basiri
Department of Computer Engineering, Shahrekord University, Iran

Arman Kabiri
Department of Computer Engineering, Shahrekord University, Iran

Abstract

Opinion mining is a subfield of data mining and natural language processing that concerns with extracting users' opinion and attitude towards products or services from their comments on the Web. Persian opinion mining, in contrast to its counterpart in English, is a totally new field of study and hence, it has not received the attention it deserves. Existing methods for opinion mining in the Persian language may be classified into machine learning– and lexicon-based approaches. These methods have been proposed and successfully used for polarity-detection problem. However, when they should be used for more complex tasks like rating prediction, their results are not desirable. In this study, first an exhaustive investigation of machine learning– and lexicon-based methods is performed. Then, a new hybrid method is proposed for rating-prediction problem in the Persian language. Finally, the effect of machine learning component, feature-selection method, normalisation method and combination level are investigated. The experimental results on a large data set containing 16,000 Persian customers' review show that this proposed system achieves higher performance in comparison to Naïve Bayes algorithm and a pure lexicon-based method. Moreover, results demonstrate that this proposed method may also be successfully used for polarity detection.

Keywords

Data mining; machine learning; natural language processing; opinion mining; Persian Language; sentiment analysis

1. Introduction

With the rapid development of social media, customers' comments have been converted to a valuable source of knowledge. These comments are now emerging for more different types of products than before. For example, digital equipment, sport products, books and cosmetics are from the most reviewed products by customers. Such customers' comments are beneficial for users to make the purchase decision and for business managers to make business decision by analysing customers' feedbacks. Therefore, mining such textual sources is of great importance from different points of view.

Opinion mining is a subfield of data mining (DM), natural language processing (NLP) and information retrieval (IR) that aims to extract valuable information from customers comments on the Web [1,2]. Although the unstructured nature and informal style of writing make opinion mining problems more complicated than other text processing problems, its vast spectrum of applications has converted opinion mining to a hot research topic [2].

Todays' opinion mining applications range from social media [3] and electoral campaign [4] monitoring to multimodal [5,6] and group-centric recommender systems [7]. In addition to these applications, predicting wins in sport games [8], predicting happiness [9] and terrorism detection [10] are among the newer applications of opinion mining.

Regardless of which application is selected, the polarity detection [11,12] and rating prediction [2,13] are two vital tasks in opinion mining. Categorising an opinionated text into either positive or negative categories is the main goal of

Corresponding author:
Mohammad Ehsan Basiri, Department of Computer Engineering, Shahrekord University, Rahbar Bolvar, Shahrekord, 64165478, Iran.
Email: basiri.sku@gmail.com

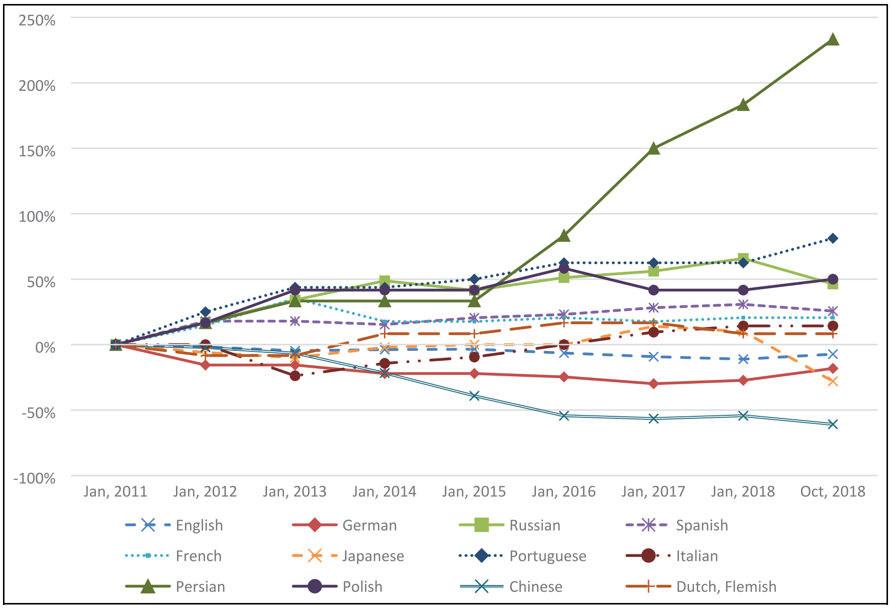


Figure 1. The growth rate of main 12 languages on the Web from 2011 to October 2018 [17]. Each point shows the growth with respect to the starting point of the chart.

Table 1. Trends of the usage of content languages for websites since January 2011 [17].

	English	German	Russian	Spanish	French	Japanese	Portuguese	Italian	Persian	Polish	Chinese	Flemish
January, 2011	57.6%	7.7%	4.1%	3.9%	3.4%	5.0%	1.6%	2.1%	0.6%	1.2%	4.6%	1.2%
October, 2018	53.4%	6.3%	6.0%	4.9%	4.1%	3.6%	2.9%	2.4%	2.0%	1.8%	1.8%	1.3%

the former, while assigning a real value to an opinionated text is the purpose of the latter. Most of existing studies for opinion mining in the Persian language have concentrated on the polarity-detection problem [14,15].

Starting at 2012, Persian opinion mining has a short history [16], and it has not received the attention it deserves. The Persian language is now spoken by more than a 100 million speakers around the world, mainly in Iran. Persian is used by 2% of all the websites whose content language is known, and it is now the ninth most widely used language in the Web [17]. Moreover, as depicted in Figure 1, the growth rate of the Persian language on the Web is much faster than other main languages. Table 1 compares the usage percentage of 12 most widely used languages of the Web for January 2011 and October 2018. As can be seen in the table, Persian is now the ninth most commonly used language, while in 2011, it was the 14th language after Turkish and Arabic languages which are not among the top-10 languages in 2018.

With regard to the classification method they employ, opinion mining approaches may be classified into machine learning (ML)- and lexicon-based methods [1,18,19]. For Persian opinion mining, both approaches have been successfully used. However, when the training data set is small or when the lexicon is imprecise, the performance might not be satisfying. In the first case, the small size of the training data set makes the classifier general, so that the performance degrades when it faces new unseen instances. For resolving the second problem, creating precise lexicons is of great importance. However, the more precise the lexicon is, the more likely it is that the system becomes over-fitted.

For the Persian language, as will be described in the next section, both ML- and lexicon-based methods are recently exploited. However, for the first category, the existing works including Shams et al. [16], Bagheri et al. [15] and Hajmohammadi and Ibrahim [20] are only applicable for polarity-detection problem and suffer from the small size of the training data set. For the second category, the works by Basiri et al. [19], Golpar-Rabooki et al. [21] and Alimardani and Aghaei [22] were all designed for polarity detection and were evaluated on small data sets.

In order to address the aforementioned problems, in this study, we first investigate different lexicons and analyse their strength and shortcoming to create a precise lexicon. Then, we proposed a new hybrid method, HOMPer, which employs opinionated lexicon terms besides common n-gram features in a ML-based method. This proposed method is distinguished from previous Persian opinion mining methods as follows:

- A new hybrid method is proposed in which both lexicon-based and linguistic features are used.
- A large document-level data set is created and used for training the proposed system.
- A new lexicon is created and used in the hybrid method.
- Two feature-selection methods are used in the HOMPer.
- Different combination levels are tested in the process of designing HOMPer.

The rest of the article is organised as follows. A complete review of Persian opinion mining studies is presented in section 'Related work'. The proposed system is introduced in section 3. The experimental setup and results are described in section 4. Finally, the conclusion and future work are discussed in section 5.

2. Related work

In this section, first a brief overview of multi-lingual opinion mining studies is presented and then, a comprehensive overview of Persian opinion mining is given.

2.1. Multi-lingual opinion mining

Research on opinion mining, also known as sentiment analysis, started in early 2000s and most of the reported studies have addressed the problem for widespread languages such as English, Chinese and Arabic [7,11,23]. Instead of providing a detailed review of the field, in this section, a review of multi-lingual studies is presented.

In designing multi-lingual opinion mining systems, the following approaches may be followed [24]:

- Translating the unseen test texts into the source language and using a classifier trained on the source language to classify them.
- Translating the training corpus and building a classifier trained on the translated corpus.
- Translating the lexicon of the target language and building a lexicon-based classifier to use it.

Denecke [25] used the first two of the aforementioned methods for determining polarity of German movie reviews, and Cieliebak et al. [26] proposed a new corpus for German opinion mining from 10,000 tweets. Ishijima et al. [27] analysed sentiment towards the Japanese economy, and Miyakawa et al. [28] proposed a quality table–based method for sentiment expression and word identification in the Japanese language.

As one the first studies on Arabic opinion mining, Abbasi et al. [29] utilised languages stylistic and syntactic features to classify movie reviews and Web forum postings. Aldayel and Azmi [30] recommended a hybrid of semantic orientation and ML techniques to classify Arabic tweets. Similar approaches have been proposed for other languages such as Turkish [31], Urdu [32] and Chinese [7].

2.2. Persian opinion mining

Shams et al. [16] were the first who reported the study on Persian opinion mining in 2012. Their suggested method was an unsupervised latent Dirichlet allocation (LDA)-based method evaluated on three manually created data sets. They compared their proposed method with a baseline algorithm and reported a 9% improvement. However, this method was only applicable for polarity detection, and it did not consider language-related problems of Persian opinion mining.

Later, Bagheri et al. [15] proposed a Naïve Bayes (NB) model for Persian opinion mining. A mutual information-based feature-selection method was also used in their model, and it was evaluated on a manually gathered collection of cell-phone reviews. This study only considered the polarity-detection problem and did not address the Persian language directly, so it suffered from the same limitations as that of Shams et al. [16].

Later on, using support vector machine (SVM) and NB, Hajmohammadi and Ibrahim [20] proposed another ML-based strategy and compared their models on a data set of online Persian movie reviews. Like previous studies, this study is also restricted to the polarity-detection problem.

As the first lexicon-based approach, Basiri et al. [19] proposed a framework for Persian opinion mining. This study was the first Persian opinion mining study, in which some of the Persian text processing difficulties were addressed [19]. Compared with the NB, sequential minimal optimization (SMO) and J48, this lexicon-based method showed a better

Table 2. Existing studies for Persian opinion mining from 2012 to 2018.

Author	Year	Sentiment classification method	Problem addressed
Shams et al. [16]	2012	Machine learning	Polarity detection
Bagheri et al. [15]	2013	Machine learning	Polarity detection
Hajmohammadi and Ibrahim [20]	2015	Machine learning	Polarity detection
Bagheri [18]	2014	Machine learning	Polarity detection
Basiri et al. [19]	2014	Lexicon based	Polarity detection
Golpar-Rabooki et al. [21]	2015	Lexicon based	Polarity detection
Alimardani and Aghaei [22]	2015	Lexicon based	Polarity detection
Basiri and Kabiri [1]	2017	Lexicon based	Polarity detection
Amiri et al. [14]	2015	Lexicon based	Polarity detection
Asgarian et al. [36]	2018	Machine learning	Polarity detection
Basiri and Kabiri [37]	2018	Machine learning + lexicon-based	Rating prediction + polarity detection
Basiri and Kabiri [35]	2017	Lexicon based	Rating prediction + polarity detection
Dashtipour et al. [38]	2016	Lexicon based	Polarity detection

performance in terms of mean absolute error (MAE) and F-score [19]. However, this study was also limited to the polarity-detection problem.

In order to improve their previously proposed ML-based method, Bagheri [18] investigated different feature-selection methods to address some Persian text processing difficulties. However, the first limitation of this study was that it only utilised NB learning algorithm. Another limitation was that although this was a ML-based method, the data set on which it was evaluated was too small and domain-specific.

Recently, Golpar-Rabooki et al. [21] proposed a feature extraction method for mining Persian reviews. In this study, after creating a lexicon, some pre-processing steps were performed on the reviews. Then, two different feature extraction methods were used. Similar to the previous reported research on Persian opinion mining, this study is also limited to the polarity detection. Moreover, the small size (only 340 reviews) of the data set used for evaluation is another limiting factor of this study.

More recently, using the combination of Persian SentiWordNet and three ML algorithms, Alimardani and Aghaei [22] proposed a new method for Persian opinion mining. The first step in this method was creating a Persian SentiWordNet using the existing English SentiWordNet [33] and Persian WordNet [34]. Then, this Persian SentiWordNet was used to weight the features. However, this study was also limited to polarity-detection problem.

Basiri and Kabiri [1] proposed a sentence-level sentiment analysis method in Persian, which is considered as the first sentence-level method for Persian opinion mining. They showed that direct translation of lexical resources is not sufficient for Persian opinion mining [35]. In a more recent study, Asgarian et al. [36] investigated the impact of different features on sentiment polarity detection of Persian reviews. They showed that sentiment lexicon quality has a critical role in the overall quality of sentiment classification. Finally, Basiri and Kabiri [37] improved their previous method by lexicon refining and emphasised the results of Asgarian et al. that the quality of sentiment lexicon has a great impact on the quality of both polarity-detection and rating-prediction systems. A summary of most important studies on Persian opinion mining is shown in Table 2.

The main difference between the existing ML-based studies for Persian opinion mining and this study is that we exploit lexicon features in the ML-based component while the previous studies mainly rely on the n-gram features. Another difference is that the proposed method uses a lexicon-based component in which different lexicon terms are used. The main difference between the existing lexicon-based methods and the proposed method is that existing methods usually utilise the lexicon to assign the polarity or rating score of a review, while the proposed method uses lexicon terms as training features for the ML component.

3. Proposed system

The baseline method that is improved in this study is the lexicon-based method for opinion mining in Persian language recently proposed by Basiri and Kabiri [35]. In order to improve the performance of the baseline method, three types of solutions are used as follows: addressing language-related problems, utilising the advantages of lexicon-based approaches and exploiting the benefits of ML-based methods. The details of these steps are shown in Figure 2.

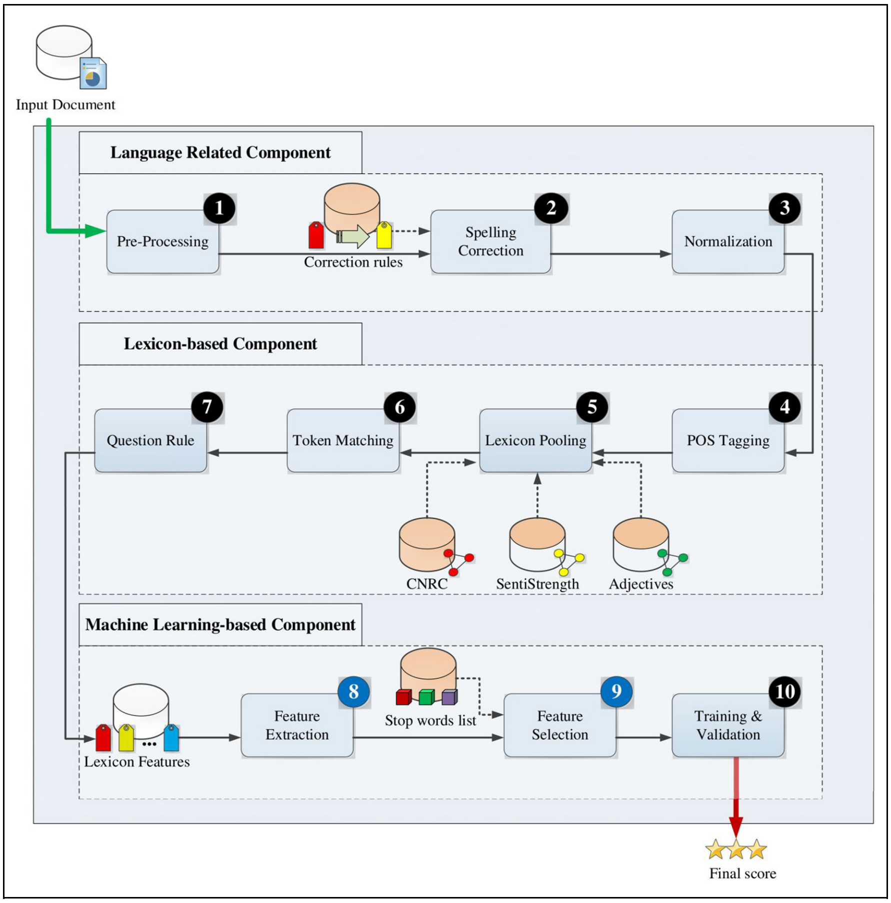


Figure 2. The overall view of the proposed system.

3.1. Solving language-related problems

The following language-related problems of Persian opinion mining are addressed in this study. Although there are some available tools for Persian text processing, they are not suitable for solving the problems addressed in this study. Hence, we implemented the following solutions in Java language.

- Different forms of writing: this problem arises either when imported Arabic letters and sounds are used, or when some ambiguous Unicode characters exist in the document [19]. As an example of the problem of imported Arabic letters, the 'Tanvin' and 'Hamza' characters may be considered [39]. These characters may be written or ignored in a word containing them. This increases the possible number of ways of writing a word to 48 in some cases [19]. Module 1 in Figure 2 is used to address this problem as well as two more pre-processing steps: filtering other languages' words and filtering illegal characters. These filtering steps reduce the final feature space of the ML component of the proposed system and hence, decrease its computational complexity.
- Spelling correction: In order to correct different spelling errors, module 2 in Figure 2 is used. In this module, three tasks are performed; correction of repetitive chars (used for emphasis), correction of repeated punctuation, and consolidation of multi-forms characters. Multi-form characters are those characters, in contrast to Arabic, that are written interchangeably. For example, the 'T' character has two possible forms, while each of the 'S' and 'Z' have four possible forms of writing. This module affects the quality of the token-matching module of the lexicon-based component of the proposed system and hence, increases its accuracy.
- Normalisation is necessary, as Persian Web texts frequently contain non-standard spellings resulted from the fact that Persian Web users tend to write words as they are pronounced in daily conversations [1]. An example of such informal non-standard spelling was given in Basiri et al. [39]. Although the review in that example is a short text containing only two sentences, nine informal words exist in the review. This shows the importance of the

normalisation module in the proposed system. Correcting such informal texts is necessary, since existing lexicons usually contain formal words.

3.2. Lexicon-based component

In the first module of the lexicon-based component (i.e. Module 4), part-of-speech (POS) tags are used to specify verbs. These verbs are then used as indicators of sentence boundaries. The rationale behind this choice is that usually each verb has an independent meaning. In order to find the POS tags, NLPTool is used in this study [40].

The next module in the lexicon-based component is the lexicon pooling component. In order to utilise the benefits of lexicon-based methods in the proposed system, we analyse and exploit three lexicons for Persian opinion mining. The first lexicon, National Research Council Canada (CNRC), is the corrected version of National Research Council (NRC) [41]. CNRC has 2698 terms, while the original size of NRC is 9450. This decrease in the size is obtained by applying the following three steps on the NRC:

- Pointless words removal: in this step, all non-opinionated words of the Persian language are removed.
- Long phrases removal: in this step, translated long phrases in Persian are removed, as they are never matched with any phrase in the data set.
- Semantic filtering: in this step, all terms with incorrect labels are corrected.

The second lexicon, SentiStrength [13], has been vastly used for English opinion mining [13,42–44]. The Persian version of this lexicon contains 2765 words. The third lexicon, Adjective [19], was created by extracting the adjectives from about 160,000 short comments in Persian [19]. It has been previously shown that adjectives are one of the most important signs of opinion [45,46]. These lexicons are intersected and the maximum scores of common terms are used in the resulted lexicon.

In the token-matching module, for each sentence, all words are searched against the resulted lexicon of the previous step. Then, the scores of the matched words are averaged and used as the overall score for the sentence. In the matching process, partial matching is used instead of full-matching. Partial matching separates the pre-known suffixes from the main word and the separated word is then looked up in the lexicon. This process is similar to what usually is performed in a typical stemming phase. The partial matching seems appropriate for languages like Persian, in which numerous suffixes associate with words. This vast use of suffixes decreases the chance of full matching of opinionated words with lexicon terms.

The final step in the lexicon-based component of Figure 2 is using the question rule module. This module is used to filter sentences with the indicative mood. The main reason for applying this filter is that usually interrogative sentences are not reliable sources on which we can base our decision. For instance, although the sentence ‘Do you think this camera is a reliable camera?’ contains a positive adjective, ‘reliable’, its writer does not intend to express a positive idea about that camera, and it is just a question. In order to specify the mood of a sentence, we consider punctuation as a reliable sign.

3.3. ML component

In this component, a feature-level combination of ML and lexicon-based method is used. Specifically, in the feature-extraction module, bigrams are extracted from the data set and combined with the lexicon terms obtained from the lexicon-based component.

Bigram features have been previously shown to be efficient features for opinion mining tasks, especially when combined with unigram features [47]. However, in this study, we hypothesise that replacing unigrams by lexicon terms can improve the performance of the ML component. The rationale behind this choice is that in contrast to non-sentiment bearing general unigrams, lexicon terms are determinant factors for the overall sentiment of a review.

By extracting bigrams and lexicon-based features, the size of feature space is still too large for a typical ML algorithm. This large feature space not only makes the training phase of the proposed hybrid method slow but also reduces the overall accuracy of the system [15,48]. In order to overcome these difficulties, a feature-selection step is necessary [48–50].

In the feature-selection module, the following steps are used as follows:

- Frequency filter: in this step, features are filtered according to their frequency of occurrence. Specifically, those features with frequency less than 10 are considered as rare features and hence are removed. The main goal of this filter is reducing the size of the feature space, as well as increasing the precision of the proposed hybrid system.

Table 3. The contingency table for computing TP, TN, FN and FP values for polarity detection.

		Label in the data set	
Classifier's decision		Yes	No
	Yes	TP	FP
	No	FN	TN

TP: true positive; TN: true negative; FP: false positive: FN: false negative.

- Stop-word filter: stop words are those frequent words that do not contribute to any opinion. Therefore, inclusion of such words in the feature space may decrease the overall performance of the system. Using a predefined list of Persian stop words, we remove such words in module 9.

The final module is a necessary step for each ML method. In the proposed system, Naïve Bayes algorithm is used as the classifier component of the hybrid method [6]. This algorithm has been shown to outperform other ML algorithm for opinion mining [13,44].

4. Experiments and results

In order to show the effectiveness of the proposed HOMPer system, both polarity-detection and rating-prediction problems are considered. The polarity-detection problem is a binary classification problem, in which a positive or negative label is assigned to each review. Rating prediction, on the other hand, is a multi-class classification problem in which a five-star rate is assigned to each review. The former is clearly an easier task, and solutions to it have higher accuracy in comparison to those proposed for solving the latter.

In the experiments, a large manually labelled data set containing about 16,000 customer comments from Digikala.com (http://www.digikala.com) [51], collected since July 2016 to February 2017, is used [1]. This data set contains different types of customer reviews from cell-phones to computer peripheral.

4.1. Evaluation criteria

Different metrics are used in opinion mining studies from which the following criteria are of more interest [2,15,44,52]; Precision (π), recall (ρ), F-measure, accuracy and MAE. The following equations show the relationship between these measures

$$\pi = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\rho = \tfrac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F\text{−measure} = \frac{2 \times \pi \times \rho}{\pi + \rho} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |p_i - r_i|}{n} \tag{5}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively [44]. These measures, for the polarity-detection problem, are obtained according to Table 3. For the rating-prediction problem, each measure is calculated separately for each class and then their average is used in equations (1)–(5). For MAE, n is the number of test comments, and p_i and r_i are predicted and real rate of the ith test comment, respectively.

Table 4. Comparison of different lexicons for polarity-detection problem.

Lexicon	Recall	Precision	F-Measure	Accuracy	MAE
SentiStrength	0.56	0.56	0.56	0.76	0.24
CNRC	0.55	0.55	0.55	0.75	0.25
Adjective	0.55	0.56	0.56	0.77	0.23

MAE: mean absolute error; CNRC: national Research Council Canada.
Note: NRC is the National Council Canada and CNRC is a data set (the corrected version of NRC).

Table 5. Comparison of different lexicons for rating-prediction problem.

Lexicon	Recall	Precision	F-Measure	Accuracy	MAE
SentiStrength	0.27	0.32	0.29	0.37	0.93
CNRC	0.26	0.26	0.26	0.37	1.12
Adjective	0.27	0.29	0.28	0.36	1.03

MAE: mean absolute error; CNRC: national Research Council Canada.
Note: NRC is the National Council Canada and CNRC is a data set (the corrected version of NRC).

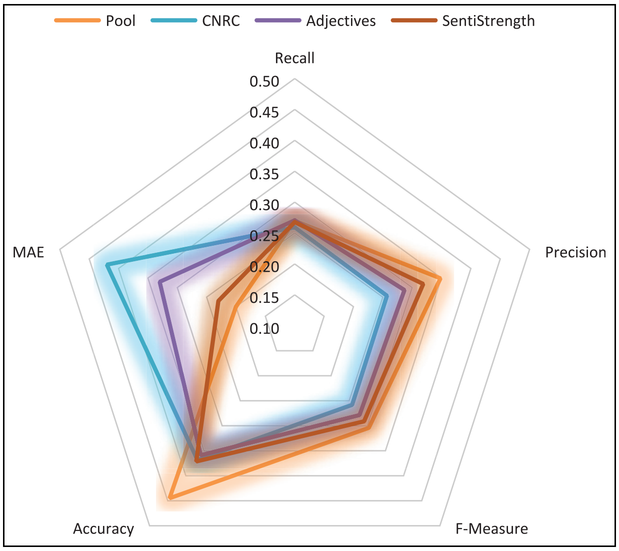


Figure 3. Comparison of performance of three different lexicons and their union (Pool) for rating-prediction problem (MAE is normalised).

4.2. Lexicon-based component results

In order to show the differences between three lexicons described in 'Lexicon-based component', we compare their performance on the aforementioned data set in Tables 4 and 5.

As can be seen in Tables 4 and 5, the overall performance of all lexicons for polarity detection is higher than their performance on rating-prediction problem. This is expected, since binary classification of opinions is a simpler task than multi-class prediction. Previous studies on opinion mining have also shown that [2,44].

Another point in Table 4 is that SentiStrength has a slightly better performance with respect to all performance measures and for both tasks. However, this does not imply that other lexicons are not useful because there may be some opinionated words in one lexicon while the other does not contain that word or the label assigned to a common word in one lexicon may be not as accurate as that of another lexicon. In order to show this, the performance of three lexicons are compared with the performance of their combination (module 5 of Figure 1) in Figure 3.

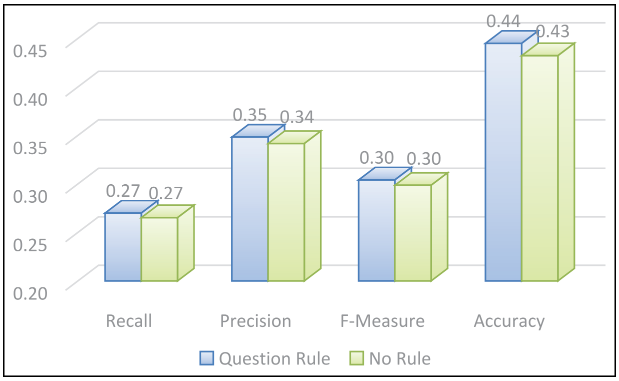


Figure 4. The effect of using question rule module on the overall performance of the lexicon-based component of the proposed system for rating-prediction problem.

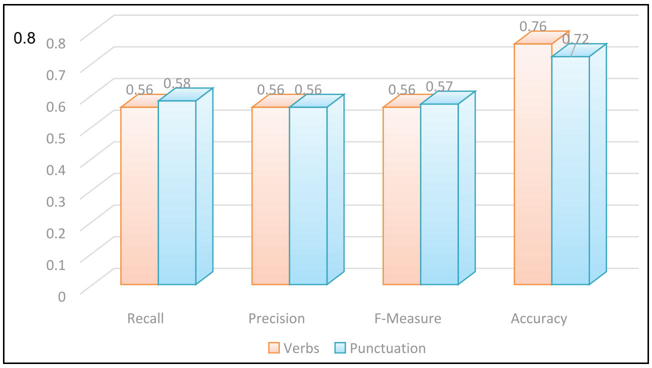


Figure 5. Comparison of using punctuation instead of verbs as sentence boundary indicator for polarity-detection problem in the lexicon-based component of the proposed HOMPer system.

As can be seen in Figure 3, with respect to all five measures, the Pooled lexicon outperforms all lexicons. In other word, when the lexicons are aggregated, their performance is higher than that of their isolated one. This shows that there is useful information in each of three lexicons that can be exploited when they are aggregated. In other words, the positive effect of using non-unique lexicon words and smoothing the rate of common words is higher than the negative effect of ignoring unique words.

The final module of the lexicon-based component in the proposed system is used to filter sentences with indicative mood. As mentioned earlier, this module is useful, since usually interrogative sentences are not reliable sources on which we can base our decision, and filtering such sentences can improve the overall performance of the system. In order to show the utility of using this module, we compare the performance of the lexicon-based component with and without using this module in Figure 4.

As shown in Figure 4, although the effect of using the question rule module is not very high, it is useful and with respect to all measures filtering interrogative sentences improves the performance of the system.

Another problem which may affect the overall performance of the system is the indicator used for detecting the boundaries of sentences as described in ‘Lexicon-based component’. In order to show this, Figures 5 and 6 show the comparison of using punctuation instead of verbs as sentence boundary indicator for polarity-detection and rating-prediction problems, respectively.

As could be seen in the figures, for polarity-detection problem, both verbs and punctuations have similar effects though due to their higher accuracy, verbs are preferred to punctuations. For the rating-prediction problem, the use of verbs as sentence boundary indicators is clearly a better choice with respect to all measures. This could be the result of the fact that users usually ignore or incorrectly use punctuations in informal texts. Hence, verbs are better indicators for opinion mining in informal text.

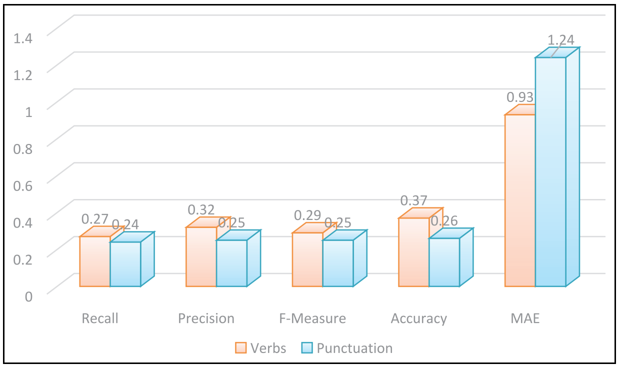


Figure 6. Comparison of using punctuation instead of verbs as sentence boundary indicator for rating-prediction problem in the lexicon-based component of the proposed HOMPer system.

Table 6. Comparison of lexicon-based and machine learning (ML) methods on polarity-detection problem.

Lexicon	Recall	Precision	F-measure	Accuracy
Lexicon based	0.56	0.56	0.56	0.76
ML using unigram	0.57	0.66	0.61	0.82
ML using bigram	0.55	0.63	0.59	0.82

ML: machine learning.
The Naïve Bayes algorithm is used in the ML classification module.

Table 7. Comparison of lexicon-based and machine learning (ML) methods on rating-prediction problem.

Lexicon	Recall	Precision	F-measure	Accuracy	MAE
Lexicon based	0.27	0.32	0.29	0.37	0.93
ML using unigram	0.41	0.41	0.42	0.56	0.79
ML using bigram	0.37	0.42	0.39	0.60	0.68

ML: machine learning: MAE: mean absolute error.
The Naïve Bayes algorithm is used in the ML classification module.

4.3. Hybridization results

Lexicon-based methods are preferred to ML-based methods when training data are not available, and the difference in performance of these approaches are negligible [2,44]. However, as shown in Tables 6 and 7, pure lexicon-based method has a considerably lower performance in comparison to pure ML-based method.

The results for the ML method in Tables 6 and 7, was obtained using the NB classifier that has been previously shown to be an appropriate choice for opinion mining tasks [2,19]. As can be seen in Tables 6 and 7, the ML method clearly outperforms the lexicon-based method with respect to all performance measures. This higher performance is obtained regardless of the feature type employed by the algorithm, although using bigram features leads to higher accuracy and lower MAE. These results emphasise the fact that ML methods outperforms lexicon-based methods when they are trained on a relatively large data set.

In this study, we hypothesised that replacing unigrams by lexicon terms can improve the performance of the ML component. In fact, in the proposed method, a feature-level combination of ML and lexicon-based methods is proposed. In order to verify this hypothesis, the performance of the proposed hybrid method is compared with that of the ML method using different features in Figure 7.

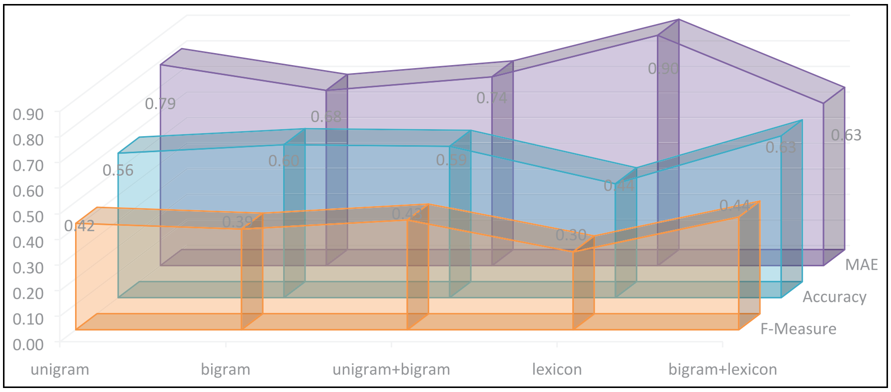


Figure 7. Comparison of the performance of the proposed hybrid method, HOMPer, (indicated with bigram + lexicon) with lexicon-based and machine learning (using unigram and bigram features) methods on rating-prediction problem.

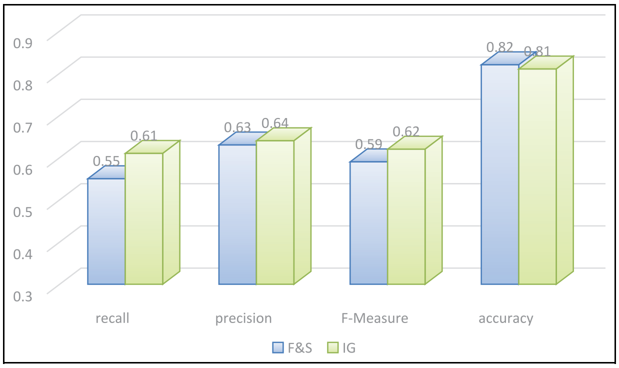


Figure 8. The effect of using feature-selection method on polarity-detection problem.
The original number of features was 305020 and using information gain (IG) and frequency and stop-word filters (F&S) it is reduced to 15,251 and 3000, respectively.

As can be seen in Figure 7, pure lexicon-based method has the lowest performance, while the proposed hybrid method obtains the highest performance. Moreover, unigram features have an acceptable performance compared with other types of features. This is an important result from an implementation point of view because extracting these features is simple and is not resource intensive.

As shown in Figure 7, the combination of unigram and bigram features leads to better performance than using them in isolation. Another point in Figure 7 is that, as we hypothesised, lexicon features increase the overall performance when they are combined with bigrams. In other words, lexicon features are more informative than unigrams.

4.4. The effect of feature-selection method

In order to show the utility of using feature-selection method, we compared two different settings. In the first setting, two feature-reduction methods, namely frequency filter and stop-word filter, are used as described in 'ML component'. In the second setting, the well-known information gain (IG) strategy is used. The results of comparing these two settings on the polarity-detection and rating-prediction problems are shown in Figures 8 and 9, respectively.

It should be noted that the total number of features in this study was originally 305,020, while after application of frequency filter and stop-word filter, this number is reduced to 150,000 and 3000, respectively. On the other hand, using IG, the total number of features is 15,251.

As could be seen in Figures 8 and 9, with respect to almost all measures, the IG method outperforms the two-step filtering feature-selection method (F&S) described in section 'ML component'. This could be due to the nature of the IG that 'measures the increase in information about the class gained by including the feature' [53].

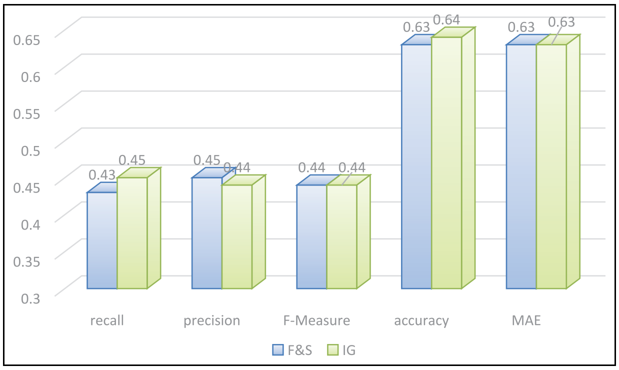


Figure 9. The effect of using feature-selection method on rating-prediction problem.
The original number of features was 305020 and using information gain (IG) and frequency and stop-word filters (F&S) it is reduced to 15,251 and 3000, respectively.

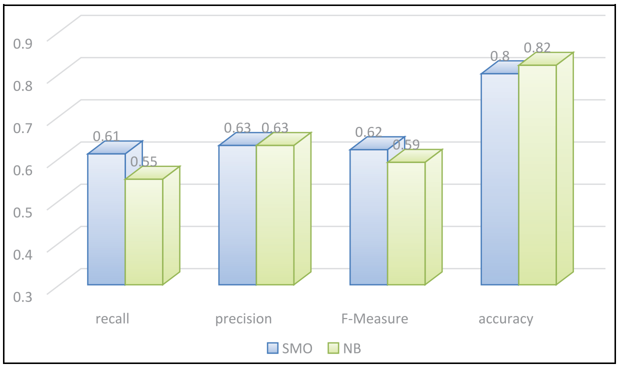


Figure 10. Comparison of using NB and SMO in the proposed system for polarity detection, where the combination of bigrams and lexicon terms is used as features and frequency filtering and stop-word filtering are used as feature-selection method.

4.5. The effect of ML algorithm

Previous studies have shown that NB and SVM are state-of-the-art ML algorithms for opinion mining in several languages [11,23]. Figures 10 and 11 show the comparison between using NB and SMO implementation of SVM in the proposed system. In these figures, the combination of bigrams and lexicon terms is used as features and frequency filtering, and stop-word filtering are used as feature-selection method.

As could be seen in Figures 10 and 11, with respect to almost all measures in polarity-detection problem, the SMO method outperforms the NB algorithm. However, in rating prediction, the NB method outperforms the SMO. This can lead to the use of the SMO method for polarity detection and the NB method for rating prediction in the proposed HOMPer system.

4.6. The effect of normalisation

In order to show the effect of the normalisation module described in 'Solving language-related problems', we compared the amount of improvement obtained using this module for pure lexicon-based method (named as Pool), pure ML method and the proposed HOMPer system (hybrid) in Figure 12.

As could be seen in Figure 12, the normalisation module has the highest impact on the lexicon-based system. This could be justified by the fact that the most effective part of a lexicon-based method is its term-matching module, and the normalisation module improves the matching ability of this module. The low impact of the normalisation module on the ML-based systems, however, could be justified using the fact that supervised learning methods learn the important terms

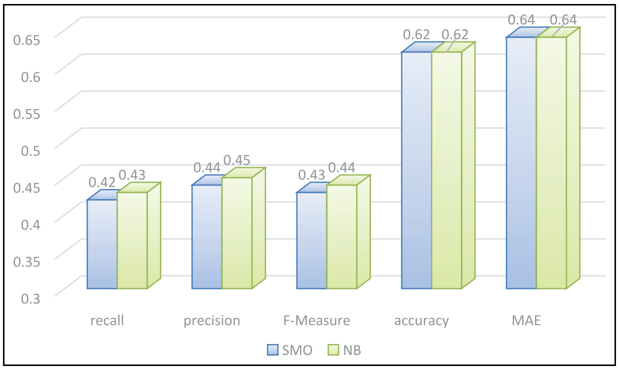


Figure 11. Comparison of using NB and SMO in the proposed system for rating prediction, where the combination of bigrams and lexicon terms is used as features and frequency filtering and stop-word filtering are used as feature-selection method.

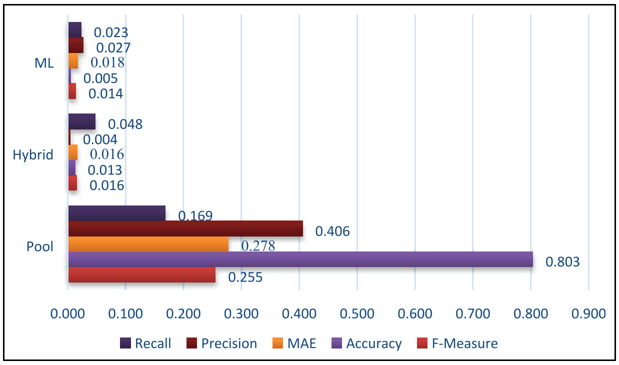


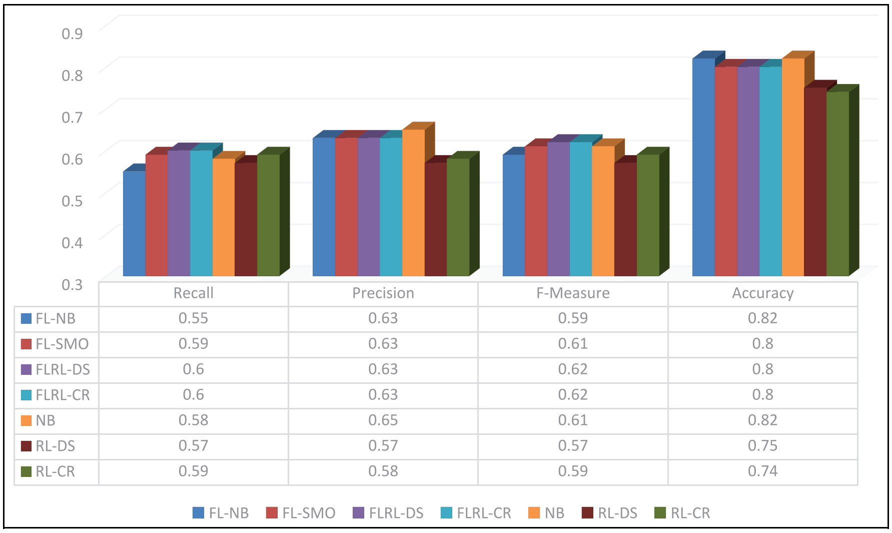
Figure 12. Comparison of the amount of improvement obtained using the normalisation module for pure lexicon-based method (Pool), pure machine learning method (ML) and the proposed HOMPer system (hybrid).

in the training phase regardless of the way in which the words are written, and hence the normalisation module could not help these methods very much.

4.7. The effect of combination level

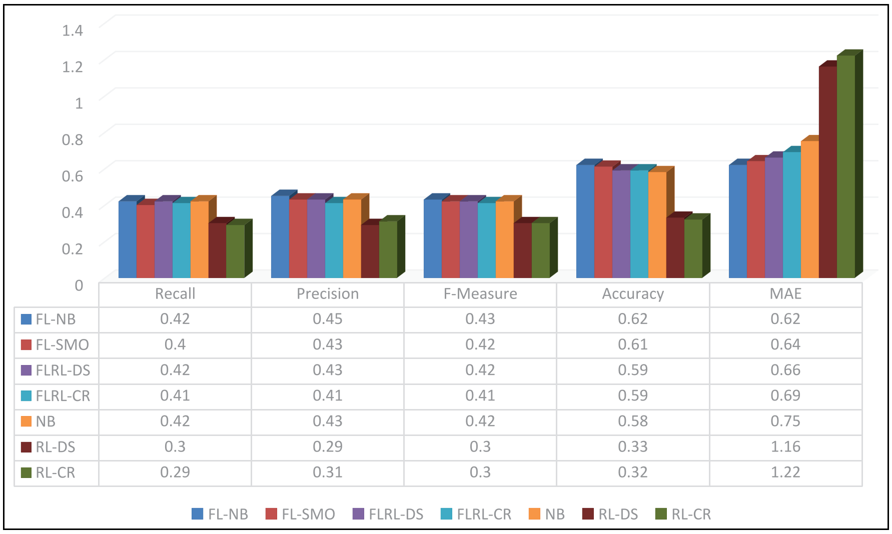
The combination method and the level at which the combination is applied is an important factor in the overall quality of any opinion mining system [1,54]. In this study, the combination could be performed at feature or review level. The feature-level combination was described in 'ML component' and 'Hybridization results'. At this level, lexicon-based terms are combined with n-gram features in a supervised process to train the classifier. At the review level, on the other hand, the outputs of lexicon-based and ML-based methods are combined using an aggregation method [2,5,6]. In order to show the impact of the combination level on the output of HOMPer, four settings are tested as follows:

- ML-based system, in which unigram features are used to train the NB classifier (denoted as NB in Figures 13 and 14).
- Feature-level combination as described in 'ML component' and 'Hybridization results' (denoted as FL-NB and FL-SMO in Figures 13 and 14).
- Both feature-level and review-level combinations (denoted as FLRL-DS and FLRL-CR Figures 13 and 14): in this combination, the results of feature-level combinations (i.e. FL-NB and FL-SMO) are combined using the Dempster–Shafer (DS) [55] combination rule as described in Subrahmanian and Reforgiato [46] and cross-ratio (CR) uninorm [56] as described by Basiri and Kabiri [54]. These methods are powerful combination methods recently used for opinion mining [57,58].



	Recall	Precision	F-Measure	Accuracy
FL-NB	0.55	0.63	0.59	0.82
FL-SMO	0.59	0.63	0.61	0.8
FLRL-DS	0.6	0.63	0.62	0.8
FLRL-CR	0.6	0.63	0.62	0.8
NB	0.58	0.65	0.61	0.82
RL-DS	0.57	0.57	0.57	0.75
RL-CR	0.59	0.58	0.59	0.74

Figure 13. Comparison of different combination levels in polarity detection.
FL-NB and FL-SMO are feature-level combination as described in 'ML component' and 'Hybridization results', FLRL-DS and FLRL-CR use both feature-level and review-level combinations using Dempster–Shafer (DS) and cross-ratio (CR) combination methods, NB is a pure machine learning classification method, and RL-DS and RL-CR use only review-level combination.



	Recall	Precision	F-Measure	Accuracy	MAE
FL-NB	0.42	0.45	0.43	0.62	0.62
FL-SMO	0.4	0.43	0.42	0.61	0.64
FLRL-DS	0.42	0.43	0.42	0.59	0.66
FLRL-CR	0.41	0.41	0.41	0.59	0.69
NB	0.42	0.43	0.42	0.58	0.75
RL-DS	0.3	0.29	0.3	0.33	1.16
RL-CR	0.29	0.31	0.3	0.32	1.22

Figure 14. Comparison of different combination levels in 5-star rating prediction.
FL-NB and FL-SMO are feature-level combination as described in 'ML component' and 'Hybridization results', FLRL-DS and FLRL-CR use both feature-level and review-level combinations using Dempster–Shafer (DS) and cross-ratio (CR) combination methods, NB is a pure machine learning classification method, and RL-DS and RL-CR use only review-level combination.

- Only review-level combination (denoted as RL-DS and RL-CR Figures 13 and 14): in this setting, the results of ML classification using NB and pure lexicon-based method are combined using the above-mentioned DS and CR combination methods.

As could be seen in Figures 13 and 14, the performance of feature-level combinations (FL-NB and FL-SMO) and two-level combinations (FLRL-DS and FLRL-CR) are higher than those of review-level combinations (RL-DS and RL-CR). This shows the effectiveness of HOMPer system which exploits feature-level combination. Moreover, for rating prediction, HOMPer is preferred to two-level combination systems (i.e. FLRL-DS and FLRL-CR), as it achieves a higher F-measure and accuracy.

5. Conclusion

In this article, we proposed a novel hybrid method utilising the benefits of both ML and lexicon-based approach for Persian opinion mining. In the proposed method, three layers exist for addressing language-related problems, utilising advantages of lexicon-based method and exploiting benefits of ML. Different language-specific problems such as different forms of writing, informal writing, spelling correction and normalisation are addressed in the language-related layer. Moreover, lexicon pooling is used to utilise three available lexicons in the lexicon-based layer. Finally, a feature-level combination of ML and lexicon-based method is used in the proposed system. To the best of our knowledge, this is the first study to examine both ML and lexicon-based method for Persian opinion mining at different combination levels. Most importantly, the experimental results on a large real data set of Persian customer reviews show that the proposed method is effective in polarity-detection and rating-prediction problems. For future work, we will use other types of combination methods and different combination levels. Another direction may be improving the lexicon-based component of the proposed system. In addition, exploiting nature-inspired feature-selection methods and investigating their effect on the quality of the proposed system may be another line for future research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This work was financially supported by the research deputy of Shahrekord University (grant 96GRD1M1874).

ORCID iD

Mohammad Ehsan Basiri https://orcid.org/0000-0003-2893-3892

References

1. Basiri ME and Kabiri A. Sentence-level sentiment analysis in Persian In: *Proceedings of the 2017 3rd international conference on pattern recognition and image analysis (IPRIA)*, Shahrekord, Iran, 19–20 April 2017, pp. 84–89. New York: IEEE.
2. Basiri ME, Ghasem-Aghaee N and Naghsh-Nilchi AR. Exploiting reviewers' comment histories for sentiment analysis. *J Inf Sci* 2014; 40: 313–328.
3. Groh G and Hauffa J. Characterizing social relations via NLP-based sentiment analysis. In: *Proceedings of the 5th international AAAI conference on weblogs and social media*, Barcelona, 17–21 July 2011, pp. 502–505. New York: AIAA.
4. Ceron A, Curini L and Iacus SM. Using sentiment analysis to monitor electoral campaigns: method matters – evidence from the United States and Italy. *Soc Sci Comput Rev* 2015; 33: 3–20.
5. Nemati S and Naghsh-Nilchi A. An evidential data fusion method for affective music video retrieval. *Intell Data Anal* 2017; 21: 427–441.
6. Nemati S and Naghsh-Nilchi AR. Incorporating social media comments in affective video retrieval. *J Inf Sci* 2016; 42: 524–538.
7. Zhang Y. GroRec: a group-centric intelligent recommender system integrating social, mobile and big data technologies. *IEEE Trans Serv Comput* 2016; 9: 786–795.
8. Schumaker R, Jarmoszko A and Labedz C. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decis Support Syst* 2016; 88: 76–84.
9. Neidhardt J, Rümmele N and Werthner H. Predicting happiness: user interactions and sentiment analysis in an online travel forum. *Inf Technol Tour* 2017; 17: 101–119.
10. Iskandar B. Terrorism detection based on sentiment analysis using machine learning. *J Eng Appl Sci* 2017; 12: 691–698.
11. Liu B. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press, 2015.
12. Pang B and Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2008; 2: 1–135.
13. Thelwall M, Buckley K and Paltoglou G. Sentiment strength detection for the social web. *J Am Soc Inf Sci Technol* 2012; 63: 163–173.
14. Amiri F, Scerri S, Khodashahi MH et al. Lexicon-based sentiment analysis for Persian text. In: *Proceedings of recent advances in natural language*, 2015, pp. 9–16, http://aclweb.org/anthology/R15-1002
15. Bagheri A, Saraee M and de Jong F. 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)*, Mashhad, Iran, 14–16 May 2013, pp. 1–6. New York: IEEE.
16. Shams M, Shakery A and Faili H. 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 2012)*, Shiraz, Fars, Iran, 2–3 May 2012, pp. 216–221. New York: IEEE.

17. W3Techs – World Wide Web Technology Surveys, https://w3techs.com/ (accessed 1 January 2017).
18. Bagheri ASM. Persian sentiment analyzer: a framework based on a novel feature selection method. *Int J Artif Intell* 2014; 122: 115–129.
19. Basiri M, Nilchi A and Ghassem-Aghaee N. A framework for sentiment analysis in Persian. *Open Trans Inf Process* 2014; 1: 1–14.
20. Hajmohammadi MS and Ibrahim R. A SVM-based method for sentiment analysis in Persian language. In: *International Conference on Graphic and Image Processing (ICGIP 2012)*, vol. 8768, p. 876838. International Society for Optics and Photonics, 2013.
21. Golpar-Rabooki E, Zarghamifar S and Rezaeenour J. Feature extraction in opinion mining through Persian reviews. *J Artif Intell Data Min* 3: 169–179.
22. Alimardani S and Aghaei A. Opinion mining in Persian language using supervised algorithms. *J Inf Syst Telecommun* 3: 135–141.
23. Liu B. Sentiment analysis and opinion mining. *Synth Lect Hum Lang Technol* 2012; 5: 1–167.
24. Liu B and Zhang L. A survey of opinion mining and sentiment analysis. In Aggarwal CC and Zhai CX (eds) *Mining text data*. Boston, MA: Springer, pp. 415–463.
25. Denecke K. Using SentiWordNet for multilingual sentiment analysis. In: *Proceedings of the IEEE 24th international conference on data engineering workshop, 2008 (ICDEW 2008)*, Cancun, Mexico, 7–12 April 2008, pp. 507–512. New York: IEEE.
26. Cieliebak M, Deriu J, Egger D et al. A Twitter corpus and benchmark resources for German sentiment analysis. In: *Proceedings of the 4th international workshop on natural language processing for social media*, Valencia, April 2017, pp. 45–55. Stroudsburg PA: Association for Computational Linguistics.
27. Ishijima H, Kazumi T and Maeda A. Sentiment analysis for the Japanese stock market. *Glob Bus Econ Rev* 2015; 17: 237.
28. Miyakawa S, Saitoh F and Ishizu S. A quality table-based method for sentiment expression word identification in Japanese In: *Proceedings of the LNCS*, Vol. 10289, pp. 48–59, https://link.springer.com/chapter/10.1007/978-3-319-58637-3_4
29. Abbasi A, Chen H and Salem A. Sentiment analysis in multiple languages. *ACM Trans Inf Syst* 2008; 26: 1–34.
30. Aldayel HK and Azmi AM. Arabic tweets sentiment analysis – a hybrid scheme. *J Inf Sci* 2016; 42: 782–797.
31. Vural AG, Cambazoglu BB, Senkul P et al. A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish In: Lent R and Gelenbe E (eds) *Computer and information sciences III*. London: Springer, 2012, pp. 437–445.
32. Mukhtar N and Khan MA. Urdu sentiment analysis using supervised machine learning approach. *Int J Pattern Recognit Artif Intell* 2017; 1851001.
33. Baccianella S, Esuli A and Sebastiani F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the LREC*, Malta, 17–23 May 2010, pp. 2200–2204, http://www.lrec-conf.org/proceedings/lrec2010/slides/769.pdf
34. Shamsfard M, Hesabi A, Fadaei H et al. Semi automatic development of Farsnet; the Persian wordnet. In: *Proceedings of the 5th global wordnet conference*, Mumbai, India, 31 January–4 February 2010.
35. Basiri ME and Kabiri A. Translation is not enough: comparing lexicon-based methods for sentiment analysis in Persian. In: *Proceedings of the 2017 international symposium on computer science and software engineering conference (CSSE)*, Shiraz, Iran, 25–27 October 2017.
36. Asgarian E, Kahani M and Sharifi S. The impact of sentiment features on the sentiment polarity classification in Persian reviews. *Cognit Comput* 2018; 10: 117–135.
37. Basiri ME and Kabiri A. Words are important: improving sentiment analysis in the Persian language by lexicon refining. *ACM Trans Asian Low-resource Lang Inf Process* 2018; 17: 26.
38. Dashtipour K, Hussain A, Zhou Q et al. PerSent: a freely available Persian sentiment lexicon, https://link.springer.com/chapter/10.1007/978-3-319-49685-6_28
39. Basiri ME, Ghasem-Aghaee N and Reza A. Lexicon-based sentiment analysis in Persian. *Curr Futur Dev Artif Intell* 2017; 155–184.
40. Asgarian E, Saeedi R, Stiri A et al. https://wtlab.um.ac.ir (accessed 1 July 2016).
41. Mohammad SM, Kiritchenko S and Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: *Proceedings of the 7th international workshop on semantic evaluation exercises (Semeval-2013)*. Atlanta, GA, 14–15 June 2013.
42. Thelwall M, Wilkinson D and Uppal S. Data mining emotion in social network communication?: gender differences in MySpace. *J Assoc Inf Sci Technol* 2010; 61: 190–199.
43. Thelwall M, Buckley K, Paltoglou G et al. Damping sentiment analysis in online communication: discussions, monologs and dialogs, https://link.springer.com/chapter/10.1007/978-3-642-37256-8_1
44. Basiri ME, Naghsh-Nilchi AR and Ghasem-Aghaee N. Sentiment prediction based on Dempster–Shafer theory of evidence. *Math Probl Eng* 2014; 2014: 361201.
45. Benamara F, Irit S, Cesarano C et al. Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: *Proceedings of the ICWSM*, Boulder, CO, pp. 1–4, https://pdfs.semanticscholar.org/1133/455c8b743073160dd439361c4aaf8de0faad.pdf

46. Subrahmanian VS and Reforgiato D. AVA: adjective-verb-adverb combinations for sentiment analysis. *IEEE Intell Syst* 2008; 23: 43–50.
47. Wang S and Manning CD. Baselines and bigrams: simple, good sentiment and topic classification. In: *Proceedings of the 50th annual meeting of the association for computational linguistics*, Jeju, Republic of Korea, 8–14 July 2012, pp. 90–94. New York: Association for Computational Linguistics.
48. Nemati S and Basiri ME. Particle swarm optimization for feature selection in speaker verification, https://link.springer.com/chapter/10.1007/978-3-642-12239-2_39
49. Nemati S, Ehsan M, Ghasem-aghaee N et al. A novel ACO – GA hybrid algorithm for feature selection in protein function prediction. *Exp Syst Appl* 36: 12086–12094.
50. Aghdam MH, Tanha J, Naghsh-Nilchi AR et al. Combination of ant colony optimization and Bayesian classification for feature selection in a bioinformatics dataset. *J Comput Sci Syst Biol* 2009; 2: 186–199.
51. Digikala, http://www.digikala.com (2017, accessed 15 February 2017).
52. Saraee M and Bagheri A. Feature selection methods in Persian sentiment analysis, https://link.springer.com/chapter/10.1007/978-3-642-38824-8_29
53. Flach P. 2012 *Machine learning: the art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.
54. Basiri ME and Kabiri A. Uninorm operators for sentence-level score aggregation in sentiment analysis. In: *Proceedings of the 2018 4th international conference on web research (ICWR)*, Tehran, Iran, 25–26 April 2018, pp. 97–102. New York: IEEE.
55. Shafer G. *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press, 1976.
56. Yager RR and Rybalov A. Uninorm aggregation operators. *Fuzzy Sets Syst* 1996; 80: 111–120.
57. Appel O, Chiclana F, Carter J et al. Cross-ratio uninorms as an effective aggregation mechanism in sentiment analysis. *Knowledge-based Syst* 2017; 124: 16–22.
58. Schouten K and Frasincar F. Survey on aspect-level sentiment analysis. *IEEE Trans Knowl Data Eng* 2016; 28: 813–830.