Social Network Analysis and Mining (2022) 12:9
https://doi.org/10.1007/s13278-021-00840-1

ORIGINAL ARTICLE

Extending persian sentiment lexicon with idiomatic expressions for sentiment analysis

Kia Dashtipour[1] · Mandar Gogate[1] · Alexander Gelbukh[2] · Amir Hussain[1]

Received: 5 March 2021 / Revised: 30 October 2021 / Accepted: 5 November 2021 / Published online: 25 November 2021

Abstract
Nowadays, it is important for buyers to know other customer opinions to make informed decisions on buying a product or service. In addition, companies and organizations can exploit customer opinions to improve their products and services. However, the Quintilian bytes of the opinions generated every day cannot be manually read and summarized. Sentiment analysis and opinion mining techniques offer a solution to automatically classify and summarize user opinions. However, current sentiment analysis research is mostly focused on English, with much fewer resources available for other languages like Persian. In our previous work, we developed PerSent, a publicly available sentiment lexicon to facilitate lexicon-based sentiment analysis of texts in the Persian language. However, PerSent-based sentiment analysis approach fails to classify the real-world sentences consisting of idiomatic expressions. Therefore, in this paper, we describe an extension of the PerSent lexicon with more than 1000 idiomatic expressions, along with their polarity, and propose an algorithm to accurately classify Persian text. Comparative experimental results reveal the usefulness of the extended lexicon for sentiment analysis as compared to PerSent lexicon-based sentiment analysis as well as Persian-to-English translation-based approaches. The extended version of the lexicon will be made publicly available.

1 Introduction

Nowadays, millions of people share their opinions and feedback on products and services via comments, reviews, and social media. Positive and negative feedback impacts the sales of products and services as well as the organization public perception. However, the rate at which the information is generated makes it impossible to manually analyse this information. Sentiment analysis offers a solution to automatically process these data by computationally understanding and classifying subjective information from source materials. Companies have successfully deployed such a system, as a part of business intelligence and big data analytic technologies (Cavallari et al. 2019; Yang et al. 2018; Guellil et al. 2021), to use this feedback for improving their products (Cambria et al. 2019; Dragoni et al. 2019; Hussain et al. 2021).

Due to informal and highly colloquial nature of such reviews and comments uploaded by general Internet users, the source text often contains idiomatic expressions. An idiomatic expression is a language construction that conveys a meaning distinct from the literal meaning of the words it is formed of, or a multi-word expression with non-compositional meaning. While idiomatic expressions often make speech more expressive, they are difficult to understand if not listed in the lexicon (Zikopoulos et al. 2011; Dashtipour et al. 2021). In addition, they are highly language-dependent and unpredictable: knowing idiomatic expressions in one language often provides no help in interpreting idiomatic expressions in another language (Wang et al. 2013; Li et al. 2021; Dashtipour et al. 2021). Finally, a lexicon-based sentiment analysis framework often incorrectly classifies the sentences consisting of idiomatic expression. Therefore, the inclusion of idiomatic expressions in the lexicon could help accurately understand a wide range of opinions available on the Internet.

✉ Kia Dashtipour
k.dashtipour@napier.ac.uk

Mandar Gogate
m.gogate@napier.ac.uk

Amir Hussain
A.Hussain@napier.ac.uk

[1] School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh EH10 5DT, Scotland, UK

[2] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico

Although Persian language is an official language of Iran, Afghanistan (variety called Dari) and Tajikistan (variety called Tajik), with about 130 million of speakers across these countries, there are currently very limited tools available to automatically summarize the overall opinion (Ling Lo et al. 2016; Dashtipour et al. 2019, 2017, 2018, 2017; Ieracitano et al. 2018; Jiang et al. 2021; Dashtipour et al. 2021); in particular, Persian sentiment analysis techniques suffer from the lack of resources to detect and interpret idiomatic expressions in the reviews.

Persian language has a number of interesting sociolinguistic peculiarities. A notable difference between Persian and English language is the way of writing using a complex, right-to-left Persian script (or a variant of Cyrillic script in Tajikistan), often not compatible with existing keyboards or devices, that leads to much greater than in English variation in spelling and code switching. In addition, some letters of the Persian script, such as T, S and Z ت، س، ز are written in different shape depending on the context, which creates more spelling variants in informal user-contributed texts often typed on phone keyboards: for example, the word طبعیت ("environment") can be written in various forms (Basiri et al. 2019; Nezhad et al. 2019; Dashtipour et al. 2017, 2020; Gogate et al. 2020; Ahmed et al. 2021; Gogate et al. 2019, 2017).

Particularly important for this work is the fact that the Persian language has many idiomatic expressions, metaphors, slang and swear words and profanity expressions, with much more complicated system of social acceptance and taboos than English does, leading to a more complex system of nuances, double meaning and wordplay. We attribute this mostly to an interplay of sociolinguistic factors such as education level, general standard of living, religion and dialectal variation. Persian speakers often prefer not to express their sentiment explicitly but use idiomatic expressions or euphemisms to convey their thoughts (Khoshnevisan 2019; Gogate et al. 2020, 2020; Dashtipour et al. 2021; Gogate et al. 2019).

In addition, the source text consisting of many informal words, as well as transliteration or literal translations of English words, makes it difficult to analyse the text polarity (Mullen and Malouf 2006). The process of extracting features such as idioms for sentiment analysis of Persian texts is more complex than extracting traditional features, and it is sometimes difficult to assign polarity to such features. We found that nearly 8% of the movie review dataset contained idioms. Not considering them has negative impact on the classification of the overall polarity of the text (Mansouri 2015).

In this paper, we present PerSent lexicon (Dashtipour et al. 2016) with 14000 idioms in the Persian language, along with their sentiment polarity, and show its value for the sentiment analysis task. In addition, we integrate the extended lexicon and the idiom analysis engine into our lexicon-based Persian sentiment analysis framework. We evaluate the usefulness of the obtained novel resource via application to sentiment analysis using machine learning applied to a dataset of reviews of movies and products. We show that our extended PerSent-based sentiment analysis outperforms state-of-the-art sentiment analysis approaches as well as Persian-to-English translation-based approaches. The extended version of the lexicon will be made publicly available for research purposes.

The main contribution of this work is the first-of-its-kind publicly available Persian Sentiment lexicon for accurately classifying source text consisting of idiomatic expressions. We provide an algorithm to detect Persian idioms in a sentence. We also provide a methodology for using the lexicon in a machine learning-based sentiment analysis framework and show that the proposed model outperforms the baseline lexicon-based algorithm that simply counts the average polarity of the words and expressions in the document. In this way, we illustrate a sentiment analysis framework for a resource-poor language, to which more advanced corpus-based methods, in particular, deep learning-based techniques, are not applicable. In addition, we show that for idiomatic expression-based sentiment analysis, less pre-processing of the input text is required and a simpler algorithm can be used in comparison with phrase-based and concept-based sentiment analysis, since idioms are fixed expressions with no or very little syntactic variation, while detecting phrases or concepts in the text may require paraphrase detection.

The paper is organized as follows. In Sect. 2, we discuss related work. In Sect. 3, we describe the methodology used for compilation of our lexicon. In Sect. 4, we describe our procedure for the evaluation of the obtained lexicon and the datasets used for evaluation. In Sects. 5 and 6, we present experimental results and discussion, accordingly. Finally, Sect. 7 concludes the paper and outlines the directions of future work.

2 Related work

An idiom is a phrase or fixed expression that conveys meaning different from the combination of literal meanings of its constituent words, such as figurative meaning: for example, "give me a hand" means in English "help me"; "hold on a second" means in English "wait for a short time" (Langlotz 2006). Thus, it is difficult or impossible to deduce its meaning without knowing it beforehand, even if the meaning of the individual words is known.

Idioms exist in all known languages; it is estimated that more than 20,000 idiomatic expressions exist in the English language, and Persian language is no exception. For

example, مثل موش و گربه هستند ("as mouse and cat") in Persian refers to two people who keep fighting with each other. The following properties characterize idioms (Fraser 1970):

- *Conventionality*: literal meaning of the phrase does not coincide with its idiomatic meaning;
- *Inflexibility*: the syntax of idioms is very restrictive;
- *Figuration*: idioms often present figurative meaning;
- *Informality*: idioms often contain informal words;
- *Affect*: idioms often have non-neutral (i.e. positive of negative) polarity.

A correct interpretation of idioms is crucial for the understanding of the meaning of a text. To identify idioms in a given sentence, various linguistic resources, such as lists of fixed expressions, phrases, cliches and proverbs, need to be considered (Passaro et al. 2019).

Non-native speakers experience difficulties in understanding idioms, which can significantly affect their personal and professional communication (Nippold and Martin 1989). Therefore, most of the second-language classes focus on idioms as an important part of language acquisition (Liu 2003; Erman and Warren 2000). Correct interpretation of idioms is very important in sentiment analysis. Williams et al. (2015) collected more than five hundred English idioms and used web-based crowdsourcing to assign their polarity. They evaluated the usefulness of their resource on a movie reviews dataset, using an SVM classifier for sentiment analysis. In their experiments, they obtained an accuracy of 0.70. However, they showed that their approach was not suitable for document-level sentiment analysis, so it was applied for sentence-level analysis.

Liang et al. (2018) built a sentiment lexicon SlangSD that contained informal words. They used online resources to collect idioms and slang words. The performance of the classifier using their resource was boosted from the baseline 0.73 to 0.87. A shortcoming of their approach was that the lexicon was relatively small, and hence, it could not effectively detect informal words in texts. Ibrahim et al. (2015, 2015) created an Arabic lexicon to identify idioms in the text. They trained an SVM to evaluate the performance of the features extracted from text, including idioms. Once the idioms have been extracted, a corpus of 1000 positive and 1000 negative tweets was used to identify their polarity. The overall accuracy achieved was 0.986. The technique that they used does not generalize to Persian because they developed their lexicon for Arabic dialects and colloquial Arabic language.

Gul (2014) proposed an approach for detecting idioms in Urdu texts. In this work, an idiom lexicon of 2500 Urdu idioms along with their polarity was developed. A classifier was trained on a dataset that was manually annotated with positive and negative polarity (Raj and Kajla 2015). The technique that they used does not generalize to Persian because their lexicon consists of Urdu words.

Table 1 summarizes the most important existing resources and approaches to the use of idioms for sentiment analysis in various languages. We included in the table several recently developed approaches, as well as some approaches, published several years ago that have not been outperformed by later approaches. The table also shows the accuracy of polarity classification with and without detection of idioms. For example, the figures for Wang et al. (2010) suggest that idioms alone provide high accuracy for the Chinese dataset they used for evaluation. In all cases, adding idioms gives a significant boost in sentiment analysis accuracy.

3 Persian idiom lexicon

In order to build our Persian idiom lexicon for sentiment analysis, we extracted idioms from a website with a list of 925 Persian idioms. This website provides the most widely used Persian idioms, though there are many more idioms not widely used in daily communication and not understandable

Table 1 Approaches for idiom detection in text

Reference	Language	Classifier	Dataset	Accuracy without idioms	Accuracy idioms	Total idioms	Multi-word idioms
Gul (2014)	Urdu	SVM	Product Reviews	0.684	0.745	2500	N/A
Wang et al. (2010)	Chinese	Lexicon-based	Online Reviews	0.5	0.827	9200	1200
Xie and Wang (2014)	Chinese	Lexicon-based	Movie Reviews	0.623	0.815	3000	N/A
Wu et al. (2016)	English	Lexicon-based	Twitter	0.651	0.848	96462	0
Williams et al. (2015)	English	Naive Bayes	Movie Reviews	0.64	0.7	580	200
Verma and Vuppuluri (2015)	English	SVM	Online Reviews	N/A	0.95	N/A	N/A
Citron Francesca et al. (2016)	German	Lexicon-based	Movie Reviews	0.658	0.67	619	N/A
Djemaa et al. (2016)	French	Lexicon-based	Tweets	0.805	0.828	872	N/A

for many native speakers; the treatment of such rarer idioms is a topic of our future work.

Three annotators manually assigned polarity to the idioms in the form a number between − 1 (very negative) and + 1 (very positive), with one decimal point precision, such as + 0.7. In their work, they could consult Internet or other sources for usage examples as they felt appropriate. In the resulting lexicon, the values of the annotators were averaged and, for simplicity of presentation, rounded to one decimal digit, since these estimates were very subjective anyway. The annotators were educated native speakers of Persian, two of them between 50 and 60 years old and one 30 years old. The agreement between the annotators was substantial, with Fleiss of 0.73 (Fleiss et al. 2020).

In some cases, the annotators did not agree on whether the polarity is positive or negative; examples of such idioms are نان و کباب ("bread and kebab") or حزب باد ("wind party": a person who keeps changing political views). In such cases, these idioms were removed. In addition, for some idioms, annotators were not sure as to their polarity; such idioms were also removed from the lexicon.

In about 20% of cases, the annotators agreed on the sign of polarity but the difference of the assigned numerical values was more than 0.5, such as + 0.9 versus + 0.3. In such cases, we used TextBlob, a Python library for natural language processing, based on the widely used Natural Language Toolkit (NLTK). It includes an automatic polarity detector for English sentiment analysis, pre-trained with supervised learning using naive Bayes as the classifier, which can identify the polarity for a word, multi-word expression, or sentence. To determine the sentiment of a Persian phrase, we used Google translator to automatically obtain an idiomatic translation of the phrase into English. Then, we manually corrected these translations where the translation provided by Google was not idiomatic enough. Finally, we fed it into TextBlob's sentiment detector do obtain the automatic sentiment value judgement. For example, for بد شکل ("bad face") the polarity obtained in this way was − 0.6; for صورت زشت ("ugly face") it was -0.7. The Persian dictionary has been used, which consists of 14,000 phrases. We used three Persian language experts to remove unrelated phrases. Our lexicon consists of 925 Persian idioms, 326 negative and 374 positive idioms (total 700), and 225 of the idioms were neural.

Finally, of the 925 idioms that have been annotated manually, those idioms that were annotated with neutral polarity were not included in our lexicon; by neutral, we understood either the label "neutral" assigned by the annotators or the value of zero assigned by TextBlob. We did not use any threshold for the exclusion of neutral idioms, because, as Fig. 1 illustrates, the majority of the idioms that had nonzero polarity value were strongly opinionated, so there were very few cases of "almost neutral but still not neutral" annotations. We also decided to remove from the lexicon idioms that contained strong profanity, because they might create ethical, religious or legal problems in annotating and sharing the lexicon. This did not affect our evaluation because the review datasets on which we evaluated our lexicon did not contain strong profanity due to the policy of the corresponding websites. Investigating the impact of profanity on sentiment analysis using a suitable dataset will be a topic of our future work.

This resulted in 326 negative and 374 positive idioms. These figures look a bit unexpected because in any language negative expressions are more numerous than positive ones. We attribute the inverse relation in our figures to the removal of the idioms that contained strong profanity, which was mostly negative. Figure 1 shows the number of idioms by assigned polarity. Table 2 shows the numbers of different types of idiom, such as dialectal, light swear words and slang. Table 3 gives examples of idioms in our lexicon along with their idiomatic translation and polarity.

4 Evaluation methodology

In order to evaluate the performance of our idiom lexicon, we used our lexicon and the idiom lexicon to assign polarity to sentences with various classification algorithms. It is to be noted that we used the original PerSent lexicon and its extension with the idioms lexicon. The algorithms we used varied from direct counting of the average polarity of the words and expressions in the sentence to the use of machine learning techniques to extend the sentiment values from the words present in our lexicon to other words that co-occur with them in texts.

Figure 2 shows the framework we used to evaluate the performance of our idiom lexicon. The left-hand part of this figure represents the process of compilation and annotation of the lexicon described in the previous section. In this section, we describe the corresponding processing steps we followed for evaluation of the obtained lexicon, shown in the right-hand part of the figure.

4.1 Pre-processing and feature extraction

Pre-processing Datasets collected from online sources are noisy. At the pre-processing step, the noise and uninformative parts of the text were removed. This speeded up the rest of the process. In addition, normalization was done at this stage. We used Persian normalization algorithm JHAZM to normalize the words and phrases. For example, فیلم عَاَاَالّلی بود ("The movie was greatttttt") was changed to فیلم عَالی بود ("The movie was great") (Dashtipour et al. 2016; Nourian et al. 2015). This might possibly cause false positives in our

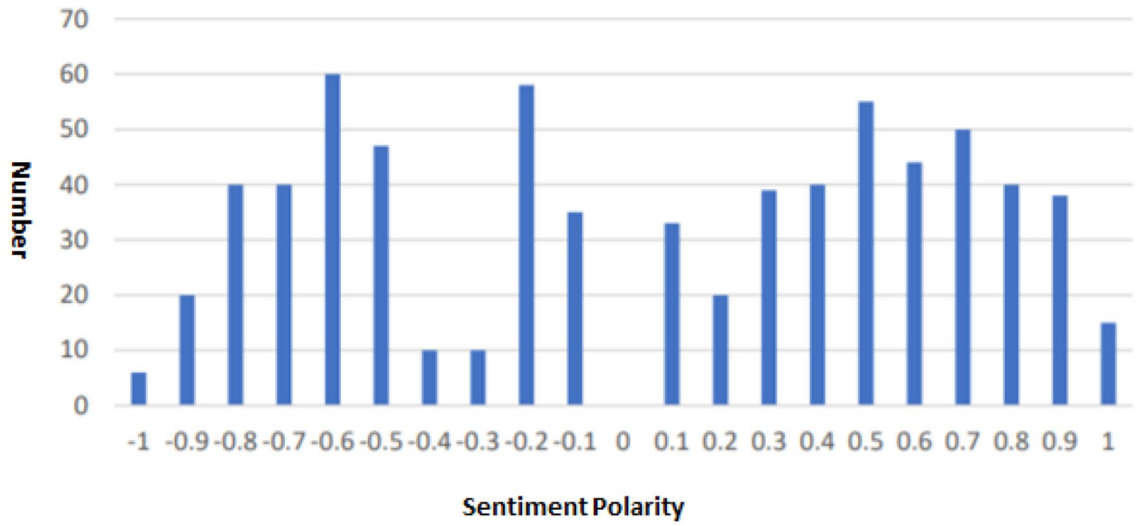

Fig. 1 Number of idioms in our lexicon by polarity

Table 2 Statistics of idioms in our lexicon by type

Type of idiom	Frequency
Dialectal	442
Slang	220
Light swear words	38

idiom detection procedure; in our future work, we plan to study the effect of pre-processing on idiom detection.

N-gram feature extraction In the experiments that involved machine learning algorithms, we used trigram and four-gram features. For example, فیلم بسیار بدي بود ("It was really bad movie") was transformed into two trigrams: فیلم بسیار بدي and بسیار بدي بود (Lopez-Gazpio et al. 2019).

Bag of words We also used the bag-of-words (unigram) features to reflect the frequency of occurrence of words for training the classifier (Ayadi et al. 2019). The bag-of-words features for the sentence الّه وکیلي فیلم بدي بود ("God Lawyer was bad movie") were الّه وکیلي ("God-Lawyer"), بود ("was"), فیلم ("movie"), and بدي ("bad"), each one with term frequency 1. We used adjectives, adverbs, verbs and nouns as features (Deshpande et al. 2019).

Idiom detection algorithm The idiom detector automatically detects the idioms in Persian sentences. First, the sentences are tokenise, after tokenization, the algorithm automatically identify the idiom, and it fed into machine learning classifiers for evaluation the performance of the approach.

Algorithm 1 presents the procedure that we used to detect idioms in a Persian text. As any lexicon-based algorithm, our procedure detects idioms present in the lexicon, but does not discover new idioms from the texts. Some idioms contain discontinuous phrases or words that can change, for example, "pull up his socks" or "pull up her socks"; a Persian example is قدم شمّا روي چشم (lit. "step on my eyes") or قدم تو روي چشم (lit. "step on his eyes"), which means

Table 3 Examples of idioms

Idiom	Idiomatic translation	Example of usage	Polarity	Assignment
تو دل برو (One in the heart)	Cute	She is so cute چه تو دل برو بود	+ 0.3	Manual
دمّاغ کسي بر خوردن (To eat someone's nose)	Did not like it (He / she did not like the situation)	John did not like what we told him. دمّاغش بر خورداز حرفمون بدش امد به	− 0.2	Automatic
به پیسي خوردن (To eat all)	Poor (pejorative)	They do not have money و اصلا پول ندارن پیسي خوردن	− 0.4	Automatic
زشت و بي ریخت (Unshaped)	Ugly (pejorative)	David is ugly دیوید بي ریخت هست	− 0.7	Automatic
به درك وَاصل شدن (Go to the hell)	Go to the hell (pejorative)	He go to hell به درك وَصل شدن	− 0.9	Automatic

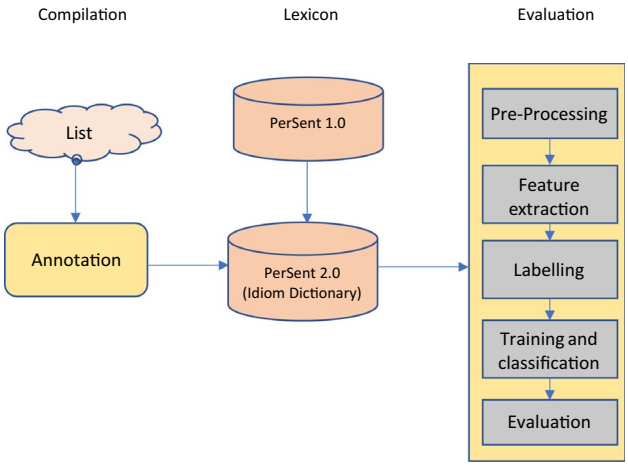

Fig. 2 Framework for compilation and evaluation of our idiom lexicon

"welcome". Our algorithm does not detect such types of idioms, which are statistically very rare in Persian, because in Persian verbs and nouns have no grammatical gender (Chen et al. 2014).

Persian idioms are usually located in the middle of the sentence. Persian sentences start with a noun and end with a verb, and the idiom is usually located in the sentence between noun and verb. The idioms can be detected automatically in any part of the sentence. The algorithm detects the idioms in the sentence even if multiple idioms are present in different parts of the sentence. For example, in the following sentence, فیلم مسخره و اعصاب خورد کني بود ("It is a mockery movie and it drives me crazy when I see it"), our algorithm identifies the idioms by tagging the idiom in the sentence: فیلم مسخرهو اعصاب خورد کنيبود (shown here by underlining instead of the idiom tag).

Example with positive polarity: وقعا دست مريزاد ميگم به تمام عوامل فیلم ("I would like to hurray bravo the whole team of the film"). The algorithm detects the idiom in the sentence and returns به تمام عوامل فیلم وقعا دست مريزادميگم..

Example with negative polarity: با دیدن این فیلم عمرم هدر دادم ("Seeing this film, I wasted my life"). The algorithm detects the idiom in the sentence and tags it as follows: با دیدن این فیلم عمرم هدر دادم

Example with multiple idioms: The algorithm can detect multiple idioms in the sentence: فیلم بود مزخرفي و اشغال ("the movie was trash and fiddle and faddle"). The algorithm tags the idioms in the sentence as فیلم اشغالو مزخرفيبود..

4.2 Labelling with average polarity

The PerSent Persian lexicon contains 1500 Persian words along with their polarity and part-of-speech tag: noun, adjective, adverb or verb (Dashtipour et al. 2016). As a variant of classification technique, we used the PerSent lexicon and our idiom lexicon to identify the average polarity of the features. For example, in فیلم بسیار خوبي بود ("it was great movie"), the word خوبي ("great") was searched in the PerSent lexicon to identify its polarity. If the idiom was detected in the sentence, the idiom lexicon was used to assign polarity to it. For example, for من فیلم دوست دارم ولي بازي بازیگرها چنگي به دل نميزد (lit. "I really love the movie but the acting did not play guitar with my heart"), the output was + 1 for the positive word and -0.4 for the negative idiom, and thus + 0.6 overall. We calculated the average polarity of the review as

$$AP = 1/n \sum_{i}^{n} P(w_i) \tag{1}$$

where AP denotes the average polarity, P denotes the polarity of the word, w_i denotes the ith word in the review and N denotes the number of words in the review. We used this value for automatic labelling of texts used as training examples for supervised classification algorithms, as explained below.

4.3 Direct and machine learning-based classification

The purpose of our classification experiments was twofold: to evaluate the usefulness of the developed lexicon for Persian sentiment analysis and to identify the most accurate machine learning algorithm for sentiment analysis of Persian texts. To show that our framework performed better with the PerSent 2.0 lexicon that includes the idiom lexicon than with the original PerSent 1.0 lexicon alone, we conducted various experiments in document-level polarity classification of Persian texts. In these experiments, we compared the results obtained with two versions of the lexicon: the old PerSent 1.0 (without idioms) and the new PerSent 2.0 (with idioms) on the task of document-level polarity classification. We also compared the techniques that rely on our lexicon with a baseline technique that does not rely on it.

We used three classification methods: a direct method, based on the average polarity from equation (1); a machine learning-based method that implicitly extended the annotation from the existing lexicon to a greater number of features, and a baseline experiment, based on automatically translation into English without using our PerSent lexicon or the idiom lexicon.

Direct classification: We used the average polarity from equation (1) to assign positive or negative polarity to texts. Namely, we assigned the negative label to the text if AP was negative; otherwise, we assigned positive polarity. Note that we never assigned neutral polarity, even when AP was zero.

Machine learning classification: We also tried a more sophisticated procedure that internally involved training a supervised classifier, with a distant-supervision approach. Namely, we first used the unsupervised direct classifier to identify the polarity in a large unlabelled dataset, and then trained a supervised classifier on these examples obtained with the unsupervised classifier. In order to train SVM, we have used RBF kernal, for naive Bayes the sample weight is equal to none used.

We did not experiment with using purely supervised approaches since the aim of our experiments was not to show how well one can classify Persian texts when one has a large enough corpus of manually labelled examples. Here we only to show how useful our lexicon is for classification of Persian texts in lexicon-based (distant-supervision) manner, without any manually labelled examples at all. A more complete distant-supervision approach could use for training both use manually labelled dataset and a much larger corpus automatically annotated with the help of our lexicon; however, currently, we do not have at our disposal such a large corpus. In addition, our current experiments are sufficient to demonstrate the value of our idiom lexicon.

In order to train the supervised classifier, the features were extracted from the reviews and the lexicon was used to assign polarity to the sentences. Then, we used this automatically labelled dataset to train a supervised classifier. We evaluated the trained classifier on our manually annotated dataset for which the polarity of the reviews was known. (The main reason for using binary classification was its efficiency.) We used five different classifiers: support vector machine (SVM), naive Bayes classifier, k-nearest neighbour classifier (kNN), decision tree and convolutional neural network (CNN). For the CNN, we used deeplearning4j, an open-source Java library (Nicholson and Gibson 2017).

The whole process was still unsupervised since no manually labelled examples were used for training; however, this process allowed us to extend implicitly the labels present in the lexicon to words that are not listed in the lexicon but co-occur with the words present in the lexicon, in order to improve the results. We did not export the learnt data in the form of a separate, larger lexicon with these automatically obtained sentiment values, but this is certainly possible. Such a list could be used in our future work, for example, for manual revision of the assigned sentiment values in order to extend the original lexicon.

Due to the small size of our manually labelled dataset and the lack of data in Persian language, in our experiments, we used the same dataset as a source of unlabelled data for the distant-supervision procedure. Namely, we used a tenfold cross-validation procedure, with the training portion of the dataset used only as a source of unlabelled data. For each fold, denote by A the test set with manually assigned polarity labels and by B a corpus obtained from the training set by removing the labels. Then we automatically annotated B using equation (1), trained a classifier on the obtained dataset and tested it on the manually annotated set A. Note that, in spite of using a training procedure internally, the whole process is lexicon-based and unsupervised since it does not use any manually labelled examples.

Classification via translation In this experiment, the whole dataset was translated into English using the Google translator. The translated sentences were passed to the TextBlob, and the average polarity was calculated. However, the TextBlob software was not able to identify the polarity for some idioms, because Google translator provided non-idiomatic translation (Loria et al. 2014).

4.4 Dataset used

To evaluate the performance of the framework, we used the following three datasets.

Movie Reviews dataset: Due to the lack of available resources for the Persian language, we had to compile our dataset. For this, we collected more than 1000 movie reviews from two popular movie review sites. We annotated these reviews as positive or negative and selected 500 positive and 500 negative reviews. The movie reviews in our dataset are on comedy and action movies, from 2014 to 2016. This dataset is rich in colloquial language, slang and idioms; for example, يك فيلم بَا لوده بَازي فرَاوَان (“It is movie with lots of zany acting”), however, as we have explained above, it does not contain profanity.

Persian VOA dataset: We used widely used benchmark Persian Voice of America (VOA) news dataset, which contains 500 positive and 500 negative news headlines. The language of headline news is quite formal; they contain much fewer informal or colloquial language expressions, slang words or idioms than the movie reviews (Mirsarraf et al. 2013).

Amazon reviews dataset: In order to compare the result, we used the Amazon reviews dataset, which contains more than one million English-language reviews on movies and TV. We used the bag of words of the reviews to compare the result with the Persian datasets. Table 4 summarizes the statistics of the datasets used. It includes the number of idioms detected in each dataset with our lexicon.

5 Experimental results

We measured the performance in terms of accuracy: the ratio of the number of correctly classified documents to the total number of documents in the dataset. In addition, to estimate the bias of the classifier, we report separately the recall of the positive and negative examples on the test set: the ratio of the correctly classified positive (respectively, negative)

texts to the total number of positive (respectively, negative) texts in the dataset. The difference between recall on the two classes measures the bias of the classifier.

Tables 5 and 6 show the results on the two datasets we used, with and without the use of our idiom lexicon, obtained with tenfold cross-validation procedure. All reported figures are averaged over the tenfold.

Table 7 shows that on the Movie Review dataset, the translation technique outperforms the original PerSent lexicon without the idiom lexicon, but is outperformed by our PerSent lexicon with the idiom lexicon, which again demonstrates the usefulness of our idiom lexicon for dealing with highly informal texts. For the VOA Persian dataset, translation showed comparable result, because this dataset contains much fewer informal and slang words, for which either translation would fail or TextBlob would not have data; see Table 4. Indeed, for formal texts, TextBlob is more complete than our small PerSent lexicon with only 1500 Persian words. Comparison with a very large English-language Amazon corpus suggests that our results are good enough and can be considered closed to the state of the art, though we cannot quantify this because results on different languages are incomparable. There are lots of studies used to translate Persian sentences into English and apply English lexicon. Therefore, we translate the Persian sentences into English, then apply machine learning classifiers results. The experimental results illustrate that the translated sentences achieved lower performance as compared to Persian sentences.

Table 4 Datasets used

Dataset	Language	Positive	Negative	Persian idioms
Movie Reviews dataset	Persian	500	500	524
Persian VOA news dataset	Persian	500	500	118
Amazon Reviews dataset	English	1000	1000	_

6 Discussion and further analysis

We applied five different classifiers to the Movie Review dataset and the Persian VOA dataset. Comparison of the results shows that CNN outperformed the other classifiers. This classifier gave better performance with our idiom lexicon than with the original PerSent lexicon alone.

Since the language in the Persian VOA dataset is more formal, the idiom lexicon was not as useful for it as it was for the movie review dataset, which includes many idioms, slang words and informal words. Our idiom lexicon is particularly useful for highly informal texts, which is especially important because the majority of texts on Internet to which opinion-mining techniques are typically applied are highly informal. The CNN classifier typically outperforms other classifiers in terms of accuracy; however, the CNN takes increasingly more time to train the model for large dataset (Iqbal et al. 2019).

In particular, in our experiments, CNN was trained to evaluate the performance of our lexicon. However, it showed superior performance in comparison with more traditional approaches such as SVM. We attribute this to our noisy data, which can cause low performance. At the pre-processing step, we did not remove all stop words, because this could affect automatic detection of idioms in the sentence. Various other factors affect the performance of the CNN classifier, such as the choice of feature extraction techniques, which is difficult to adapt to different types of dataset. The performance of the CNN improves when we increased the number of convolutional layers: five layers was enough to construct the model while increasing the number of pooling layers can deteriorate the results (Hassan and Mahmood 2017).

For the convenience of the reader, Figs. 3 and 4 show in the graphical form the data from Tables 5 and 6, respectively: the average overall accuracy of the five classifiers on the datasets using the original PerSent lexicon and after adding our idiom lexicon to it. One can see that the direct classifier that just uses the average polarity gave good results on both datasets. This shows that the average polarity can be used in practice to assign polarity labels to texts. However, the addition of distant supervision further improved the results, with the CNN classifier providing the best accuracy.

Table 5 Performance of different classifiers on the movie reviews dataset. R stands for Recall

Classifier	Without idioms			With idioms		
	R positive	R negative	Accuracy	R positive	R negative	Accuracy
Direct	0.651	0.642	0.646	0.762	0.754	0.758
Decision tree	0.667	0.654	0.66	0.775	0.768	0.771
Naive Bayes	0.672	0.662	0.667	0.752	0.748	0.75
SVM	0.675	0.667	0.671	0.778	0.771	0.774
kNN	0.681	0.671	0.676	0.749	0.746	0.747
CNN	0.707	0.692	0.699	**0.801**	**0.775**	**0.788**

Table 6 Performance of different classifiers on the Persian VOA dataset. R stands for Recall

Classifier	Without idioms			With idioms		
	R positive	R negative	Accuracy	R positive	R negative	Accuracy
Direct	0.674	0.668	0.671	0.732	0.726	0.729
Decision Tree	0.68	0.673	0.676	0.701	0.697	0.699
Naive Bayes	0.686	0.676	0.681	0.719	0.714	0.716
kNN	0.688	0.684	0.686	0.711	0.708	0.709
SVM	0.696	0.689	0.692	0.734	0.727	0.73
CNN	0.7	0.693	0.696	**0.755**	**0.746**	**0.75**

Table 7 Accuracy of classification via translation. R stands for recall

Dataset	R positive	R negative	Accuracy
SVM on translated Movie Reviews dataset	0.722	0.717	0.719
SVM on translated Persian VOA dataset	0.738	0.721	0.729
SVM on Amazon Reviews dataset	0.711	0.698	0.704

On both corpora, the use of the idiom lexicon yielded better results, though for the highly informal Movie Review dataset improvement was much greater than for the formal Persian VOA dataset. In particular, improvement on the Movie Review dataset was statistically significant for all classifiers, while improvement on the Persian VOA dataset was statistically significant only for the direct classifier.

The fact that the figures of the recall on the positive polarity class and the negative polarity class are similar shows that the classifiers are not biased since the datasets are balanced. Figures 3 and 4 show that direct classifier with idioms performed better than some of machine learning algorithms. We can attribute such cases to wrong generalizations made from co-occurrences of some unigrams or n-grams by chance with some idioms, due to the very small size of the available raw corpus for training the supervised classifiers.

Figures 5 and 6 show the distribution of all examples and the correctly classified examples by average polarity obtained with the equation (1) with the original PerSent lexicon and after we added our new idiom lexicon to it. Note that the figures show absolute numbers, not percentages. The figures suggest that the addition of the idioms made the analysis more detailed. Indeed, while the number of documents assigned almost neutral polarity or extreme polarity (completely positive or completely negative) has decreased with the addition of the idioms lexicon, the number of documents assigned moderate polarity (positive or negative) has increased. This indicates a more balanced and specific estimation of the sentiment conveyed by those documents since very few real texts convey extremely positive or extremely negative feelings.

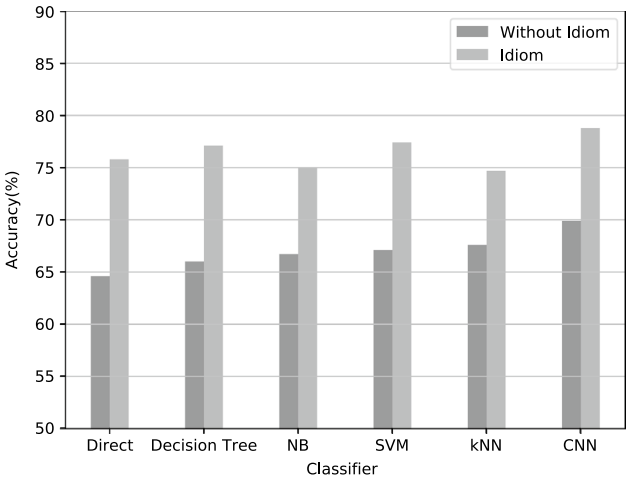

Fig. 3 Comparison of the result for five different classifiers for the Movie Reviews

7 Conclusions and future work

We have extended the PerSent Persian sentiment lexicon with 1000 idiomatic expressions. The resulting lexicon is not only useful for detect idioms in Persian texts but also for accurately classifying Persian texts. We used several algorithms to evaluate the performance of our lexicon on the polarity detection task and have shown that it improves the performance of the classifiers, especially on user-contributed contents rich in informal language.

We have also shown that the use of deep learning algorithms, especially CNN, to extend implicitly the annotation from the sentiment lexicon to the words not included in it in a distant-supervision manner improves the results. This

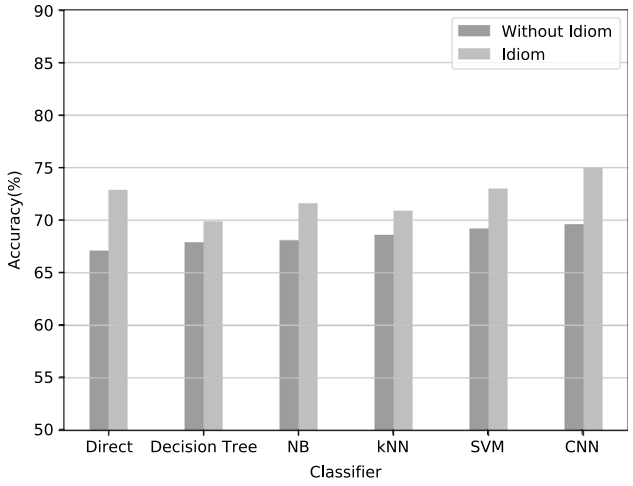

Fig. 4 Comparison of the result for five different classifiers for the Persian VOA dataset without idiom and with our idiom lexicon

leaves the whole labelling process lexicon-based and unsupervised since the training data for the machine learning algorithms are obtained automatically with a lexicon-based algorithm; no manually annotated examples are involved in training.

As part of our future work, we plan to develop a multilingual idiom detection framework for English and Persian languages. We also plan to overcome certain shortcomings of our idiom detection method: to enable it to classify code-mixed texts, to identify multiple meanings of words and to deal with cultural and regional language variation. Namely, there are words in Persian that have several meanings: e.g. خورشید (sun) can be used to say بازیگران مثل خورشید میدرخشیدند ("the male actor was shining like sun"). Our current sentiment classification method is unable to detect these peculiarities in the text. We plan to extend it to be able to distinguish such figures of speech as, in this case, metaphor.

Another scenario where our framework needs enhancement is code-mixed text, which includes a mixture of Persian

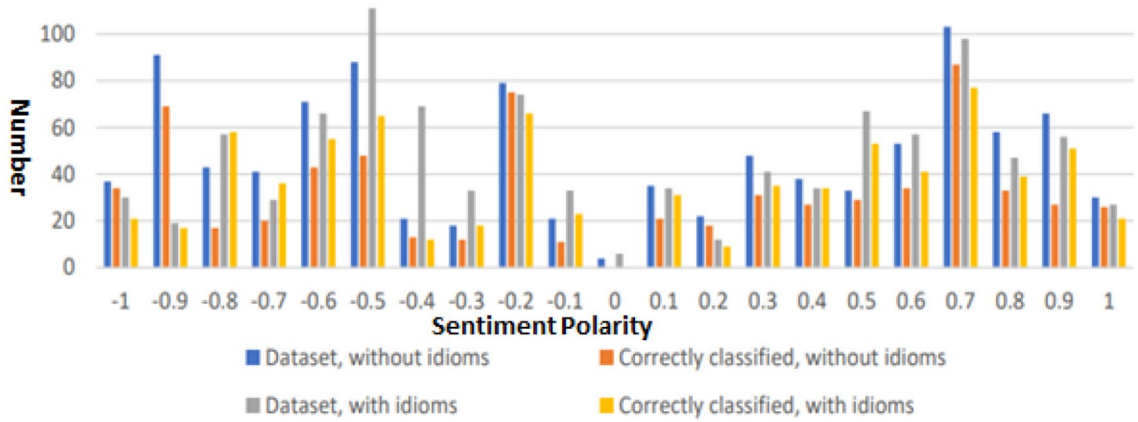

Fig. 5 Number of documents in the Movie Review dataset by average polarity according to equation (1) before and after adding idioms to the PerSent lexicon

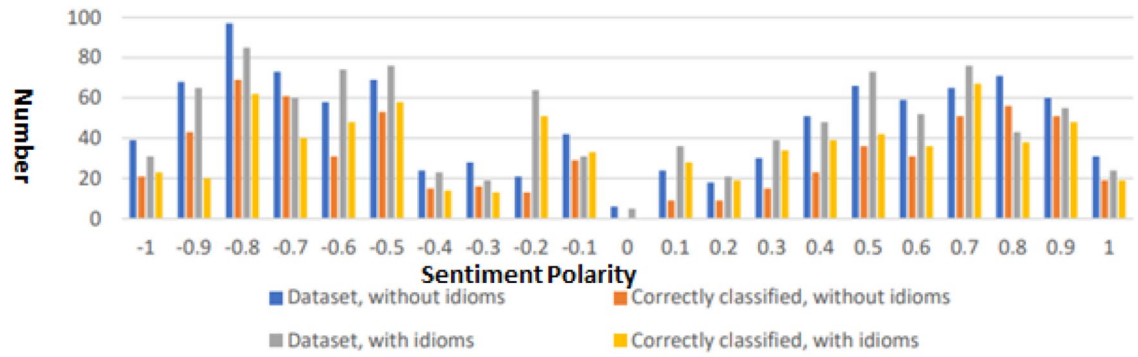

Fig. 6 Number of documents in the Persian VPA dataset by average polarity according to equation (1) before and after adding idioms to the PerSent lexicon

and English idioms. For example, "Dude, تو کار بزرگي انجام دادي ("Dude, you have done a great job").

Some of the slang words are culture-specific or region-specific. For example, "Laila and Majnun" is an ancient Persian love story. It is currently difficult for idiom detectors to recognize that ليلي و مجنون ("Laila and Majnun") can be translated as "lovers". Our idiom detection framework will be enhanced to detect such culture-specific or region-specific slang. Similarly, our method is to be adapted to handle different dialects of the Persian language. These goals can be achieved both by including in our lexicon manually annotated region- and dialect-specific idioms and by automatically identifying idioms already included in our lexicon as region- or dialect-specific. The latter can be done using dialect-specific corpora and corpora with geo-localization information.

Finally, our algorithm currently allows detection of only those idioms that are manually included in our lexicon. However, new idiomatic and slang expressions constantly appear in language, especially in the language of Internet, microblogging and social networks. Automatic detection of new slang and idiomatic expressions from raw texts and automatic discovery of their sentiment value is an ambitious goal, which would probably involve deep learning techniques and semantic analysis of the text. This is particularly challenging for a resource-poor language such as Persian.

References

Cavallari S, Cambria E, Cai H, Kevin C-CC, Vincent WZ (2019) Embedding both finite and infinite communities on graphs. IEEE Comput Intell Mag 2:1069

Yang H-C, Lee C-H, Chun-Yen W (2018) Sentiment discovery of social messages using self-organizing maps. Cogn Comput 10(6):1152–1166

Guellil I, Adeel A, Azouaou F, Benali F, Hachani A-E, Dashtipour K, Gogate M, Ieracitano C, Kashani R, Hussain A (2021) A semi-supervised approach for sentiment analysis of arab (ic+ izi) messages: application to the algerian dialect. SN Computer Science 2(2):1–18

Cambria E, Poria S, Hussain A, Liu B (2019) Computational intelligence for affective computing and sentiment analysis [guest editorial]. IEEE Comput Intell Mag 14(2):16–17

Dragoni M, Federici M, Rexha A (2019) Reus: a real-time unsupervised system for monitoring opinion streams. Cognit Comput 53:1–20

Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, Ali A, Sheikh A (2021) Artificial intelligence-enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the UK and the USA: observational study. J Med Internet Res 23(4):e26627

Zikopoulos P, Eaton C et al (2011) Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, London

Dashtipour K, Taylor W, Ansari S, Gogate M, Zahid A, Sambo Y, Hussain A, Abbasi Q, Imran M (2021) Public perception towards fifth generation of cellular networks (5G) on social media. Front Big Data 2:1036

Wang Q-F, Cambria E, Liu C-L, Hussain A (2013) Common sense knowledge for handwritten chinese text recognition. Cogn Comput 5(2):234–242

Li J, Jiang F, Yang J, Kong B, Gogate M, Dashtipour K, Hussain A (2021) Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing 465:15–25

Dashtipour K, Gogate M, Gelbukh A, Hussain A (2021) Persian sentence-level sentiment polarity classification. In: ICOTEN

Ling LS, Cambria E, Chiong R, Cornforth D (2016) A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. Knowledge-Based Syst 105:236–247

Dashtipour K, Gogate M, Li J, Jiang F, Kong B, Hussain A (2019) A novel hybrid persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. Neurocomputing (In-Press)

Dashtipour K, Gogate M, Adeel A, Algarafi A, Howard N, Hussain A (2017) Persian named entity recognition. In: 2017 IEEE 16th international conference on cognitive informatics & cognitive computing (ICCI* CC), pp 79–83. IEEE

Dashtipour K, Gogate M, Adeel A, Ieracitano C, Larijani H, Hussain A (2018) Exploiting deep learning for persian sentiment analysis. In: International conference on brain inspired cognitive systems, pp 597–604. Springer

Dashtipour K, Gogate M, Adeel A, Hussain A, Alqarafi A, Durrani T (2017) A comparative study of persian sentiment analysis based on different feature combinations. In: International conference in communications, signal processing, and systems, pp 2288–2294. Springer,

Ieracitano C, Adeel A, Gogate M, Dashtipour K, Morabito FC, Larijani H, Raza A, Hussain A (2018) Statistical analysis driven optimized deep learning system for intrusion detection. In: International conference on brain inspired cognitive systems, pp 759–769. Springer

Jiang F, Kong B, Li J, Dashtipour K, Gogate M (2021) Robust visual saliency optimization based on bidirectional markov chains. Cogn Comput 13:69–80

Dashtipour K, Gogate M, Gelbukh A, Hussain A (2021) Adopting transition point technique for persian sentiment analysis. In: ICOTEN

Basiri ME, Kabiri A (2019) Homper: A new hybrid system for opinion mining in the persian language. J Inf Sci 3:0165551519827886

Nezhad ZB, Deihimi MA (2019) A combined deep learning model for persian sentiment analysis. IIUM Eng J 20(1):129–139

Dashtipour K, Hussain A, Gelbukh A (2017) Adaptation of sentiment analysis techniques to persian language. In: International conference on computational linguistics and intelligent text processing, pp 129–140. Springer

Dashtipour K, Gogate M, Li J, Jiang F, Kong B, Hussain A (2020) A hybrid persian sentiment analysis framework: integrating dependency grammar based rules and deep neural networks. Neurocomputing 380:1–10

Gogate M, Dashtipour K, Adeel A, Hussain A (2020) Cochleanet: a robust language-independent audio-visual model for real-time speech enhancement. Inf Fus 63:273–285

Ahmed R, Gogate M, Tahir A, Dashtipour K, Al-Tamimi B, Hawalah A, El-Affendi MA, Hussain A (2021) Deep neural network-based contextual recognition of Arabic handwritten scripts. Entropy 23(3):340

Gogate M, Hussain A, Huang K (2019) Random features and random neurons for brain-inspired big data analytics. In: 2019 international conference on data mining workshops (ICDMW), pp 522–529. IEEE

Gogate M, Adeel A, Hussain A (2017) Deep learning driven multimodal fusion for automated deception detection. In: 2017 IEEE symposium series on computational intelligence (SSCI), pp 1–6. IEEE

Khoshnevisan B (2019) Spilling the beans on understanding English idioms using multimodality: an idiom acquisition technique for Iranian language learners. Int J Lang Transl Intercult Commun 8:128–143

Gogate M, Dashtipour K, Bell P, Hussain A (2020) Deep neural network driven binaural audio visual speech separation. In: 2020 international joint conference on neural networks (IJCNN), pp 1–7. IEEE

Dashtipour K, Gogate M, Adeel A, Larijani H, Hussain A (2021) Sentiment analysis of persian movie reviews using deep learning. Entropy 23(5):596

Gogate M, Adeel A, Dashtipour K, Derleth P, Hussain A (2019) Av speech enhancement challenge using a real noisy corpus. arXiv preprint arXiv:1910.00424

Mullen T, Malouf R (2006) A preliminary investigation into sentiment analysis of informal political discourse. In: AAAI spring symposium: computational approaches to analyzing weblogs, pp 159–162

Mansouri M (2015) Idiomatic expressions in the frame work of minimalist approach: evidence from Persian. Lang Related Res 6(3):271–292

Dashtipour K, Hussain A, Zhou Q, Gelbukh A, Hawalah AY, Cambria E (2016) Persent: a freely available persian sentiment lexicon. In: International conference on brain inspired cognitive systems, pp. 310–320. Springer

Langlotz A (2006) Idiomatic creativity: a cognitive-linguistic model of idiom-representation and idiom-variation in English, vol 7. John Benjamins Publishing, London

Fraser B (1970) Idioms within a transformational grammar. Found Lang 5:22–42

Passaro L, Silvio GSM, Lenci A (2019) Do idioms have a heart? the side (sentiment of idiomatic expressions) project. In: 11th international conference on the mental lexicon, pp 1–4. CAN

Nippold MA, Martin ST (1989) Idiom interpretation in isolation versus context: a developmental study with adolescents. J Speech Lang Hear Res 32(1):59–66

Liu D (2003) The most frequently used spoken American English idioms: a corpus analysis and its implications. TESOL Q 37(4):671–700

Erman B, Warren B (2000) The idiom principle and the open choice principle. Text-Interdiscip J Study Discourse 20(1):29–62

Williams L, Bannister C, Arribas-Ayllon M, Preece A, Spasić I (2015) The role of idioms in sentiment analysis. Expert Syst Appl 42(21):7375–7385

Liang W, Morstatter F, Liu H (2018) Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. Lang Resour Eval 52(3):839–852

Ibrahim HS, Abdou SM, Gheith M (2015) Sentiment analysis for modern standard arabic and colloquial. arXiv preprint arXiv:1505.03105

Ibrahim HS, Abdou SM, Gheith M (2015) Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis. arXiv preprint arXiv:1506.01906

Gul N (2014) Sctur: a sentiment classification technique for urdu text. Int J Comput Commun Syst Eng 1(3):97–101

Raj S, Kajla T (2015) Sentiment analysis of swachh bharat abhiyan. Bus Anal Intell 5:32

Wang L, Yu S (2010) Construction of Chinese idiom knowledge-base and its applications. In: Proceedings of the 2010 workshop on multiword expressions: from theory to applications, pp 11–18

Xie S, Wang T (2014) Construction of unsupervised sentiment classifier on idioms resources. J Cent South Univ 21(4):1376–1384

Wu L, Morstatter F, Liu H (2016) Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification. arXiv preprint arXiv:1608.05129

Verma R, Vuppuluri V (2015) A new approach for idiom identification using meanings and the web. In: Proceedings of the international conference recent advances in natural language processing, pp 681–687

Citron Francesca MM, Cristina C, Michael K, Luna B, Markus C, Jacobs Arthur M (2016) When emotions are expressed figuratively: psycholinguistic and affective norms of 619 idioms for German (panig). Behav Res Methods 48(1):91–111

Djemaa M, Candito M, Muller P, Vieu L (2016) Corpus annotation within the french framenet: a domain-by-domain methodology. In: Tenth international conference on language resources and evaluation (LREC 2016)

Fleiss A, Han C, Sasha S, DiPietro DM (2020) Constructing equity portfolios from sec 13f data using feature extraction and machine learning. J Financial Data Sci 2(1):45–60

Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AYA, Gelbukh A, Zhou Q (2016) Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput 8(4):757–771

Nourian A, Rasooli MS, Imany M, Faili H (2015) On the importance of ezafe construction in persian parsing. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, Vol 2: Short Papers, pp 877–882

Lopez-Gazpio I, Maritxalar M, Lapata M, Agirre E (2019) Word n-gram attention models for sentence similarity and inference. Expert Syst Appl 132:1–11

Ayadi W, Elhamzi W, Charfi I, Atri M (2019) A hybrid feature extraction approach for brain mri classification based on bag-of-words. Biomed Signal Process Control 48:144–152

Deshpande A, Aneja J, Wang L, Schwing Alexander G, Forsyth David (2019) Fast, diverse and accurate image captioning guided by part-of-speech. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10695–10704

Chen K-H, Huang G-S, Chia-Tung Lee R (2014) Bit-parallel algorithms for exact circular string matching. Comput J 57(5):731–743

Nicholson AC, Gibson A (2017) Deeplearning4j: Open-source, distributed deep learning for the jvm. Deeplearning4j.org,

Loria S, Keen P, Honnibal M, Yankovsky R, Karesh D, Dempsey E et al. (2014) Textblob: simplified text processing. Second TextBlob Simplified Text Process

Mirsarraf MR, Dehghani N (2013) A dependency-inspired semantic evaluation of machine translation systems. In: International conference of the cross-language evaluation forum for European languages, pp 71–74. Springer

Iqbal F, Hashmi JM, Fung BCM, Batool R, Khattak AM, Aleem S, Hung PCK (2019) A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access 7:14637–14652

Hassan A, Mahmood A (2017) Deep learning approach for sentiment analysis of short texts. In: 2017 3rd international conference on control, automation and robotics (ICCAR), pp 705–710. IEEE