PerSent: A Freely Available Persian Sentiment Lexicon

Kia Dashtipour[1(✉)], Amir Hussain[1], Qiang Zhou[2],
Alexander Gelbukh[3], Ahmad Y.A. Hawalah[4], and Erik Cambria[5]

[1] Department of Computing Science and Mathematics,
University of Stirling, Stirling FK9 4LA, Scotland
kd28@cs.stir.ac.uk
[2] Tsinghua University, Beijing, China
[3] CIC, Instituto Politécnico Nacional, 07738 Mexico City, Mexico
[4] Taibah University, Madina, Saudi Arabia
[5] School of Computer Engineering, Nanyang Technological University,
Singapore, Singapore

Abstract. People need to know other people's opinions to make well-informed decisions to buy products or services. Companies and organizations need to understand people's attitude towards their products and services and use feedback from the customers to improve their products. Sentiment analysis techniques address these needs. While the majority of Internet users are not English speakers, most research papers in the sentiment-analysis field focus on English; resources for other languages are scarce. In this paper, we introduce a Persian sentiment lexicon, which consists of 1500 words along with their part-of-speech tags and polarity scores. We have used two machine-learning algorithms to evaluate the performance of this resource on a sentiment analysis task. The lexicon is freely available and can be downloaded from our website.

1 Introduction

In recent years, with fast growing of Internet people all around the world share their opinions on different topics. This huge amount of unstructured data available online in different languages is very useful for companies and organisations to improve their products and services (Poria et al. 2014).

The corresponding field of science and technology is called sentiment analysis (SA). SA techniques involve a number of tasks, among them identification of the polarity (positive/negative) or emotion (happy, sad, angry, etc.) expressed the text or in a sentence (Turney 2002). Sentiment polarity can be binary or can involve multiclass classification, such as strongly positive, positive, neutral, negative, and strongly negative. Most of research is focused on the binary polarity classification, though identifying at least the neutral opinion in the sentence is more helpful (Tang et al. 2009).

In the recent years, sentiment analysis has been a very active area of research. There have been compiled numerous lexical resources and datasets for English language. However, much less effort has been devoted to the development of lexical resources in other languages, which makes is difficult for researchers to analyze the text in

languages other than English because of lack of available lexical resources (Dashtipour et al. 2016).

In particular, there is no well-known dataset or lexicon available for Persian language (Neviarouskaya et al. 2011). In this paper, we present PerSent, a Persian polarity lexicon for sentiment analysis, which contains words and phrases along with their polarity and part-of-speech tag. We evaluate its quality and performance via applications to a sentiment analysis task using different features such as POS-based features, the presence and frequency of sentiment words, average polarity of words, etc., and two machine-learning algorithms: SVM and Naïve Bayes.

The lexicon is freely available for the research community and can be downloaded from the URL http://www.gelbukh.com/resources/persent.

This paper is organized as follows. Section 2 discusses related work on lexicons in languages other than English; Sect. 3 presents PerSent, our Persian sentiment lexicon; Sect. 4 describes our evaluation methodology, and Sect. 5 gives the evaluation results. Section 6 concludes the paper.

2   Related Work

Data analysis is important for small and large companies. They gather opinions from texts available in Internet. Analysis of such opinions has great impact on customer relationships. Companies use customer comments about negative features of products to improve their products (Cambria et al. 2016). Moreover, sentiment analysis is not only restricted to product reviews but is also used in other fields such as politics, sport, etc.

In this section, we give some background on sentiment classification and discuss related work.

2.1   Types of Sentiment Analysis

Approaches to sentiment classification can be divided into three groups: statistical approaches, knowledge-based approaches, and hybrid approaches.

Statistical approaches use machine-learning algorithms such as SVM or Naïve Bayes to classify text. They can use supervised or unsupervised learning methods. Supervised methods use labelled data to classify the text, while unsupervised ones use only raw data (Maynard and Funk 2012). Statistical approaches are usually used, for example, to detect sentiments holders and target (Cambria et al. 2013).

Knowledge-based approaches classify the text by affect categories based on the presence of unambiguous affect words such as *sad*, *happy*, *afraid*, or *bored* (Cambria 2016). They use lexicons to calculate the statistics of positive and negative words in the given text: for example, the word *good* is known to be positive and the word *bad* negative. The lexicon can contain single words or phrases. The advantage of knowledge-based approaches is that they do not require trained data; the main disadvantage is lack of scalability.

Hybrid approaches combine statistical and knowledge-based methods to improve performance and accuracy (Maynard and Funk 2012; Cambria 2016). Pak and Paroubek

(2010) developed a dataset that contains positive and negative documents; for classification, they calculate the cosine similarity between the given document and the documents with known polarity. They evaluated their method using the Naïve Bayes algorithm.

2.2   Knowledge-Based Approaches

Various lexicon-based approaches have been used for sentiment classification of documents in different languages; see Table 1. Most of the lexicon-based approaches used adjectives to identify the polarity of the text. There have been suggested different methods to develop sentiment lexicons, such as manual, corpus-based, and dictionary-based compilation. Manual construction is time consuming; it is usually combined with other methods to improve performance.

Corpus-based methods use lists of sentiment words along with their polarity and syntactic patterns to find more sentiment words and their polarity. For example, Hatzivassiloglou and McKeown (1997) developed graph-based technique for learning lexicons; they identified polarity of adjectives using conjunctions. They used a clustering algorithm to divide words into positive and negative. They achieved 82% of accuracy.

Table 1. Existing sentiment lexicons for various languages

Reference	Lexicon name	Language	Comments
(Elarnaoty et al. 2012)	MPQA Subjectivity Lexicon	Arabic	Arabic news. POS tags
(Elhawary and Elfeky 2010)	Arabic Lexicon for business reviews	Arabic	Word polarity. 200 positive, 250 negative
(Abdul-Mageed and Diab 2014)	SANA	Arabic and dialects	3,325 adjectives, 617 positive, 550 negative, 2158 objective
(Dehkharghani et al. 2015)	SentiTurkNet	Turkish	15,000 synsets positive, negative and neutral. POS tags
(Sidorov et al. 2013)	SEL	Spanish	Emotion lexicon. 2036 words on different emotions such as joy, sadness, surprise, disgust
(Remus et al. 2010)	SentiWortSchatz	German	1818 positive, 1650 negative. Adjectives, adverbs, nouns, verbs. POS tags
(Waltinger 2010)	German Polarity Clues	German	10,141 features associated with positive, negative and neutral
(de Albornoz et al. 2012)	SentiSense	English	5,496 words and 2,190 synsets labelled with an emotion
(Mahyoub 2014)	ArabicWordNet	Arabic	23481 words: 15,890 nouns, 6,048 verbs, 1,234 adjectives, 264 adverbs

Dictionary-based approaches do not require pre-compiled lists of sentiment words. They are used to collect sentiment words and their orientation manually and look up synonyms and antonyms in a dictionary. The main disadvantage of this method is that it unable to find sentiment words with domain-specific orientation: sentiment words can be positive in one domain and negative in another. For example, the word *large* is positive when it refers to a computer screen, but negative when it refers to a mobile phone (Hu and Liu 2004).

2.3   Persian Language

Persian uses 32 letters, which cover 28 Arabic letters. Its writing system includes special signs and diacritic marks that can be used in different forms or omitted from the word. Short vowels are not indicated in writing. There are letters with more than one Unicode encoding. Some words have more than one spelling variants. Spelling of some words changes with time. All this increases the number of both homographs and synonyms, which presents problems in computational treatment of Persian (Karimi 1989; Seraji et al. 2012).

Saraee and Bagheri (2013) proposed a method for feature selection in Persian sentiment analysis able to calculate the co-occurrence of Persian words in different classes. They used customer reviews to evaluate the performance of the approach. Naïve Bayes algorithm has been used in evaluation. The overall accuracy of their approach was 75%. The advantage of this approach is its simplicity; a disadvantage is the need of a great amount of training data.

Chen and Skiena (2014) proposed a lexicon for major languages such as English, Arabic, Japanese, and Persian. The English data has been collected online. They used Google translator to translate data into different languages and WordNet to gather synonyms and antonyms for English; these words and phrases were translated into different languages. They used Wikipedia pages to evaluate the performance of their lexicon, and obtained the overall performance of 45.2%. An advantage of this approach is its ability to develop lexicons for 136 languages; a disadvantage is that the lexicons for most of these languages were only of less than one hundred words and phrases.

3   PerSent Persian Sentiment Lexicon

Many researchers note that the main problem of the multilingual sentiment analysis is the lack of resources. To overcome this issue, we developed a Persian lexicon of 1500 Persian words along with their polarity and part of speech tag, which we called Per-Sent. Table 2 shows some examples.

Most of the previous research on sentiment used adjectives to identify the polarity of sentences (Hu and Liu 2004). Some researchers used adverbs and adjectives together to build a lexicon (Benamara et al. 2007); some used adjectives, adverbs, and verbs (Taboada et al. 2011). For our Persian sentiment lexicon, we used adjectives, adverbs, verbs, and nouns, because all these words and phrases are useful to determine the polarity of the sentence.

Table 2. Examples from our Persian sentiment lexicon

Word	Translation	POS	Polarity	Score
خوب	good	adjective	positive	+0.7
بد	bad	adjective	negative	−0.69
زشت	ugly	adjective	negative	−0.7
آزردن	annoying	verb	negative	−0.1689

Table 3. Statistics by POS

Part of speech	Words
Adjective	556
Adverbs	51
Nouns	371
Verbs	522

A lexicon can be developed in different ways, such as manually or using existing lexicons such as SentiWordNet (Esuli and Sebastiani 2006) or General Inquirer (Stone et al. 1966). The words and phrases used in our lexicon were taken from different resources such as movie review website, weblogs, and Facebook. There were four different categories of sources, namely, websites related to movies, news, mobile phones, and computers.

We manually assigned polarity between −1 and +1 to each word and phrase. The degree of intensity was indicated: e.g., "خوشحال" (*happy*), "بشاش" (*cheerful*), and "شاد" (*delighted*) have different positive values. In order to assign polarity manually to some words and phrases, we used the TextBlob Python package, used to assign polarity to words, phrases, and sentences in English (Yang 2015. For this, we translated Persian words into English. We also manually assigned a part of speech (POS) tag to each word or phrase. Table 3 shows the distribution of the POS tags in the lexicon.

4 Evaluation Methodology

In order to evaluate the performance of our lexicon, we used two classification algorithms; we used our lexicon to assign polarity to the features extracted from the dataset. Figure 1 shows the general framework we used to evaluate the performance of our lexicon. Below we describe each processing steps.

Pre-processing. The pre-processing step consisted of four parts, tokenisation, normalisation, stop-word removal, and stemming. Normalization was used to remove noise from the text. Stemming was used to remove inflection of infected forms: the lexicon only provides the base form of the words.

Feature selection. The purpose of the feature selection was to remove unnecessarily features, which improved the performance and efficiency of the classification.
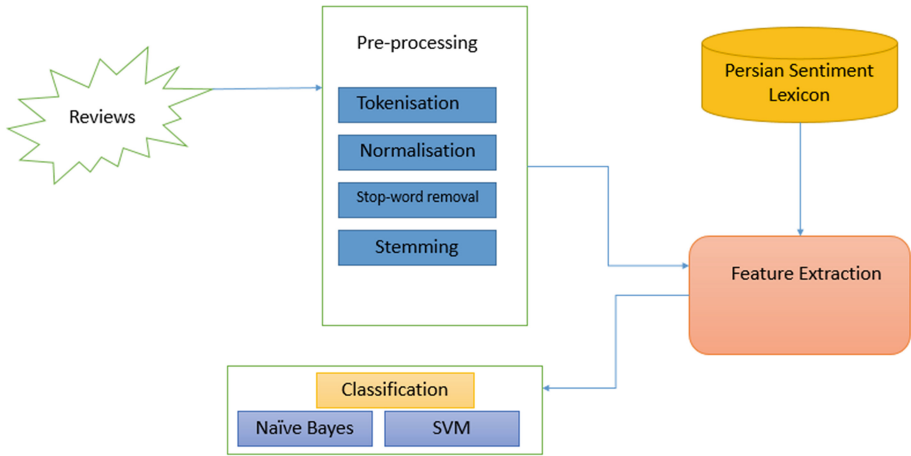
Fig. 1. The Persian framework

Table 4. Features used

Category	Features
Presence and frequency of sentiment words	Frequency of positive words
	Frequency of negative words
	Presence of positive words
	Presence of negative words
POS-based features	Frequency of positive and negative adjectives
	Frequency of positive and negative adverbs
	Frequency of positive and negative verbs
	Frequency of positive and negative nouns
Word polarity	Overall polarity of negative words
	Overall polarity of positive words

The features we used were based on word polarity, POS tag, and presence and frequency of sentiment words; see Table 4.

Presence and frequency of sentiment words. The sentiment words identify the overall polarity for sentiment classification. Example of positive words in Persian are "زیبا" (*beautiful*) and "عالی" (*excellent*), and of negative words are "زشت" (*ugly*) and "بد" (*bad*). The features of presence of positive and of negative words (two different features) are binary, without considering the number of occurrences of a given word, while the other two features are integer and indicate the number of occurrences of positive and of negative works, correspondingly.

POS-based features. Our lexicon contains words along with their POS tag, such as adverb, verb, noun, or adjective. Most of the previous research used only adjectives and nouns to identify the polarity of sentences (Kouloumpis et al. 2011), but we consider

eight different features: the frequencies of positive and of negative adjectives, adverbs, verbs, and nouns, correspondingly.

Word Polarity. Our lexicon gives polarity for words. As two different features, we used the overall polarity of negative and of positive words, correspondingly.

5   Experimental Results

We applied simple baseline approaches to sentiment analysis using our lexicon to the Persian VOA (Voice of America) news corpus, which contains 500 positive and 500 negative news headlines. We then measured the performance in terms of accuracy:

$$\text{Accuracy} = \frac{\textit{number of data classified correctly}}{\textit{total number of data}}$$

5.1   Results

We used support vector machine (SVM) and Naïve Bayes classifier for evaluation. The support vector machine gave better results than Naïve Bayes; see Table 5. In this experiment all the features were used.

Table 5. Performance of different classifiers with all features

Classifier	Accuracy
SVM	**69.54%**
Naïve Bayes	65.02%

We also compared the effectiveness of different features in order to determine their importance; see Table 6. The accuracy varied from 46% to 63%. SVM gave uniformly better results than Naïve Bayes did. The experiment showed that the mere presence of opinion words gives better performance than their frequency.

Table 6. Performance of the frequency features

	SVM	Naïve Bayes
Frequency of positive words	54.28%	53.74%
Frequency of negative words	52.01%	46.02%
Presence of positive words	61.12%	52.39%
Presence of negative words	59.98%	46.17%
All frequency features	**63.00%**	59.23%

We also compared the POS features, such as the frequency of positive and negative adjectives, adverbs, verbs, and nouns, correspondingly; see Table 7. SVM again almost uniformly outperformed Naïve Bayes.

Table 7. Performance of the POS features

	SVM	Naïve Bayes
Frequency of positive and negative adjectives	65.01%	61.23%
Frequency of positive and negative adverbs	63.83%	59.07%
Frequency of positive and negative verbs	64.25%	61.04%
Frequency of positive and negative nouns	61.49%	63.34%
All POS features	**69.29%**	63.19%

Table 8 shows the results for overall polarity of negative and of positive words. Positive words outperformed negative words, and SVM outperformed Naïve Bayes.

Table 8. Performance of the overall polarity feature

	SVM	Naïve Bayes
Overall polarity of negative words	57.23%	56.49%
Overall polarity of positive words	66.02%	64.97%

5.2　Discussion

Based on related work on lexicon-based methods, we expected that PerSent lexicon would perform better. Classification of news into positive and negative is, however, a difficult task, because most of the bad news do not contain any subjective terms that would help to classify them as negative.

The main problem of our lexicon is its relatively small size: 1500 words are not enough for Persian because it has many dialects and actively us idiomatic expressions, and thus requires a larger lexicon, development of which would take time and effort (He and Zhou 2011).

Another problem is that our simple application did not properly handle sarcasm. A much more sophisticated system should be developed to be able to identify sarcasm in the texts. Further study is required to detect ironic and sarcastic sentences. Sarcasm should be studied independently and another tool needs to be developed to handle sarcasm in order to improve our classification performance.

Similarly, our simple testing application did not properly handle code switching between Persian and English: some sentences used a mixture of Persian and English words.

Adjectives gave better results in comparison with other parts of speech, because the examination of adjectives in a sentence is easier as compared with other words. For example, in "عکس زیبایی است", which means "It is beautiful picture", the adjective clearly indicates the sentiment.

Rather not surprisingly, all features together gave better results than individual features separately, because in this way the algorithm had access to more information.

6 Conclusions

We have developed a new lexicon for Persian language, which can be used for Persian sentiment analysis. The lexicon contains 1500 Persian words along with their polarity on a numeric scale from −1 to +1 and the part of speech of tag. The majority of the values were assigned manually. The new lexicon is freely available for download from the URL http://www.gelbukh.com/resources/persent.

Our experiment results show that our lexicon is a useful tool to determine the polarity of sentences in Persian. In the experiments, we used two classifiers: SVM and Naïve Bayes, of which SVM gave better results.

As future work, we plan to extend our lexicon, try computer-assisted methods of its compilation, as well as to apply our lexicon to a wider variety of tasks and corpora. In addition, we will combine knowledge-based methods with deep textual features for sentiment classification (Poria et al. 2015a). An end-to-end Persian sentiment analysis framework based on the linguistic patterns and common-sense knowledge is another important work to be done (Poria et al. 2015b, 2012; Cambria et al. 2015). Aspect-based sentiment analysis (Poria et al. 2016) and disambiguating sentiment words (Pakray et al. 2011a, 2011b, 2010) will play a major role in such a framework.

References

Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst. (TOIS) **26**(3), 12 (2008)

Abdul-Mageed, M., Diab, M.T.: SANA: a large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In: LREC, pp. 1162–1169 (2014)

Benamara, F., Cesarano, C., Picariello, A., Recupero, D.R., Subrahmanian, V.S.: Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: ICWSM (2007)

Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**(2), 102–107 (2016)

Cambria, E., Howard, N., Xia, Y., Chua, T.S.: Computational intelligence for big social data analysis. IEEE Comput. Intell. Mag. **11**(3), 8–9 (2016)

Cambria, E., Poria, S., Bisio, F., Bajpai, R., Chaturvedi, I.: The CLSA model: a novel framework for concept-level sentiment analysis. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 3–22. Springer, Heidelberg (2015). doi:10.1007/978-3-319-18117-2_1

Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)

Cambria, E., Speer, R., Havasi, C., Hussain, A.: SenticNet: a publicly available semantic resource for opinion mining. In: Common-sense Knowledge, AAAI Fall Symposium series, vol. 10 (2010)

Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: ACL, vol. 2, pp. 383–389 (2014)

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y., Gelbukh, A., Zhou, Q.: Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn. Comput. **8**, 1–15 (2016)

Dehkharghani, R., Saygin, Y., Yanikoglu, B., Oflazer, K.: SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. Lang. Resour. Eval. **50**, 1–19 (2015)

de Albornoz, J.C., Plaza, L., Gervás, P.: SentiSense: an easily scalable concept-based affective lexicon for sentiment analysis. In: LREC, pp. 3562–3567 (2012)

Elhawary, M., Elfeky, M.: Mining Arabic business reviews. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1108–1113. IEEE (2010)

Elarnaoty, M., AbdelRahman, S., Fahmy, A.: A machine learning approach for opinion holder extraction in Arabic language. arXiv preprint arXiv:1206.1011 (2012)

Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of LREC, Vol. 6, pp. 417–422 (2006)

Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 174–181. Association for Computational Linguistics (1997)

He, Y., Zhou, D.: Self-training from labeled features for sentiment analysis. Inf. Process. Manage. **47**(4), 606–616 (2011)

Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)

Karimi, S.: Aspects of Persian syntax, specificity, and the theory of grammar. University of Washington (1989)

Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: the good the bad and the omg!. In: ICWSM, vol. 11, pp. 538–541 (2011)

Mahyoub, F.H., Siddiqui, M.A., Dahab, M.Y.: Building an Arabic sentiment lexicon using semi-supervised learning. J. King Saud Univ. Comput. Inf. Sci. **26**(4), 417–424 (2014)

Maynard, D., Funk, A.: Automatic detection of political opinions in tweets. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 88–99. Springer, Heidelberg (2012). doi:10.1007/978-3-642-25953-1_8

Neviarouskaya, A., Prendinger, H., Ishizuka, M.: SentiFul: a lexicon for sentiment analysis. IEEE Trans. Affect. Comput. **2**(1), 22–36 (2011)

Pak, A., Paroubek, P.: Twitter based system: using Twitter for disambiguating sentiment ambiguous adjectives. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 436–439. Association for Computational Linguistics, July 2010

Pakray, P., Neogi, S., Bhaskar, P., Poria, S., Bandyopadhyay, S., Gelbukh, A.: A textual entailment system using anaphora resolution. In: System Report, Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook, November 2011a

Pakray, P., Pal, S., Poria, S., Bandyopadhyay, S., Gelbukh, A.: JU_CSE_TAC: textual entailment recognition system at TAC RTE-6. In: System Report, Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook (2010)

Pakray, P., Poria, S., Bandyopadhyay, S., Gelbukh, A.: Semantic textual entailment recognition using UNL. Polibits **43,** 23–27 (2011b)

Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of EMNLP, pp. 2539–2544 (2015a)

Poria, S., Cambria, E., Gelbukh, A.: Aspect extraction for opinion mining with a deep convolutional neural network. Knowl.-Based Syst. **108**, 42–49 (2016)

Poria, S., Cambria, E., Gelbukh, A., Bisio, F., Hussain, A.: Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Comput. Intell. Mag. **10**(4), 26–36 (2015b)

Poria, S., Cambria, E., Winterstein, G., Huang, G.B.: Sentic patterns: dependency-based rules for concept-level sentiment analysis. Knowl.-Based Syst. **69**, 45–63 (2014)

Poria, S., Gelbukh, A., Das, D., Bandyopadhyay, S.: Fuzzy clustering for semi-supervised learning–case study: construction of an emotion lexicon. In: Mexican International Conference on Artificial Intelligence, pp. 73–86, October 2012

Remus, R., Quasthoff, U., Heyer, G.: SentiWS – a publicly available german-language resource for sentiment analysis. In: LREC, May 2010

Saraee, M., Bagheri, A.: Feature selection methods in Persian sentiment analysis. In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, vol. 7934, pp. 303–308. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38824-8_29

Seraji, M., Megyesi, B., Nivre, J.: A basic language resource kit for Persian, In: LREC, pp. 2245–2252 (2012)

Shi, H.X., Li, X.J.: A sentiment analysis model for hotel reviews based on supervised learning. In: 2011 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 3, pp. 950–954. IEEE (2011)

Sidorov, G., et al.: Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. In: Batyrshin, I., González Mendoza, M. (eds.) MICAI 2012. LNCS (LNAI), vol. 7629, pp. 1–14. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37807-2_1

Stone, P., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The general inquirer: a computer approach to content analysis. J. Reg. Sci. **8**(1), 113–116 (1968)

Subrahmanian, V.S., Reforgiato, D.: AVA: adjective-verb-adverb combinations for sentiment analysis. IEEE Intell. Syst. **23**(4), 43–50 (2008)

Tang, H., Tan, S., Cheng, X.: A survey on sentiment detection of reviews. Expert Syst. Appl. **36**(7), 10760–10773 (2009)

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)

Taghva, K., Beckley, R., Sadeh, M.: A stemming algorithm for the farsi language. In: ITCC, vol. 1, pp. 158–162, April 2005

Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)

Waltinger, U.: GermanPolarityClues: a lexical resource for german sentiment analysis. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC) (2010)

Yang, Y.: Application of Latent Dirichlet Allocation in Online Content Generation. Ph.D. thesis, University of California, Los Angeles (2016)