A Comparative Study of Persian Sentiment Analysis based on different Feature Combinations

Kia Dashtipour[1], Mandar Gogate[1], Ahsan Adeel[1], Amir Hussain[1], Abdulrahman Alqarafi[1], Tariq Durrani[2]

[1]CogBID Lab, Dept. of Computing Science and Mathematics, University of Stirling, Stirling, UK
[2]University of Strathclyde, Glasgow, UK

{kda, mgo, aad, ahu, aaq}@cs.stir.ac.uk
durrani@strath.ac.uk

Abstract. In recent years, the use of internet and correspondingly the number of online reviews, comments and opinions have increased significantly. It is indeed very difficult for humans to read these opinions and classify them accurately. Consequently, there is a need for an automated system to process this big data. In this paper, a novel sentiment analysis framework for Persian language has been proposed. The proposed framework comprises three basic steps: pre-processing, feature extraction, and support vector machine (SVM) based classification. The performance of the proposed framework has been evaluated taking into account different features combinations. The simulation results have revealed that the best performance could be achieved by integrating unigram, bigram, and trigram features.

Keywords: Sentiment Analysis, Persian, Feature Selection, N-gram

1 Introduction

Analyzing customer relationship has become increasingly important for companies and organizations because now-a-days potential customers decide to whether buy the product/service or not based on other customers' reviews. There is a huge potential to automatically analyze the unstructured data like text, videos available on social media and summarize the customer likes/dislikes (Pradhan et al.., 2016). Sentiment analysis is the process of classifying unstructured data like reviews based on polarity and emotion (Hussein, 2016).

Sentiment analysis (Rao et al., 2016) has been used in wide variety of application ranging from sports and business to politics and tourism (Martin et al., 2016).

Sentiment analysis has number of challenges, most of current approaches has been developed for English language and other languages such as Persian has been less developed, the other issue for sentiment analysis is lack of tools and resources in other languages most of the current tools have been developed for English (Dashtipour et al.., 2016). In order to, overcome this challenges, in the paper, the framework has been developed to extract features such as POS and Ngram from Persian movie reviews and support vector machine has been used to evaluate the performance of the feature selection for Persian movie reviews.

In literature, extensive research has been carried out to identify best features combination for English language. However, to the best of our knowledge, Persian feature selection for sentiment analysis is not yet well researched (Lo et al.., 2016).

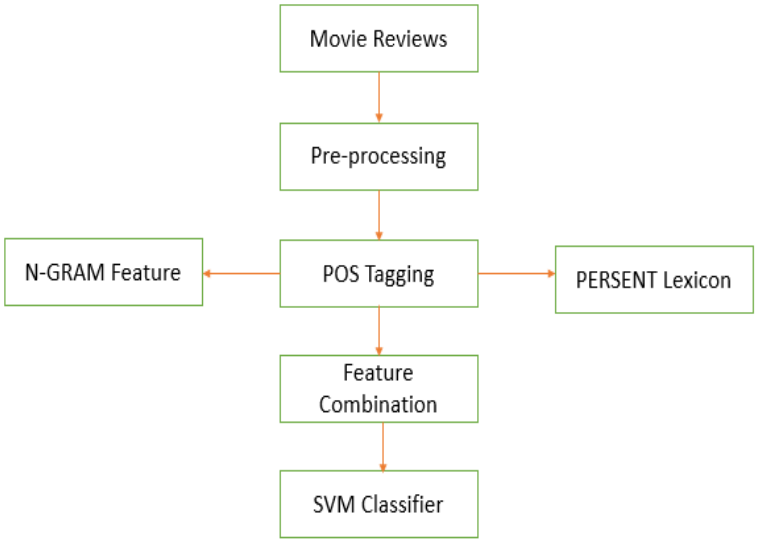
In English, different techniques have been proposed to classify unstructured data. The sentiment analysis is performed using four approaches: keyword spotting, lexical affinity, statistical method and concept level techniques (Cambria, 2016).

- **Keywords spotting**: Keyword spotting classifies the text by affect categories based on the presence of unambiguous affect such as sad, happy, afraid and bored (Cambria et al.., 2016).
- **Lexical Affinity:** The lexical affinity is a more sophisticated approach compare to keyword spotting. This approach assigns arbitrary words a probable affinity for specific emotion. For example, 75% accidents have indicated the negative effect as a “car accident” or “damage by the accident” (Cambria, 2016).
- **Statistical method:** Statistical approach leverage the power of machine learning such as Support vector machines (Cambria et al., 2017) to identify sentiment.
- **Concept based approach:** This approach is focused on semantic analysis of text by use of semantic networks. For example, in a sentence does not mention emotion adjectives, but it can be linked to other concept (Cambria et al., 2014).

The rest of the paper is organized as follows: Section 2, the proposed framework is presented. Section 3 presents the experimental results. Section 4 concludes the work and future work is presented.

2 Methodology

In this section, the proposed framework shown in Fig. 1 is discussed. Different features such as N-gram and part-of-speech (POS) tags (adjective, noun, verb and adverbs) could be used to identify sentiment in Persian Movie Reviews. The polarity of features is determined using PerSent lexicon (Persian lexicon). Finally, the performance of feature combinations is evaluated using a Support Vector Machine (SVM) classifier. More details are presented in the following subsections.



1. *Figure 1. Proposed framework*

Pre-processing: The Persian movie review dataset has been collected from Internet, the data is unstructured and noisy (Ghosh et al., 2017). The noisy and inconsistent information has been removed in the pre-processing stage.

For example, some of the words have apostrophe which might lead to incorrect part-of-speech (POS) tagging. These apostrophes have been removed. Moreover, in the review dataset some words are spelled incorrectly, for instance "فیلم عالیییییی بود" (The movie was greatttttt) or "ازش منتفرم" (I hateeeee it). The spelling has been corrected by removing the repeating character (Nirmal et al., 2015).

In addition, the variants of characters leading to the incorrect tagging of the sentence are corrected. For example, in Persian there are different “S” (س، ص، ث) and “T” (ت، ط). Both “S” and “T” (Desai et al., 2016) variants are mixed together.

Part of Speech (POS) features: The grammatical structure of sentences is analyzed to tag words with POS. Persian part-of-speech includes: noun, verb, adjective, adverb. These features have been extracted from the data (Shelke et al., 2017) to identify the polarity of the sentence.

N-gram: N-gram is sequence of *n* consecutive words in the text. For example, “ فیلم بسیار بدی بود” (The movie was bad), the unigram for sentence is “فیلم”, “بسیار”, “بدی”, “بود” (Tiwari et al., 2017).

Method	**Feature**	**Feature (English Translation)**
Unigram	فیلم بسیار عالی بود	The Movie is great
Bigram	فیلم بسیار بسیار عالی عالی بود	The movie Movie is Is great
Trigram	فیلم بسیار عالی بسیار عالی بود	The movie is Movie is great
POS Features	فیلم (Noun) بسیار (Adjective) عالی (Adjective) بود (Verb)	The (Determiner) Movie (Noun) Is (Verb) Great (Adjective)

Table 1: The example of various feature selection

PerSent Lexicon: The PerSent is a Persian lexicon consist of 1500 words along with their part-of-speech tag and polarity between -1 to +1. The polarity of the extracted features is identified using the PerSent lexicon (Dashtipour et al., 2016).

3 Experimental Results

Dataset: The Persian Movie reviews have been collected manually from www.caffecinema.com. The dataset consists of 500 positive and 500 negative movie reviews. The movie reviews have been collected from 2014 to 2016 movies.

Methodology: JHAZM part of speech tagger is used to tag Persian sentences. N-gram features like unigram, bigram and trigram are extracted using Java programming language. LIBSVM library was used to train the SVM classifier using combinations of n-gram and POS tag features. The performance of extracted features from movie reviews is evaluated using SVM classifier.

Results: The results are summarized in Table 2. Among n-gram features, classifier trained on trigram achieved better performance in term of accuracy. In addition, frequency of positive and negative adjective feature achieved better accuracy with respect to other POS features. Overall, the integration of unigram, bigram and trigram acquired best performance with 88.36 % accuracy.

Feature Combination	Accuracy	Feature Combination	Accuracy
Unigram	71.26%	Unigram + Bigram	74.85%
Bigram	74.37%	Unigram + Trigram	77.52%
Trigram	76.32%	Bigram + Trigram	78.65%
Unigram + Bigram + Trigram	88.36%	Frequency of positive and negative adjective + Frequency of positive and negative Adverb	75.89%
Frequency of positive and negative adjective	74.52%	Frequency of positive and negative adjective + Frequency of positive and negative noun	75.29%
Frequency of positive and negative adverb	72.78%	Frequency of positive and negative adjective + Frequency of positive and negative verb	75.01%
Frequency of positive and negative noun	72.45%	Frequency of positive and negative adverb + Frequency of positive and negative noun	74.39%
Frequency of positive and negative verb	70.23%	Frequency of positive and negative adverb + Frequency of positive and negative verb	73.65%
POS (Adjective + Adverb + Noun + Verb)	78.52%	Frequency of positive and negative noun + Frequency of positive and negative verb	73.49%

Table 2: Results

The bigram and trigram performed better in comparison with unigram because the unigram features contain lots of noise which affect the performance of the classifier. In addition, the multiword expressions suffer from problem of sparsity. Figure 2 summarizes the results of the experiments.

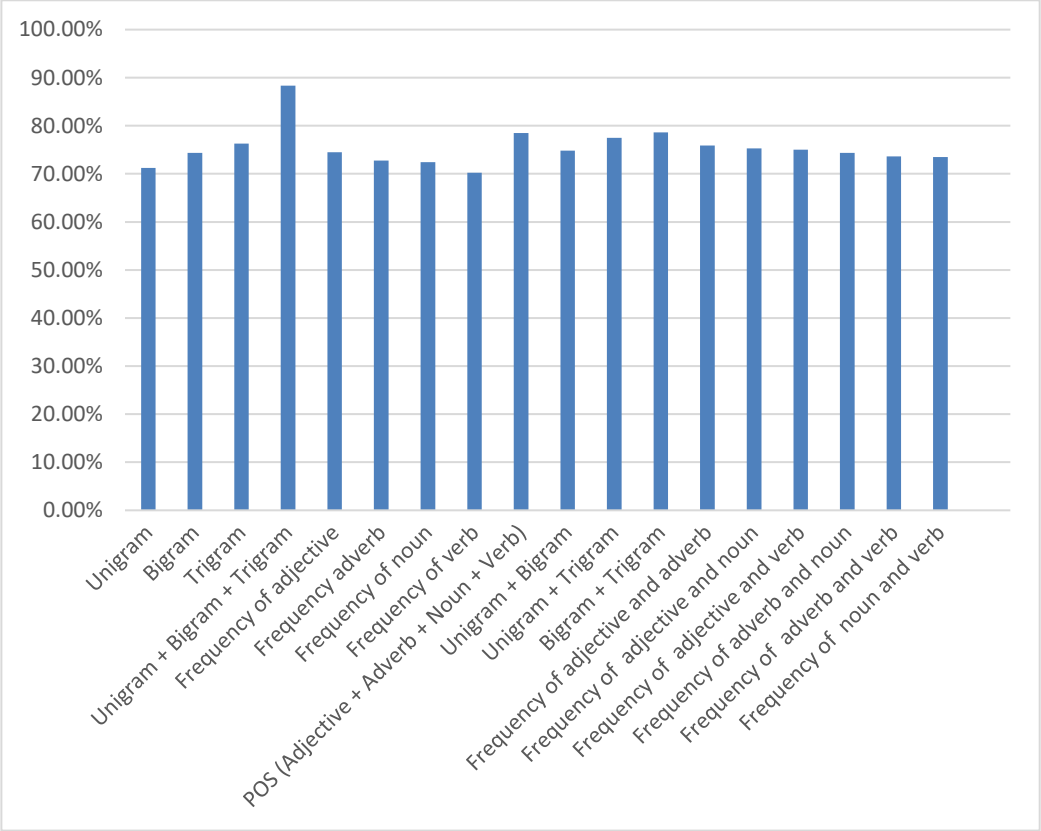


Figure 2: Feature Combination

4 Conclusion and Future work

In this paper, a novel sentiment analysis framework for Persian language has been proposed. A comparative study of Persian sentiment analysis based on different features combination is presented. The performance of the proposed framework has been evaluated by combining different features in terms of classification accuracy. Among all features combination, the integration of unigram, bigram and

trigram achieved best performance. In future, we intend to compare the performances of manually extracted features with deep learning based automated feature extraction to identify the best feature for Persian sentiment analysis.

Acknowledgements

The authors are grateful to the anonymous reviewers for their insightful comments and suggestions which helped improved the quality of the paper. This work was part-supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/M026981/1 (AV-COHGEAR)

References

1. Agarwal, B., Poria, S., Mittal, N., Gelbukh, A. and Hussain, A., 2015. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation, 7(4), pp.487-499.
2. Cambria, E., Das, D., Bandyopadhyay, S. and Feraco, A., 2017. Affective Computing and Sentiment Analysis. In A Practical Guide to Sentiment Analysis (pp. 1-10). Springer International Publishing.
3. Cambria, E., Schuller, B., Xia, Y. and White, B., 2016. New avenues in knowledge bases for natural language processing. Knowledge-Based Systems, 108(C), pp.1-4.
4. Cambria, E., 2016. Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), pp.102-107.
5. Cambria, E., Schuller, B., Xia, Y. and Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), pp.15-21.
6. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y., Gelbukh, A. and Zhou, Q., 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognitive Computation, 8(4), pp.757-771.
7. Dashtipour, K., Hussain, A., Zhou, Q., Gelbukh, A., Hawalah, A.Y. and Cambria, E., 2016. PerSent: A Freely Available Persian Sentiment Lexicon. In Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8, pp. 310-320. Springer International Publishing.
8. Desai, M. and Mehta, M.A., 2016, April. Techniques for sentiment analysis of Twitter data: A comprehensive survey. In Computing, Communication and Automation (ICCCA), 2016 International Conference on, pp. 149-154. IEEE.
9. Ghosh, M. and Sanyal, G., 2017. Preprocessing and Feature Selection Approach for Efficient Sentiment Analysis on Product Reviews. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, pp. 721-730. Springer, Singapore.
10. Hussein, D.M.E.D.M., 2016. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences.
11. Lo, S.L., Cambria, E., Chiong, R. and Cornforth, D., 2016. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artificial Intelligence Review, pp.1-29.

12. Martin, V.M.A., David, K. and Bhuvaneswari, R., 2016. A Survey on Various Techniques for Sentiment Analysis and Opinion Mining. Data Mining and Knowledge Engineering, 8(3), pp.78-82.
13. Nirmal, V.J. and Amalarethinam, D.G., 2015. Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data. International Journal of Fuzzy Mathematical Archive, 6(2), pp.149-159.
14. Pradhan, V.M., Vala, J. and Balani, P., 2016. A Survey on Sentiment Analysis Algorithms for Opinion Mining. International Journal of Computer Applications, 133(9).
15. Priyanka, C. and Gupta, D., 2013, August. Identifying the best feature combination for sentiment analysis of customer reviews. In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, pp. 102-108. IEEE.
16. Rao, S., 2016, August. A Survey on Sentiment Analysis and Opinion Mining. In Proceedings of the International Conference on Advances in Information Communication Technology & Computing, p. 53. ACM.
17. Shelke, N., Deshpande, S. and Thakare, V., 2017. Domain Independent Approach for Aspect Oriented Sentiment Analysis for Product Reviews. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, pp. 651-659. Springer, Singapore.
18. Tiwari, P., Mishra, B.K., Kumar, S. and Kumar, V., 2017. Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7(1), pp.30-41.
19. Tripathy, A., Agrawal, A. and Rath, S.K., 2016. Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57, pp.117-126.
20. Trupthi, M., Pabboju, S. and Narasimha, G., 2016, March. Improved feature extraction and classification—Sentiment analysis. In Advances in Human Machine Interaction (HMI), 2016 International Conference on, pp. 1-6. IEEE.