Feature Selection Methods in Persian Sentiment Analysis

Mohamad Saraee [1], Ayoub Bagheri [2]

[1]School of Computing, Science and Engineering, University of Salford, Manchester, UK
m.saraee@salford.ac.uk
[2]Intelligent Database, Data Mining and Bioinformatics Lab,
Electrical and Computer Engineering Dep., Isfahan University of Technology, Isfahan, Iran
a.bagheri@ec.iut.ac.ir

Abstract. With the enormous growth of digital content in internet, various types of online reviews such as product and movie reviews present a wealth of subjective information that can be very helpful for potential users. Sentiment analysis aims to use automated tools to detect subjective information from reviews. Up to now as there are few researches conducted on feature selection in sentiment analysis, there are very rare works for Persian sentiment analysis. This paper considers the problem of sentiment classification using different feature selection methods for online customer reviews in Persian language. Three of the challenges of Persian text are using of a wide variety of declensional suffixes, different word spacing and many informal or colloquial words. In this paper we study these challenges by proposing a model for sentiment classification of Persian review documents. The proposed model is based on stemming and feature selection and is employed Naive Bayes algorithm for classification. We evaluate the performance of the model on a collection of cellphone reviews, where the results show the effectiveness of the proposed approaches.

Keywords: sentiment classification, sentiment analysis, Persian language, Naive Bayes algorithm, feature selection, mutual information.

1 Introduction

In the recent decade, with the enormous growth of digital content in internet and databases, sentiment analysis has received more and more attention between information retrieval and natural language processing researchers. Up to now, many researches have been conducted sentiment analysis on English, Chinese or Russian languages [1-9]. However on Persian text, in our knowledge there is little investigation conducted on sentiment analysis [10]. Persian is an Indo-European language, spoken and written primarily in Iran, Afghanistan, and a part of Tajikistan. The amount of information in Persian language on the internet has increased in different forms. As the style of writing in Persian language is not firmly defined on the web, there are too many web pages in Persian with completely different writing styles for the same words [11, 12]. Therefore in this paper, we study a model of feature selection in sentiment classification for Persian language, and experiment our model on a Persian product review

dataset. In the reminder of this paper, Section 2 describes the proposed model for sentiment classification of Persian reviews. In Section 3 we discuss important experimental results, and finally we conclude with a summary in section 5.

2 Proposed Model for Persian Sentiment Analysis

Persian sentiment analysis suffers from low quality, where the main challenges are

- Lack of comprehensive solutions or tools
- Using of a wide variety of declensional suffixes
- Word spacing
    - In Persian in addition to white space as inter-words space, an intra-word space called pseudo-space separates word's part.
- Utilizing many informal or colloquial words

In this paper, we propose a model, using n-gram features, stemming and feature selection to overcome the Persian language challenges in sentiment classification.

2.1 Sentiment Classifier

In this paper, we consider Naive Bayes algorithm which is a machine learning approach as the sentiment classifier [13]. In the problem of sentiment classification we use vector model to represent the feature space. For the feature space we extract n-gram features to deal with the conflicting problem of space and pseudo-space in Persian sentences. Here we use unigram and bigram phrases as n-gram features. Therefore in this model, the sequence of the words is important. Experiments show using n-gram features could solve the problem of different word spacing in Persian text.

2.2 Feature Selection for Sentiment Analysis

Feature Selection methods sort features on the basis of a numerical measure computed from the documents in the dataset collection, and select a subset of the features by thresholding that measure. In this paper four different information measures were implemented and tested for feature selection problem in sentiment analysis. The measures are Document Frequency (DF), Term Frequency Variance (TFV), Mutual Information (MI) [14] and Modified Mutual Information (MMI). Below we discuss presented MMI approach.

Mutual Information and Modified Mutual Information

In this paper we introduce a new approach for feature selection, Modified Mutual Information. In order to explain MMI measure, it is helpful to first introduce Mutual Information by defining a contingency table (see Table 1).

Table 1. Contingency table for features and classes

	c	$\bar{c}$

f	A	B
$\bar{f}$	C	D

Table 1 records co-occurrence statistics for features and classes. We also have that the number of review documents, N = A+B +C +D. These statistics are very useful for estimating probability values [13, 14]. By using Table 1, MI can be computed by equation (7):

$$MI(f,c) = \log \frac{P(f,c)}{P(f)P(c)} \quad (1)$$

Where $P(f,c)$ is the probability of co-occurrence of feature *f* and class *c* together, and $P(f)$ and $P(c)$ are the probability of co-occurrence of feature *f* and class *c* in the review documents respectively. Therefore by Table 1, MI can be approximated by Equation (8):

$$MI(f,c) = \log \frac{A*N}{(A+B)*(A+C)} \quad (2)$$

Intuitively MI measures if the co-occurrence of *f* and *c* is more likely than their independent occurrences, but it doesn't measure the co-occurrence of *f* and $\bar{c}$ or the co-occurrence of other features and class *c*. We introduce a Modified version of Mutual Information as MMI which consider all possible combinations of co-occurrences of a feature and class label. First we define four parameters as the following:

- $p(f,c)$: *Probability of co-occurrence of feature f and class c together.*
- $p(\bar{f},\bar{c})$: *Probability of co-occurrence of all features except f in all classes except c together.*
- $p(\bar{f},c)$: *Probability of co-occurrence of all features except feature f in class c.*
- $p(f,\bar{c})$: *Probability of co-occurrence of feature f in all classes except c.*

We calculate MMI score as Equation (10):

$$MMI(f,c) = log \frac{p(f,c)*p(\bar{f},\bar{c})}{p(f)*p(c)*p(\bar{f})*p(\bar{c})} - log \frac{p(\bar{f},c)*p(f,\bar{c})}{p(f)*p(c)*p(\bar{f})*p(\bar{c})} \quad (3)$$

Where $P(f)$ and $P(c)$ are the probability of independent occurrence of feature *f* and class *c* in the review documents respectively. $p(\bar{f})$ is the number of review documents which not contain feature *f* and $p(\bar{c})$ is the number of documents with the classes other than class *c*. Based on Table 4, MMI can be approximated by Equation:

$$MMI(f,c) = \frac{A*D - C*B}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (4)$$

3 Experimental Results

To test our methods we compiled a dataset of 829 online customer reviews in Persian language from different brands of cell phone products. We assigned two annotators to

label customer reviews by selecting a positive or negative polarity on the review level. After annotation, the dataset reached to 511 positive and 318 negative reviews.

3.1 Comparative study

In our experiments, first we evaluated Persian sentiment classification in two phases:

Phase 1. *Without n-gram features and stemming*
Phase 2. *With n-gram features and stemming*

Table 2 shows the F-score results for the two phases. From the results we can observe that using of n-gram features and stemming for sentiment classification has 4% and 0.3% improvements for negative and positive classes respectively.

Table 2. F-scores for phases 1 and 2, Without and with n-gram features and stemming

Phase	Class	F-score
1	Negative	0.7480
	Positive	0.8570
2	Negative	0.7880
	Positive	0.8600

In this work we applied four different feature selection approaches, MI, DF, TFV and MMI with the Naive Bayes learning algorithm to the online Persian cellphone reviews. In the experiments, we found that using feature selection with learning algorithms can perform improvement to classifications of sentiment polarities of reviews.

Table 3 indicates Precision, Recall and F-score measures on two classes of Positive and Negative polarity with the feature selection approaches.

Table 3. Precision, Recall and F-score measures for the feature selection approaches with naive bayes classifier

Approach	**Class**	Precision	Recall	F-score
MI	Negative	0.4738	0.8356	0.6026
	Positive	0.8130	0.4260	0.5538
DF	Negative	0.8148	0.7812	0.7962
	Positive	0.8692	0.8898	0.8788
TFV	Negative	0.8226	0.7800	0.7996
	Positive	0.8680	0.8956	0.8814
MMI	Negative	0.7842	0.8568	0.8172
(Proposed Approach)	Positive	0.9072	0.8526	0.8784

The results from Table 3 indicate that the TFV, DF and MMI have better performances than the traditional MI approach. In terms of F-score, MMI improves MI with 21.46% and 32.46% on Negative and Positive classes respectively, DF overcomes MI with 19.36% and 32.5% better performances for Negative and Positive review documents respectively and TFV improves MI with19.7% and 32.76% for Negative and Positive documents respectively. The reason of poor performance for MI is that of MI only uses the information between the corresponding feature and the corresponding class and does not utilize other information about other features and other classes. When we compare DF, TFV and MMI, we can find that the MMI beats both DF and TFV on F-scores of Negative review documents with 2.1% and 1.76% improvements respectively, but for the Positive review documents DF and TFV have 0.04% and 0.3% better performance than the MMI, respectively.

To assess the overall performance of techniques we adopt the macro and micro average, Figure 1 shows the macro and micro average F-score.

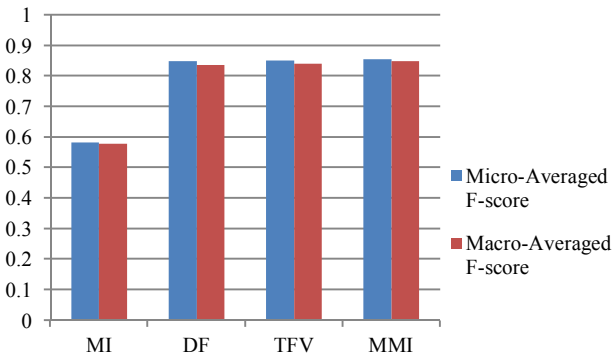

Fig. 1. Macro and micro average F-score for MI, DF, TFV and MMI

From this Figure we can find that the MMI proposed approach has slightly better performance than the DF and TFV approaches and has significant improvements on MI method. The basic advantage of the MMI is using of whole information about a feature, positive and negative factors between features and classes. MMI in overall can reach to 85% of F-score classification. It is worth noting that with a larger training corpus the feature selection approaches and the learning algorithm could get higher performance values. Additionally the proposed approach – MMI – is not only for Persian reviews and in addition can be applied to other domains or other classification problems.

4 Conclusion and Future Works

In this paper we proposed a novel approach for feature selection, MMI, in sentiment classification problem. In addition we applied other feature selection approaches, DF, MI and TFV with the Naive Bayes learning algorithm to the online Persian cellphone reviews. As the results show, using feature selection in sentiment analysis can improve the performance. The proposed MMI method that uses the positive and negative

factors between features and classes improves the performance compared to the other approaches. In our future work we will focus more on sentiment analysis about Persian text.

References

1. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. Mining Text Data. pp. 415-463, (2012)
2. Pang, B., Lee, L., Vaithyanathan S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical methods in natural language processing, Vol. 10, pp. 79-86. ACL (2002)
3. Moraes, R., Valiati, J. F., Gavião Neto, W. P.: Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert Systems with Applications. (2012)
4. Cui, H., Mittal, V., Datar, M.: Comparative experiments on sentiment classification for online product reviews. In: Proceedings of National Conference on Artificial Intelligence, vol. 21, no. 2, p. 1265. Menlo Park, Cambridge, London, (2006)
5. Yussupova, N., Bogdanova, D., Boyko, M.: Applying of sentiment analysis for texts in russian based on machine learning approach. In: Proceedings of Second International Conference on Advances in Information Mining and Management. pp. 8-14. (2012)
6. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. (2005)
7. Zhu, J., Wang, H., Zhu, M., Tsou, B.K., Ma, M.: Aspect-based opinion polling from customer reviews. IEEE Transactions on Affective Computing, vol. 2(1), pp. 37-49, (2011)
8. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of Conference on World Wide Web, pp. 342-351. (2005)
9. Turney, P. D., Littman, M. L., Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council Canada, (2002)
10. Shams, M., Shakery, A., Faili, H.: A non-parametric LDA-based induction method for sentiment analysis. In: Proceedings of 16th IEEE CSI International Symposium on Artificial Intelligence and Signal Processing. pp. 216-221. (2012)
11. Farhoodi, M., Yari A.: Applying machine learning algorithms for automatic Persian text classification. In: Proceedings of IEEE International Confernce on Advanced Information Management and Service, pp. 318-323. (2010)
12. Taghva, K., Beckley, R., Sadeh, M.: A stemming algorithm for the Farsi language. In: Proceedings of IEEE International Conference on Information Technology: Coding and Computing, ITCC. vol. 1, pp. 158-162, (2005)
13. Mitchell, T.: Machine Learning. second edition, McGraw Hill, (1997).
14. Duric, A., Song, F.: Feature selection for sentiment analysis based on content and syntax models. Decision Support Systems. (2012)