University College of Takestan

Available online at **http://UCTjournals.com**

UCT Journal of Research in Science, Engineering and Technology

UCT. J. Resea. Scien. Engineer.Techno. (UJRSET)

Volume 4,Issue1

11-15 (2016)

ISSN 2382-977X

Algorithm for Persian Text Sentiment Analysis in Correspondences on an E- Learning Social Website

Anahid Rais Rohani[1] and A'zam Bastanfard[1]*

1 Department of Computer, College of Mechatronic, Karaj Branch, Islamic Azad University, Alborz, Iran

Original Article:
Received 22 Aug. 2015 Accepted 26 Jan. 2016 Published 30 Jan. 2016

ABSTRACT

By 2000, sentiment analysis had been only studied based on speech and changes in facial expressions. Since then, studies have been focused on text. Concerning Persian text mining, studies have been conducted on the methods for extracting properties for classification and examination of opinions on social websites with an aim to determine text polarity. The present research was aimed to prepare and implement an algorithm for Persian text sentiment analysis based on the following six basic emotional states: happiness, sadness, fear, anger, surprise, and disgust. In this research, sentiment analysis was carried out using the unsupervised lexical method. Lexicons are divided into four categories, namely the emotional, boosters, negation, and stop lists. The algorithm was written in six different ways using different properties. In the first method, the algorithm was capable of identifying an emotional word in a sentence. The sentiment of the sentence was determined based on the given emotional word. However, it should be noted that the text itself is also important for sentiment analysis because in addition to the emotional words, other factors (such as boosters and negating factors) are also present in the sentence and affect the text sentiment. Hence, the algorithm was enhanced in the subsequent methods to detect the boosters and negating words. Results of running the algorithm using different methods indicated that the algorithm accuracy increased with an increase in the number properties involved. In the sixth method, an algorithm capable of identifying emotional, boosters and negative words was applied to two data samples including sentences written by typical users and sentences written by university students on an electronic learning social website. The accuracy of the algorithm with 100 data samples from typical users and 100 data samples from university students was 80% and 84%, respectively.

Keyword:
- ✓ sentiment analysis
- ✓ opinion mining
- ✓ text mining

*** Corresponding author:**
bastanfard03@gmail.com (A'zam Bastanfard),
anahid_r@yahoo.com (Anahid Rais Rohani)

Peer review under responsibility of **UCT Journal of Research in Science, Engineering and Technology**

INTRODUCTION

Since the data used in this research was not structured data, the proposed algorithm was implemented based on the lexical method. Moreover, to increase the accuracy of the proposed method, in addition to emotional words, other properties (such as boosters and negative factors) were also used to enable the algorithm detect the text sentiment based on the six basic emotional states. To assess the words, weights were assigned to the words in the lexicons and the words were put in different emotional categories. The 6 basic sentiment categories included happiness, sadness, anger, fear, disgust and surprise. In this paper, first previous research was reviewed and then the procedure for preparing lexicons was explained. A description of the different phases of the proposed algorithm was also presented. Finally, the discussion and conclusions on implementation of the algorithm on Persian texts were presented.

2- Previous Work

Sentiment analysis, as a branch of text mining, is a contemporary concern and studies on this topic have been started since 2000 in foreign languages. [1] Diman Ghazi et al. (2014) used the wordlist method to examine the six classifications of texts. They examined the basic model, SVM and SMO methods in Weka (software). Results indicated that SVM performed better with large datasets, whereas LR was more effective than SVM with a bag of words. However, [2] Emma Haddia et al. (2013) studied the effect of pre-processing on texts extracted from the Internet, and their findings improved the accuracy of SVM. For non-preprocessed data, the accuracy of the FF matrix was also enhanced, but for pre-processed data the accuracy of both FF and FP matrices increased. [3] Masashi Hadanoa et al. (2011) identified sentiments in the review documents of a game that were manually obtained from the respective website. In this research, the sentence clustering method and Bayon (which is based on the bisection algorithm) were employed, and it was concluded that the clustering method was more effective than the non-clustering method. [4] Fiorella Carla Dotti (2013) proposed functions for sentiment analysis of plain English news texts. Most of the inaccurate identifications of tokens in the proposed functions were associated with false positive sentences. They also used the demystifying module and POS tagging, which improved the overall call. [5] Alexandra Balahur et al. (2014) carried out a comparative experiment using the supervised learning and translation machine methods for a multilingual sentiment analysis. For translation purposes, they used the SMT, Moses, Bing and Google translation engines, and for classification the SVM and bagging methods were employed. Research results revealed that bagging had a positive effect on the results. [6] Carmen Banea et al. (2014) compared the multilingual and cross-lingual methods and concluded that the accuracy of the multilingual method was higher than the cross-lingual method. [7] Tomáˇs Brychcínet al. (2014) tested the HAL, COALS, BEAGLE and P&P algorithms on English, Slavic and Czech news texts. The HAL method is suitable for dense clustering, whereas COALS is useful for sparse clustering. The combination of HAL and COALS yields the best results. P&P and BEAGLE also demonstrated a slight improvement as compared to the base method. In their examination of documents of a consulting website, [8] Liang-Chih et al. (2009) compared two word-based retrieval methods (i.e. the SVM and OKAPI models) with the discourse-aware method. Their results showed that the discourse-aware retrieval model was more precise than the word-based retrieval methods. [9] Mohamad Hardyman Barawi et al. (2013) performed experiments on resource accuracy assessment. They based their comparisons of the results of description analysis on training videos and the Euclidean distance factor. They found that it is possible to interpret opinions on online training videos by drawing the curves of resources suitable for sentiment analysis. [10] Dinko Lambova et al. (2011) fused some adaptive algorithms including the SAR, MAA, BMAA and BMAADR algorithms. According to their findings, in the SAR algorithm, the best result was obtained from the manual interpretations of RIMDB (77.1%), while in the MAA algorithm the best yield was 75.6%, which was worse than the best SAR results. The highest accuracy of the BMAADR algorithm was 80% on average. The SAR algorithm also demonstrated a better performance with exclusively subjective and objective datasets (RIMDB and MPQA). The BMAADR algorithm provides the best performance with actual texts.

3- Resources and Data

In the unsupervised method, lexicons and dictionaries are used and a set of rules are also available for calculating the result [3]. In this research, a Persian lexicon was required, and since no relevant lexicon was available, lexicons were prepared as described in the following.

3-1- Lexicons

Emotional Lexicon

This lexicon contains a list of words for the 6 basic emotional states, which were labeled as happiness, sadness, fear, anger, surprise, and disgust. Then, a weight was assigned to each word based on the contribution of the given word to accurate sentiment detection.

Boosters Lexicon

This lexicon contains a list of boosters with the negative or positive polarity labels. The polarities are assigned weights, which reflect the level of their effects.

Negation Lexicon

This lexicon contains a list of words such as verbs and prefixes with negative effects. These words are known as the negative words.

Stop Lexicon

This lexicon contains a list of words and prepositions, which do not express any feeling and have no effect on the text sentiment.

3-2- Data

In sentiment analysis, it is necessary to consider the richness of the data selected for classification. Here, since the data was in the form of correspondence (messages) between people and might have been written in English in the conversation environment, Persian sentences were separated manually. A stop wordlist was also used to clean the sentences from any preposition or non-affective words.

4- Features

Feature selection is highly important in text mining. In this research, the following four categories of features were used: emotional words; parts of speech (POS), which include the booster words and negative verbs; dependencies between words; and the number of emotional words in a sentence.

Emotional Words: Properties rely on the words, and this category is suitable for simple sentences. These words are included in the sentimental lexicon with appropriate labels.

Part of Speech: It includes words such as boosters, negative words and verbs that affect the identification of a sentential sentiment. Boosters' lexicon is used to identify these words.

Words Dependency: This attribute is important when the boosters affect the sentiment of emotional words, and negating prefixes affect the subsequent words in the sentence.

Number of Emotional Words: It is used when the sentence contains several emotional words with different sentiment labels. The emotion of the sentence is determined by the maximum total weight of words in a similar category.

5- Experiences

The lexicons described in Section 3 were prepared. First, to assess the algorithm performance an editor was prepared to receive the sentence as the input entered by the user and to examine the sentiment analysis algorithm using 6 hybrid methods based on the aforementioned properties.

First Method: Emotional word features: This algorithm only identifies one emotional word in the sentence using the emotional lexicon. The prior sentiment of that word reveals the sentiment of the sentence. This method is suitable for simple sentences and the accuracy was this method was calculated to be 30%.

Second Method: Combine Emotional word features with number of words features: In this method, the number of words in the sentence is also considered in addition to the emotional word feature. In this method, the algorithm expresses the sentiment of the word with the maximum weight as the sentence sentiment. The accuracy of this method is 40%.

Third Method: Combine the emotional word features and word-dependency features: In addition to the emotional words, the dependence of boosters is also considered a feature. The algorithm identifies only one emotional and one booster word in the sentence, which reflects the sentence sentiment with polarity of the booster word. Similar to the first method, this method has a accuracy of 30%.

Fourth Method: Using the emotional word, number of words, and dependency of words simultaneously: The algorithm detects more than one emotional word and a booster word in the sentence, and considers the relationship between the boosters and emotional words. Then, it determines the sentiment with the maximum weight following sentiment calculations. In this method, the algorithm presents a higher accuracy (50%).

Fifth Method: Using the emotional word, number of words, and POS simultaneously: In this method, a negative verb and a negative prefix are involved in addition to the emotional words and number of emotional words features. The algorithm detects the negative verb in the sentence and expresses the sentiment of the category with the maximum weight due to the presence of a negative verb as the sentence sentiment after considering the effect of boosters. Concerning the effect of the negated words, if the negation word precedes an adjective, it negates the adjective and the sentence loses the sentiment, but if the sentiment is in the happiness group, the negation leads to the sadness group. There is no reversed sentiment for other categories to be able to assign the sentiment to other categories. Hence, a neutral category was assumed to include non-sentimental sentences. The accuracy of the algorithm reached 70% in this method.

Sixth Method: Using the emotional word, number of words, POS, and word-dependency simultaneously: In this method, all of the four features are used. The algorithm identifies more than one emotional word, a booster word, and the negative words in the sentence. After considering the effect of booster words and negative words on the emotional words, it stores the sentiment of the category with the maximum weight. If any negative verb is used in the sentence, the algorithm applies the verb and then expresses the sentential sentiment. The algorithm accuracy increased to 80% in this method.

6- Discussion

The proposed algorithm was examined using six hybrid methods based on the aforementioned features. Results indicated that simultaneous use of the four features yields better results. For instance, to compare the first and second methods the following sentence was analyzed: "I punished my student yesterday, and I'm sad and sorry now." In the first method, which only considers emotional words, the algorithm detects the word "punish" and identifies a sense of "anger" in the sentence. However, in the second method, which also considers the number of emotional words, the "sadness" is identified as sentiment of the sentence. The accuracy of the first and second methods was calculated to be 30% and 40%, respectively. The changes in accuracy indicate that the number of emotional words in the sentence has positive effects on the accuracy of sentiment detection by the algorithm. To compare the fourth and fifth methods the following sentence was analyzed as an example: "Thank you, but my problem wasn't solved." In the fourth method, which considers word- dependency but does not employ the POS feature to identify the sentence's verb, the sense of happiness is identified. In the fifth method, in which the algorithm identifies the verb and calculates its effect on the sentence, the sense of sadness is identified in the aforementioned sentence. The algorithm accuracy increased to 50% in the fourth method, which shows that the number of emotional words has a larger effect on the sentential sentiment. By comparing the fourth and fifth methods (with a accuracy of 70%) it is concluded that the POS property is more effective than word- dependency. The algorithm in the sixth method, which employed the four features, resulted in the highest level of accuracy (80%), and was therefore identified as the best method.

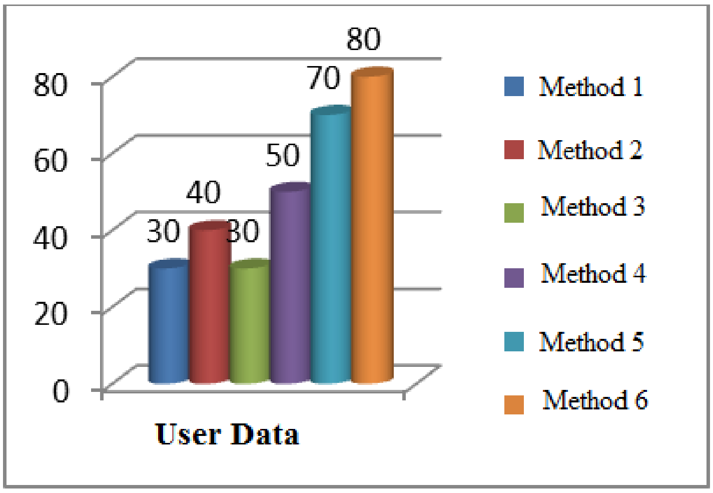

Figure (1-1): Comparing the methods on user data

The algorithm composed in method six was used for sentiment analysis on the data sample containing personal correspondence. In analyzing the 50 samples, the proposed algorithm detected the sentiment of 84% of sentences correctly and identified 4% of non-sentimental sentences as sentences with sentiment. Moreover, the sentiment of 12% of the sentences wasn't also detected correctly.

A comparison was also made between the results of the two methods using the user data and social website data. Results revealed that the accuracy of the algorithm was 80% and 84% with the user data and social website data, respectively. Therefore, the accuracy of the algorithm was higher with the social website data, because the lexicons were formed of words used on the website. Hence, it could be concluded that in the lexical method, the domain of data affects the detection of sentiment by the algorithm. There may be adjectives in the sentences that are not present in the lexicons. This is one of the disadvantages of the lexical method. The more words in the lexicons the better performance the algorithm has.

7- Conclusion and Future Work

In this research, it was claimed that the prior meaning of words is not enough for sentiment analysis as other words also affect the sentiment of a given sentence. To improve the algorithm, it was tried to detect the adverbs and negative words. In the research experiences, the designed algorithm was examined with different compositions of features The results show that using set of features all together outperform both using them separately and other combination of them.

To improve the accuracy of the algorithm in future work, it is better to add another feature such as punctuation to detect a sentence ends with dot, exclamation mark, question mark, etc. Moreover, there could be a sentence with no emotional word, but it may reflect the emotion of the author. Therefore, research should also be carried out on this possibility.

References

[١] Diman Ghazi a, Diana Inkpen a, Stan Szpakowicz , “Prior and contextual emotion of words in sentential context”, Computer Speech and Language, vol. 28 , 2014, pp. 76–92

[٢] Emma Haddia, Xiaohui Liua, Yong Shib,”The Role of Text Pre-processing in Sentiment Analysis”, Procedia Computer Science, vol. 17, 2013, pp. 26 – 32

[٣] Masashi Hadanoa, Kazutaka Shimadaa, Tsutomu Endoa,” Aspect identification of sentiment sentences using a clustering Algorithm”, Procedia - Social and Behavioral Sciences, vol. 27, 2011, pp. 22 – 31

[4] Fiorella Carla Dotti,” Overcoming Problems in Automated Appraisal Recognition: the Attitude System in Inscribed Appraisa”, Procedia - Social and Behavioral Sciences, vol. 95, 2013 , pp. 442 – 446

[5] Alexandra Balahur, Marco Turchi,” Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis”, Computer Speech and Language, vol. 28 , 2014, pp. 56–75

[6] Carmen Banea a, Rada Mihalcea a, Janyce Wiebe,” Sense-level subjectivity in a multilingual setting”, Computer Speech and Language, vol. 28, 2014, pp. 7–19

[7] Tomáˇs Brychcín, Miloslav Konopík,” Semantic spaces for improving language modeling”, Computer Speech and Language, vol. 28, 2014, pp. 192 – 209

[8] Liang-Chih Yu a, Chung-Hsien Wu b, Fong-Lin Jang,”Psychiatric document retrieval using a discourse-aware model”, Artificial Intelligence, vol. 173, 2009, pp. 817–829

[9] Mohamad Hardyman Barawi, Yet Yong Seng,” Evaluation of resource creations accuracy by using sentiment Analysis”, Procedia - Social and Behavioral Sciences, vol. 97 , 2013 , pp.522 – 527

[10] Dinko Lambova, Sebastião Paisa, Gãel Diasa,”Merged Agreement Algorithms for Domain Independent Sentiment Analysis”, Procedia - Social and Behavioral Sciences, vol. 27, 2011 , pp. 248 – 257.