International Journal of Electronics Communications and Electrical Engineering
ISSN : 2277-7040 Volume 3 Issue 1 (January 2013)
http://www.ijecee.com/ https://sites.google.com/site/ijeceejournal/

Affect Analysis in Persian Text Documents

Shokoofe Dashtbani[1], Abdol Hamid Pilevar[2]
Language Engineering Lab, Computer Engineering Dept., Bu Ali Sina University, Hamedan, Iran
[1]Sh.dashtbani@basu.ac.ir, [2]pilevar@basu.ac.ir

Abstract. in this article we review a new approach to analysis the Persian texts. This paper is studied the emotional analysis in Persian texts. Methods and features proposed in this article are compared with English methods. Recently, analysis of texts in other languages has been studied particularly in English language and many methods are presented for the rapid analysis of English texts. The method used in this paper is similar to methods used in the English language. This analysis is done by making emotional categories. The method based on emotional categorization of the semantic similarities. For building the emotional categories in Persian, we used semantic similarity and this method has been successful to emotional analyze of the Persian texts about 80 percent of the cases.

Keywords: Semantic similarity, Word Similarity, Affect category, Degree of density, Affect lexicon, Emotional text.

1 Interoduction

Homology, resemblance or similarity is a concept that is currently has been used widely and various explanations have been proposed for it. The definition is presented in this article; we tried to provide a particular application of the knowledge that provided by it. The meaning of Similarity in here is similar to the feeling that knowledge of the concept gives us. Here we propose a special model to obtain seminaries. In this article a probability based model is used to measure the similarity Natural language processing is a branch of artificial intelligence, Persian language processing in this paper has been investigated. Emotional analysis of the text structure is important because it can somewhat reduce the uncertainties.

Extensive research literature on emotional analysis was performed. Analysis for an Opinion Mining Application [1] and Automatic Spoken Affect Analysis and Classification [2] , are related works is done in this area.

One of the basic techniques used in this paper is using similar words in Persian language, this techniques has been used to find similarity among the words in Persian text and emotional categories.

---

2 The corresponding author

Solutions for the English language is available such as to obtain the similarity using Feature vector [3], Edge-based (Distance) Approach [4], Information concept [5], Using Fuzzy Logic[6] and Semantic Similarity Based on Corpus Statistics and Lexical[7] also calculated.

To making emotional categories, we need to precise study in Persian language or we must use a specialist in Persian literature. In this paper we will introduce 40 emotional categories of Persian language using grammar Persian books [8]. These categories will be used to analyze the input text after parsing the input text into its constituent words.

Text uncertainties will handle by allowing a single lexicon entry to belong to multiple emotional categories. In addition to resolve this ambiguity, emotional categories are also expressed feelings and words also expressed in natural language [9]. Imprecision is handled, not only via multiple category assignments, but also by degree of density between lexicon entries and their various categories. All emotional words with density numeric and the number that represents the amount of similarity between the lexicon entries and emotional categories from input text put in an attachment file. Finally, according to the attachment file, the input text is analyzed. Recently, some emotional calculations have also provided in the English language [10].

Finally, we suggest some strategies to choose appropriate way to expedite the proposed method.

2 Proposed methods

In English, several methods for obtaining similarity is presented, For example, to determine the similarity between two words can consider the root of the words. If the word trigrams were the same in English language, two words are similar. But in the Persian language, this may not be the same because for example ‘علم’ [elm] and ‘علوم’ [olum] have the same root in Persian but their trigram is not the same. Among the methods available, the method can be used for the Persian language is semantic similarity method , this is why the word net software has also been used.

In this article, for constructing emotional categories, the method is used with a some changes. To begin, we need a comprehensive dictionary of words that contain emotional weights or having such emotional weights in the sentence. Dictionaries that we've collected including about one thousand words in Persian language, the selected words have feels or it does not carry any feeling by itself but in the text can be modified to take emotional. The words were collected under the supervision of Persian literature specialists as much the dictionary becomes more complete, the accuracy of results increases.

The next step is to build the tree of the words in the dictionary. The tree which is presented here is based on emotional words in Persian language. To build this tree, the root node contains all of the words in the dictionary. Branching from the root node based on placements of the words and their categorizations.

Tree is constructed after a vast amount of investigation, including four levels that the final level is the emotional classification. In fact, each leaf contains words that are

emotionally similar. Some class or leaves maintain concepts that are implicitly a sense of emotion or they maintain concepts that they have not directly emotion. Build a powerful sense of trees, has a direct relationship with correctly clustering words in Glossary. The classification of words will manage the ambiguity of the words of the text. In this paper 20 categories are considered, also a word can be repeated in the other leaves of tree. For example death is a word that can feel pain and discomfort and fear is felt. Label that represents feel of bunch is selected among all the words that belong to a class. The word has most strongly feel in the expression those feelings, Tagged as label class.

2.1. Similarity calculation

For analysis of the text first should be made the tree. All of the words in input text examined. Therefore for each lexicon entry is required to calculate the amount of "similarity" with the emotional categories.

For each input text, one table is constructed that all words (except prepositions, turning letters and other letters have not emotion) are in this table. In this table there are 22 columns, for any of these words we will search the tree. Searching on this tree starts from the leaves that are emotional categories and goes upward. We calculate amount of similarity with every emotion word in categories. For this purpose, we select the label of word which has the highest intensity among the words in that class.

A word that belongs to several affect categories will generally have different intensities from category to category.

Therefore the amount of similarity is the numerical level of the lexicon entry, this number show the similar characteristics with the label of emotional categories. Amount of similarity is calculated with Formula 1:

$$\text{Sim}(X,X') = \frac{2 \log p(C'')}{\log p(C) + \log p(C')} \quad (1)$$

The word X is a member of category C and the label category studied X' while X' is in the class C'. Assume that X and X 'are both independent. P(C) is probability of randomly selecting a word in class C. P (C ') is probability of randomly selecting a word in class C '. C´´ is a class which X and X´ are belong it, it is the father of C and C´. P(C´´) is probability of randomly selecting a word in class C´´.

Similarity is in the numerical range [0,1], range is from 0 to 1 by increments of 0.1. If both words X and X´ are in the same class similarity degree is equal to 1.0. For example, the similarity degree for two words like "pain" and "happiness" is calculated as follows:

$$\text{Sim}(\text{happiness, pain}) = \frac{2 \log p(\text{ Emotional})}{\log p(\text{pain}) + \log p(\text{happiness})} \quad (2)$$

Similarity degree is equal to 0.0528 in this example. It's a low-intensity degree. Assigning category labels and membership degrees to lexicon entries is a very subjective process.

Another example is the words that are very resemble, consider "happiness" and "satisfaction":

$$\text{Sim(happiness, satisfaction)} = \frac{2 \log p(\text{ Pleasant})}{\log p(\text{satisfaction}) + \log p(\text{happiness})} \quad (3)$$

Similarity degree is equal to 0.7; it's a high-intensity degree.

In this process, categories with the highest similarity degrees are grouped together. We have constructed a table of 20 categories from lexicon entries and similarity degrees. For example: the similarity degree of 8 words with "interest" class is shown in Table 1:

Table 1..Similarity degree of 8 words with 'interest'

Lexicon entries	Similarity degree
tired	0
satisfaction	0.02
insanity	0
Grief	0
Death	0
love	1
pain	0

2.2. Density calculation

After the document is parsed and tokens (individual words) are generated one at a time and for lexicon entries numerical similarity values assigned, which represent the similarity of the affect level described by that entry.

Another column in this table is density for given lexicon entries. In fact, density is calculated because in input text, similarity degree of one word may be the same for two classes.

In this case, the density of any of the words here will have a decisive role. For example, the word "pleasure" has the same similarity to "satisfactory" and "happiness".

So no matter how the words categories "satisfactory" in the input text are more frequent.

Finally, the sense of the text will be "satisfactory".

Managing the uncertainties that mentioned at the beginning of the article, here are represented. In other words, one word can be a member of several emotional categories. After computing similarity and density and arranging elements finally we assigned words to the affect classes. As a result, the ambiguity in the notion of words is removed.

3 Test and result

After the affect lexicons in a document are tagged and the overall score for the document is calculated by using the awarded percentage to each category, with respect to the similarity of the column and densities. Finally, each column containing a higher percentage will show the dominant feeling of the text.
As an example, consider the following paragraph:

'*Do you think that being happy is elusive and away dream? Do you believe that external events, give you the happiness as a gift? If you suspect that your joyful is hidden in your thought and mind, Opportunity to build a life full of happiness and satisfaction is ahead of you, happiness, gratitude and appreciation is necessary. So that each day is better be grateful for the least thing that you have.* '

Table 2. Similarity degree for five categories for entries

JOYANCE	TORNADO	DISORDER	SATISFACTION	LOVE	
0.5587142	0.28831	0	1	0.27834463	**happy**
0.2883142	0.28831	0.054498509	1	0.27834463	**dream**
0	0.05308	0.27834463	0.057034928	0.05135125	**event**
0.2883142	0.28831	0	1	0.27834463	**gift**
1	0.88842	0	1	0.25984831	**joyful**
0.5587142	0.28831	0	1	0.27834463	**satisfaction**
0.5587142	0.28831	0	1	0.27834463	**gratitude**
0.5587142	0.28831	0	1	0.27834463	**appreciation**
0.5587142	0.28831	0	1	0.27834463	**grateful**

Table 1. Text analysis for the five categories

Result	LOVE	SATISFACTION	DISORDER	TORNADO	JOYANCE
percent	17%	53%	25%	25%	25%

Results in Table 3 in addition to the similarity are calculated based on the density of input words in the text. As we can see the results has expressed as percentages and they show amount of emotional for the input text. On the other hand, they show weights of the emotional categories and we can find an affect category that has more weights in the document. The feeling of the input text can be understood as intuitive. For this example in Persian language, the most feel is satisfied.

4 Conclusion

To manage uncertainties in the Farsi language, semantic similarity is a useful method. This method has higher accuracy compare to similar methods while it is much faster. In any case, efforts have been reported in this article, Could be a prelude to further studies and research in the Persian language. Our plans in the immediate Future include:

- Develop a wider range of emotional categories to to increase the ability of expressing the emotion details more in the input text.
- Using other techniques in artificial intelligence to increase the accuracy in determining the emotional categories.

References

1. G.Grefenstette, Yan Qu, J. G. Shanahan, D. A. Evans, “Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application”, Gegory Clairvoyance Corporation, 5001 Baum Bd, Suite 700, Pittsburgh, PA, 15213-1854, USA
2. Deb Roy and Alex Pentland, “Automatic Spoken Affect Analysis and Classification”, MIT Media Laboratory, Perceptual Computing Group, 20 Ames St. Cambridge, MA 02129 USA.
3. Dekang Lin, “An Information-Theoretic Definition of Similarity”, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2
4. Jay J. Jiang and David W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, 1997, Taiwan
5. [Resnik, 1995b] Resnik, P. (1995b). “Using information content to evaluate semantic similarity in a taxonomy”, In Proceedings of IJCAI-95, pages 448–453, Montreal, Canada.
6. Pero Subasic and Alison Huettner ,”Affect analysis of text using fuzzy semantic typing”, CLARITECH Corporation, Justsystem Group, 5301 Fifth Avenue, Pittsburgh, PA 15232, USA
7. Jay J. Jiang, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, David W. Conrath, MGD School of Business, McMaster University, Hamilton, Ontario, Canada L8S 4M4.

8. H. Givi, H. Anvari, " Dastor zaban Farsi"
9. Lotfi A. Zadeh, "Fuzzy Logic Computing with Words", IEEE Transactions on Fuzzy Systems, 2, 103-111, 1996.
10. Rosalind W. Picard, "Affective Computing", MIT Press, 1997.