5th International Conference on
15 September 2018
Information Technology,Computer&Telecommunication

Study of emotions and thoughts of the society and the effect of various events on them by social media

Amin Karimi Monsefi[1]
Researcher at Shahid Beheshti University
Behzad Zakeri
Researcher at University of Tehran
Babak Darafarin
Researcher at University of Tehran

Abstract

Nowadays, social media is the best place to argument topical issues of the societies. From simple events to the most complicated problems, thousands and thousand contextual contents share among social media. These data show the proportion of social satisfaction about discussed topics. In this research, by considering data which were produced by Twitter's users and weighting them by fuzzy methods, these data compared and satisfaction level of people related to specific subjects was calculated.For this purpose, reference databases were built by web clawlers to extract data from sport, economic and social websites.

Keywords: Data Mining, Text Mining, Social Media, System Fuzzy, Clasification

[1] Corresponding Author

1. Introduction

Presence of the internet and social media have caused that mankind experience new ways of communication. Fast data transmission is one of the most important features of this environment which caused the production of the enormous amount of data related to social issues every day. These data may contain simple and daily issues of a country, or follow the trend of significant social events.

By developing the communication systems, the role of designing of new methods for fast analysing of information trends become more and more important. For this reason, different methods and machines with computational and analytical methods have been used to accelerate this process, and Data mining is one of them which present an effective method for analyzing data [1]. Data mining is trying by defining the problem based on data and categorization and analysing meaningful data, recognise available aims and needs on the internet. Suggested methods in this paper are based on available Contextual content on social media. While texts are the considerable part of the existing content on the internet, they are more comprehensive on target subjects, and for this reason, they have higher priority to use compare to other forms of contents.

In this research, one of the most popular social media which is twitter have been used as target society. Everyday huge amount of tweets are released about social events by users of Twitter [2], and the popularity of Twitter has made it a powerful tool for analysing data. For this purpose, three categories of tweets which are social, economic, and sports have been studied.

The first step is the categorization of tweets into three discussed categories. For this purpose, each contextual data is compared to the database of words and phrases of each category. The result of this comparison is an output of FSAW[2]. Also, another categorization based on satisfaction has been applied on data. In order to simplification, this categorization only contains two states: satisfaction and dissatisfaction. For instance, the content of an input can be about a social issue which shows the author's satisfaction about a specific event.

To assess and categorization data, a database of benchmark terms is needed. For building this database, a set of websites with similar topics (social, economic, sports) have been considered, and web crawlers were used to extract contents from topical sites. By integration of existing words in different websites which are related to our subjects, and neglect to redundant words, such as names, pronouns, and..., a list of relevant words with higher probability to use in related subjects has been achieved. Then, by taking the advantage of other phrases which show the level of satisfaction, the process of comparing and weighting has been done.

[2] Fuzzy Simple Additive Weighting Method

2. Methods

In this part, firstly learning data generation is introduced, Some features of the target community are analysed.

2.1. Generating of Learning Data

For data generating several websites related to our main three subjects were considered, and Recent contents for all these sites were extracted by web crawlers [3].
Firstly, a database with 1300 words and prepositions related to social, economic and sports subjects was produced.
Then, all existing contacts in all websites were extracted and then were broken into single words. By using our database, redundant words from extracted contexts were filtered. For example, words like "sir" or "they" was removed.
although all of the remaining words are meaningful, some of them are not the main infinitive and should change to infinitive. For this reason, all words was changed to the main infinitive [4]. As an example, a word like “playing” was exchanged to “play”.
For each subject, related words linked together, and by this way all of the data categorised into our three target groups. Also, the value of each word in its category calculated by (Eq-1):

$$f(word) = \frac{len(word)}{len(all\ word)}$$

(Eq-1)

$$if\ \ len(word) > \checkmark:\quad f(word) = 1$$

Due to the (Eq-1), it is clear that the value of each word is calculated by the proportion of the number of usage of that word compare to all the words. However, for words which were used more than the particular time (✓), the value was assumed to be 1. Also, the value of (✓) should be chosen carefully, and it depends on the number of whole words[5].
for instance, a word like "football" in each category has the different value. while in the sports category has the value of 1, its value in econimic and social categories is 0.0009 and 0.0078 respectively.
For analyzing the proportion of satisfaction compared to the dissatisfaction with the society, for each category, a database with frustration and unsatisfactory words, and also a database with hopeful and satisfactory words was built.

2.2. Data Extraction Method from Tweets

In order to data extraction from tweets, Twitter provides a series of APIs for programmers, and by using them thay can access to part of tweets. For this process, by using the library of TweePy in Python language it is possible to build a socket among Tweets and APIs. In the following, by

taking the advantages of Spark and spring source which were produced by socket it is possible to have an access to considerable part of Tweets all around the world in the form of json.

2.3. Categorizing of Tweets

After changing words to infinitive from which discussed in section 2.1, words valuation by (Eq-2) has been done for all three main categories[6].

$$val(set) = \sum_i f(word_i) * len(word_i)$$

(Eq-2)

The largest value of the (Eq-2) for all three categories is considered, and if that value was larger than ✔, that Tweet will belong to that category. For assessment of satisfaction, it is just needed to count the number of words in both satisfactory and unsatisfactory databases.

2.4. Determining the Parameters

in previous sections, two parameters called ✔and ✔were defined which choosing the correct value for each, is a vital part of the research. For calculating the parameters some of the Tweets were chosen and for each one, a category was allocated manually. Therefore, a specific value was assigned to parameters and each step one of the Tweets was analysed, and it is tried to change the parameters in a way that maximum of Tweets categorize correctly [7].

3.Results and Discussions

By analyzing Tweets of Iranians, it was understood that during world cup 2018 games which were accompanied by turmoil economical conditions, people rather than complaining about the economic problems were willing to show their emotions about satisfactory national team results. By contrast, after finishing the national team games situation was changed and again people start to complain about economic issues. For this reason, Tweets of the community after all Iran's game in the world cup and after some days after sudden inflation in Iran's economy was analysed.

	Social unsatisfactory	Social satisfactory	Economocal unsatisfactory	Economical satisfactory	Sports unsatisfactory	Sports satisfactory
Iran vs Morocco	1.8%	2.6%	9.7%	0.2%	10.6%	57.6%
Iran vs Spain	0.9%	0.6%	12.8%	1.2%	24.3%	42.3%
Iran vs portugal	4.1%	1.2%	24.3%	0.02%	8.2%	56.1%
After Iran's economic inflation	2.4%	1.1%	54.1%	0.08%	5.1%	31.1%

(Table.1) Tweets analyzing

In many cases, tweets were not countable or were or relevant to our three main categories, and for this reason, the sum of events is not equal to 100%.

4.Conclusion

There is no doubt that social media is a strong tool for analyzing people's thoughts, and provide useful information from societies.

References

[1] H. Han,W. Jeong, and D.Wolfram. Log analysis of academic digital library: user query patterns. In iConference 2014 Proceedings, pages 1002–1008, 2014
[2] S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. A transaction log analysis of a digital library. International Journal on Digital Libraries, 3(2):152–169, 2000
[3] H.-R. Ke, R. Kwakkelaar, Y.-M. Tai, and L.-C. Chen. Exploring behavior of e- journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect onsite in Taiwan. Library & Information Science Research, 24(3): 265–291, 2002
[4] D. Lindberg. Internet access to the National Library of Medicine. Effective Clini- cal Practice, 3(5):256–260, 2000
[5] X. Niu, B. M. Hemminger, C. Lown, S. Adams, C. Brown, A. Level, M. McLure, A. Powers, M. R. Tennant, and T. Cataldo. National study of information seeking behavior of academic researchers in the United States. Journal of the American Society for Information Science and Technology, 61(5):869–890, 2010
[6] S. Pontis and A. Blandford. Understanding "influence:" an exploratory study of academics' processes of knowledge construction through iterative and interactive information seeking. Journal of the American Society for Information Science and Technology, 66(8):1576–1593, 2015
[7] S. Pontis, A. Blandford, E. Greifeneder, H. Attalla, and D. Neal. Keeping up to date: An academic researcher's information journey. Journal of the American Society for Information Science and Technology, 68(1):22–35, 2017