ANALYZING EMOTIONS FROM TEXT CORPUS USING WORD SPACE

Mohsin Manshad Abbasi, Professor Anatoly P. Beltiukov

Department of Theoretical Foundations of Computer Sciences
Udmurt State University
Izhevsk, Russian Federation
e-mail: mohsinmanshad@gmail.com, belt.udsu@mail.ru

***Abstract**: Emotion identification and analysis is a topic of growing interest. It is because of abundant amount of data available online and offline in different languages. In this paper, we propose a new methodology called Word space. In this methodology, we use the frequency of words to analyze the text corpus before its classification. Word space is the concept derived from Metric Space, which is defined as a set for which the distance between all of its members, is defined. Those distances are when taken together called a metric on the set. In Word space distances between the words and their occurrences are measures. The emotion carry words with relatively high frequency and less distance between their occurrences are strong emotions. Whereas less frequent emotions that occurs far from, each other are weak emotions in the text. Using this concept, we will analyse the document, categorize and can summarize its emotions. In the conclusion section, we explain the interesting results that we observed using this methodology.*

1. Introduction

Emotion Analysis is one of the most pursued research topics in recent times. Many researchers and companies have explored the area of opinion identification and its analysis. With the increase in the number of Internet users, there is a proliferation of opinions available on the web. This phenomenon has a huge impact on various applications such as product review summarization and the public opinion monitoring systems.

There is a need and possibility for parallel and distributed computing use in the sphere of NLP (Natural Language Processing) [1]. Various techniques and methodologies are used for the identification of Emotions from text. Word space is relatively a modern concept based on the concepts of Metric Space in Mathematics. In Word Space, the distance between the first and the next occurrence of a word is measured. Each word and its occurrences are save using a vector. In our research, we will use the similar concept for words that express emotions using the exemplary text and equations explained below.

A lot of people are happy and satisfied with the service of online shopping sites and it is because of they are secure and safe. Online shopping sites are selling good products therefore the people are happy and satisfied with them. They have been able to build trust among the different customer. It is the simplest and easiest way to shopping.

Though number of online shoppers is increasing, but still there is one biggest pitfall adding to its limitation factor, that is the risk factor it entails. Starting from credit card scams to fake product deliveries, we have clearly heard of online shopping going wrong in many ways. The risk factor is also on its peak. The products are fake and the website are not secure increases the risk to buy online. Still most of recent surveys on internet shows that people are highly satisfied buying the essential things online. The reason behind the popularity of online shopping is: it is easy and fast. It's the most simplest way of shopping from sitting at home or wherever you are and therefore, customers are happy and enjoying new shopping experience these days. It seems like online shopping stores are cheap and more efficient than the local stores we visit. Online shopping sites play significant role making human life easier. It's safe if you care some important things

Blog taken from https://www.quora.com/Is-online-shopping-safe-or-not

The text is taken from an online blog where the people express their opinion about online shopping. The words expressing positive emotion are highlighted with green marker whereas the negative are with red marker. Some of those words occurred more frequently while other are less frequent. In our work, we will analyze the similar blog expressing opinion of people resulting in big data. To analyze the data we are using some equation and formulas as explained below.

In first equation, we have C_1 that represent the number of occurrence of a word expressing an emotion in text. For every word expressing an emotion in text, the number of its occurrence counted resulting in an array C carrying the occurrence count of all emotion-carrying words from the text. Similarly F_1 represents the frequency of occurrence of a word by dividing its word count C_1 with the total number of words T_m in the text and on applying it to all the members of Word count array C, we have a new array F represent the frequency of all the emotion carrying words in the text.

$$C_1 = \text{Count (Word Occurrence)} \quad (1)$$

$$C = [C_1\ C_2\ C_3\ C_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots C_n] \quad (2)$$

$$F_1 = \left(\frac{C_1}{T_m} \times 100\right)\% \quad (3)$$

$$F = [F_1\ F_2\ F_3\ F_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots F_n] \quad (4)$$

The primary objective of this study is to analyze emotions from text. The emotion carrying words and their occurrences extracted in the form of a vector. This vector values are then use to identify the frequency of occurrence of the words expressing emotions in the text. On the basis of these calculations, the emotions expressed in the text can be categorize as frequent, less frequent, most frequent and least frequent emotion. It is also observed that while extracting the words carrying emotions, there should not be any negation such as may not, cannot, don't before the emotion carrying word. If it occurs then the word polarity will be change. Positive emotion will become negative and negative emotion will become positive emotion.

Table 1 Describes emotion reverses on applying negations

Operator	**Original Emotion**	**Resulting Emotion**
Negation	Positive	Negative
Negation	Negative	Positive

2. Applications

The analysis of text corpus based on emotions can be used for summarizing the text [2]. We can identify important emotions that occur more frequently in the text and are more similar to the words mentioned in the title of the text. For this purpose, two properties of the words are considered. The Local property and the Global property. Local property is the frequency of occurrence of the word in the text while Global property is the maximum semantic similarity between a word and the title of Text [3]. Using this method the words carrying important information or emotions will not be missed while summarizing the text.

It can be used to identify the occurrence of the words and to avoid its repetition. It has been observed that the occurrence of same words expressing emotions frequently results in losing the interest of the reader in the text. For making the sentences of good quality and attract the attention of the readers, the identification of word repetitions is very important. The similar idea was used on twitter data written in Indonesian Language. The idea was to remove the repeatedly characters and words so that the data to be processed in accordance with the needs [4].

It can be used to identify the sentence similarity measure for paraphrase identification. It is more difficult to distinguish between precise and loose paraphrases than between loose paraphrases and non-paraphrases [5].

It can be very helpful for making the prediction of the next occurrence of an emotion in the text based on calculations of previous occurrence of a word in the text. A similar methodology was applied for text-based emotion prediction problem using supervised machine learning. The experiment was performed on children's fairy tales and the results were convincing [6].

It can be used for approximate string matching. Such as online search engines that takes a query from the user and match each word of this query with the database of text available online. However, this matching is based on word to word matching without taking in consideration the emotions in the text. In future, the emotion expressing words can get the priority in search engines for performing the text matching based on emotions.

3. History of Emotion Analysis

The development of the General Inquirer System (1966) (Stone, [7]) by Philip Stones in Harvard was probably the first milestone to identify textual emotions. The system usually counts the positive or negative emotion instances.

After this, a lot of work has been done for identification of emotions from text in different languages. An important among them was the contributions of Jaynce Wiebe, Peter Turney and Vasileios Hatzivassiloglou during early 90's. Jaynce Wiebe in 1990 (Wiebe, [8]) defines the term "Subjectivity" for Information Retrieval research. Later on in the year of 1997 (Hatzivassiloglou et. al., [9]) identified the semantic orientation of adjectives. After a few years Peter Turney (2002, [10]) came up with his revolutionary approach of Thumbs Up and Thumbs Down for positive and negative review classification. Pang (et. al., 2002, [11]) has suggested the building of sentiment lexicon manually for a domain. (Denecke, 2009, [12]) reported an interesting study on multiple domains to demonstrate the usefulness of the prior polarity scores from the SentiWordNet. Clustering had been observed as a technique based on generalizations of graph partitioning that don't require pre specified ad hoc distance functions and is capable of automatically discovering document similarities or associations [13].

In Russia studies devoted to sentiment analysis in Russian before 2011 are not very numerous. In (Ermakov, 2009, [14]) a sentiment analysis system extracting opinions about cars from a Russian blog community is presented. The patterns are language-dependent and domain dependent, which means that patterns must capture the lexical, syntactic and stylistic features of the analyzed text. It is not possible to directly translate or map the English pattern base into Russian pattern [15].

Emotion analysis from text in Russian language appears mainly in multilingual experiments. Zagibalov in (Zagibalov et al., 2010, [16]) compare corpora of reviews related to the same books in English and Russian. In (Steinberger et al., 2011, [17]), construction of general emotion vocabularies for several languages is described. Chetviorkin and Loukachevitch (2012, [18]) described, the generation of the Russian sentiments and emotional vocabulary for the generalized domain of products and services and so on.

In bilingual experiment on Russian and Romanian languages, it is observed that the word spelling can be considered as a word phonetic equivalent. This feature allowed limiting the search to letter-based representations [19]. For emotion identification and analysis in Russian language, most of the researchers are using Rule based techniques which is quite necessary as the rules are key to develop a grammar that can be used to develop stemmer and then semantic and syntax analyzer [20]. It has been observe that the methods of machine learning performs better in text categorization and classification.

4. Related work

In 1977 (Halliday and Hasan, [21]) classified lexical cohesion into two categories: reiteration category and collocation category. Reiteration category considers repetition, synonym, and hyponyms, while collocation category deals with the co-occurrence between words in text document. After this, the major effort was the work on the surprising behavior of Distance Metrics in High Dimensional Space [22]. In this research, it was explained that the fractional distance metric provides more meaningful results both from the theoretical and empirical perspective.

The main factors that affect the efficiency when searching metric spaces are intrinsic dimensionality of the space and the search radius [23].

(Edgar C. et al. 2001 [24]) in their work present a technique to perform search in Spaces using the distance. They proposed a concept for working with general cases to model similarity with a distance function that satisfies the triangle inequality, and the set of objects.

Kruengkrai and Jaruskululchi try to determine text title [25]. Their approach takes advantages of both the local and global properties of sentences. They used clusters of significant words within each sentence to calculate the local property of sentence and relations of all sentences in document to determine global property of text document.

In 2012, Maryam Kiabod et al. [26] present an approach for summarizing the text by identifying significant words from the Text. They used the extractive method to select the subset of the sentences that contains main concept of the text. The algorithms for preprocessing the text before analyzing the emotions improves the efficiency and effectiveness of the complete process [27].

5. Methodology

In this method, the list of words carrying emotions are extracted as shown in equation below. W represent list of all words expressing emotions. D1 represents the average distance between the different occurrences of emotion carrying word W_1. On applying it to all words carrying emotion we have a list D that represents the average distances of all emotion expressing words. The strength S_1 of a word is directly proportional to its word count C_1 and inversely proportional to the average distance D_1.

$$W = [W_1\ W_2\ W_3\ W_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots W_n] \quad (5)$$

$$D_1 = \frac{1}{N}\sum_{n=1}^{n}[W_1 D_n] \quad (6)$$

$$D = [D_1\ D_2\ D_3\ D_4 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots D_n] \quad (7)$$

$$S_1 \propto C_1 \quad (8)$$

$$S_1 \propto \frac{1}{D_1} \quad (9)$$

Before calculating the average distance occurrences, the maximum and minimum distance values were identified and then normalized. For this purpose the maximum value is divided by the total number of occurrences of that word and the minimum value is multiplied.

$$Norm = \max(D)\ ,\ min\ (D) \quad (10)$$

$$Norm = \frac{\max(D)}{C}\ ,\ \min(D) \times C \quad (11)$$

The strength of an emotion is the total occurrence of it Ci divided by the average distance of occurrence D and multiplied by the total number of emotions in the textT_E. On calculating the strength, average distance and occurrence, the strong, very strong and weak emotions can be identified and the text corpus can be categorized as positive or negative.

$$S(W) = \frac{\sum_{i=1}^{n} Ci}{D} \times T_E \quad (12)$$

6. Results and Discussion

The data obtained from the blogging corpus for the purpose of experiment showing the comments of people from October 1, 2013 to September 23, 2016. More than 25 people showed their opinion using around 6000 words. They used a total of 78 emotion oriented words. Among them, 40 words are representing positive emotions, 32 are negative and 6 with inversed polarity. The ten most frequent positive and negative emotions are shown in the tables below.

Table 2 *Show most frequent positive emotions in text*

Emotion	Total Occurrence	AVG Distance	Percentage
Safe	29	66	14.5
Secure	16	81	8.04
Good	12	223	6.03
Trust	9	77.3	4.52
Easiness	6	67.6	3.01

Table 3 *Show most frequent positive emotions in text*

Emotion	Total Occurrence	AVG Distance	Percentage
Hack	7	45.7	3.51
Trouble	6	95.8	3.01
Steal	6	52.6	3.01
Unsafe	2	27.5	1.005
Bad	3	359	1.50

For analysis, an online blog on the topic, "is online shopping safe or not" has been used. The five most frequent positive and negative emotions in the text corpus are discussed here. The most frequent positive emotion is safe. It occurs almost after every 66 words and has a high percentage of 14.5. It is probably because of the reason that it is the part of the topic and most of respondents try to answer it with the same word. Then the words secure and good are the emotion that show the user feels secure and the quality of products they buy are good. They also trust and feel easiness in doing shopping online.

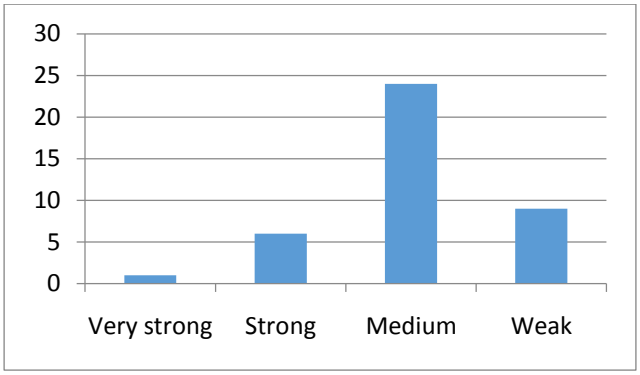
On the other hand, the most frequent negative emotion presented by users is Hack. It is because they are scared of their credit card information theft online. This emotion is shown on average after every 45 words which can be considered has an important and strong emotion. Then the second frequent negative emotion is steal. It is related with the previous emotion that describes the security threats while doing shopping online. People can steal the personal information of the buyer. This will create trouble for the buyer and the shopping can be unsafe.

It has been observed that in general the people response to online purchase is very satisfying and positive. This result is concluded by observing the positive emotion, their high frequency, average distance and the percentage of occurrence. The date of writing the blog is also very important. People now trust more online shopping than before. It is because of the modern technology. However, they are still concern about technology challenges in the form of hacking and steal.

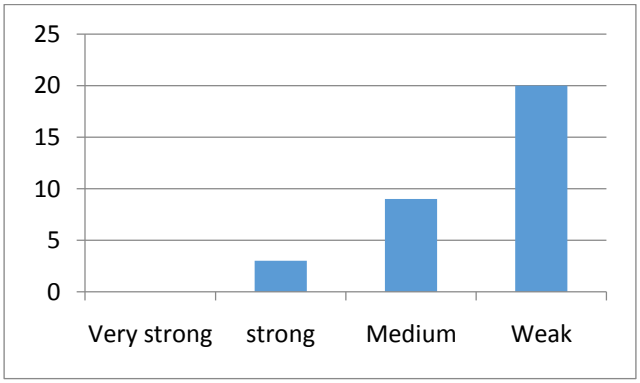
Table 4 *Comparison of positive and negative emotions from text*

Emotion Type	Total	Very Strong	Strong	Medium	Weak
Pos+	40	1	6	24	9
Neg-	32	0	3	9	20

The above table describes the text contains more positive emotion than negative emotions. The positive emotions are more strong whereas then negative emotions are mostly weak emotions.



Graph 1 *Representing Categories of Positive Emotion*



Graph 2 *Representing Categories of Negative Emotions*

The graphs are representing different strengths of positive and negative emotions. It is observable that negative emotions are mostly weak emotions with an average occurrence of less than 3 % in the text whereas the positive emotion are mostly medium strength emotions along with strong and even very strong emotions. It elaborates that text is representing positive emotions and is positive text.

7. Future Work

In future, we will apply our methodology on social media and tweets. The size of the data will be increased. The results obtained from this approach will be applied for summarizing the text based on the emotions represented in the text. We will observe other advantages of the methodology for sentiment analysis.

8. References

1. Solovyev V, Polyakov V, Ivanov V, Anisimov I, Ponomarev A. "An approach to semantic natural language processing of Russian texts". *Research in Computing Science*, 2013, pp. 65-73.
2. Ashmita S, Ruhi B. "Auto Text Summarization with Categorization and Sentiment Analysis". *International Journal of Computer Applications*, Vol. 130. November 2015, pp. 07.
3. Maryam K, Mohammad N. D, Sayed M. S. "A Novel Method of Significant Words Identification in Text Summarization". *Journal of Emerging Technologies in Web Intelligence*, Vol. 4. August 2012.
4. Fachrian A, Arif D. "Improving the Performance of Repeated Character Preprocessing in Recognizing Words in the Indonesian Sentiment Classification". *Journal of Basic and Applied Scientific Research*, 2017, pp.1-9.
5. Pronoza E, Yagunova E. "Comparison of sentence similarity measures for Russian paraphrase identification", In: *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, 2015, pp.74-82.
6. Cecilia O.A, Dan R, Richard S. "Emotions from text: machine learning for text-based emotion prediction". In: *Proc. of the 5th conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Canada, 06-08 October 2005, pp. 576-586.
7. Stone, P.J, Dunphy, D.C, Smith, M.S. "The General Inquirer: A Computer Approach to Content Analysis". MIT Press - Cambridge, 1966.
8. Wiebe, Janyce M. "Identifying Subjectivity characters in Narrative". In: *Proc. of the 13th International Conference on Computational Linguistics*, Helsinki Morristown NJ: Association of Computational Linguistic, 1990, pp. 401-408.
9. Vasileios H., Kathleen R.M. "Predicting the Semantic Orientation of Adjectives" *EACL '97* In: *Proc. of the eighth conference on European chapter of the Association for Computational Linguistics*, Madrid, Spain, July 07-12, 1997.
10. Peter D.T. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". *In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 2002, pp. 417-424.
11. Pang B, Lee L. "Thumbs up? Sentiment Classification using Machine Learning Techniques" *In: Proc. of the Conference on Empirical Methods in Natural Language*, Philadelphia, 2002, pp.79-86.
12. Denecke K. "Are SentiWordNet scores suited for multi-domain sentiment classification?". *In: Proc. Of the fourth International Conference on Digital Information Management*, University of. Michigan, USA, 1-4 November 2009.
13. D. Boley, M. Gini, R. Gross. "Partition based clustering for web document categorization". *Elsevier Journal for Decision Support Systems*, Vol. 27, Issue 3, December 1999, pp. 329-341.
14. Ermakov A. "Knowledge extraction from text and its processing: Current state and prospects". In*: Proc.: of the Computational Linguistics and Intellectual Technologies*, 2009. pp. 50–55.
15. Pivovarova L, Du M, Yangarber R. "Adapting the PULS event extraction framework to analyze Russian text". In: *Proc. of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, 8-9 August 2013, pp. 100–109.
16. Zagibalov, T , Belyatskaya et al. "Comparable English-Russian Book Review Corpora for Sentiment Analysis" , Edition 2010.
17. Steinberger J, Lenkova P, Kabadjov M., et al. "Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora", In: *Proc. of Recent Advances in Natural Language Processing*, Bulgaria, 12-14 September 2011, pp. 770–775.
18. Chetviorkin I, Braslavskiy P, Loukachevich N. " Sentiment Analysis Track at ROMIP 2011" Computational Linguistics and Intellectual Technologies: *In: Proc. of the International Conference* (Dialog 2012), Bekasovo, pp. 1–14.
19. Sokolova M, Bobicev V. "Classification of emotion words in Russian and Romanian languages". *International Conference RANLP*, Borovets, Bulgaria 2009, pp. 416–420.
20. Abbasi M.A, Beltiukov A.P. "Analysis of sentiment and emotion from text written in Russian language*" 5th All Russian Conference on Information technology for intelligent decision making support (ITIDS)*, Ufa, Russian Federation, May 16-19, 2017.
21. Halliday M, Hasan R. "Cohesion in English". London: Longman, 1975.
22. Aggarwal C.C, Hinneburg A, Keim D.A. "On the surprising behavior of distance metrics in high dimensional space". *International Conference on Database Theory*, 2001, pp. 420-434.
23. Cha E, Navarro G, Baezayates R, Jose Y, Marroquin L. "Searching in metric spaces" *ACM Computing Surveys*, Vol. 33, Issue 3, September 2001, pp. 273–321.
24. Edgar C, Gonzalo N, Richardo B.Y, Jose L.M. "Searching in Metric Spaces". *ACM Computing Surveys*, Vol. 33, No. 3, September 2001, pp. 273–321.
25. Jaruskululchi C, Kruengkrai. "Generic text summarization using local and global properties of sentences". *IEEE/WIC international conference on web intelligence*, October 2003, pp.13-16.
26. Kiabod M, Naderi M, Sharafi S.M. "A novel method of significant words identification in text summarization". *Journal of Emerging Technologies in Web Intelligence*, Vol. 4, Issue 3, August 2012.
27. Abbasi M.A, Beltiukov A.P. "Механизм предварительной обработки текста перед анализом настроений*". 6th All Russian Conference on Information technology for intelligent decision making support (ITIDS)*, Ufa, Russian Federation, May 28-31, 2018.