LEXER: LEXicon Based Emotion AnalyzeR

Shikhar Sharma(✉), Piyush Kumar, and Krishan Kumar

Department of Computer Science and Engineering, National Institute of Technology Uttarakhand, Srinagar Garhwal, India
{shikhar01.cse14,kpiyush26.cse14,kkberwal}@nituk.ac.in

Abstract. The huge population of India poses a challenge to government, security and law enforcement. What if we could know beforehand the consequences of any events. Social spaces, such as Twitter, Facebook, and Personal blogs, enable people to show their thoughts regarding public issues and topics. Public emotion regarding future and past events, like public gatherings, governmental policies, shows public beliefs and can be deployed to analyze the measure of support, disorder, or disrupted in such situations. Therefore, emotion analysis of Internet content may be beneficial for various organizations, particularly in government, law enforcement, and security sectors. This paper presents an extension to state-of-art-model for lexicon-based sentiment analysis algorithm for analysis of human emotions.

Keywords: Emotion detection · Natural language processing · Security · Prediction

1 Introduction

Social spaces, the Internet is one of the most important personal view exchange portals in the current era. The social media, such as Reddit, Blogs, and Pinterest etc. and the Internet, supply a very suitable platform for communicating with each other and sharing views with everyone. Accordingly, the part of the Internet in crime prevention and investigation has promptly increased. The Internet is rapidly becoming a way of preventing chaos by providing information for warning systems in concern of public safety. Sentiment study has already been used in various areas [1], non-secure area to forecast and monitor common views. Grabner et al. [2] deployed a domain based lexicon for classifying Twitter reviews. Sentiment examination is also executed on Twitter for forecasting box-office collections of movies [3]. Along with increasing interest in "Affective Computing", the task of "Emotion Detection" using text has also gained more attention during recent years. However, very little efforts are done in the detection of multiple emotion at the same time. Instead, most of the previous works assumed that emotions are mutually exclusive and focused on multi-class classification. But, the human emotions are much more complicated than that: emotions have connections, some occur together simultaneously, while some are opposite of each other, while resonate and create other emotional states [4,5].

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 373–379, 2017.
https://doi.org/10.1007/978-3-319-69900-4_47

The focus of the above work is mostly classifying emotion of document sources or processing the tweets around the event into a single type of emotion. However, they do not provide insights into how to characterize person's multiple emotion, which is the main contribution of this work. Perhaps the closest work to us is [5–8]. In contrast, we provide a fully automated and principled solution. The salient features of our work are stated as follows:

- LEXER technique can be employed to analyze the emotions behind the tweets (*anger, disgust, sadness, surprise, fear, joy, neutral*) over the Internet.
- A fuzzy set function is used to complement the emotional value of a negated word. This in comparison to polarity reversal is more realistic and reliable. Therefore, it can help to prevent the public outrages, communal riots in stipulated time.
- From the point of view of tweets, analysis, an efficient multi-emotion analyzer has not been applied to the best of our understanding. The proposed principled solution still has good emotion analyzing capabilities in comparison to previous works.

The manuscript is structured as follows, details of the work are discussed in Sect. 2, results are shown in Sects. 3 and 4 concludes the work.

2 Proposed Method

The approaches in sentiment examination can be grouped into two categories. Using lexicon is one of them. It demands calculation of the sentiment based on the semantic orientation of words or expression that happen in the text. From this perspective, a vocabulary comprising of negative and positive expressions is used, Moreover, a value is allotted to each word that can be either negative or positive called sentimental value [7]. In our model, we tried to extend this approach to analyze the emotions of the users.

Instead of having a dictionary of negative and positive words, we created a dictionary has different emotion values for words. Normally saying, the lexicon-based perspective uses a snippet of text that can be understood as **bag of words** [8]. Ensuring this understanding, emotion values from the vocabulary are allotted to every expression that being used in the text. The different values are combined together using a function known as combining function, such as average or sum, to make the end most prognosis relating to the comprehensive emotion for the text. Apart from the emotion estimate, the thought of the local context of the expression is also important, like intensification, inversion and downtoning.

The availability of labeled training set is very scarce. Thus, the work resolute in implementing a lexicon-based expertise. The dependency on the labeled data is the main drawback of machine learning algorithms. Also, the sufficiency and correctness of labeled data are extremely difficult to ensure. Besides this, the ease of modification and understanding for the human in case lexicon-based procedure gives an advantage over traditional machine learning expertise.

Thus, this can be contemplated as a notable merit of our work. We discovered that there is an ease in the generation of an efficient vocabulary in comparison to the collection and labeling relevant corpus. Moreover, lexicon based approaches are easily transformable into different languages.

2.1 Emotion Lexicon

The emotion lexicon constructed to consist of 3000 words. It is manually generated using movie reviews as the baseline. Each element of the vocabulary is given six values depicting different emotions, i.e. Anger, disgust, sadness, surprise, fear, joy. The values vary in the range of 0 (no dominance) to 100 (most dominating). From the knowledge of human psychology, it knows that sometimes we human exhibit emotions which are a mixture of different emotions like anger and disgust. For example, sentence *Corruption in India is increasing day by day* represents anger as well as disgust. It may be wrong in such cases to decide between anger and disgust. To overcome such issues, we tried to classify the emotion into multiple emotions.

2.2 Intensitifiers and Downtoners

Intensifiers are the words like definitely, really, too, etc. These words can be defined as words that increase the dominance of a particular emotion over the other. They can be classified into two categories [8,9], namely downtoners (rarely, never) and amplifiers (really, too) as decreasing and increasing the intensity of emotions respectively. In our work, all the intensifiers were sorted on the basis of frequency and the 25 most frequently used intensifiers were selected and then, these intensifiers were subdivided into two classes, namely downtoners and intensifiers. By means of experimentation, we concluded that downtoners can decrease the value of the emotion by half.

2.3 Inversion

The most widely used technique for handling inversion in lexicon-based expertise is to reverse the polarity of an item in the vocabulary. It applies to words that are preceded by a negator in a sentence [10,11], for example, happy: 87 and not happy: −87. In our work, we decided to use a different procedure for inversion. Instead of reversing the emotion value, we employed a complementing function which complements the value of recognized words, for e.g. happy: 80 and not happy: 20. At first, a lexicon comprising of 20 negating words is created manually. Following this, we used the Twitter corpus was used to select the most frequent inversion of adjectives and verb expressions. Then, we applied the concept of complementing a fuzzy set. Since each emotion value can be treated as membership value, its complement can solve the problem.

The main merit of using the complementing function instead of the polarity reversion is better accuracy in allotting the emotional values of expressions.

For example, in the sentence: *I don't enjoy the ride.*, the emotion which will be allotted to the sentence using the traditional polarity reversal procedure would be sadness (opposite of joy is sadness) and the sentence will be the dominance of sadness as emotion. In fact, it will be same as *I felt this ride saddening.*, which is contradictory to real world scenarios. But using the complement function as shown in Eq. 1 would give better results.

$$F_n = 100 - F_e \tag{1}$$

Moreover, no intensifiers and negators are included in emotion lexicon. Also, if they are surrounded only by neutral emotion, they are considered as neutral words as proposed by Jurek et al. [8].

2.4 Combining the Results: Combining Function

After the identification of all the expressions in the text, the local context of these expressions is verified. The combining function is then applied to obtain the endmost value. In our work, there is the requirement of a function that can be applied to the single expression. It returns the absolute value of each emotion from the text normalized to $0 - 100$. This resulted in better efficiency to analyze the emotion with respect to intensity. Accordingly, it also determined how strongly emotion dominates in a sentence. So, we deployed the function from [8]. Firstly, an average is calculated for each emotion within a message. Then, the value of each emotion is calculated as shown in Eq. 2.

$$F_e = min\{\frac{A_e}{2 - log(3.5 \times W_e)}, 100\} \tag{2}$$

Where, F_e denotes the value of that emotion, A_e stands for the average emotion value for each emotion, W_e stands for the count of that emotion.

2.5 Emotion Classification

After calculation of endmost value between 0 and 100 for each class of emotion, the dominant emotions are identified. If the value of the emotion is more than 30, it returns that emotion or 0. If there are words only pertaining to a specific emotion in the message, then it is selected as final emotion. If the two or more emotion values are at a sufficient distance (the difference between the values is sufficient) then, the emotions having values greater than 30 are selected. Otherwise, the text is treated as neutral [12,13].

3 Experiment and Discussion

The proposed method was tested for various real-time inputs (self-made Facebook comments dataset), movie reviews dataset (manually labeled) as well as the Twitter dataset. As proposed by Jurek et al. [8], the accuracy of the model increased dramatically with normalization of the values. The same trends are obtained in the proposed method too. The following three sub-sections show in details the result of the method.

3.1 Qualitative Analysis

Table 1 depicts the accuracy of the model for the Twitter dataset (self-labeled), movie reviews dataset (the part which is not used in labeling) and the self-made dataset by using the Facebook comments. The model also classified texts having more than one emotion correctly. For e.g., *Stop it!* can be interpreted having disgust, anger in the voice of the speaker. Our model produces the positive results regarding the issue in lesser time as compared to machine learning models.

Table 1. Accuracy over different labeled datasets

Dataset	Accuracy (%)
Twitter dataset	69.1
Movie reviews dataset	65.7
Self made Facebook dataset	67.2

3.2 Quantitative Analysis

Table 2 shows the confusion matrix for the Twitter dataset (all the texts were labeled having one emotion only). It contains 100 anger, 100 disgust, 100 sadness, 100 surprise, 100 fear, 100 joy manually labeled samples.

Table 2. Confusion matrix for different emotions

Assigned emotion	Labelled emotion						
	Anger	Disgust	Sadness	Surprise	Fear	Joy	Neutral
Anger	69	12	09	15	05	18	10
Disgust	11	70	11	08	03	12	05
Sadness	03	04	72	07	02	00	03
Surprise	07	06	01	68	11	01	07
Fear	02	05	02	00	62	02	05
Joy	00	01	05	01	09	65	08
Neutral	08	02	00	01	08	02	62

3.3 Computational Complexity

The proposed model was implemented on the standard desktop computer with 2.7 GHz dual core CPU has 4 GB RAM. The time required by the method to compute the results is shown in Table 3.

Table 3. Time requirement for different datasets

Dataset	Time (Sec)
Twitter dataset	5.12
Movie reviews dataset	6.67
Self made Facebook dataset	4.89

4 Conclusion

Social spaces, the Internet is one of the most important personal view exchange portals in the current era. The social media, such as Reddit, Blogs, and Pinterest etc. and the Internet, supply a very suitable platform for communicating with each other and sharing views with everyone. Accordingly, the part of the Internet in crime prevention and investigation has promptly increased. The Internet is rapidly becoming a way of preventing chaos by providing information for warning systems in concern of public safety. So, we deployed a model to automatically predict the emotional state of the user. This could prevent many negative events which can cause a great loss of life and wealth. Public outrages and riots in a country having such a large population can be disruptive. Hence, a model like ours can become a great aid in time of need.

References

1. Maite, T., et al.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
2. Gräbner, D., Zanker, M., Fliedl, G., Fuchs, M.: Classification of customer reviews based on sentiment analysis. In: Fuchs, M., Ricci, F., Cantoni, L. (eds.) Information and Communication Technologies in Tourism 2012, pp. 460–470. Springer, Vienna (2012)
3. Krauss, J., et al.: Predicting movie success and academy awards through sentiment and social network analysis. In: European Conference on Information Systems (2008)
4. Ovesdotter, A., et al.: Emotions from text: machine learning for text-based emotion prediction. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2005)
5. Duc-Anh, P., et al.: Multiple emotions detection in conversation transcripts. PACLIC **30**, 85 (2016)
6. Garcia, A., et al.: A lexicon based sentiment analysis retrieval system for tourism domain. Expert Syst. Appl. Int. J. **39**, 9166–9180 (2012)
7. Efstratios, K., et al.: Ontology-based sentiment analysis of twitter posts. Expert Syst. Appl. **40**(10), 4065–4074 (2013)
8. Anna, J., et al.: Improved lexicon-based sentiment analysis for social media analytics. Secur. Inf. **4**(1), 9 (2015)
9. Thelwall, M., Buckley, K.: Topic-based sentiment analysis for the social web: the role of mood and issue-related words. J. Am. Soc. Inf. Sci. Technol. **64**(8), 1608–1617 (2013)

10. Casey, W., et al.: Using appraisal groups for sentiment analysis. In: 14th ACM International Conference on Information and Knowledge Management. ACM (2005)
11. Andrius, M., et al.: Combining lexicon and learning based approaches for concept-level sentiment analysis. In: ACM Sentiment Discovery and Opinion Mining (2012)
12. Erik, C., et al.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)
13. Awais, A.: Sentiment analysis of citations using sentence structure-based features. In: ACL 2011 Student Session (2011)