ISSN:0975-9646

Clustering of Emotions Expressed in Microblogging Websites

Sonali Gupta

Abstract- In today's world there is significant role of knowing the emotions or the sentiment of the people for various social aspects like politics, war, products launched, elections and many more. But the task in our research work is to how to know and cluster that data which have been collected from various micro blogging sites and perform clustering analysis on tweets and know the sentiment of the people regarding certain topic. In this research paper we have done work on collecting data from various micro blogging sites like tweeter. We collect the tweets of various people and then create the score matrices based on tweets. By using these score matrices we can analyze that what are the opinion of the people regarding certain topic by applying clustering on tweets. For example if we are collecting topic regarding certain government policy. Then firstly we will collect the tweets of people that have tweeted regarding that topic and then we will cluster the similar tweets regarding that topic. From this clustered data we will find the similar words and then score them and create the score matrices based by using rule base engine on number of word that have occurred again and again in the tweets . A word that have occurred in the score matrices repeatedly depicts that what the people think about that thing such as if people are happy with that thing then the score of that particular word is high in the matrices. Using this technique we will we be efficiently able to know the sentiment of the people on various topic. It is also use full in commercial environment where the companies can know the opinion of the people based data collected from the micro blogging sites and can make changes to the product according to the needs of the people. In today's worlds where everything is web based this technique will play important role in knowing the opinion of people on various social aspects.

Keywords: MBEWC, IWD, EDR, NWD.

1. INTRODUCTION

Emotions express humans feeling and experiences on some subject matters. They are typically recognized in text, speech, body gestures, and some visual information. Microblogging is a form of online communication by which users broadcast brief text updates, also known as tweets, to the public or a selected circle of contacts. A variegated mosaic of microblogging uses has emerged since the launch of Twitter in 2006: daily chatter, conversation, information sharing, and news commentary, among others. Regardless of their content and intended use, tweets often convey pertinent information about their author's mood status. As such, tweets can be regarded as temporally-authentic microscopic instantiations of public mood state. Emotion mining is crucial for many applications, including customer care (Gupta, Gilbert, and Fabbrizio 2010), sale prediction (Liu, et al. 2007), game animation (Bernhaupt et al. 2007), and robot simulation (Becker, Kopp, and Wachsmuth 2004). Capturing people's feelings, predicting their reactions to events, and generating suitable emotions are typical tasks in emotion mining.

In this article we have tried to develop a model to gauge the sentiments of the people by clustering the tweets of the various people and creating the score matrices based on these tweets. The score matrices will be created by using a rule based engine; similar words in the tweets will be identified and scored. A word that repeatedly appears in the score matrices depicts the sentiment of the people at that instance.

The rest of this paper is organized as: Section 2 discusses previous work related to emotion studies. Section 3 introduces the Plurk social network and describes the extraction of the dataset. Section 4 discusses how emotions from reader and writer perspectives are analyzed. Section 5 describes the SVM classifier, along with the feature set. Section 6 details the performance of the prediction tasks, and discusses and compares the usefulness of different types of features. The final section concludes the paper.

2. LITERATURE REVIEW

Previous studies (e.g., Yang, Lin, and Chen 2007; Yang, Lin, and Chen 2008) have used an emotion-tagged weblog corpus to investigate the ways in which people express their emotions, trying to detect writers' affective status with textual contents they have written. While these studies aimed to perform emotion analysis and detection from the writer's perspective, a few papers have studied reader emotion generation (Lin, Yang, and Chen 2007; Lin and Chen 2008) using an emotion-tagged news corpus, modeling how readers react to articles on news websites. To study how writer emotion affects readers' feelings, Yang, Lin and Chen (2009) used the Yahoo! Kimo Blog and Yahoo! Kimo News to produce a dataset annotated with both writer and reader emotions.

Mishne (2005) adopts mood taggings in LiveJournal articles to train a mood classifier on document-level with SVM. Mishne and Rijke (2006) use a blog corpus to identify the intensity of community mood during some given time intervals. Jung, Choi, and Myaeng (2007) also focus on the mood classification problem in LiveJournal. Yang, Lin, and Chen (2007a) use Yahoo! Kimo Blog as corpora to build emotion lexicons. A collocation model is proposed to learn emotion lexicons from weblog articles. Emotion classification at sentence level is experimented by using the mined lexicons to demonstrate their usefulness. Yang, Lin, and Chen (2008) further investigate the emotion classification of weblog corpora using SVM and conditional random field

(CRF) machine learning techniques. The emotion classifiers are trained at the sentence level and applied to the document level. Their experiments show that CRF classifiers outperform SVM classifiers.

Lin, Yang and Chen (2007) pioneer reader emotion analysis with an emotion-tagged Yahoo! Kimo news corpus. They classify documents into reader emotion categories with SVM and Naïve Bayes classifiers (Lin, Yang and Chen, 2008). Besides classification, Lin and Chen (2008) propose pairwise loss minimization (PLM) and emotional distribution regression (EDR) to rank reader emotions. They show that EDR is better at predicting the most popular emotion, but PLM produces ranked lists that have higher correlation with the correct lists. Yang, Lin, and Chen (2009) further introduce the application of emotion analysis from both the writer's and reader's perspectives. The relationships between writer and reader emotions are discussed in their works.

Besides long articles, some studies also deal with emotion detection of short messages from microblogs and news headlines. Strapparava and Mihalcea (2007) focus on the emotion classification of news headlines. Go, Huang, and Bhayani (2009) use distant supervision for sentiment classification of Twitter messages. In their study, SVM outperforms Naïve Bayes and Maximum Entropy, and has the accuracy of 82.2%. Sun et al. (2010) focus on the Plurk microblogging platform, using text content and the NTU Sentiment Dictionary to build their feature set. These studies all focus on writer's emotions rather than reader's emotions.

Bollen *et al* we perform a sentiment analysis of all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. For every day in the timeline, we extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of the Profile of Mood States (POMS), a well-established psychometric instrument. We compare our results to fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. We find that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. We speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators.

In order to identify the emotion expressed in corpus and to estimate the feelings conveyed by micro-blog data, in this paper, we categorize emotional words, build attitudinal words weight dictionary (WD) which consists of 1342 words, and construct self-defined negative words dictionary (NWD), degree words dictionary (DWD) and interjection words dictionary (IWD). We then process classified statistics on 2213 micro-blog items, finding that items whose first sentence express the main idea account for 23.8% of all the complex sentences, and items whose last sentence express the main idea 51.3%, respectively. We first calculate the weights of each clause in a micro-blog item, then take special treatment to the first and last sentence, finally we add up all the weights to get the emotional index (EI) of the item. Test results from the micro-blog emotion weight calculator (MBEWC) developed in C are cross-checked and reach an accuracy rate of 80.6%.

3. METHODOLOGIES

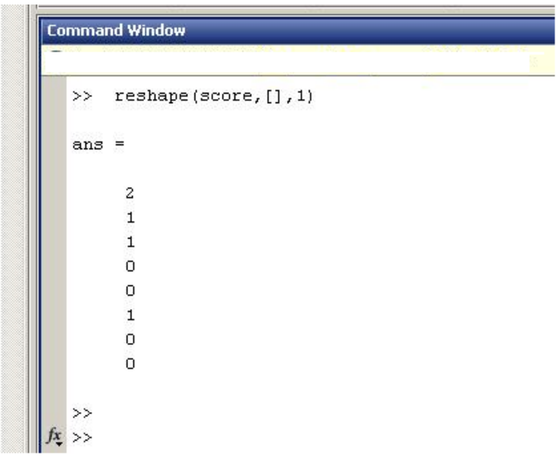
For the purpose of our study we extracted the tweets from the twitter between the time span to gauge the mood of the users. Twitter API (Application Process Interface) was used to extract the posts. Since twitter permits 70 requests per 60 minutes, the extracted data was continuously stored in MySQL. For the sake of convenience, in our study we confined our emotional lexicon database to five emotional states namely 'joy', 'sad', 'fear', 'anger' and 'surprise' and their most commonly used synonyms as well as emoticons.

The score matrix for these emotional expressions was created based on matching between the emotional lexical database and the tweet. Suppose there is a word in the emotional lexical database/dictionary and same word appears in the tweet, then the score matrix for that particular emotional state is updated by the number of times the word has appeared in the tweet. The score matrix will look something as shown in table 1.

Table 1

Lexical data base → Tweets ↓	Joy	Sad	Anger	Fear	Surprise	Explanation
1	1	0	0	0	1	Same tweet may have expressions of joy as well as surprise.
2	0	1	1	1		Same tweet may have expressions of sadness, anger as well as fear.
3	2	0	0	0	0	Expression of joy may appear twice in a particular tweet

A screen shot of the MATLAB window showing emotional state 'joy' for different tweets.



```
>>  reshape(score,[],1)

ans =

     2
     1
     1
     0
     0
     1
     0
     0

>>
>>
```

4. Result and Discussions

K –mean algorithm using MATLAB was run for clustering of people based upon the emotional state. The following color coding was used for different emotional states:

i. Sad = dark blue
ii. Anger = red
iii. Fear = pink
iv. Joy = green
v. Surprise = light blue

Emotional states at different time intervals are depicted in the snap shots 1 to 5.

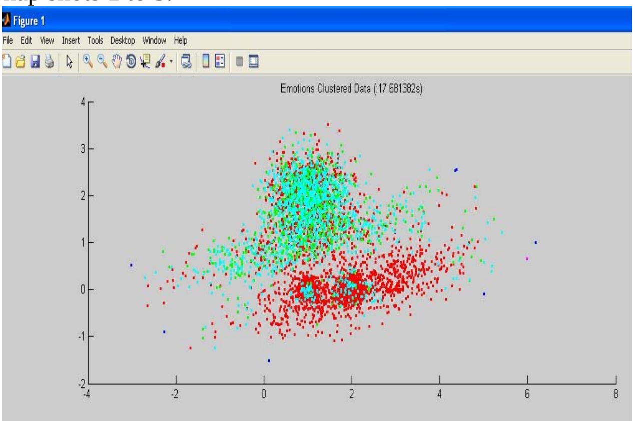
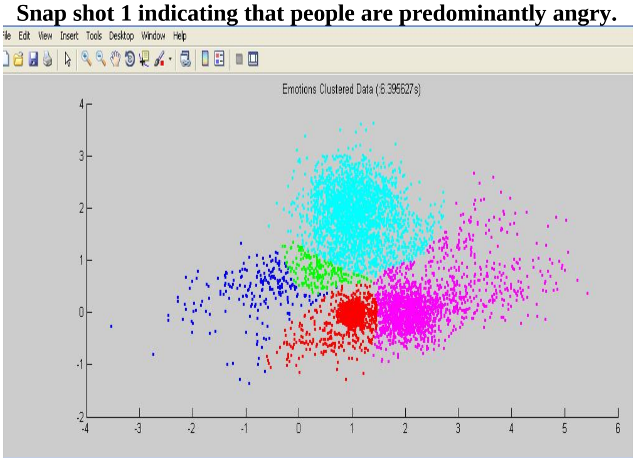


Figure 1
File Edit View Insert Tools Desktop Window Help
Emotions Clustered Data (:17.681382s)
4
3
2
1
0
-1
-2
-4
-2
0
2
4
6
8

Snap shot 1 indicating that people are predominantly angry.


File Edit View Insert Tools Desktop Window Help
Emotions Clustered Data (:6.395627s)
4
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
5
6

Snap shot 2 indicating that people are predominantly surprised.

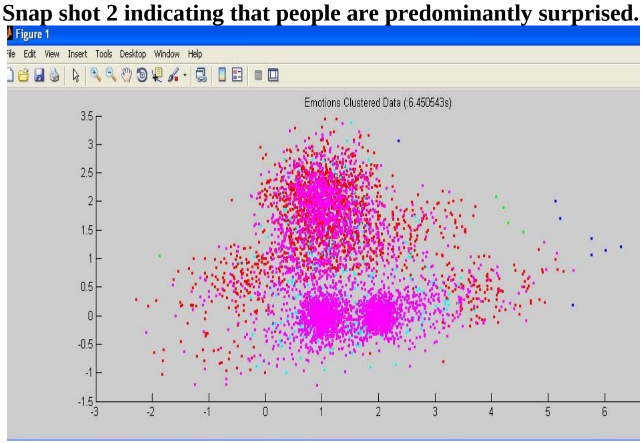


Figure 1
File Edit View Insert Tools Desktop Window Help
Emotions Clustered Data (:6.450543s)
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
4
5
6

Snap shot 3 indicating two predominant emotional states – fear and anger.

5. Conclusion

Human beings are different form all other beings because we can express our feelings and since we do not live in isolation, how we feel has an impact on the people around us. Emotions have a ripple effect in a society and therefore capable of generating extreme reactions. Social networking sites besides providing for an easy outlet to these emotions also facilitates seamless propagation of these sentiments across the globe. It is because of these reasons that social networking sites are increasingly being used by different organizations and individuals to further their interests. This is precisely the reason that it is of utmost importance to understand and analyze the feelings of a society. Uses of such analysis are endless – Government may use the information to gauge the mood of the nation, a marketer may use the information to gauge the response of a product etc.

The present study is an attempt to develop a clustering model for accumulating information pertaining to emotional states of the microbloggers and to draw a meaning full picture from the information so extracted. In our research work we collected the tweets of people at various instants. We created the score matrix for particular set of tweets based upon that how many times a particular word has appeared in the tweet and then matching with the lexical database we created the score for that particular tweet. We have created graphs for those particular data sets and showed which emotion people are having at that instant of time by using color code combination for various emotions.

6. Future scope

We can further work on various dimensions of human sentiments. In our research work we have taken emotions like sad, happy, fear, joy and surprise. But in future we can work upon various dimensions like to analyze the sentiments of wishes, comparison and preferences.

REFERENCES

Emotion mining research on micro-blog by Yang Shen Yang Shen, Shuchen Li Shuchen Li, Ling Zheng Ling Zheng, Xiaodong Ren Xiaodong Ren, Xiaolong Cheng Xiaolong Cheng. 009 1st IEEE Symposium on Web Society (2009) Publisher: Ieee, Pages: 71-75

Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena by Johan Bollen, Alberto Pepe, Huina Mao. Computer (2009) Volume: cs.CY, Issue: arXiv:0911.1583v0911[cs.CY], P ublisher: ACM Press,Pages: 17-21.