DESIGN, ANALYSIS AND IMPLEMENTATION OF GEOLOCATION BASED EMOTION DETECTION TECHNIQUE OVER TWITTER DATA

Prarthana Kumari[1], Sudheep Elayidom[2]

[1]*M.Tech Student, Department of Computer Science, CUSAT, Kerala, India*
[2]*Department of Computer Science, CUSAT, Kerala, India*

Abstract

It has been a topic of utmost importance to researchers that emotions of public has a direct impact on various social science problems such as politics, online business and so on. With emotion analysis, we can bring sensitivity to analytics and stay attuned to the feelings of customers during chat sessions, track social media reactions to a press releases, or gauge the public outlook on financial news. In order to meet these need we create a system for analyzing moods of tweets on any topic trending on twitter.com. We collected 1. 3 × 10^3 emotional tweets, and then these were annotated for emotion, geographic location. Bayes classifier has been used for analysis.

Keywords: *Emotion Analysis, Twitter, Geographic Distribution*

----------***----------

1. INTRODUCTION

In order to get response of public on various issues like politics, online business, disastor, social media, data has been the single strongest research focus for the past several years. Traditional way of collecting these data are usually through surveys, which sometimes does not provide the real state of mind of an individual and many of the times people are forced to fill feedback form so it does not reflect their actual opinion.

On the other hand, social networking sites like facebook, twitter etc. provides a standard platform for users to express their opinion, thoughts on an issue. Since these social networking sites does not force user to give their feedback, it naturally comes from them, so it is more real. Twitter is one platform which provide text for analysis.

2 SYSTEM DESIGN

2.1 Data Collection

We bulit a system that is basically used for analysis, in real time of the tweets expressed by user on twitter. Twitter provides users to post or read 140-character text. In recent times around 19% [6] of online users use twitter to express their opinion. Since an unregistered user can also access tweets, that makes twitter available for majority of people. This is not the case with other social network platforms which have different privacy settings.

Twitter public API can be used for streaming tweets from twitter.com to a local system. Tweets are collected as they occur, and they are filtered by a keyword. Keywords are basically used as filter in consumption of tweets. The standard twitter API provides provides a 1% random sample of all posted tweets.

2.2 Data Classification

The most daunting task in emotion analysis is to identify emotional tweets. For that we have used a large vocabulary of emotion terms that has been collected from various sources, like Affective Norms for English Words (ANEW) [7] and the Linguistic Inquiry and Word Count (LIWC) [8]. ANEW provides a set of approximately 1000 English words with emotional ratings attached with them. LIWC is text analysis software that calculates different categories of words used by people across the globe. Each emotional word has been classified into five emotion categories of joy, surprise, anger, sadness and fear.

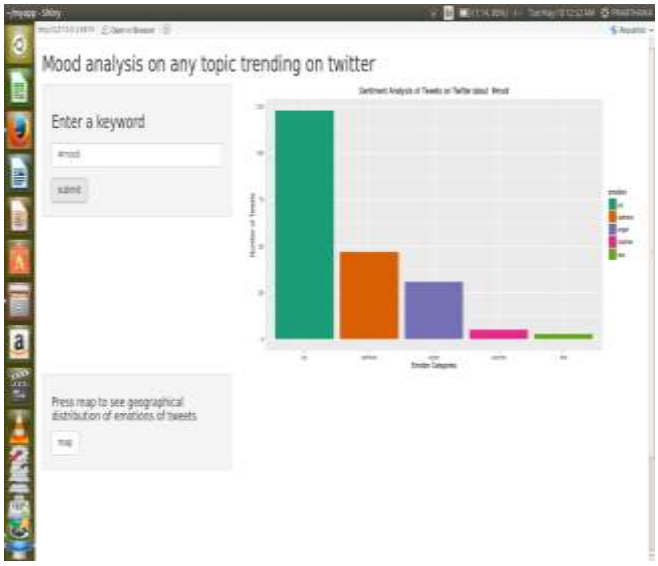

Chart -1: Barplot of emotions of tweets

The proposed system not only classify tweets according to different emotions categories but also classify tweets on locations and distribute it geographically. Location of tweets can be obtained from users profile but problem

occurs when user does not specify physical location in their profile. A user can geotag their tweets on twitter but it is not used often. In order to get physical location of a user, the time zone specified in a user profile is selected as proxy marker.

2.3 Data Visualization

Data visualization is the most important part of the system as it helps people to understand the significance of data by placing it in a visual effect. An interactive map has been used to display various emotions analyzed from text. Emotions obtained from tweets has been plotted on map as shown in figure 1.

Fig -1: Geographic distribution of emotions

3. MODELS AND RESULTS

Tweets were collected using twitter search rest API. Search keyword can be a english word that is used to filter tweets containing that keyword. These tweets are then analyzed for emotion analysis distributed geographically over the map.

We tagged a subset of 4,000 messages by hand, using five tags (sadness, joy, anger, fear, surprise).

The final classifier uses the following features:
Words (unigram tagging): The standard bag-of-words feature for Bayesian classification models.

Word pairs (bigram tagging) : Some wordpairs does not make sense if they are seprated. So, to get emotions from those words bigram tagging is used.

Special features: Twitter messages can contain certain features like hashtags (words preceded by # to indicate subject), replies (usernames preceded by @ to indicate a reply to that user), or links to external sites. We recognize when a message contains one of these features and separate out the relevant information: what site is being linked, who the reply is to, what the hashtag is. This information is used as a feature, while the original string is removed from the message and replaced with a placeholder. This maintains the meaningfulness of messages like Im at a party with @someone, since with @ is a feature with a positive tendency; if @someone were removed completely the useful bigram would be removed as well.

Stopword elimination: Usually, stopwords does not have emotions, so they are eliminated before analysis.

Message length and capitalization: We implemented this feature to look for any correlation between length, capitalization and emotional content. Both features were measured by the quartile into which the message fell. When implemented, the classifier did indicate possible usefulness of combining length and capitalization into one feature, especially among short messages, but the accuracy of the classifier suffered, as demonstrated in the Results section. This drop in accuracy could easily indicate over fitting for the training data which was not generalizable to the test data. Additionally, if no words in a message strongly indicate emotional content, allowing the classifier to use a weak length-emotion correlation to classify the message seems likely to result in inaccuracies.

3.1 Results

The following table demonstrates the performance of the classifier using different combinations of features:

The classifier consistently outperforms the baseline value for success 50.7% given the proportion of neutral messages in the data set. In presenting the three different feature combinations, the tradeoffs in adding features on top of unigram and bigram tagging become clear: adding features for a variety of special features (hashtags, etc.) causes a small increase in accuracy, but adding features for message length and capitalization decreases accuracy, as these features are biased by the training data. With or without special features, accuracy scores range from 60%-93%; this variability can be attributed to the breadth of words and phrases in the training data. The classifier must contend with the full complexity of the English language as well as the common idiosyncrasies of unedited content. In a random sample of messages to classify, there may easily be a large number of words the classifier hasn't encountered before, either because they are unusual, misspelled, or serendipitously not in the training set.

Table-1 : Result

Model	Accuracy (avg. 20 trials)	Standard Deviation
Bag-of-words only	73.25%	8.81
BOW, special features, stopwords	76.73%	8.97
BOW, special features, stopwords, length, capitalization	73.02%	9.03

4. CONCLUSION

We set out to create a classifier to indicate the mood of Twitter messages based on a corpus of messages taken from Twitter data feeds. Using a Nave Bayes classifier and bag-of-words feature extractor optimized for Twitter messages, relatively high accuracy was obtained over a representative subset of all Twitter messages. The classifier is thus fairly successful in the goal of tagging Twitter accurately in conditions as realistic as possible.

REFERENCES

[1]. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnella, and J. N. Rosenquist. (2010). Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter .Available: http://www.ccs.neu.edu/home/amislove/twittermood

[2]. D. Milne, C. Paris, H. Christensen, P. Batterham, and B. O'Dea. (2014). The We Feel emotion explorer. Available: http://wefeel.csiro.au

[3]. Holzman L. and Pottenger W. 2003. Classification of emotions in internet chat: An application of machine learning using speech phonemes, Technical Report LU-CSE-03-002, Lehigh University.

[4]. Kim, Elsa and Sam Gilbert. "Detecting Sadness in 140 Characters: Sentiment Analysis and Mourning Michael Jackson on Twitter" Web Ecology Project, August 2009. http://www.webecologyproject.org/2009/08/detecting sadness-in-140-characters/

[5]. Yessenov, Kuat, and Sasa Misailovic. "Sentiment Analysis of Movie Review Comments" Massachusetts Institute of Technology, Spring 2009. http://people.csail.mit.edu/kuat/courses/6.863/report.pdf

[6]. Pew Internet Research. (2014). Social Networking Fact Sheet. Available: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

[7]. M. M. Bradley and P. J. Lang, "Affective norms for English words(ANEW): Instruction manual and affective ratings," The Center for Research in Psychophysiology, University of Florida Technical Report C-1, 1999.

[8]. J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," Mahway: Lawrence Erlbaum Associates, vol. 71, p. 2001, 2001.