ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Emotion Analysis of Social Media Data

[1] Sonia Xylina Mashal, [2] Kavita Asnani
[1][2] Department of Computer Engineering, Goa College of Engineering, Goa University, Goa-India

***Abstract* - Emotion analysis in text is a trending field of research that is closely related to Sentiment analysis. With the increase of Internet connectivity, all the human interactions has moved to the social media platforms hence a lot of data can be tapped for research purposes. Emotion analysis is carried to study the writer's state of mind. In this paper social media data is classified into five emotion categories (happy, sad, anger, fear and surprise) using supervised machine learning techniques(Naïve Bayes and Support Vector Machine).**

***Index Terms:*— emotion analysis, Naïve Bayes, SVM, Twitter.**

I. INTRODUCTION

Emotion is the behavioral tendency of the brain after encountering a sudden situation. Emotion impacts our daily lives and is also important in all aspects of our daily living. It influences our social relations, memories and even day-to-day decision making. Even though, the study of emotion began way back in the 1980s, computation of emotions from text is a recent field. In linguistics, the automatic computation and detection of emotions from text is becoming one of the highly researched topics.

Emotion classification is done on various documents, news snippets, short messages and tweets. Social media data such as tweets is been largely used for emotion analysis but the drawbacks of these analysis is usage of a small-sized dataset, smaller sizes of datasets provide an insufficient coverage of emotion categories. Nowadays there is ample amount of data available on the internet which can be used but the challenges faced are :

(i) Manual Data Annotation: Most of the research work have used human experts for manually annotating the datasets, this can lead to biased or inaccurate annotations.
(ii) Length: Social media data such as tweets have a length of 140 words. Emotion analysis can be hampered in these situations because the author tries to express all his emotions within that short length.
(iii) Lexicons: The increasing content of slang and short forms have rendered the available lexicons resources less useful for the data analysis task.

In this paper the focus is on supervised emotion classification using machine learning techniques like Naïve Bayes and Support Vector machines(SVM) using Twitter data. The Twitter data used was automatically collected by using emotion hashtags for five emotion categories namely happy, sad, angry, fear and surprise. After testing the results of the classifiers are compared.

II. LITERATURE SURVEY

The following subsections provide a description of the main topics of concern related to this paper work.

A. Emotion

Emotion, in layman language, is any relative conscious experience which can be characterized by high and instant mental activity. Emotion is a strong feeling which derives from one‟s method of forming judgment about circumstances, event or relationships with others .Emotions are complex. Many dimensional models of emotion have been studied , researched and developed, but only a few amongst them remain as the dominating models. Some of these dimensional models can be used for emotion classification in text, which can be document, sentence, short message or tweets. The models are used for data collection in emotion classification. The Circumplex model does emotion tagging by placing 28 words in eight different categories over a two-dimensional circular space[1]. Emotion tagging is done on the varying levels of monoamine neurotransmitters and the eight basic emotions as labeled by Tomkins[2]. Ekman‟s model provides six discrete emotion categories namely happiness, sadness, anger, disgust fear and surprise[3].

B. Classification of emotions

Liza Wikarsa and Sherly Novianti Thahir developed a text mining application to detect emotions of Twitter users that are classified into six emotions, namely happiness, sadness, anger, disgust, fear and surprise. Preprocessing, processing and validation are the three main phases of text mining that were used in this application.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Tasks such as case folding, cleansing, stop-word removal, emoticons conversion, negation conversion, and tokenization of the training data as well as the test data were conducted in the preprocessing phase. Weighting and classification using Naïve Bayes algorithm was done in the processing phase and a ten-fold cross validation was carried out in the validation phase for measuring the accuracy generated by the application. This model obtained 83% accuracy for 105 tweets. The authors focused on social media data rather than text documents and also emphasized on the need of proper preprocessing steps to obtain useful results.[4]

Li Yu, Zhifan Yang, Peng Nie, Xue Zhao and Ying Zhang tackles the task of multi-source emotion tagging for online news. A new classification model is proposed with two layer logistic regression, this approach takes output from a basic classifier and combines them in a new classifier, hence providing a more accurate prediction. This research was considered as an initial step towards multi-source tagging, the results can further be improved by using emotion dictionary and feature selection.[5]

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan and Amit P. Sheth proposed the idea of overcoming a bottleneck of lack of coverage of emotional situation in datasets used for emotion identification tasks. They experimented with various feature combinations and the effect of varying the size of training data. The combinations of features addressed are n-grams(unigrams,bigrams, and trigrams) along with emoticons and punctuations, POS tagging, emotion lexicon(LIWC,WordNet Affect, MPQA), and n-gram positions. [6]

III. IMPLEMENTATION DETAILS

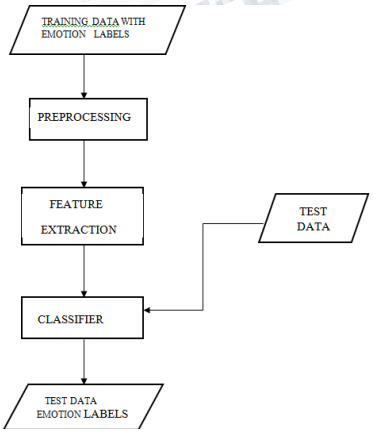

Figure: Architecture of Emotion Classification Process

A. Dataset Collection

The dataset was obtained from [6]. For creating the dataset, first seven different emotion words for seven emotion categories were collected from existing psychology theory, then Twitter Streaming API was used for collecting tweets which had at least one of the emotion word in the form a hashtags.[6]The collected dataset was then divided into testing and training sets. The training sets were used to train the classifiers so that the test data is correctly labeled.

B. Preprocessing

Preprocessing of the collected data is of utmost importance because the tweets are only 140 words long in length and hence there is a presence of slang, URL"s, user-mentions which do not contribute in any manner to the classification process, infact the presence of such elements can mislead the classifier.

The preprocessing steps include the following[7]:

- Lower casing all the words.
- Replacing user mention with @user.
- Replacing letters and punctuations that are repeated more than twice with two same letters(Eg.happpy□ happy).
- Removing hashtags.

C. Feature Extraction

After all the data is preprocessed, the stopwords are removed from the tweets because they usually constitute large amount of words which do not provide useful information. After stopword removal the feature extraction process is done. In this paper bag of words is used for training both the classifiers that are used.

D. Classifier

The two classifiers that are used for classification purpose in this paper are Naïve Bayes and Support Vector Machine(SVM) both these classifiers are found to be very efficient in handling large amount of text data.

IV. EXPERIMENTAL RESULT

The dataset used for training consists of 622 training instances with an approximate 120 instances each belonging to the categories happy, sad, fear, anger and surprise. The test dataset consists of 50 instances. TABLE 1 gives the accuracy, F1 score, recall and precision of Naïve Bayes and SVM classifiers used for predicting class labels of the test data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Table1: Results for SVM and Naïve Bayes for accuracy, F1score, Recall and Precision.

	Accuracy	F1 score	Recall	Precision
SVM	0.94	0.9395	0.94	0.9484
Naïve Bayes	0.94	0.9407	0.94	0.9484

From the results it can be seen that both the classifiers provide the same accuracy. This result implies that bag of words is a good method of feature selection because it covers most of the words that may occur in the test data hence better classification is provided.

CONCLUSION

Emotion analysis on social media data(tweets) was conducted using Naïve Bayes and SVM classifier , both the classifiers are found to be very effective for text classification and can also be used for large datasets. The classifiers can be can be checked with other methods of feature selection and an increased size of dataset . This classification and analysis process can further be extended to detect specific topics that are causing certain emotion outbreak.

REFERENCES

[1] James A. Russell, "Emotion in Human Consciousness Is Built in Core Affect", Journal of Consciousness Studies,12,No.8-10,pp 26-42, 2005.

[2] James A. Russell, A Circumplex Model of Affect,Journal of Personality and Social Psychology , Vol. 39, No. 6, 1161-1178,1980.

[3] Nancy A. Remington, Leandre R. Fabrigar, Penny S. Visser, "Reexamining the Circumplex Model of Affect", Journal of Personality and Social Psychology , Vol. 79, No, 2, 286-300, 2000.

[4] Liza Wikarsa, Sherly Novianti Thahir, "A text mining application of Emotion Classifications of Twitters Users using Nave Bayes Method", IEEE 2015.

[5] Li Yu, Zhifan Yang, Peng Nie, Xue Zhao, Ying Zhang, "Multi-Source Emotion Tagging for Online News", 12th Web Information System and Application Conference 2015.

[6] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth," Harnessing Twitter „Big Data" for Automatic Emotion Identification", 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.

[7] Sanket Sahu, suraj Kumar Rout, Debasmit Mohanty,"Twitter Sentiment Analysis:A more enhanced way of classification and scoring", 2015 IEEE International Symposium on Nanoelectronic and Information Systems.