IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 17-20
www.iosrjournals.org

Emotion Analysis of Social Media Data using Machine Learning Techniques

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

Abstract: *Analysis of emotion in text is a very young field in the area of computational linguistics. Analysis of emotions involves evaluation and classification of text into emotion classes based on certain levels as defined by emotion dimensional models which are described in the theory of psychology. Emotion analysis and classification is performed to identify the expressions of emotion in a given text. In this paper a huge dataset of social media data (tweets) is classified into five emotion categories (happy, sad, anger, fear and surprise) using machine learning techniques(Naïve Bayes and Support Vector Machine).*
Keywords: *classification, emotion, emotion analysis, Naïve Bayes, SVM.*

I. Introduction

The rise of social media sites gave birth to the method of communication technique we now know as instant messaging where there is a lot of communication taking place using emoticons and short text forms. Text based representations of feelings are used in texts. Most of the human conversation have moved from been face-to-face to online, hence there is a lot of social media data available which can be used for analysis of emotions to identify an user's state of mind when he/she had written that post. Emotion analysis is important because it affects our daily decision making capabilities.

In linguistics, the automatic computation and detection of emotions from text is becoming one of the highly researched topics, even in terms of many applications, automatic detection of emotion would be considered as a plus point. Even though, the study of emotion began way back in the 1980s, computation of emotions is a recent field. With the availability of ample of data from social media, the task of classification faces issues like, having a standard dataset for lexical analysis of the available data. Annotating the data has issues of its own because if a human expert is used he can be biased and collecting annotated data from online resources is difficult. Emotion analysis on Twitter data pose problem of length since the author tries to express all his emotions within that short length and also the increasing content of slang increases the preprocessing required. Emotion classification systems use supervised, unsupervised or hybrid learning techniques to classify emotions expressed in different forms of text into emotion categories which are defined according to various psychological studies[1].

This paper focuses on the task of emotion analysis and classification using machine learning techniques (Naïve Bayes and SVM). The data used is from Twitter. The data 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. The datasets used for classification tasks are huge, hence providing a better coverage of the vocabulary that may occur in the Twitter or any other social media data. The work also highlights the effect of varying the size of the training data.

II. Literature Survey

The literature survey discusses topics relating to emotion classification and description about emotions.

a. Emotion

Emotion is a strong feeling which derives from one's method of forming judgment about circumstances, event or relationships with others .Emotions are complex. The study of emotion in psychology started in the 1970s and till date many different theories have being proposed, studied and scrutinized. 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. For example the Circumplex model is a 2D model which was developed based on 8 emotions categories and 28 emotion words were placed in these 8 categories[2] and the Positive Activation and Negative Activation model(PANA) is yet another model which classifies emotion words on the scale from high positive activation to low positive activation[3], these two models can be used for data collection for emotion analysis task. But models like Lovheims cube[4] cannot be used for data collection or analysis because emotion

in that is defined on levels of monoamine transmitters from the brain and we cannot detect levels of these transmitters from text. Most of the research work carried out in the field of emotion classification mainly uses Ekman's model for classification of emotions. Paul Ekman carried out a study to identify emotion based on different facial expressions between different cultural people, this study was done to find out whether people from different cultures have same facial expressions to represent certain set of basic emotions. Ekman's model provides six discrete emotion categories namely happiness, sadness, anger, disgust fear and surprise[5].

b. Emotion Classification

Emotion classification is a task wherein the aim is to detect and recognize types of feelings through the expression of texts, such as *anger, disgust, fear, happiness, sadness,* and *surprise*. Emotion detection may have different applications such as finding out how happy the citizens are, providing better services to an individual and suggestions in helping an individual who is in anxiety which can be identified through the outgoing texts and emails. Many authors have done noteworthy work in this field.

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. 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[6].

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 classifiers 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[7].

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth proposed the idea of overcoming a bottleneck of lack of coverage of emotional situations in datasets used for emotion identification tasks. They experimented with various feature combinations and also with the effect of the size of the training data. The combinations of features addressed are n-gram *(unigrams, bigrams and trigrams)* along with emoticons and punctuations, POS tagging, emotion lexicons *(LIWC, WordNet Affect, MPQA)* and n-gram positions[8].

K Dhanasekaran and R Rajeswari have presented a text feature classification approach for effective information extraction in [9]. In that paper, discriminative sequence analysis is used to analyze various features.

It is seen that Emotion classification/analysis started from just classifying emotions into categories to experimenting with different types of features, data, size of data and use of emotion lexicons.

III. Implementation Details

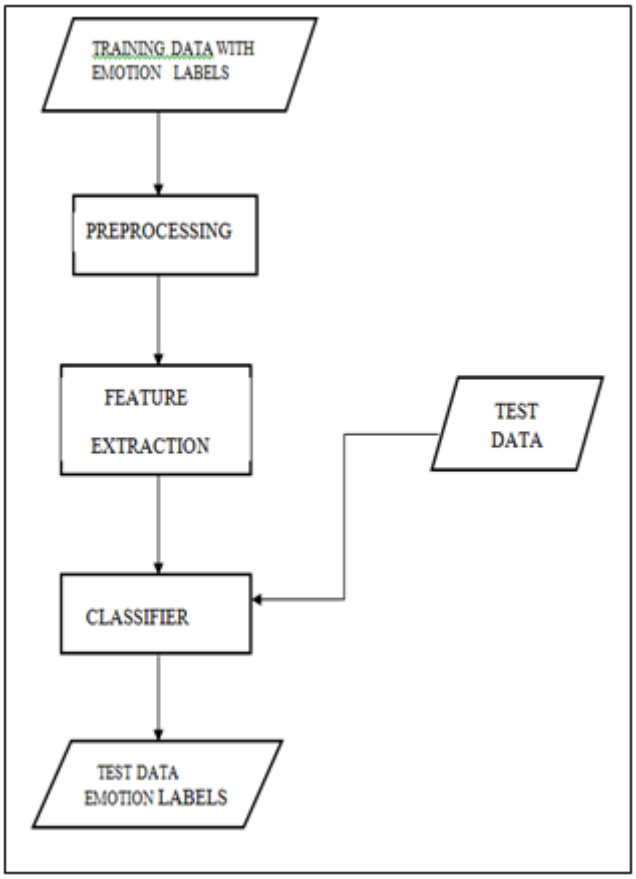

Figure 1: Architecture of Emotion Classification Process

3.1 Dataset Collection

The dataset was obtained from [10]. 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 hashtag[8]. 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.

3.2 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, in fact the presence of such elements can mislead the classifier.

The preprocessing steps include the following[11]:

- 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.

3.3 Feature Extraction

After data preprocessing, the stopwords are removed from the tweets by using a stopword file, this is done because they usually constitute large components of sentences and they do not provide useful information. After stopword removal the feature extraction process is done. In this paper bag of words is considered for training both the classifiers that are used.

3.4 Classifier

Naïve Bayes and Support Vector Machine(SVM) are used for classification purpose because SVM is found to be very useful in matters of handling sparse data and Naïve Bayes works effectively for text classification.

IV. Experimental Results

The results were taken by experimenting with the size of the dataset used for training. The training data with 600 records approximately, with equal records belonging to the five classes of emotion that is *happy, sad, angry, fear and surprise* is used for taking results given in the TABLE 1 and for TABLE 2 a training set of 1000 records is considered.

	Accuracy	**F1 score**	**Recall**	**Precision**
SVM	0.94	0.9395	0.94	0.9484
Naïve ayes	0.94	0.9407	0.94	0.9484

TABLE 1: Results for SVM and Naïve Bayes for accuracy, F1 score, Recall and Precision with an input of 600 training.

	Accuracy	**F1 score**	**Recall**	**Precision**
SVM	0.89	0.8967	0.8932	0.91
Naïve Bayes	0.86	0.867	0.861	0.92

TABLE 2: Results for SVM and Naïve Bayes for accuracy, F1 score, Recall and Precision with an input of 1000 training.

V. 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 results show that for larger datasets the results do not degrade. The classifiers 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]. Yoad Lewenberg, Yoram Bachrach, Svitlana Volkova," Using Emotions to Predict User Interest Areas in Online Social Networks", 2015 IEEE

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

[3]. DavidWatson; Auke Tellegan, "Towards a consensual structure of Mood", Psychological Bulletin, Vol. 98, No. 2. 219-235,1985.

[4]. Hugo Lövheim, A new three-dimensional model for emotions and monoamine neurotransmitters, Medical Hypotheses 78 (2012) 341–348. Processing and its Applications.Research in Computing Science 46, 2010, pp. 131-142.

[5]. Paul Ekman, "Universals and Cultural Differences in Facial Expressions of Emotion", Nebraska Symposium on Motivation, Vol 19,1971.

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

[7]. 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.

[8]. 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.

[9]. K Dhanasekaran and R Rajeswari, "Text feature classification approach for effective information extraction via discriminative sequence analysis", International Journal of Applied Engineering Research, Vol. 10 (1), pp.2067-2079, 2015.

[10]. http://knoesis.org/projects/emotion.

[11]. 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.