Mood Classification Of Social Media Text

L.Divya Vani
Andhra University
divyavani91@gmail.com

D.Suneetha
Gitam University
suneethadwarapu@yahoo.com

Abstract

The research work investigates the problem of feature selection for text categorization in the domain of social media data analysis. Specifically, large repositories of textual content expressed by the users in social media in the form of tweets/blogs is analysed based on ‘Sentiment’ lexicon. Naïve Bayes classifiers are used to determine the emotion based on mood words and polarity based on normal words of a given tweet and the results are combined to classify the tweet into one of the mood patterns. The data set used to implement the mood classification is ‘Twitter Data’ which is a live streaming content.

Keywords:
Mood Classification   Mood Analysis   Mood Patterns

1. Introduction

Social media now-a-days is extensively used for two way communication; users can not only read the content but can also respond to it. It has evolved from a simple broadcast media for publishing news articles into a much sophisticated media that facilitates commenting by the users to share their views. The adoption of user contribution has led to the emergence of social media like YouTube, Facebook, Twitter, etc through which communities of various types transact via online communication media. Social media constitute sentiment-laden corpora. Therefore user generated content is usually sentiment-associated, facilitating a company/organization to gain insight into customer opinions about their products and those of its competitors.

Thus to identify and monitor opinion sources on web is a growing research aspect, called opinion mining[4]. Opinion mining mainly focuses on Subjectivity Detection, Sentiment Classification, Sentiment Analysis and Opinion Summarization. In this context, we focus on ‘Mood’ which is a popular form of sentiment. Mood is a strong form of sentiment expression which conveys the state of mind like being happy, sad or angry. Social media content has a lot of textual data and is rich in sentiment, hence we discuss several fundamental issues related to mood sensing from these texts and its applications.

There are many potential applications for Text-based mood classification which is a sub-problem of opinion and sentiment mining. Mood classification can be used to filter search results to gain detailed insight into patterns of how bloggers behave and relate to one another in a blogosphere [1]. Text based mood analysis has additional challenges than standard text categorization and clustering. A feature set that works without requiring supervised feature training is desired to classify mood in the blogosphere which is vast and is growing day-by-day.

Mood formulation depends on the specific content of the users comments/blogs including the order of linguistic components and usage of mood words. We use predefined package named ‘Sentiment’ which intersects Psychology and Linguistics to identify mood words specifically for English. Naïve Bayes classifier is used to classify the textual data and provide the appropriate moods by considering the training data of Carlo Strapparava and Alessandro Valitutti, “WordNet-Affect: an affective extension of WordNet”.

The remaining part of the paper is organized as follows. Related work of the paper is discussed in the second section. The methodology of our paper is explained in detail in section 3. Followed by section 4 consisting experimentation and results. Our entire work has been concluded in section 5.

2. Related Work

2.1 Feature selection for mood classification

For text categorization, many feature selection methods in machine learning are studied. Term statistics is an alternative approach to term class interaction in selecting features. Term Frequency (TF) and Document Frequency (DF) are commonly used for feature reduction in data mining. The joint term
frequency and inverse document frequency (TF.IDF) is also used in text mining.

2.2 Mood Analysis

Mood analysis is a subset of sentiment analysis and opinion mining where related information is identified, extracted and utilized in real-world applications. Sentiment bearing representation in user feedback facilitates public opinion monitoring for governments and business organizations. A method was proposed by Feng et al [1] to group blogs into sentiment clusters, using Chinese sentiment lexicon for sentiment-bearing representation.

2.3 Mood Patterns

In addition to classification and analysis phase,[3] clustering of mood into patterns is also an important task as it provides a proper human emotion structure. Thus we discover intrinsic patterns in mood structure using unsupervised learning approaches. We made use of predefined source of mood words like ‘WORDNET’ to classify and extract the mood patterns from the textual data.

2.4 Naïve Bayes Classifier

It is a simple probabilistic classifier with naïve (strong) assumptions between the features. Naïve Bayes is a simple technique for constructing classifiers. This also plays a major role in classifying our text into their specified class labels.

3. Methodology

The entire work has been formulated as a framework containing three phases. The initial phase deals with the problem of mood classification after conducting a text based feature selection. For identifying emotion and polarity reflected by a given tweet-phase 2 analyses the classified output of previous phase to reduce the uncertainty regarding mood sensing. The final phase extracts mood patterns for different levels of moods from a very large data set. The whole work is represented schematically as given below in figure 1 and is discussed in detailed as follows.

3.1 Mood Classification

This paper deals with the problem of mood estimation from text using supervised classification technique. Social media data like Twitter data set is used for the task of mood classification. It is a Livejournal data set. Live journal allows people to express their mood in terms of tweets, therefore providing an excellent source of ground truth data for sentiment analysis.

To extract tweets from twitter live streaming data, we need to create a twitter app in our account and get authentication to retrieve the data. The textual data extracted is firstly preprocessed to remove all unwanted stop words from the content and the filtered text, i.e. tweets are taken as input for our framework. Each tweet is represented as a list of unigrams. A unigram represents a mood word or normal word. These tweets are further processed to extract the moods.

We denote the corpus of all blog posts by B and the set of all mood categories by M= {sad, happy, anger…….}. Each blog post d ∈ B is also labeled with a mood category C_d ∈ M in standard feature selection and the objective is to extract from d a feature vector X_d. The provided moods range diversely in the emotion spectrum.

The problem of generic text document classification has been investigated by the text mining researchers. The simple Naïve Bayes classifier is extensively used for text classification based on occurrence of terms/keywords. Due to high dimensionality of text corpus term based feature selection methods are used to improve classification performance greatly. The feature set considered in this task consists of unigrams representing mood words and normal words.

Specifically,[5] the occurrence of a term is measured in terms of Term Frequency TF(v,d) representing the number of times the term appears in a tweet (d). Document Frequency DF(v) is the number of blogposts/tweets containing the term v. In Text Mining TF.IDF(v,d) improves discriminative power where TF.IDF(v,d)=TF(v,d)*IDF(v) where IDF(v)=|D|/DF(v) is the inverse document frequency. A term will be selected if it has high DF(v) value, or high average values of TF.IDF(v,d) across all documents d.

Sentiment Lexicon, which is a set of predefined mood words, is used in inferring the general text representation. It is a set of sentiment conveying words which act as a standard for studies in cognition and emotion. It contains two sub sources emotion and subjectivity. At first all moods are categorized into six classes: Joy, Surprise, Anger, Sadness, Fear and Disgust using emotion in the lexicon with Naïve Bayes classifier [6]. All the words in our tweets may not belong to the set of words in the lexicon resulting
unknown values. In order to decrease the unknown values and increase the mood estimation results, we calculate the polarity of the tweets using subjectivity in the lexicon. Polarity is categorized into three classes: positive, negative and neutral.

3.2 Mood Analysis

We analyze the results of emotion and polarity of the tweets. Initially the output is stored in a dataframe where the tweet, its emotion and polarity are mentioned in columns. Here we perform proper analysis of the emotion and polarity [2] like, If Emotion= unknown and Polarity = Positive then it is much likely to be classified as Happy and If Emotion= unknown and Polarity= Negative then it falls under Sad or Anger, the remaining combinations with unknown fall under neutral category. Now, the number of unknowns which is high previously has drastically reduced by analyzing proper mood category which the tweet is most likely to fall in. In order to compare our results more efficiently, we restrict our classification in this experiment to three popular moods {sad, happy, anger}.

3.3 Mood Patterns

Mood patterns of emotion and polarity are obtained using unsupervised approaches. A cloud visualization of mood words analysed in this data set is also plotted. After proper classification and analysis we combine the emotion and polarity to draw the mood graph of the tweets with three popular moods.

4. Experimentation And Results

The data set we considered is taken from the Social media content (i.e., twitter data).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Twitter data</th>
<th>Live streaming content</th>
<th>25000-50000 tweets/blogs</th>
</tr>
</thead>
</table>

The classification of emotion results in 6 different moods for the considered data. The results are plotted in the below graph.

![Figure 4: Mood graph representing 6 clusters](image-url)

As we have observed in figure 4 the unknowns are very high. So as to decrease the unknowns and increase the classification performance we calculate the polarity of the tweets and is plotted as in the following graph.

![Figure 3: Process flow diagram of the system](image-url)

![Figure 2: Representing Emotion and Polarity](image-url)

![Figure 1: Process flow diagram of the system](image-url)
Finally, the emotion and polarity in figure 4 and 5 are combined together to analyse the tweets into their respective mood category. We restrict our mood categories to three popular moods for better analysis.

Word cloud of the mood words which are present in the tweets to classify and analyse them into a particular mood category are also plotted as shown in figure 7.

5. Conclusion

Large scale mood analysis of social media text is proposed. The whole work is formulated as a framework which deals with a series of steps. Firstly, the problem of mood classification for general textual data has been investigated. Followed by, mood analysis which reduces the uncertainty realized in the previous phase. Finally, we extract mood patterns for different levels of mood from a very large data set. Similar approach may be applied to improve the classification accuracy of opinion mining and user feedback analysis in various organizations.

6. References