Applying Sentiment Analysis to Twitter Movie Reviews

Pallavi Sharma

Department of Computer Science

North Dakota State University

pallavi.sharma@ndus.edu

Abstract

In this project an attempt has been made to classify the polarity of movie reviews posted on Twitter for the movie "Dunkirk" as positive and negative. For this purpose, reviews are first fetched from Twitter and cleaned to be useful for further analysis. Bayes' classification is used to classify polarity of the reviews. Additionally, the reviews are also classified according to seven basic emotions. Later word cloud is generated to determine the most frequent terms in different categories of emotions. Accuracy of 80.07% and 84.53% was achieved for classifying positive and negative sentiments respectively for the reviews.

Keywords: Sentiment analysis, Bayes' classification, word cloud

1. Introduction

Twitter is a popular social media platform for micro blogging. Users from various walks of life and from around the world, broadcast short burst messages of 140 characters to let the world know of their whereabouts, activities, feelings, news etc [1]. Many a times, it is used to post reviews related to restaurants, movies, customer services etc. Since it is a public platform, it is quite influential and thus, can create greater hold for customer market. A common public opinion can influence the market for a product and to an extent, can decide the future of the product. Therefore, it is desired that a product gets positive reception which helps to build its market. If the product starts to get negative reception, it is advantageous if one gets to know about it in time and can then turn around the marketing campaign to address issues or make changes. To be able to do so, it is required that enough coverage is done for these reviews and they are analyzed properly. This is done through Sentiment Analysis. Sentiment Analysis is the process of determining the emotional tone of sentences, to gain understanding of emotions, attitude and opinions [2]. The Obama administration used sentiment analysis to know public opinion to policy announcements and campaign messages ahead of 2012 presidential election [2]. One of the major industries which make use of online reviews is the movie industry. Users posting movie reviews online affect the movie ratings, its public reception majorly. Sentiment Analysis when applied to movie reviews can help gauge the public opinion regarding which set of viewers liked the movie, which set disliked it, what was the major emotion related to the review etc. This can later help to predict the major trends related to particular genre of movies, directors, actors. In this project, sentiments are analyzed related to the movie "Dunkirk" which is based during times of World War II. This movie received mixed reviews and was highly criticized for its portrayal of certain historical facts.

2. Methodology

The aim of the current project is classify the sentiments of the movie reviews pulled from Twitter for the movie "Dunkirk" as positive and negative. This objective is accomplished using methodology described in detail in following section. The methodology adopted in this project involves three steps: 1) Data Acquisition from Twitter, 2) Data Cleaning and 3) Classification based on Bayes' Classifier

2.1 Data Description

Prior to application of classification method, there are two steps that need to be completed: 1) Data Acquisition from reliable source and 2) Data Cleaning using appropriate tools.

2.1.1 Data Acquisition

Data Acquisition is the process of gathering data from the relevant sources to find answers for the problem statement. Data is generally acquired through primary sources, where data is collected through surveys, interviews etc, and secondary sources, where data is readily available through web sources, books, public libraries etc [3]. In this project, data (movie reviews) was pulled from secondary source, Twitter [4] using Twitter API. To collect data, a Twitter API was needed. For this purpose, a Twitter profile was created and an application was registered [5]. To authenticate the Twitter account to use this application, a number of authentication ways are available. For this project, authentication using OAuth method was used where access tokens were required. These access tokens were Consumer Key, Consumer Secret Key, Access Token and Access Token Secret Key. These can be easily generated from the application page on Twitter profile [5]. These tokens, after acquiring, were used to establish authentication setup. This authentication setup and further procedures in the project were implemented in RStudio® [6]. Once direct authentication was set up, tweets related to movie "Dunkirk" were retrieved from Twitter in English language using certain hash tags. The hash tags were "#Dunkirk" and "#ChristopherNolan".

2.1.2 Data Cleaning

Data Cleaning is the process of correcting raw data by first analyzing the quality of data and then correcting it to generate meaningful information to be of any further use [7]. In this project, after acquiring tweets, a number of steps were followed to generate useful data for classification. First, the text related to reviews in the tweet was separated from all the other attributes present in a tweet. Then, retweet entries ("RT") and tags ("@people") were removed. This was followed by removal of punctuations, numbers, links, newlines, unnecessary tabs and whitespaces. After this step, the data was still in non-uniform format as some words were in uppercase and some in lowercase. All the data was converted to lowercase to maintain uniformity. Since the data was pulled out of Twitter directly, it did not have prior class label specifying which review is positive

or negative. All the reviews were considered manually and a class label specifying its polarity was added to the data. This prior knowledge of class label was necessary for classification in the model used. The explanation for the model is presented in next section. This data was now ready to be classified. As specified, data cleaning was implemented using RStudio®.

2.2 Classification

Classification refers to the task of predicting a class label for a given unlabeled point based on labeled points whose class labels are already known [8]. It is considered an instance of supervised learning [9]. In the model building phase, a classification algorithm finds relationship between different values of features associated with class labels. These features can be quantitative or qualitative in nature. Different classification algorithms choose different approaches to find relationships between features. In the model testing phase, predicted values are compared to known values in a set of unknown data and report accuracy of the model [10]. Naïve-Bayes' Classification is one of the capable data classification method that works well with both quantitative and qualitative data. Since, sentiment analysis revolves around working with text and categorical features, Naïve-Bayes' classification was an obvious choice.

Bayes' Classifer makes use of Bayes' Theorem to predict the class for a new test instance. Bayes' Theorem is given as

$$P(c_i|x) = \frac{P(x|c_i)*P(c_i)}{P(x)}$$

(1)

In Eq. 1, $P(x|c_i)$ is the likelihood that is the probability of observing x given that true class is c_i, P(ci) is the prior probability of class c_i and P(x) is the probability of observing x out of k classes [8]. It first estimates the posterior probability $P(c_i|x)$ for each class c_i and chooses the largest probability to assign new test instance.

In this project, there were seven categories of emotions, namely "sadness", "joy", "anger", "fear", "disgust", "surprise" and "unknown" which serve as our class labels. Additionally, there were three categories of polarity, namely "positive", "negative" and "neutral". This is to note that these categories of emotion and polarity are independent of each other. At first, the model was trained on training data considering the mentioned classes and then, it was tested on test

data. 10-fold cross validation was used for this purpose. Under 10-fold Cross Validation, the dataset was partitioned into 10 subsets of equal sample size. Out of these 10 subsets, 9 subsets were used as training data and 1 subset was used as validation data. This process was repeated 10 times (folds) with each 10 subsets used for validation exactly once. The results were then averaged to get a single estimation [11]. Post evaluation of classification model, an estimation of most frequent terms in each of the emotion categories was done and plotted using a Word Cloud. Word Cloud is a graphical representation of word frequency. The size of the words is proportional to their frequency. A word which has appeared more number of times than another word will have greater size in word cloud. All the words are then arranged into a cloud of words [12]. The classification is performed using RStudio®.

3. Results

3.1 Data Acquisition

A total of 3000 tweets were pulled from Twitter for movie "Dunkirk" using the hash tags "#Dunkirk" and "#ChristopherNolan".

3.2 Data Cleaning

The 3000 tweets acquired went through a number of data cleaning steps to remove unnecessary characters that were of no use for further classification. This step was accomplished in RStudio®. No tweets were deleted in this process. After data cleaning, all the tweets are manually classified as positive or negative according to their emotions. A total of 1500 tweets are classified as positive and rest 1500 as negative.

3.3 Classification

The data available after data cleaning was subjected to 5-fold cross validation where 80% data is used to train and rest 20% to test the model. The complete dataset of 3000 reviews was fed to Bayes' Classifier to calculate posterior probabilities and predict seven emotion classes. The output statistics from this step are shown in Figure 1. It can be concluded from Figure 1 that around 50% of reviews are classified as "unknown" emotion. This might be due to the fact that reviews contain lot of words that are difficult to classify in basic emotion categories. Next dominant emotion is "sadness" attributing to 40% of reviews. This might be due the fact that this

is a tragic movie based on World War II. Next noticeable emotion is “joy” which accounts for 6% of reviews. Other emotion categories “anger”, “fear”, “surprise” and “disgust” account for less than 1% of the reviews.

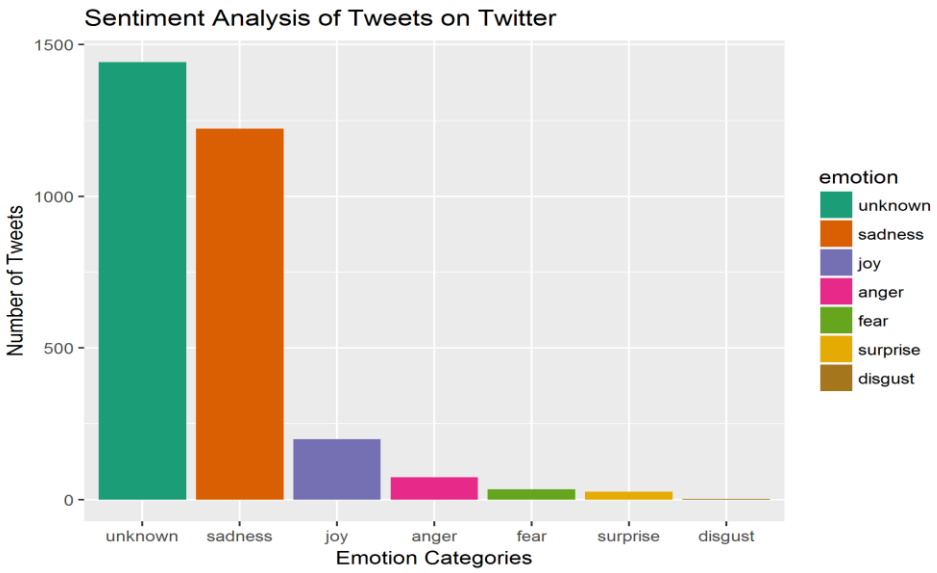

Figure 1. Sentiment Analysis of Tweets classified by emotions

This data was also used to train the model for predicting polarity of the reviews. The output statistics from this step are shown in Figure 2. A total of 1201 reviews were predicted to be positive and 1268 reviews to be negative. Rest 508 reviews were classified as neutral. It can be concluded from Figure 2 that around 42% of reviews were classified as “negative”, giving accuracy of 84.53% and around 40% reviews were classified as positive, giving accuracy of 80.07%. As mentioned earlier, this movie received mixed reviews which were not positive or negative majorly. Though the classification is performed on small dataset, it is evident from results of classification based on polarity as well.

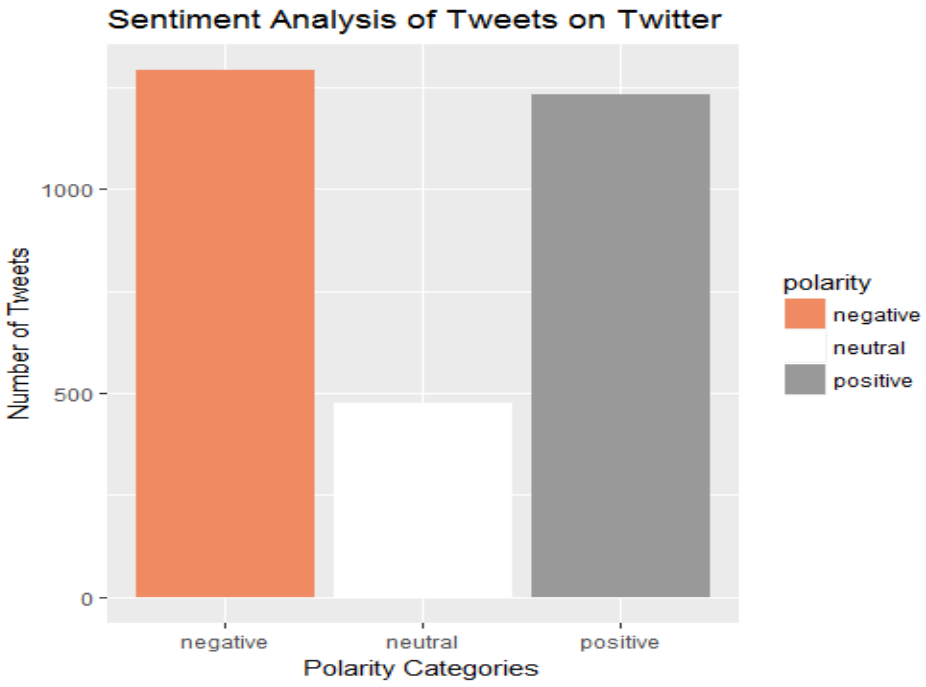

Figure 2. Sentiment Analysis of tweets by polarity

Post prediction of polarity, Word Cloud was generated to depict most frequent terms in each of the emotion categories. This is shown in Figure 3. It can be noticed from Figure 3 that frequent words in "unknown" emotion category are Tom Hardy, cold, tickets, Imax etc which linguistically do not fall under any emotions. Other notable frequent words under different categories that got correctly classified were greatest, quality, unpredictable, alarmed, unsure, deaths, best, good, sick etc.

Figure 3. Word Cloud based on emotion categories

4. Conclusion

A total of 3000 tweets were fetched form Twitter and cleaned of unnecessary characters to be classified using Bayes' Classifier. Bayes' Classifier then classified reviews in seven emotion categories, namely "sadness", "joy", "anger", "fear", "surprise", "disgust" and "unknown". 50% of total reviews are classified under "unknown" emotion, followed by 40% of reviews under "sadness". Other emotions account for less than 6% of reviews. Polarity of reviews is also predicted. Around 42% of reviews were classified as negative and 40% as positive. The accuracy of negative review prediction is 84.53% and that of positive review is 80.07%. Word Cloud is plotted to depict most frequent words in seven emotion categories. Most of the words plotted are classified correctly in different emotion categories.

5. References

[1] https://www.lifewire.com/what-exactly-is-twitter-2483331

[2] https://www.brandwatch.com/blog/understanding-sentiment-analysis/

[3] https://businessjargons.com/data-collection.html

[4] Twitter site

[5] https://themepacific.com/how-to-generate-api-key-consumer-token-access-key-for-twitter-oauth/994/

[6] Team R. RStudio: integrated development for R. RStudio, Inc, Boston, MA URL http://www rstudio com. 2015.

[7] https://docs.microsoft.com/en-us/sql/data-quality-services/data-cleansing

[8] Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, May 2014. ISBN: 9780521766333

[9] https://en.wikipedia.org/wiki/Statistical_classification

[10] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746

[11] https://www.openml.org/a/estimation-procedures/1

[12] https://datavizcatalogue.com/methods/wordcloud.html