RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis of Twitter tweets using supervised classification technique

Pranav Waykar, Kailash Wadhwani, Pooja More, Archana Kollu
Department Of Computer Engineering, DR.D.Y.Patil Institute Of Engineering And Technology,Pune.

ABSTRACT
Making use of social media for analyzing the perceptions of the masses over a product, event or a person has gained momentum in recent times. Out of a wide array of social networks, we chose Twitter for our analysis as the opinions expressed their, are concise and bear a distinctive polarity. Here, we collect the most recent tweets on users' area of interest and analyze them. The extracted tweets are then segregated as positive, negative and neutral. We do the classification in following manner: collect the tweets using Twitter API; then we process the collected tweets to convert all letters to lowercase, eliminate special characters etc. which makes the classification more efficient; the processed tweets are classified using a supervised classification technique. We make use of Naive Bayes classifier to segregate the tweets as positive, negative and neutral. We use a set of sample tweets to train the classifier. The percentage of the tweets in each category is then computed and the result is represented graphically. The result can be used further to gain an insight into the views of the people using Twitter about a particular topic that is being searched by the user. It can help corporate houses devise strategies on the basis of the popularity of their product among the masses. It may help the consumers to make informed choices based on the general sentiment expressed by the Twitter users on a product.
Keywords - Data Mining, Feature extraction Naïve Bayes Classifier, Natural language Processing, Twitter, Unigram

I. INTRODUCTION

Twitter is a popular micro blogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics. We propose a method to automatically extract sentiment (positive or neutral or negative) from a tweet. This is very useful because it allows feedback to be aggregated without manual intervention. Consumers can use sentiment analysis to do a research on products or services before making a purchase. Marketers can use this to research public opinion of their company and products, or to analyse customer satisfaction. Organizations can also use this to gather critical feedback about problems in newly released products. There has been a large amount of research in the area of sentiment classification. Traditionally most of it has focused on classifying larger pieces of text, like reviews. Tweets (and micro blogs in general) are different from reviews primarily because of their purpose: while reviews represent summarized thoughts of authors, tweets are more casual and limited to 140 characters of text. Generally, tweets are not as thoughtfully composed as reviews. Yet, they still offer companies an additional avenue to gather feedback. Previous research on analysing blog posts by Pang et al. [3] have analysed the performance of different classifiers on movie reviews. The work of Pang et al. has served as a baseline and many authors have used the techniques provided in their work across different domains. In order to train a classifier, supervised learning usually requires hand-labelled training data.

With the large range of topics discussed on Twitter, it would be very difficult to manually collect enough data to train a sentiment classifier for tweets. Hence, we have used publicly available Twitter datasets. However, this dataset consist only of positive and negative tweets. For neutral tweets, we have used the publicly available neutral tweet dataset provided. We run the machine learning classifiers Naïve Bayes trained on the positive and negative tweets dataset and the neutral tweets against a test set of tweets.This can be used by individuals and companies that may want to research sentiment on any topic.

II. BACKGROUND

Defining the sentiment

For the purpose of this work, we define sentiment as a positive or negative inclination of the expression stated by the author. If the expression doesn't bear any polarity, it is marked as a neutral sentiment.

Table 1: Example Tweets

Sentiment	Keyword	Tweet
Positive	Weather	The weather is pretty good this morning!
Negative	Work	Dammnn…. I hate this clerical work
Neutral	Bus	The bus arrives at 8 in the evening.

Related Work

Topics related to the one discussed in this work, have been researched before. Alec Go, Richa Bhayani and et al [4] classify tweets using unigram features and the classifiers are trained on data obtained using distant supervision. Radha N and et al [5] shows that using emoticons (distant supervision) as labels for positive and sentiment is effective for reducing dependencies in machine learning techniques and this idea is heavily used in [4]. Pang and Lee [3] researched the performance of various machine learning techniques in the specific domain of movie reviews.

III. METHODOLOGY

A. *Pre-processing*

The Twitter language model has many unique properties. These properties can be used to reduce the feature space:

1) Usernames

In order to direct their messages users often include Twitter usernames in their tweets. A de facto standard is to include @ symbol before the username (e.g. @towardshumanity). A class token (AT_USER) replaces all words that begin with @ symbol.

2) Usages of links:

Users very often include links in their tweets. To simplify our further work, we convert a URL like "http://tinyurl.com/cmn99f" to the token "URL".

3) Stop words:

There are a lot of stop words or filler words such as "a", "is", "the" used in a tweet which does not indicate any sentiment and hence all of these are filtered out.

4) Repeated letters:

Tweets contain very casual language. For example, if you search "hello" with an arbitrary number of 'o's in the middle (e.g. helloooo) on Twitter, there will most likely be a nonempty result set. I use pre-processing so that any letter occurring more than two times in a row is replaced with two occurrences. In the samples above, these words would be converted into the token "hello".

B. *Feature Vector*

After pre-processing the tweets, we get features which have equal weights.

Unigram

Features which are individually enough to understand the sentiment of a tweet is called as unigram. For example, words like 'good', 'happy' clearly express a positive sentiment.

C. Classification

For the purpose of classification of tweets, we make use of Naïve Bayes classifier. Naïve Bayes is a probabilistic classifier based on Bayes' theorem. It classifies the tweets based on the probability that a given tweets belongs to a particular class.

We consider three classes namely, positive, negative and neutral. We assign class c* to tweet d where,

$$c* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

In this formula, *f* represents a feature and $n_i(d)$ represents the count of feature *fi* found in tweet d. There are a total of m features. Parameters *P(c)* and *P(f|c)* are obtained through maximum likelihood estimates, and add-1 smoothing is utilized for unseen features. We have used the Python based Natural Language Toolkit library to train and classify using the Naïve Bayes method.

IV. EVALUATION

A. *Training data*

There are publicly available data sets of Twitter messages with sentiment indicated by [4]. We have used a combination of these two datasets to train the machine learning classifiers. For the test dataset, we used 20 tweets collected run-time during the execution.

B. *Experimental Setup*

The Twitter API has a parameter that specifies which language to retrieve tweets in. We always set this parameter to English (en). Thus, our classification will only work on tweets in English because the training data is English-only.

We build a web interface which searches the Twitter API for a given keyword for the past one day or seven days and fetches those results which is then subjected to pre-processing. These filtered tweets are fed into the trained classifiers and the resulting output is then shown as a graph in the web interface.

V. RESULTS

When a keyword was entered into the search box, the tweets about the entered topic were collected and classified. For the purpose of testing the application, we searched tweets about "Donald Trump". 20 tweets were shown to the user and classified using Naïve Bayes classifier. The result of the classification was displayed in the form of a pie-chart as follows:

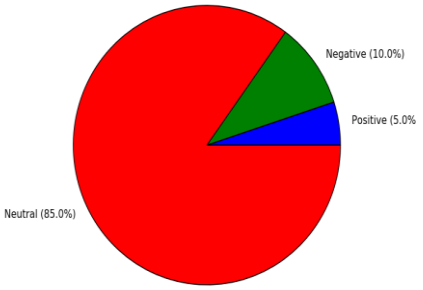

Figure 1: Pie-chart

Once the results were displayed, the user was asked if he wants to see the segregated tweets for the sake of justification. The accuracy of the results depend on the number of training tweets being fed to the classifier. Higher the number of tweets greater is the accuracy.

Now, this result about "Donald Trump" can be used by voters and political analysts alike. Voters can use the data to see the positive as well as negative aspects of Mr.Trump whereas the political analysts and psephologists can use it to make their predictions.

VI. FUTURE WORK

Machine learning techniques perform well for classifying sentiment in tweets. We believe the accuracy of the system could be still improved. Below is a list of ideas we think could help the classification:-

A. Semantics

The polarity of a tweet may depend on the perspective you are interpreting the tweet from. For example, in the tweet "Federer beats Nadal :)", the sentiment is positive for Federer and negative for Nadal. In this case, semantics may help. Using a semantic role labeler may indicate which noun is mainly associated with the verb and the classification would take place accordingly. This may allow "Nadal beats Federer :)" to be classified differently from "Federer beats Nadal :)".

B. Internationalization

Currently, we focus only on English tweets but Twitter has a huge international audience. It should be possible to use our approach to classify sentiment in other languages with a language specific positive/negative keyword list.

VII. CONCLUSION

A live Twitter feed is collected under the keywords entered by the user. The feed is stored locally in a json file. The data is pre-processed to remove unnecessary spaces, symbols and useless features. It still requires further work to remove as much noise as possible. 20 tweets are then stored as a csv file for analysis. A number of Lexicon based methods are utilised on individual tweets from the file to assess their usefulness. The chosen classifier for this work is a Naive Bayes Classifier utilising the text processing tools in NLTK and their capacity to work with human language data. It is trained on tagged tweets and then used to analyse the sentiment in the tweets about the searched topic. The result is represented in the form of a pie diagram which shows the percentage of users who have positive opinion on the searched topic as compared to the ones have negative opinion or are neutral.

REFERENCES

[1]. Adam Tsakalidis, Symeon Papadopoulos, Alexandra Cristea, Yiannis Kompatsiaris, "Predicting Elections for Multiple Countries Using Twitter and Polls)," IEEE. 2015.

[2]. Gayo-Avello, Daniel, A meta-analysis of state-of-the-art electoral prediction from Twitter data, Social Science Computer Review, 2013.

[3]. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[4]. Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford Digital Library Technologies Project, 2009.

[5]. Jiawei Han, Micheline Kamber, "Data mining: concepts and techniques", Morgan Kaufmann Publisher, second edition, pages 310-317.

[6]. Publicly available Twitter dataset - http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip.

[7]. Steven Bird, Ewam Klein, Edward Loper, "Natural Language Processing with Python", O'Reilly, 2009.

[8]. Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", AAAI Conference on Weblogs and Social Media.