Survey: Sentiment Analysis on Twitter Data using machine learning classification techniques

Fr. Puneet1 Dr. Vinay Goyal2 Er. Rajdeep Kaur3
Department of CSE, Chandigarh University Gharuan

Abstract: Large no of persons sharing their views daily on different blogging sites which contains semi structures data. Classification techniques are required to identify people’s opinion about object or product. We will discuss different machine learning classification techniques using data mining approaches. Positive, negative or neutral are the basic classification for post, message or tweets because these are very valuable for the different and specific product brands or the customers who want to search the opinion from others about product before purchase. Basic steps to analyze twitter data also discussed in this paper.

Introduction
Twitter is most popular social network websites with the extensive growth in the usage of online social media its growth rate is very speedy. Through the twitter platform users share opinion or information about products, personalities, politicians, companies and events. Twitter is a micro blogging facility for the consumers where they produce category messages which are also known as tweets. These tweets basically provide the communicate opinions about diverse topics. We will discuss different techniques and approaches used for automatically take out sentiment positive or negative from tweet. Sentiment analysis basically used to find out products or services before making a purchase. Every minute opinion and reviews are being expressed online and a customer and users relies on these opinion or reviews or feedback given by various other users to make decisions according to buying a product or developing a software, when it comes to an society or group that provides services. Critical feedback and analyses can also be found out from newly product that can be very helpful for organization or companies. To find out whether a message or text that can be review is positive or negative or neutral, Sentiment analyses is used. It’s basically provide the customer approaches or belongings regarding that particular brand via examine various parameters such as emotion, situation, context etc. So surveyor uses this sentiment analysis to examine opinion regarding their product and company by the various consumers and also examine the consumer satisfaction from the feedback provided by them. These scrutinizes are also used by society or group to capture significant feedback in newly released products and also useful to solve particular problems faced by the society related to that product.

Basic Taxonomy of text classification

Tokenization: It is the process of chopping up words or sentences into smaller pieces. There are number of techniques to segment the sentences and to solve various issues that may encounter.
**Word Normalization:** Word Normalization is the reduction of each word to its base and stem form for example Capital letters should be normalized to lowercase till it occurs in the middle which indicate the name of a writer, brand, place etc.

**Bag-of-words:** This is the easiest language unigram model used in Natural Language Processing (NLP) to keep the track of how many times that particular word is occurred. So in this model we count how many times that particular word is occurred and after that assign a specific subjectivity score to each word. In this model text has been segmented into sentences and then sentence segmented into words. After that words have been normalized, tokenized and subjectivity score is associated with each word which can be looked up in a sentiment lexicon. Then the overall score is evaluated, and if the overall score is negative then text will be nominated as negative and if the overall score positive than the text will be nominated as positive.

**Classifier Evaluation:** For formative the precision of a single Classifier, or analyze the outcome of various Classifier, the F-score is mainly used. This F-score is expressed by

\[ F = \frac{2pr}{p + r} \]

Where \( p \) is precision and \( r \) is the recall respectively.

The precision ‘\( p \)’ is the ration of number of properly classified examples divided by the total examples (classified examples).

The recall ‘\( r \)’ is the ratio of number of correctly classified examples by the number of training set examples.

**Naive Bayes**

Naive Bayes classifier is a powerful algorithm for the classification. Naive Bayes approach is basically useful for large data sets for training and even with millions of records with some attributes. It gives great results when we use it for textual data analysis such as Natural Language Processing.

The Naive Bayes classifier uses the prior probability of each label and calculate the number of times that particular label is encounter in training set. So Bayes classifier is based on frequency It also consider the involvement from each label which represent particular quality. The frequency can be checked, for ‘positive’ and ‘negative’.

For example if a particular word such as ‘fascinating’ appears in 1 of 7 of the positive tweets and none of the negative tweets than that means probability of the ‘positive’ label. So probability of occurrence will be multiplied when that particular word is seen as part of the input.
Basic steps to analyze twitter data:

1. Gathering tweets: Number of the tools are used to gather tweet data

   i. **Twitter API:** Twitter APIs Python wrapper for performing API requests such as downloading tweets and searching tweets. It provides a simple Python interface (API Interface) that allows multiple users or group of users to connect via the web, IM, and SMS. Twitter exposes a web services API which intended to make it even easier for Python programmers to use.

   ii. **MongoDB:** It’s an open source document storage database and is the go-to “NoSQL” database. Working with a database feel like working same as working with JavaScript and transactions in MongoDB will feel just like transactions developers are familiar with from relational databases.

   iii. **PyMongo:** A Python wrapper for interfacing with a MongoDB instance. This library will connect your Python scripts with your database and read or insert records.

   iv. **Cronjobs:** It’s a Job scheduler which basically work on time way. It’s allowed to run JavaScript at specific time interval.

2. Preprocess Tweets: We first need to remove noise from the tweet and to remove noise we have to do the listed steps for pre-processing the tweets:

   i. Firstly we need the tweet message in lower cases. So by using various techniques we have to convert the message into lower case.

   ii. Get rid of all types of URLs either via using regular expression matching or reinstate with generic word URL.

   iii. Matching or replace with generic words.

   iv. Replace hash-tags with proper substitute.

   v. Eliminate punctuation and extra white spaces from the start and ending of the message / tweet.

3. Analyzing tweets for Sentiment: Sentiment Analysis: is the process of formative whether a piece message is positive or negative. Number of tools in the market that provide automated sentiment analysis solutions. Monkey-Learn is one of tremendous scalable tool to perform sentiment analysis and automates text classification. We are able to get results quickly with built-in public modules in MonkeyLearn, without machine learning knowledge.

4. Categorizing tweets: In the field of data mining information retrieval etc classification is one of the major researches filed, and its solutions are at the heart of numerous technology. In these fields various application start from the automatic cataloging of newspaper pages and web pages to the management of incoming e-
mails and from the annotation of DNA genome sequences to sentiment analysis of tweets.

5. **Visualizing the results:**
An open source Twapper Keeper was chosen after analyze the reviews of various online tools that could collect and manage tweets. Twapper Keeper is a web application designed to collection social media data via Twitter to allow for long-term archival and analysis.

**Conclusion**

Twitter is a micro blogging service which has been specifically built to discover and analyze the occurrence of events or task at any split second of time or wherever place in the world. In the number of surveys, it is found that social media related features can be used to predict sentiment in Twitter. We discuss basic classification methods for twitter analysis. This paper concludes the sentiments of tweets which are extracted from twitter. The difficulty increases with complexity of opinions expressed. Music, books, art and movies are more difficult to express in sentiment analysis. We discuss the basic steps for the twitter analysis and tools required to perform the different steps.

**References**


MindTrek Conference, 6th–8th October, Tampere, Finland