INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

Twitter Sentiment Analysis Using Vader

Vipul Kumar Chauhan
vipul.chauhan705@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Ashish Bansal
ashish.bansal3007@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Dr. Amita Goel
amitagoel@rediff.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

ABSTRACT

Twitter is a micro-blogging site used by people to express their opinions on various topics. Sentiment Analysis is the process of extracting meaningful customer insight from the text in terms of sentiment score.

Twitter Sentiment analysis is an application of sentiment analysis, on the twitter data (tweets).But today it has become difficult to analyze tweets because of the changed and challenging formats of the tweets. The increase in the use of various slangs, emoticons, abbreviations, and puns in tweets, has made it difficult to analyze tweets in the same ways as before.

In this paper, we aim to review some papers regarding research in sentiment analysis on Twitter, describing the methodologies adopted and models applied; along with describing Vader Sentiment Analysis which is a Python-based approach.

Keywords: *Sentiment Analysis; Methodologies; Natural Language Processing; Sentiment Analysis with Python; Python; Application; Conclusion.*

1. INTRODUCTION

Twitter has emerged as a major micro-blogging website, has over 100 million users generating over 500 million tweets every day. With such large opinion based audience, Twitter has become an informative source for many organizations, institutions, and companies for information regarding their product or services which customer use.

With millions of tweets coming up every day, companies are able to model in customer insight in terms of graphs and tables based on the sentiment reflected in their tweets.

On Twitter, users share their opinions in the form of tweets, using only 140 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons etc. Along with this, people also use sarcastic and polysemy language in their tweets. Hence it is well understood that Twitter language is unstructured. In order to extract meaningful information from tweets, sentiment analysis is used which gives result in terms of percentage sentiment on a particular scale. The results from this can be used in many areas like analyzing and monitoring changes of sentiment with an event, sentiments regarding a particular brand or release of a particular product, analyzing public view of government policies etc.

In this paper, we describe sentiment analysis along with the new evaluating tool VADER. VADER has the benefits of traditional sentiment lexicons along with improved ones, which can be easily used and extended. VADER sentiment lexicons are of a much higher standard because they have been validated by humans. VADER distinguishes itself from others in terms that it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domains.

2. ABOUT SENTIMENT ANALYSIS

Sentiment analysis is a process of calculating sentiment of a particular statement or sentence. It's a classification technique which derives opinion from the tweets and formulates a sentiment score which reflects the sentiment based opinion of the text.

Sentiments are subjective to the topic of interest. We are required to formulate that what kind of features we decide to extract from the text. For example, we can have two-class tweet sentiment classification (positive and negative) or three class tweet sentiment classification (positive, negative and neutral). The dimension of the sentiment class is a crucial factor in deciding the efficiency of the model. As what we want to calculate has to be programmed in terms of a class in the sentiment calculator. The greater the efficiency of the program the better and refined are the results.

Sentiment analysis approaches can be broadly categorized into two classes – lexicon based and machine learning based. Lexicon based approach is unsupervised as it proposes to perform analysis using lexicons and a scoring method to evaluate opinions. Whereas machine learning approach involves the use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification either using simple computation or else machine learning approaches

2.1 TWITTER SENTIMENT ANALYSIS

The aim of Twitter sentiment analysis is to categorize the tweets into three common classes, namely, positive sentiment class, negative sentiment class, neutral sentiment class. Also further on more calculation can be done and more features like most used words, common words, most used emoticon and average sentiment core of the complete data can also be calculated.

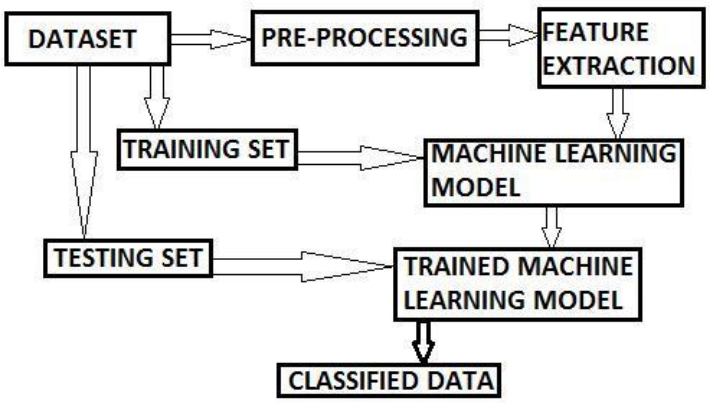
Performing sentiment analysis is challenging on Twitter data, as we mentioned earlier. Here we define the reasons for this:

- **Limited Tweet size**: with just 140 characters in hand, compact statements are generated, which results in a sparse set of features. Also along with the use of slang, abbreviations, emoticons in the tweets make it difficult to analyze the data.
- **Use of Slang**: these words are different from English words and it can make an approach outdated because of the evolutionary use of slangs.
- **Twitter Features**: it allows the use of hashtags, user reference, emoticons, and URLs. These require different processing than other words.
- **User Variety**: the users express their opinions in a variety of ways, some using different language in between, while others using repeated words or symbols to convey an emotion. Some only use series of emoticons to represent a visual picture, on the other hand, some use sarcastic statements which seem something else but actually mean something else.

Above are the set of problems mostly faced in the stage of pre-processing of the data. Apart from these, we also face another set of problems which are the result of the inadequate information.

3. METHODOLOGY

In order to perform sentiment analysis, we are required to collect data from the desired source (here Twitter). This data undergoes various steps of pre-processing which makes it more machine sensible than its previous form. Given below is flow- process representation of the pre-processing of the tweets.



3.1 TWEET COLLECTION

The very first step is to actually collect the twitter data on desired or relevant topic. This collection involves the use of twitter-API to extract real-time data from Twitter.

Twitter's streaming API like Tweepy (python-twitter library) is most generally used for collecting tweets. The format of the retrieved text is converted as per convenience (for example JSON), in our case in CSV format.

In case of a machine learning model, we have to divide the collected data in terms of test and training data. But here in VADER we first clean the data as per the need and compute the resulting tweet by tweet.

3.2 PRE-PROCESSING OF TWEETS

The preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the tweets, removal of undesired data form the tweets and other objectification of any other desired feature. The steps taken in cleaning should be done so to make the data easier to work with. Below are a few steps used for pre-processing of tweets:

- **Removal of re-tweets.**
- **Converting upper case to lower case:** In case we are using case-sensitive analysis, we might also take into account case change of two same words as different.
- **Stop word removal:** Stop words are those which do not affect the meaning of the sentence, thus are better removed.
- **Stemming:** Replacing words with their roots, reducing different types of words with similar meanings. This helps in reducing the dimensionality of the feature set.
- **Special character and digit removal:** Digits and special characters don't convey any sentiment. Thus have to be removed.
- **Expansion of slangs and abbreviations.**
- **Spelling correction:** This can sometimes be important as a wrong spelling could affect the sentiment but the reflect is noticeable only when the frequency of such misspelt word is very high.
- **Generating a dictionary of words that are important or for emoticons.**
- **Part of speech (POS) tagging:** It assigns a tag to each word in the text and classifies a word to a specific category like a noun, verb, adjective etc. POS taggers are efficient for explicit feature extraction.
- **URL and Username removal:** Usernames and URLs are not important from the perspective of future processing, hence their presence is futile.

3.3 FEATURE EXTRACTION

A feature is a piece of information that can be used as a characteristic which can assist in solving a problem (like prediction). The quality and quantity of features are very important as they are important for the results generated by the selected model. Given below are the most common types of features extracted:

- **Unigram Features:** One word is considered at a time and decided whether it is capable of being a feature.
- **N-gram Features**: More than one word is considered at a time.
- **External Lexicon**: Use of a list of words with predefined positive or negative sentiment.

Frequency analysis is a method to collect features with highest frequencies used in. This is the most commonly used method for collecting different types of features from the data. The feature result calculated is divided into two categories: Common Features and Tweet Specific Features.

3.4 SENTIMENT CLASSIFIERS

There are different types of Sentiment Classifiers which can be used for calculating and then classing the sentiments in appropriate classes. Some of them are probability based while others are machine learning based. Also, we have described VADER sentiment classifier.

- **Naïve Bayes:** It is a probabilistic classifier with strong conditional independence assumption that is optimal for classifying classes with highly dependent features.
- **Support Vector Machine Algorithm:** Support vector machines are supervised models with associated learning algorithms that analyze data used for classification and regression analysis, It makes use of the concept of decision planes that define decision boundaries.
- **Artificial Neural Network:** the ANN model used for supervised learning is the Multi-Layer Perceptron, which is a feedforward model that maps data onto a set of pertinent outputs. Training data given to input layer is processed by hidden intermediate layers and the data goes to the output layers. The number of hidden layers is a very important metric for the performance of the model.
- **Vader Sentiment Analysis**: VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. Introduced in 2014, VADER text sentiment analysis uses a human-centric approach, combining qualitative analysis and empirical validation by using human raters and the wisdom of the crowd

4. SENTIMENT ANALYSIS WITH PYTHON

4.1 PYTHON

Python is a high level, interpreted programming language, created by Guido van Rossum. The language is very popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks.

Python provides a large standard library which can be used for various applications, for example, natural language processing, machine learning, data analysis etc.

For complex projects like Sentiment analysis Python is very useful because of the no. of the libraries and features python has to offer to eases out the programming part to a very large extent and thus provided considerable time to decide how the analysis is to be done.

4.2 NATURAL LANGUAGE PROCESSING (NLTK)

Natural Language Toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK. The NLTK library also embodies various trainable classifiers, most commonly used the naïve based classifier. NLTK library is used to create a bag of words on which the sentiment is calculated as raw analysis. Further, these results are co related to generating more accurate results.

4.3 DATA COLLECTION

We have two options to collect data for sentiment analysis. First is to use Tweepy - client for Twitter Application Programming Interface (API) and the second is download the data physically from the internet using any tweet fetching tool and then use it for the process.

To fetch tweets from the Twitter API one needs to make a developer account in twitter, doing so we will get the desired, Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret' which will be used to authenticate the pulling of the data from the twitter. These keys are then inserted into the code, which helps in dynamic collection of tweets every time we run it

The former method is slow in nature as it performs tweet collection every time we start the program. The latter approach may not provide us with the quality of tweets we require.

4.5 PRE-PROCESSING IN PYTHON

The pre-processing in Python is the same process as described above. But due to the fact that python has a rich set of features to do the same effortlessly we use a python based regular expression which does the same work on the go.

For each task of pre-processing, the related regular expression is given below:

- **Converting all upper case letters to lower case.**
- **Removing URLs:** Filtering of URLs can be done with the help of regular expression (http|https|ftp)://[a-zA-Z0-9\\./]+.
- **Removing Handles (User Reference):** Handles can be removed using the regular expression - @(\w+).
- **Removing hashtags:** Hashtags can be removed using the regular expression - #(\w+).
- **Removing emoticons:** We can use emoticon dictionary to filter out the emoticons or to save the occurrence of them in a different file. (But in our analysis we use emoticons file for their weight calculation)
- **Removing repeated characters.**

4.6 VADER SENTIMENT EXTRACTION

Various methodologies for extracting features are available to the present day. But the most common and effective format of Vader sentiment analysis is to build a dataset and then extract the required the data in desired format pre-hand. After this, we programme our own sentiment analysis model which employees the use of the Vader package for the use of its classes and sentiment functions for analysis. After this, we used the .txt file of Stop Words and Vader lexicons which has values for all sorts of words (positive and negative; along with words with tricky meaning) and emoticons, for the text matching and sentiment calculation. For accuracy, we used TextBlob for sentiment analysis at the end of the feature extraction by Vader analysis. The program splits the tweet sentence in the word bag by utilizing pythons splitting capabilities and then feeds these words to costume made logic for sentiment analysis which calculates the sentiment analysis per word, per sentence the calculates the avg. value of the whole and adds the result for the complete tweet Para and returns the result as an answer for what kind of text if present. Right now we are only interested in the negative and positive text analysis, as in general case these two play the role of extremities in any situational problem.

5. APPLICATION

- **Commerce:** Companies can make use of this research for gathering public opinion related to their brand and products. From the company's perspective, the survey of a target audience is imperative for making out the ratings of their products. As Twitter has distributed audience with rich data under the desired limit, it serves as a good platform for data collection and analysis to determine customer satisfaction.
- **Politics:** Majority of tweets on Twitter are related to politics. Due to Twitter's widespread use, many politicians are also aiming to connect to people through it. People post their support or disagreement towards government policies, actions, elections, debates etc. Hence analyzing data from it can help is in determining public view.
- **Sports Event**s: Sports involve many events, championships, gatherings and some controversies too. Many people are enthusiastic sports followers and follow their favourite players present on Twitter. These people frequently tweet about different

sports-related events. We can use the data to gather the public view of a player's action, team's performance, official decisions etc.

6. CONCLUSION

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and presenting the result under different sentiment classes. These results can be used to train a machine learning model which can be used in future to predict customer behaviour for particular product, service or expectation of people in regard to particular political party. The complete process of sentiment analyzing comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification. This topic has evolved as an important research topic during the last decade with the development of models reaching the efficiency of almost 85%-90%. But still, we have to focus on building models that have the capabilities to read between the lines, have the capabilities to understand human slangs and most importantly sarcasm. With the development in the field leaping folds by fold, it's not far when models will have extreme accuracies and would we able to understand slightest of the slightest change in the meaning of the context.

