Twitter Data Classification by Machine Learning for Sentiment Analysis

Pragya Mishra 1* and Rajesh kumar Nigam 2
1 M. Tech Scholar, Technocrats Institute of Technology Bhopal, Madhya Pradesh, India.
2 Associate Professor, Technocrats Institute of Technology Bhopal, Madhya Pradesh, India.
*Corresponding author email id: mishrapragya558@gmail.com

Date of publication (dd/mm/yyyy): 08/04/2021

Abstract – In this paper is proposed method that performs classification of tweet data sentiment on Twitter. Twitter is an online social networking website that contains a rich amount of sentiment data that can be structured, semi-structured and unstructured data that can help to analyze behavior of guy. Data classification and analyzing is big task. Here is used Naive Bayes classifier to classify data and used machine learning to analysis human behavior. Tweet analysis is work two basic mode: (A) sentiment expressed by a phrase in the context of a tweet, and (B) overall sentiment of a tweet. It is implemented with the help of python Ecosystem to improve its scalability and efficiency. In this sentiments analysis process also refer the NLP (Natural language processing). It is internal action process between human and computer, it analyzes the treasure of natural language data here we get idea how to mine twitter data to get sentiment. Here we used deep neural network and text mining algorithm to extract all text data to analysis accurate opinion of all user and increase accuracy.

Keywords – Sentiment Analysis, NLP (Natural Language Processing), Twitter data, Naive Bayes Classifier, SVM, Textblob.

I. INTRODUCTION

We live in a society where the textual data on the Internet is growing at a rapid pace and many companies are trying to use this deluge of data to extract people’s views towards their products. Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional “pulse of the nation”, and indeed the global community. A great source of unstructured text information is included in social networks, where it’s unfeasible to manually analyze such amounts of data. There is an outsized number of social network websites that enable users to contribute, modify, and grade the content, also on expressing their personal opinions about specific topics. Some examples include blogs, forums, product reviews sites, and social networks, like Twitter (http://twitter.com/). Twitter (San Francisco, CA, USA) is a microblogging site that offers the opportunity for the analysis of expressed mood, and previous studies have shown that geographical, diurnal, weekly, and seasonal patterns of positive and negative affects can be observed. Microblogging and more particularly Twitter is employed for the subsequent reasons:

- Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people’s opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Twitter’s audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it’s possible together text posts of users from different social and interest groups.
Twitter’s audience is represented by users from many countries as the audience of microblogging platforms and services grows a day, data from these sources are often utilized in opinion mining and sentiment analysis tasks. For example, manufacturing companies could also be curious about the subsequent questions: millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion, and share views about their daily lives.

Through the web communities, we get interactive media where consumers inform and influence others through forums. Social media is generating an outsized volume of sentiment rich data within the sort of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend on user-generated content online to an excellent extent for deciding. For e.g. if someone wants to shop for a product or wants to use any service, then they firstly search its reviews online, discuss it on social media before taking a choice, the amount of content generated by users is just too vast for a traditional user to research, so there’s a requirement to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA) tells the user whether the knowledge about the merchandise is satisfactory or not before they pip out, Marketers and firms use this analysis data to know about their products or services in such how that it is often offered as per the user’s requirements. Textual Information retrieval techniques mainly specialize in processing, searching, or analyzing the factual data present. Facts have an objective component but, there is another textual content that expresses subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly thanks to the large growth of obtainable information on online sources like blogs and social networks. For example, recommendations of things proposed by a recommendation system are often predicted by taking under consideration considerations like positive, or negative opinions about those items by making use of SA. Comments, reviews and opinions of the people play a crucial role to work out whether a given population is satisfied with the merchandise, services. It helps in predicting the sentiment of a good sort of people on a specific event of interest just like the review of a movie, their opinion on various topics roaming around the world. These data are essential for sentiment analysis, to get the general sentiment of population, retrieval of knowledge from sources like Twitter, Facebook, Blogs are essential. For the sentiment analysis, we focus our attention towards the Twitter, a micro-blogging social networking website. Twitter generates huge data that can’t be handled manually to extract some useful information and thus, the ingredients of automatic classification is required to handle those data. Tweets are unambiguous short text messages that are up to a maximum of 140 characters. By the use of Twitter, millions of people around the world to be connected with their family, friends, and colleagues through their computers or mobile phones. The Twitter interface allows the user to post short messages, and that can be read by any other Twitter user. Twitter contained a variety of text posts and grows every day. We choose Twitter because the source for opinion mining just because of its popularity and data processing. The Existing Database is not able to process the big amount of data within a specified amount of time. Also, this sort of knowledge base is restricted for the processing of structured data and features a limitation when handling an outsized amount of data. So, the traditional solution cannot help an organization to manage and process unstructured data. The use of Big Data technologies like Hadoop is the best way to solve Big Data challenges. Sentiment Analysis is the process of `computationally’ determining whether a bit of writing is positive, negative, or neutral. It’s also referred to as opinion mining, deriving the opinion or
attitude of a speaker. Sentiment analysis may be a term that you simply must have heard if you've got been within the tech field long enough. It is a method of predicting whether a bit of data (i.e. text, most commonly) indicates a positive, negative, or neutral sentiment on the subject in this article, we will go through making a Python program that analyzes the sentiment of tweets on a particular topic. The user will be able to input a keyword and get the sentiment on it based on the latest 100 tweets that contain the input keyword.

II. METHODOLOGY

Sentiments analysis is the invented science of psychology and sociology and both are the scientific study of people emotions, relationships, opinions, and behaviors (wiki). Psychologists apply sentiments process through the hypothesis but data scientist applies through the data. In other words, it is the computational process which identifies and categories the opinions, thoughts and ideas through the text data. The sentiments analysis process also refer the NLP (Natural language processing). It is internal action process between human and computer. It also analyzes the treasure of natural language data. Sentiments analysis is expressed in two different categories: polarity and subjectivity. The polarity measure the text data is positive (>0) or negative (<0) (or neutral (0). Classifying is a sentence as subjective or objective, known as subjectivity classification (monkeylearn.com). Subjectivity measures from (0.0 to 1.0). Where 0.0 is very objective and 1.0 is very subjective. But In this thesis we calculate only the sentiments polarity from twitter data (tweets data is in CSV format). Polarity showed three different colors positive for green color, negative in red color and neutral in blue color. Polarity calculated through the python code using library of Textblob and python module Natural Language Tool Kit (NLTK).

When we go for sentiment analysis there are many option and tools. The most popular tools are MATLAB, Python, and Java and C # and due to huge no of libraries available in python and easiest in code so mostly researcher used python because it is sensible and suitable choice. The sentiments analysis algorithm consists of 4 modules. The procedure in each model starts with importing data with pandas, since the powerfulness of pandas for processes and data preprocessing. Then used NLTK and Textblob for analyzing the text of CSV file and calculate the polarity of each text separately and output is a numeric format (-1 to +1). In this research, first collected the tweets from Twitter with given keyword and then analyze the whole text and gave the result, then Matplotlib plotting the result on the pie chat and with different colors and different formats positive, negative and neutral (greater than zero, less than zero and equal to zero). This program only those texts analyze when the required keyword is founded.

Fig. 1. System architecture working model.
2.1. Textblob Library

Textblob is the python library which process the textual data. Textblob provide API to access its methods and easily perform NLP task. The main reason behind the usage of Textblob is it’s like a python string easy to use without worrying the syntax. Textblob consist on different function like part of speech, noun phrase, sentiment analysis, tokenization, word inflection and lemmatization, wordlist, spelling correction, translation or language detection and N-gram. The textblob work and play with all kind of texts. Textblob support all kind of text formats. This also important module of python which use for sentiment analyzing and also classifying the data which part is positive of which part is negative (Steven loria, 2018). This is the main part of research and textblob is the key library for sentiments analyzing.

2.2. NLTK (Natural Language ToolKit)

Library Natural language toolkit it is also called NLTK. It is a suitcase of libraries such as symbolic and statistical natural language process which support Python English written programs. This toolkit have different classifications like sentiments, metrics, parse, tags, tokenization, chat, chunk, classify, translate, twitter, interface, draw, cluster and etc. NLTK includes graphical demonstrations and sample data. It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit, plus a cookbook (Bird, Edward, el, 2009). NLTK used the Python platform for building the programs of natural language text (human language) for using statistical natural language processing. NLTK is open source library for python which used on any platform, such as windows, Mac, Linux and many other platforms. In our thesis, we used Textblob library for sentiment analysis which import the NLTK module and sentiment analyzer is the subclass of NLTK module: Polarity results of input keyword

2.3. Sentiments Analyzer

A sentiment analyzer is a tool to implement and facilitate sentiment analysis task using NLTK features and classifiers, especially for teaching and demonstrative purposes. A sentiment analysis tool based on machine learning approaches.

2.4. Algorithm

Section A: Preparing the Test Set.

As our task of Sentiment Analysis is one that focuses heavily on textual data, one would expect there to be a lot
of text processing. This is definitely correct. In fact, both our Test and Training data will merely comprise of text. I chose to start with the Test set in order to get you all warmed up for the Training set extraction part, as it will rely more on the API. Here is a bit of an overview of what we are about to do:

1. Register Twitter application to get our own credentials.
2. Authenticate our Python script with the API using the credentials.
3. Create function to download tweets based on a search keyword.

Registering an application with Twitter is critical, as it is the only way to get authentication credentials. As soon as we get our credentials, we will start writing code. Step 3 is where the Test set lies. We will be downloading tweets based on the term that we are trying to analyze the sentiment on.

Step A.1: Getting the authentication credentials.
Step A.2: Authenticating our Python script.
Step A.3: Creating the function to build the Test set.

Section B: Preparing the Training Set.
Section C: Pre-processing Tweets in the Data Sets.
Section D: DNN Classifier.
Step D.1: Building the vocabulary.
Step D.2: Matching tweets against our vocabulary.
Step D.3: Building our feature vector.
Step D.4: Training the classifier.

Section E: Testing the Model.

III. SIMULATION RESULT

In this section, we used python to implement sentimental analysis. Some packages have utilized including tweepy and textblob. We can install the required libraries by following commands:

- pip install tweepy.
- pip install textblob.

The second step is downloading the dictionary by running the following command: python -m textblob.download_corpora. The textblob is a python library for text processing and it uses NLTK for natural language processing. Corpora is a large and structured set of texts which we need for analyzing tweets.

To connect to Twitter and query latest tweets, we need to create an account on twitter and define an application. Users need to go to the apps.twitter.com/app/new and generate the api keys. The Application settings is shown in the figure.
Following shows the sample output of the program for the ‘Corona Virus’ as a query based on the last 100 tweets from Twitter.

C:\Users\WPRICE9\AppData\Local\Programs\Python\Python37-32\python.exe D:\Twitter-Sentiment-Analysis-master\main.py
Enter Keyword/Tag to search about: coronavirus
Enter how many tweets to search: 100
How people are reacting on coronavirus by analyzing 100 tweets.
General Report:
Weakly Positive

Detailed Report:
8.00% people thought it was positive
27.00% people thought it was weakly positive
0.00% people thought it was strongly positive
3.00% people thought it was negative
11.00% people thought it was weakly negative
4.00% people thought it was strongly negative
47.00% people thought it was neutral

Figure 2 shows the sentimental analysis results based on Corona Virus. As it can be clearly seen in the figure3 the percentage of the neutral tweets are significantly high. This is also important to mention that depends on the data of the experiment we may get different results as people’s opinion may change depends on the world circumstances for example Corona Virus as it becomes the world of the year in 2020. For some queries, the neutral tweets are more than 60% which clearly shows the limitation of the current works.
The pie chart, as shown in figure 4, illustrates the data based on the results we got from this step. If we run the program in different times we may get different results, small variance, based on the tweets we fetch. We run the program three times and these results are the average of the outputs.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextBlob</td>
<td>28.4</td>
</tr>
<tr>
<td>Vader</td>
<td>31.5</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>40.9</td>
</tr>
<tr>
<td>SVM</td>
<td>41.4</td>
</tr>
<tr>
<td>FastText</td>
<td>40.6</td>
</tr>
<tr>
<td>Proposed</td>
<td>82.3</td>
</tr>
<tr>
<td>Textblob + NLTK + Machine Learning</td>
<td></td>
</tr>
</tbody>
</table>

**IV. CONCLUSION**

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model to train it and then check its accuracy, so that we can use this model for future use according to the results. The view of people can be positive or negative. An adjective plays a crucial role in identifying sentiment from part of data. It comprises of steps like data collection, text preprocessing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 90%-95%. But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don’t perform well when the number of classes is increased. Also, it’s still not tested that how accurate the model will be for topics other than the one in consideration. Twitter is large source of data which make it more attractive for performing sentiment analysis. We perform analysis on around 100 tweets so that we analyze the result. It can used for any purpose based tweets we collect with the help of keyword. Hence sentiment analysis has a very bright scope of development in future.