Twitter Sentimental Analysis Using Python

Abhay Kumar Shukla[1], Vishap Gautam[2], Sanjay Kumar Sharma[3]

[1,2,3]Department of Electronics and Communication Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow

ABSTRACT

Social media websites have Emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. Opinion of people matters a lot to analyse how the propagation of information impacts the lives in a large-scale network like Twitter. Sentiment analysis of the tweets determine the polarity and inclination of vast population towards specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. In this project, we exploited the fast and in memory computation framework 'Apache Spark' to extract live tweets and perform sentiment analysis. The primary aim is to provide a method for analysing sentiment score in noisy twitter streams. This paper reports on the design of a sentiment analysis, extracting vast number of tweets. Results classify user's perception via tweets into positive and negative. Secondly, we discuss various techniques to carryout sentiment analysis on twitter data in detail

Keywords: Big data, Social Media

1- INTODUCTION

Sentiment Analysis is process of collecting and analyzing data based upon the person feelings, reviews and thoughts [1]. Sentimental analysis often called as opinion mining asit mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data. Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it is analyzed whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment. Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e., results of one domain cannot be applied to another domain. Sentimental Analysis is used in many real-life scenarios,to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing. Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm [2].

1.1 Problem Statement:

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentences, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentences, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

Difficulty of Sentiment Analysis with inappropriate English

Difficulty of Sentiment Analysis with inappropriate English

- Sentimental Analysis of web Based Application Focus on single Tweet Only.
- Difficulty of Sentimental Analysis with inappropriate English.
- Difficulty of sentimental Analysis of Human Facial expressions.

2 - LITERATURE REVIEW

This section summarizes some of the scholarly and research works in the field of Machine Learning and data mining to analyses sentiments on the Twitter and preparing prediction model for various applications. As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing, and other promotional strategies as explained in.

The benefit of social media to know public opinions and extract their emotions are considered by authors

and explained how twitter gives advantage politically during elections. Further, the concept of the hashtag is used for text classification as it conveys emotion in few words. They suggested how previous research work suffered from lack of training set and misses some features of target data. They opted two stage approach for their framework- first preparing training data from twitter using mining conveying relevant features and then propounding the Supervised Learning Model to predict the results of Elections held in USA in 2016.
After collecting and pre-processing the tweets, training data set was created first by manual labelling of hashtags and forming clusters, next by using online Sentimental Analyzer VADER which outputs the polarity in percentage. This approach reduced the number of tweets or training set and further they applied Support Vector Machine and Naive Bayes classification algorithm to determine the polarity of tweets [3].

3- DATA CHARACTERISTICS

There are way too many social networking sites available these days, but in this paper, we are dealing with just one such site and that is twitter. Twitter is in too much fame in present because of its specific format of writing. Few of the characteristics of tweets is given below.

- Tweet length: tweets are short messages consisting of a maximum of 140 characters.
- Tweet availability: twitter is in way more fame than any other social networking site till present day. So much so that approximately 1.2 billion tweets are posted on a daily basis.
- User mentions: '@' character is used to make a mention of any user as to direct the message towards them.
- Hash tagging: '#' character is used to make the mention of the topic relating to which tweet is being written.

4- METHODOLOGY

The Proposed of method for sentiment analysis in this paper could be represented in 6 stages, each of which are listed below:

- Data Collection
- Data Pre-processing and Cleaning
- Location Geocoding
- Data exploration
- Sentimental Analysis

4.1 -Data Collection :

Data collection is the first phase for analysis as there needs to be data for us to do analysis on. In our experimentations we have used python programming language as a tool. Being that said, data collection in this particular analysis could be carried out in two ways. First way is to collect preorganized data from different sites such as Kaggle . On these sites this preorganized data is uploaded by the developers of sites themselves or is posted by different researchers for free . All one needs to do to acquire this data is to create a free account on these sites. Second way is to manually extract data from twitter using some API available for twitter. For this we have chosen tweepy as an API for extraction of tweets.
So, for using this particular API . To access tweets on twitter using API first we need to authenticate the console from which we are trying to access twitter. This could be done by following steps listed below:

- Creation of a twitter account.

- Logging in at the developer portal of twitter.
- Select "New App" at developer portal.
- A form for creation of new app appears, fill it out Fill.
- After this the app for which the form was filled out will go for review by twitter team
- Once the review is complete and the registered app is authorized then and only then the user is provided with 'API key' and 'API secret'.
- After this "Access token" and "Access token secret" are given.

These keys and tokens are unique for each user and only with the help of these can one access the tweets directly form twitter. For this paper we have extracted a large data set consisting of almost 3000 tweets.

4.3 Data Pre-processing and Cleaning

The pre-processing of data implies the processing of raw data into a more convenient format which could be fed to a classifier in order to better the accuracy of the classifier. Here, in our case the raw data which is being extracted from twitter using an API is initially totally unstructured and bogus as the availability of various useless characters seems very common in it.

For this matter we remove all the unnecessary characters and words from this data using a module in python known as Regular Expressions, are for short. This module adopts symbolic techniques to represent different noise in the data and therefore makes it easy to drop them. which need to be removed to boost the accuracy of us

resultant. These could be summoned up as follows:

- Hash tags: these are very common in tweets. Hash tags represent a topic of interest about which the tweet is being written. Hashtags look something like #topic.
- @Usernames: these represent the user mentions in a tweet. Sometimes a tweet is written and then is associated with some twitter user, for this purpose these are used.
- Retweets(RT): as the name suggests retweets are used when a tweet is posted twice by same or different user.
- Emoticons: these are very commonly found in the tweets. Using punctuations facial expressions are formed in order to represent a smile or other expressions, these are known as emoticons.
- Stop words: stop words are those word which are useless when it comes to sentiment analysis.

Pre-processing and cleaning both were performed simultaneously and are inseparable part of each other. For the better result and fast

Processing we need to take care of both the things .

4.4 Location Geocoding :

Geocoding is the computational process of transforming a physical address description to a location on the Earth's surface.

The tweets that we are collecting may contain location that will be highly unambiguous, so to locate the point of map we first need to convert this unambiguous information to some that we can use like in the form of latitude and longitude, country codes, country names, etc.

To perform this task here we used **Here** api which returns a complete and exact location on providing unambiguous location data.

To use the Here api we first need to acquire its credentials then we can use it.

In this project the geocoding section in not fully utilised because we are more focused on providing results instantly rather than taking long time for processing large amount of address and then providing data, the geocoded data will be provided when user will demand it.

4.5 Data Exploration:

Now here comes the main part that is processing large amount of data and extracting the relevant data from it.

To perform data exploration, we used Natural Language Toolkit library of python, which is specialised in processing speech text.

Processing of data involves extracting useful content, separating different parts of it and labelling it.

Let's suppose that we have a few tweets of peoples talking about someone then in this we don't completely know what is of our use but we do know that what we don't need.

On the basis of this information, we make few functions that just remove what we don't need that can be unnecessary verbs, links, emojis, phases, linking verbs, etc.

In python is very easy to do this processing because of few libraries like regular expression

. Now finally we stored this data in the form of excel and csv by formatting it in a proper way that is suitable for next step[4].

Now the next step is data visualisation in which we created be charts and graphs which will show information in a more representative way.

4.6 Sentiment Analysis

For our final project segment, we will conduct sentiment analysis on our dataset. Sentiment analysis involves determining whether a tweet is positive or negative. In this context, "positive" refers to content that is happy, cheerful, or supportive, while "negative" refers to content that expresses sadness or criticism. Various Twitter modules are available for performing this task, such as NLT, Text Blob, and Fair. Among these options, we chose Text Blob due to its simplicity, speed, and efficiency [5]. By applying sentiment analysis to all the available data, we can ultimately reach a conclusion regarding the overall sentiment of people towards the queried topic. Additionally, we will gather information on the countries where the topic is most frequently discussed, as well as identify commonly used words related to the topic.

5. CONCLUSION AND FUTURE SCOPE

This analysis of data is very useful for brand promotion, political campaign, advertising, and other organizations and individuals that want to know the feedback or thought of people on few trending topics. The is totally based on the analysis done and due to large amount of data available we can conclude many things.

The relevancy of data can be backed by the amount of data we will process, if we process more tweets then our results will be more accurate.

In future we can add more analysis and processing to extract more information from it, hence we can say this project has a lot of future scope to get more better with time.

We can also increase the quantity of tweets that we are procession to increase our accuracy.

REFERENCES

1. Varsha Sahayak, Vijaya Shete and Apashabi Pathan, "Sentiment Analysis on Twitter Data", (IJIRAE) ISSN: 2349-2163, January 2015.
2. David Zimbra, M. Ghiassi and Sean Lee, "Brand-Related "Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks", IEEE 1530-1605, 2016
3. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment Analysis of Twitter Data" Proceedings of the Workshop on Language in Social Media (LSM 2011), 2011.
4. Pak and P. Paroubek "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", vol. 10, pp. 1320-1326, 2010
5. Mondher Bouazizi and Tomoaki Ohtsuki, "Sentiment Analysis: from Binary to Multi-Class Classification", IEEE ICC 2016 SAC Social Networking, ISBN 978-1-4799-6664-6.