Opinion mining and trend analysis on twitter data

Avneesh Jha[1], Ajay Singh Chahar[2], Abhishek Singh Chauhan[3]

[1,2,3] Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Abstract

With the rise in internet users across the globe, there has been tremendous increase in the data available online. People use various social media apps and web platform to express their views and opinion regarding every possible aspects of their lives from politics to entertainment, Sports to economics etc. The project "Trend Analysis and Opinion Minion" is developed with the motive to gather all public opinions from twitter and analyze the current trends which may be helpful in determining marketing strategy, certain campaigning and spreading awareness. This can also be helpful in sensing cyber-bullying activities online.

Keywords: Trend Analysis, Opinion Mining ,Tweepy, Twitter Data.

1. INTRODUCTION

With the rise in internet users across the globe, there has been tremendous increase in the data available online. People use various social media platform to express their views and opinion regarding every possible aspects of their lives from politics to entertainment and sports to economics etc.

According to C. Aggarwal "The richness of this network provides unprecedented opportunities for data analytics in the context of social networks So the main motive of developing such project is to harvest the data available publicly over social media to perform trend analysis and opinion mining[i]. Data mining techniques may detect implicit or hidden patterns within a social networking site. This technique provides feedback to sense user opinion for political awareness during elections, shopping habits, identification of social groups , relationship among various entities and nodes and understanding the unrevealed trends in data . [1]

Trend analysis is the widespread practice of collecting information and attempting to spot a pattern. In some fields of study, the term "trend analysis" has more formally defined meanings.

Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

2. LITERATURE REVIEW

Data mining is the extraction of projecting information from large data sets, is a great innovative technology which helps corporations focus on the most important information in their data stockrooms. Data mining makes use of various statistical, machine learning and graphical methods and separate the knowledge in to a form which is very much useful for many real-world applications. Social network analysis has become a very popular field of research as it is useful for many applications. The author overviewed various data mining techniques used for social network analysis. [2]

Social network has gained remarkable attention in the last decade. Accessing social network sites such as Twitter, Facebook LinkedIn and Google+ through the internet and the web 2.0 technologies has become more affordable. People are becoming more interested in and relying on social network for information, news and opinion of other users on diverse subject matters. The heavy reliance on social network sites causes them to generate massive data characterized by three computational issues namely; size, noise and dynamism. These issues often make social network data very complex to analyses manually, resulting in the pertinent use of computational means of analyzing them. Data mining provides a wide range of techniques for detecting useful knowledge from massive datasets like trends, patterns. Data mining techniques are used for information retrieval, statistical modelling and machine learning. These techniques employ data pre-processing, data analysis, and data interpretation processes in the course of data analysis. This survey discusses different data mining techniques used in mining diverse aspects of the social network over decades going from the historical techniques to the up-to-date models.[3]

The growth in micro-blogging activity on sites over the last few years has been phenomenal. Platforms like Twitter offer an easy outlet for people to express their opinions and companies are increasingly getting interested in capturing these insights about customer behavior and preferences that could help generate more revenues. The staggering amount of data that these sites generate cannot be manually analyzed. Enter thus, Sentiment Analysis, the field where we teach machines to understand human sentiment.[4]

3. PROBLEM STATEMENT

The aim of this project is opinion mining and the analysis of the trends of the public statements gathered from different social media sources (specifically Twitter). Here Binary Sentiment analysis is performed over a test tweet based on currently fetched data from twitter over various emotional quotients. Comparison between two users based on public reaction in the form of likes, shares and number of retweets. Visualization of comparison results by plotting graphs over popularity parameters of social media(likes/retweets/shares).

For example, by analyzing the likes, shares and retweets of the tweets of two Presidential candidates Donald J Trump and Hilary Clinton and plotting a graph for better and convenient visualization of results.

4. TECHNOLOGIES AND CONCEPTS

4.1 Data collection

Data Collection is the first step and the main aim of data collection is to extract and store the required data into some meaningful structure from a very large sets of data. This can be also be classified as a data mining. The data collection should be efficient because this extracted data would serve all the post collection purpose such as cleaning, analysing more meaningful data is collected the more realistic will be the result.

4.2Data Cleaning

Once the required data set is collected from twitter, the very next step is to clean the data, i.e. to make the data more useable and more practical, in this we will remove all the smileys, full stops, punctuation marks, unknown languages words and short forms. This helps in forming a more meaningful data sets and further store the data in the a structured form (CSV file).

4.3 Trend Visualization

By Choosing a certain keyword which may be related to a famous personality, an event, any product or a movie and then by using the public statements of the user from twitter we will be analysing the trends and overall impact either positive or negative will be concluded from that by Applying tools Such as NLTK(Text Blob).

3.4 TWEEPY(Twitter API)

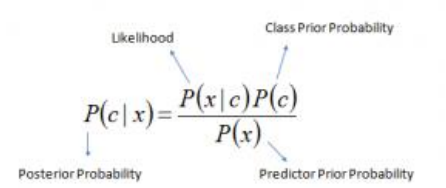
Twitter provides the way to extract the tweets by providing various API. Tweepy is the API of twitter which is developed for the python , this API class provides access to the entire twitter RESTful API methods. Each method can accept various parameters and return responses. This is used to stream the tweets of users from the twitter platform. When we invoke an API method most of the time returned back to us will be a Tweepy model class instance. This will contain the data returned from Twitter which we can then use inside our application. To acess the data we need certain keys and tokens named as (consumer key, consumer secret, Oauth_token, Oauth_secret) these needs to be passed as parameters while connecting to twitter before streaming tweets.

Consumer key ZB9n4VH2Jo1uPjqzOOsTwjWQy
Consumer secret Gi64lT7ZhTXwVdv40NJ9KD2It2VLUdV9TUBBb4KbOTSLL4RTem
OAUTH_TOKEN 318953400-DRwSuIiJI1BNr2ZsiaXzeQ3djd2MXwAie51Gdb7k
OAUTH_SECRET 6maOnmfE4FBpjwTvc6Mls5V5Qi4bEz9ltx4dzuWaQDxxw

3.4 Naïve Bayes Algorithm

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a one but a collection of algorithms which all works on same principle, i.e. every pair of features being classified is independent of each other.Naive Bayes model is quite easy to implement and works efficiently when the data set is large. The thing that makes it favourite is its less complexity and usefulness.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). [5]



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

3.4.1 Applications of Naive Bayes Algorithms

- **Real time Prediction:** Naïve Bayes is fast and easy algorithm and can be used in real time.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classification are generally used for text classification, so this makes it easier for us to apply this for spam filtering and sentiment analysis.
- **Recommendation System:** Naive Bayes Classifier builds a Recommendation System that uses machine learning and data mining techniques to predict and use the unseen pattern in the data.

5. ALGORITHM

4.1 Module 1. Opinion mining (sentiment analysis)

4.1.1Data collection

i. Get a twitter API and download Tweepy to access the twitter API through python
ii. Download twitter tweet data depending on a key word search "happy" or "sad"

4.1.2 Data representation and cleaning

iii. Format my tweets so that no capitalization, punctuation, or non ascii characters are present, as well as splitting the tweet into an array holding each word in a separate holder

iv. Create a bag of common words that appear in my tweets

4.1.3 Analysis using Naïve Bayes algorithm

v. Create a frequency table of words that have positive and negative hits

vi. Test my frequency table by using test sentences

4.2 Module 2: trend Analysis and visualization

4.2.1 Analysis to compare between two entities

i. Collection of the data using tweepy.

ii. Comparison between entities according to popularity indices of twitter(retweets, likes).

iii. Visualization of results in form of graphs using matplotlib library.

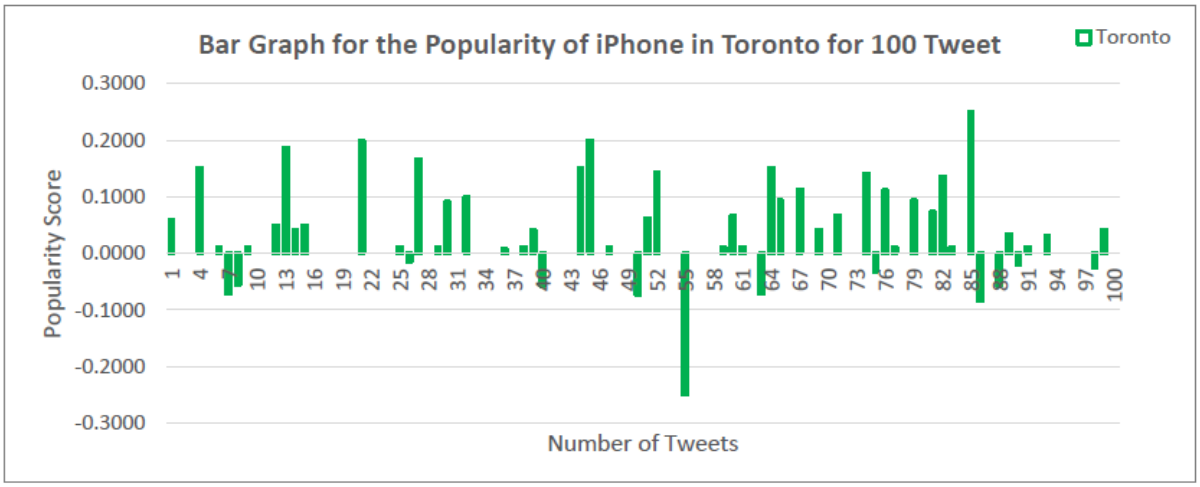


Figure 1. *Bar graph for the popularity of iPhone in Toronto for 100 tweets*

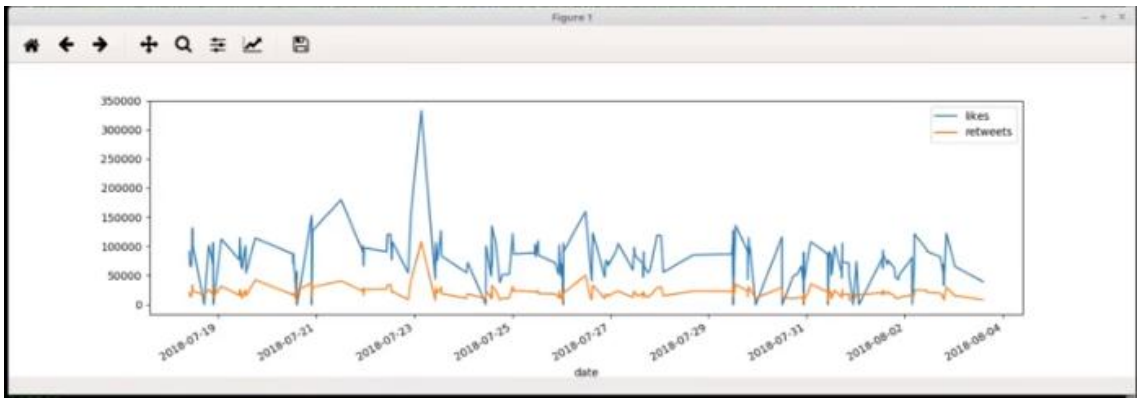


Figure 2. *Graph for likes vs retweets in a time series*

6. CONCLUSION

The emergence of social networking sites brings new ideas for data mining. People are sharing their opinions on many topics through microblogging services. Social media allows to access and analyse these opinionative messages. The work being done on the topic is vastly narrow and only addresses the issue of USER sentimental analysis and not SNA[6]. Incorporating this will be the next step in achieving better results. Also, better incorporation with social Networking sites and other Facilities and supposed android devices can help our program to achieve a more far-reaching experience. [7]

These processes consisted of data mining, extraction, visualization and decision making all of which were challenging tasks.

REFERENCES

[1]C. Aggarwal, "An Introduction to Social Network Data Analytics in Social Network Data Analytics", New York: Kluwer Academic, January 2011.

[2] Vedanayaki, M.A Study of Data Mining and Social Network Analysis

[3] A. K. Jose, N. Bhatia, and S. Krishna, "Twitter Sentiment Analysis". National Institute of TechnologyCalicut,2010.

[4] Bogdan Batrinca , Philip C. Treleaven, Social media analytics: a survey of techniques, tools and platforms.

[5] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[6] M.Rambocas, and J. Gama, "Marketing Research: The Role of Sentiment Analysis". The 5th SNA-KDD Workshop'11. University of Porto, 2013.

[7] J. Spencer and G. Uchyigit, "Sentiment or: Sentiment Analysis of Twitter Data," Second Joint Conference on Lexicon and Computational Semantics. Brighton : University of Brighton, 2008