Opinion Mining and Trend Analysis on Twitter Data

Anuj Kumar, Hoshiyar S Kanyal, Shivani Sharma, Kaushal Singh, Ayushi Dwivedi

Abstract: ***With rise in Internet use across the global, there has been a trending increase in the online data. Every person from different profession gives their view from politics to entertainment, sports or economics. The web's current evolution is a major pacesetter as it generates an effectual methodology to embed "smart data" into web pages and hence result in easy content implementation for authors. The web 2.0 has changed the way of communication on the web. Using Social Networks (SNs) they have become active participants by connecting, producing and sharing information, experiences and opinions with each other [1]. Public opinions extracted in the form of trends are interesting for researchers, sociologists, news reporters, marketing professionals and opinion tracking companies. 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 on currently fetched data from twitter over various emotional quotients. W have also performed (i) Comparison between two users based on public reaction in the form of likes, shares and number of re-tweets; (ii) Visualization of comparison results by plotting graphs over popularity of social media (likes/re-tweets/shares).***

Keywords: Smart Data, Social Networks (SNs), Web 2.0, Opinion Mining, Binary Sentimental Analysis, Emotional Quotients.

I. INTRODUCTION

Twitter is an online social media website which was developed in 2006. Currently Twitter is in top three rating and is considered to be one of those online social media websites granting a micro-blog service to write message up to 140 characters at one time, which are typically not more than 30 words. As of September 2013, Twitter has 645 million users who produced about 300 billion tweets and more than 143,199 Tweets are tweeted (i.e. transmitted or delivered) per second. Twitter uses several character symbols for performing operation. For instance, '@USER_ID' transmits direct messages, 'RT' for performing re-tweet [5]. To enforce a category or a topic discussion '#' (hash tag) is widely used. Approximately 98% users do not have enough followers.Some of the tweeter users have millions of followers but they are very few in number (as in year 2010). These users include media stars, politicians or popular news web sites (CNN, newspaper websites etc.)

Revised Manuscript Received on April 21, 2020.

***** Correspondence Author

***Anuj Kumar,** Assistant professor, Department of CSE, Hi-Tech Institute of Engg & Technology,Ghaziabad,India.Email: ap@hiet.org

Dr. Hoshiyar Singh Kanyal*, Associate Professor & HOD, Department of CSE, Hi-Tech Institute of Engg & Technology,Ghaziabad,India.Email: hkanyal1@rediffmail.com

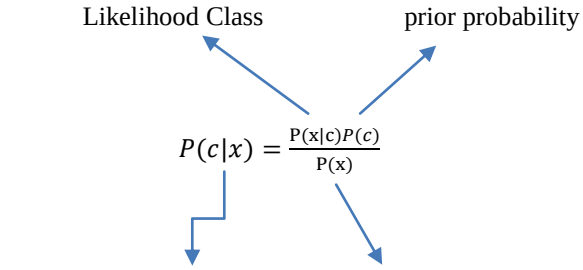
Shivani Sharma B.Tech. (CSE), MAIT, Ghaziabad , India Email ss6713899@gmailcom

Kaushal Singh, B.Tech. CSE), MAIT, Ghaziabad,India Email: kaushalsingh10a@gmail.com.

Ayushi Dwivedi, B.Tech. CSE), MAIT, Ghaziabad,India Email: ayushi.dwivedi00@gmail.com

II. PROPOSED METHODOLOGY

A. Naïve Bayes Algorithm



$P(c/x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n/c) \times P(c)$

Above,

- P(c/x) is the prior posterior probability of class (c, target) given predicator (x, attributes).
- P(c) is the prior probability of class.
- P(x/c) is the likelihood which is the probability of predicator given class.
- P(x) is the prior probability of predict.

B. Machine Learning

Computer Science came out with an emerging field in which the computers could learn without being explicitly programmed. The term "Machine Learning" was coined in 1959 by Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence. There are several challenges with Machine Learning. It is employed in a range of computing tasks [1] where designing and programming explicit algorithm with good performance is not feasible at times. For example, applications detection of network intruders or malicious network that is working towards a data breach, optical character recognition (OCR), rank learning, and computer vision. Machine learning has several advantages that include strong bonding with mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning has a unique feature of being unsupervised, thereby using it to learn and establish baseline behavioural profiles for various entities and then finding meaningful anomalies.

B. Data Mining

Data Mining is a very efficient and critical process of data pattern extraction using smart techniques. The core objective of data mining is to extract raw data from a data set and transforming it to a meaningful composition for future use.

The interesting fact is that Data Mining is the misnomer, as the major aim is pattern & knowledge extraction [2] from humungous data and not just mining the data itself.
There are six common classes of tasks involved in Data mining namely:

1. Anomaly detection (outlier/change/deviation detection).
2. Association rule learning (dependency modelling).
3. Clustering.
4. Classification.
5. Regression.
6. Summarization.

D. Twitter API (TWEEPY)

Twitter allows users to interact with its data using Twitter [6] APIs. Twitter has three APIs (Rest, Search and Stream).

- Rest API: provides simple interfaces for most Twitter functionalities
- Search API: part of Twitter's v1.1 REST API
- Stream API: family of powerful real-time APIs for Tweets and other social events [2]

For efficient data extraction server side programming/scripting language such as PHP, RUBY or PYTHON is needed. An Oath-authentication is required to access data from twitter. Four authentication parameters (consumer key, consumer secret, OAUTH_TOKEN, OAUTH_SECRET) are needed to enable API call. These keys look like following:

Consumer key "ZB9n4VH2Jo1uPjqzOOTwjWQy"

Consumer secret "Gi64lT7ZhTXwVdv40NJ9KD2It2VLUdV9TUBBb4KbOTSLL4RTem"

OAUTH_TOKEN "318953400-DRwSuIiJI1BNr2ZsiaXzeQ3djd2MXwAie51Gdb7k"

OAUTH_SECRET "6maOnmfE4FBpjwTvc6Mls5V5Qi4bEz9ltx4dzuWaQDxxw"

Once request is sent to twitter API with proper authentication it will deliver results in JSON (JavaScript Object Notation) format. JSON format is very popular and can easily be read by any program.

E. PYCHARM IDE

PYCHARM is one of the Integrated Development Environment (IDE) used especially for Python language in computer programming and is developed by the Czech company JETBRAINS.

Some of the exceptional features of PYCHARM include:

- Cross-platform with Windows, MACOS and Linux versions.
- Code Analysis
- Graphical Debugging
- An Integrated Unit Testing
- Integration with Version Control Systems (VCSs).

PYCHARM has released its Community Edition under the Apache License and a Professional Edition with extra features under a proprietary license [3].

B. Algorithms of Opinion Mining

1. Data 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".

2. Data representation and cleaning

iii. Format my tweets so that no capitalization, punctuation, or no 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.

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. Word parsing

Word Parsing is a process to split a sentence or text into parts by analysing logical syntactic components followed by the formal grammatical rules of the English Language [4].

Traditional grammatical sentence parsing usually emphasizes on subject and predicate depending on the exact meaning of sentence.

Algorithm: Word Parsing

Input: Cleaned Tweet String

Output: Word Array

1. Function df.to__csv(File name.csv, header=true, index=true, encoding="UTF-8")

2. Begin Word Parsing

3. It will generate an array of tweets from data frames.

5. Sentiment Analysis

It can be done by pattern matching or applying function for closest match. Sentiment classification [6] can be done by defining two values to each word either in positive manner or in negative manner based on the words provided by the frequency table.

C. Flow Chart

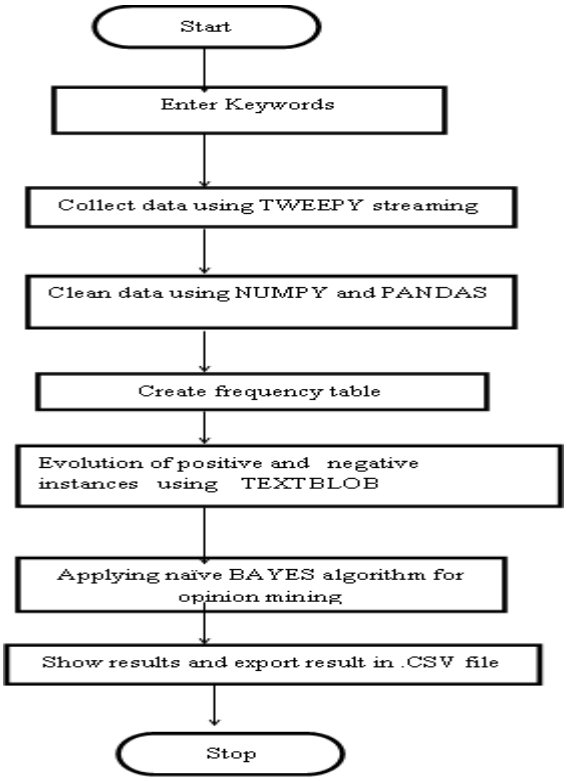


Fig 1. Flow chart of Opinion Mining Module Comparison between two entities.

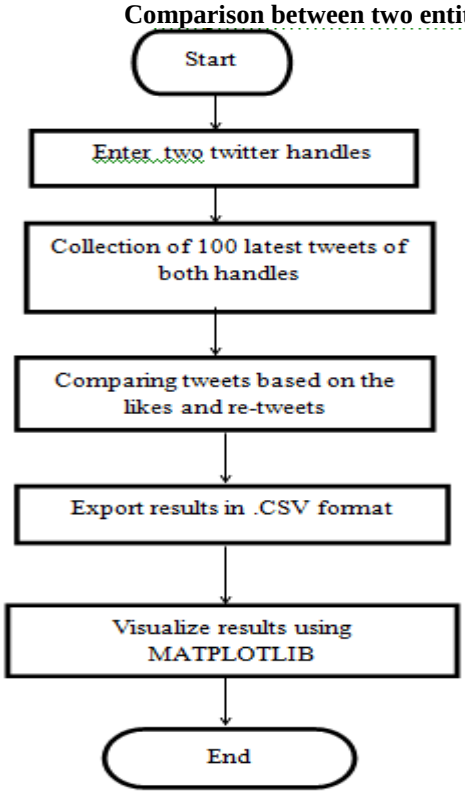


Fig 2. Flowchart of Entity Comparison Module.

i. Collection of the data using TWEEPY.

ii. Comparison between entities according to popularity indices of twitter (re-tweets, likes).

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

III RESULT ANALYSIS

we can see tweets extracted from the twitter in real time on the basis of last being tweet related to that searched keyword. Here Numerical values after each Current saved status shows the serial no of tweet stored.

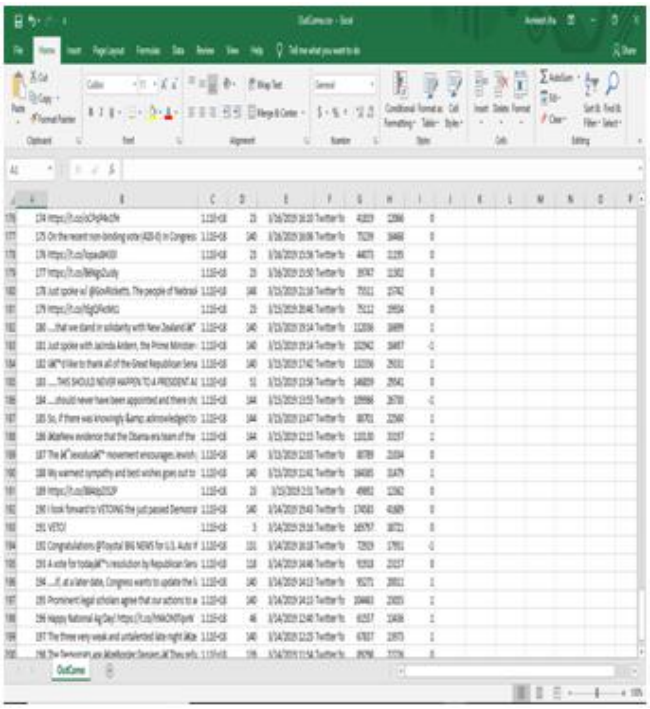


Fig:3 Fetching data based on user name then

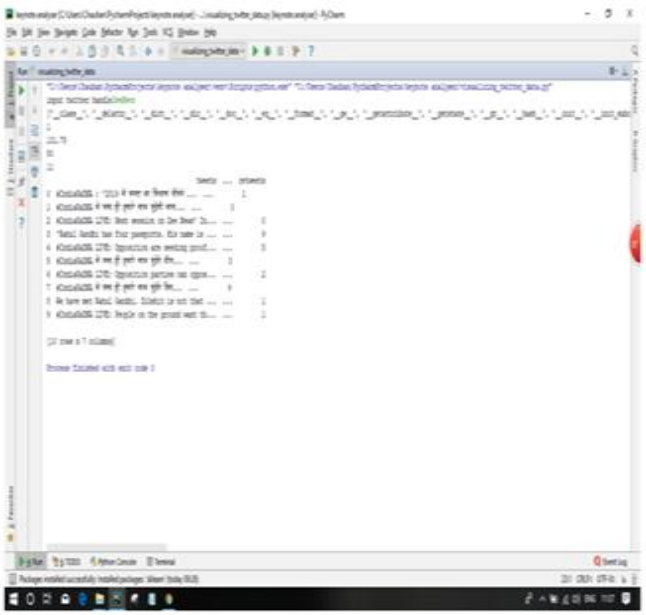


Fig: 4 Output of opinion Mining

Here the final outcome shown in terms of numeric values - 1,0,1 shows negative, neutral and positive aspects of tweets. We can see the user review in multiple system at the same time. We can search the data using user name and #tag.

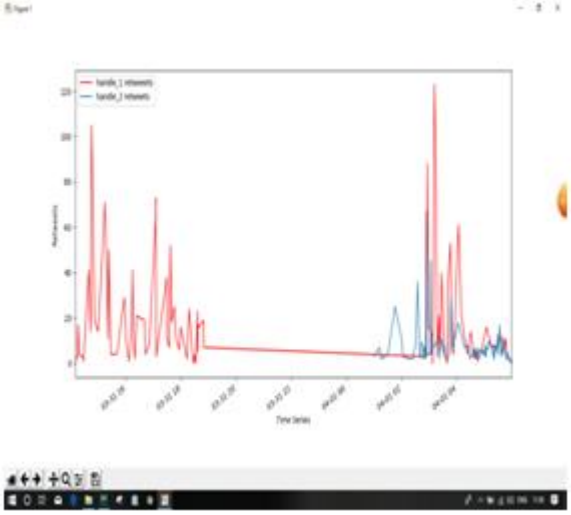


Fig:5 User 1 vs User 2 graph (on the basis of retweets)

V CONCLUSION

Social Networking Sites are rapidly emerging in today's digital era and are the major cause for innovative ideas like Data Mining. The work done so far only targets User Sentiment Analysis and not Social network analysis (SNA), therefore we aim at incorporating this analysis in order to achieve better and more proficient output. We believe that an enhanced inclusion of Social Networking Sites, other necessary facilities and supposed android supported devices would aid our program to achieve even more influential experiences.

V Future Work

At present, only Twitter has been used as a data source which limits the scope of this project and its application. Our future research aims at developing a framework integrating all the popular social networking sites like Facebook and YouTube. The limitation of the current system is that for user privacy, Twitter users do not share their geographical location; neither the Twitter APIs allow location information access such as server origin. This limitation can be handled by developing and integrating content based algorithm in order to determine user's location. This algorithm will examine user's profile, geographic region, hierarchical location, home location, travel location, city location, time zone, and zip code or postal code to predict the user's location.

REFERNCES

1. Malhar Anjaria and Ram Mohana Reddy Guddeti, " Influence Factor based opinion mining of Twitter data using supervised learning" sixth IEEE conference on COMSNETS, 6-10 Jan 2014, Bangalore, India.
2. Tuan Anh Hoang, William w Cohen, Ee-Peng Lim, Dovy Pierce, David R Redlawsk, "Politics, Sharing and emotion in Microblogs", IEEE/ACM conference on ASONAM, 2013, New York, USA.
3. D. Terrana, A. Augello, and G. Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream," in Semantic Computing (ICSC), 2014 IEEE International Conference on, 2014, pp. 128-134
4. A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text Categorization," Technimetrics, vol. 49, pp. 291-304, 2007.
5. https://en.wikipedia.org/wiki/Twitter.
6. D. N. V. S. L. S. Indira, R. Kiran Kumar, G. V. S. N. R. V. Prasad, R. Usha Rani. "Chapter 15 Detection and Classification of Trendy Topics for Recommendation Based on Twitter Data on Different Genre" , Springer Science and Business Media LLC, 2019

AUTHORS PROFILE

Anuj Kumar is currently working as a Assistant Professor and in Computer Science and Engineering departmrnt at Hi- Tech Institute of Engineering and Technology, Ghaziabad (India). He has 14 years of teaching experience. He published more than 10 research papers in different National and International journals. His area of interest is Data Mining, Data Structure, Data warehouse,Cloud computing and programming in C , C++ and Java. E-mail id: ap@hiet.org

Dr. Hoshiyar Singh Kanyal is currently working as a Associate Professor and Head of the Department of Computer Science and Engineering at Hi- Tech Institute of Engineering and Technology, Ghaziabad (India). He has more than 15 years of teaching experience. He published more than 15 research papers in different National and International journals. His area of interest is networking, Database, Data Mining and programming in C and C++. E-mail id: hsk@hiet.org

Shivani Sharma, is currently pursuing Bachelors of degree in Computer science and Engineering in Maharaja Agarsain Institute Of Technology,,Pilkhuwa, Ghaziabad, India. Her Area of interest is web development using PHP and Data base. She did course in Python and PHP. E-mail id: ss6713899@gmailcom

Kaushal, is currently pursuing Bachelors of degree in Computer science and Engineering in Maharaja Agarsain Institute Of Technology ,Pilkhuwa, Ghaziabad, India. Her Area of interest is web development using PHP and data structure. She did course in Python and PHP. E-mail id: kaushalsingh10@gmail.com.

Ayushi Dwivedi, is currently pursuing Bachelors of degree in Computer science and Engineering in Maharaja Agarsain Institute Of Technology ,Pilkhuwa, Ghaziabad, India. Her Area of interest is web deveplopment using PHP an data warehouse. She did course in Python and PHP. E-mailid: ayushi.dwivedi00@gmail.com