



Personality Prediction of User Behavior on Twitter

Sanjeev Dhawan¹, Kulvinder Singh², Priyanka Chauhan³

^{1,2,3}Department of Computer Science and Engineering
University Institute of Engineering and Technology (UIET),
Kurukshetra University, Kurukshetra-136119, Haryana, INDIA.

Abstract: Social media in this contemporary world has different influence on peoples all over the world. It becomes very sophisticated and learning information is very crucial for analyzing user behavior. In real life, posts on internet have many dissimilar effects on others and have different influence values. All posts have different effects on users. Finding the user's post effect on other user is becoming a very interesting point in this complex network. Mining and analyzing data can be difficult due to large amount of source involved and such activities are very expensive. As the usage of social sites such as Twitter have dramatically increased due to advancement in technology which provides the platform to millions of peoples to share their thoughts or opinions through tweets. A tweet is a basic way of expressing thoughts and is characteristic short. Sentiment analyses of Twitter data capture NLP (Natural Language Processing) and mining of data from content. It extracts the person thinking and way of expressing their opinions or thoughts through a particular tweet. This research on sentiment analysis on Twitter data must perform using various attribute. This paper show sentiment analysis techniques with types use to carry out origin of sentiments using tweets and relative learning of special approaches and techniques of sentimental analysis using Twitter data.

Keywords: natural language processing, sentiment analysis, social media, twitter, twitter data

I. Introduction

In current year's social network such as Twitter, Facebook, Instagram have turn out to be most well-liked intention for internet user. The activities of these users offer huge raised area in the field of researches to understand and study their online opinions of behavior personalities and preferences. Dissimilar type of personalities is connected towards development of different societal affairs and interaction behaviors on status preferences and profile. This study predicts behavior based on users social activities on Twitter. Twitter is a gold mine of data. It emerged a major micro-blogging website contrasting other social platforms. Twitter is a real-time service through which users forms connections among themselves in the process of sharing their thoughts with the help of text, images, videos and other digital artifacts.[1] With large audience, Twitter constantly attracts users to express the influences of person about an existing issue, or any other topic of interest. It becomes a major source of information. This information is very useful for individuals, government and businesses purposes. On Twitter, user is allowable to express his/her opinions with short term messages in the form of "tweets", are extremely short in comparison to regular blog posts. Every tweet is done with just 280 characters. Twitter is a popular and large social networking micro-blogging site. This will help peoples to lead compact for their statement by means of abbreviations, slangs, emotions etc. Through this, peoples easily express their view and opinions by using polysemy and sarcasm.[22] Due to easily availability of useful sources of information users shares their personal views and opinions on different subjects or topics. These may be related to current affairs, business, politics, sports and many more. Out of all other social media, choosing Twitter because 1)small length of tweets, thus less ambiguous 2)easily accessible via API 3)less source, having more information 4)unbiased 5)from various socio-cultural domains.[22]Twitter API is further easily reached and open platform in social media and becomes easy to retrieve the Twitter data as major incidents, news stories, events etc. Due to this feature twitter becomes more sympathetic to developers creating apps and tools to repossess data. Sentiment Analysis is used to extract the sentiments from tweets. This can be done in many fields similar to monitoring along with analyzing change of sentiments regarding a particular brand, events, products, government policies etc. Users easily express their opinion in the obverse of world in different form like multimedia texts. Twitter is used in predicting the behavior of the users regarding

tweets and identifying relevant user thoughts. This paper studies about the twitter sentiments analysis of existing systems to do sentiments. Many researchers work on twitter and still it is hot topic and find out how to classify tweets and analyze the behavior of users, how they respond on other's users posts. This paper has been discussing in five sections: Section II describe about sentiment analysis and its levels. Section III surveys the literature review on Sentiment analysis. Section IV elaborates on the analysis on twitter data and algorithm. Section V discusses the comparative case study of techniques having Twitter dataset. Section VI concludes with conclusion and the prospect work

II. Sentiment Analysis

Sentiment analysis is a method of computationally recognize and categorize views from section of texts and decide whether writer's thoughts towards an exact topic, tweets or product is optimistic, pessimistic or neutral. Sentimental analysis is a sorting technique which systematically identifies the opinions or views from tweets, formulates sentiments and then sentiments classification is done. This data science technique that centers around building systems that can be ratifying and extract opinions within text by usually besides just a ratifying opinions these systems can be take out attributes of appearance like polarization meaning if presenter is expressing a positive, negative opinion since publically or privately available data over the internet going exponentially huge amount of text expressing an opinion are available over few sites forums, blogs, social media. Sentimental analysis is help to grasp this un-structural information into structural data on public opinions available on products and brands. All about text information to be categorized either fact or opinion. Sentiment analysis can be moral as a categorization crisis with two sub problems need to be used off. However, no algorithm can give 100% accuracy on prediction on sentimental analysis. Sentiment analysis on twitter data may be depend upon document level and sentence level or even sub sentence level. The basic steps for performing sentiment analysis include data pre-processing, data gathering, feature extraction, sentiment detection and performing classification using machine learning or straightforward computational approaches.

Analysis levels: It is categorize in three different levels:

1. *Document Level:* It classify whether entire text give optimistic or pessimistic sentiment. Text is on particular topic is measured, which includes proportional learning cannot be measured below document level.
2. *Sentence Level:* It decides that every sentence represents estimation into optimistic, pessimistic or neutral. If there is any sentence which does not present any estimation then its mean neutral. It is correlated toward subjectivity categorization. That expresses accurate information from sentence that gives subjective features and opinions specifically, good or bad conditions.
3. *Entity Level:* This level gives all over analysis. It was previous known as feature level. It works on idea to facilitate estimation reside on thoughts and intention. Main task of this level is to do identification Prediction [29]

III. Related Work

An emergent amount of researches associated towards user's behavior in social networks and lately concerned additional awareness in the international research society. Personality detection is studied by two main disciplines: computational linguistics and Analysis of Social Network. Golbeck *et al.* [2] stated in their research that they were the first ones to look at the relationship between personality traits and social profile statistics. They created a Twitter application through which they undertook recent 2000 tweets of fifty subjects. The subjects were presented with the 45 question version of the Five Big Personality Inventory. The collection of tweets was processed with the help of two tools, first LIWC (Linguistic Inquiry and Word Count) from which they were capable to pull out a total of 79 features. Second, MRC Database which yielded 14 language features. Lee and Hwang,[4] gave the correlation between the GPS coordination of tweets and profile location on twitter and use text base federation method and apply on spatial attributes to cluster. Through this, estimation of locality of objective incident on twitter. Result shows that 50% user post their tweets in profile location and 30% users do not have locations on their tweets posted on their profiles. Furthermore, Lehmann *et al.*[5] examine how user react on received stimulus (tweet) from a friend, and explained how the 'principle of least effort' plays important role in users behavior of re-tweeting. Users seek less attention on tweets and only observed the information when it is most visible when it was on the top of Twitter stream. Specifically, capacity enumerates how a user limited interest is separated with received tweets. The study point out limited capability for development received information constrain societal contamination, and fulfilled fast decay of visibility becomes main obstacle to spread information online and likewise, Rosenman *et al.* [6] illustrate number of feature regarding Twitter. The author's concerned that only 22% relations were common on Twitter. Peoples frequently attached with persons through small intent of vigorously communicate among them. Complicated tweets seek more attention than the simple one. The structure of tweet gets more attention and liked by users and gets handful attention, most important to extensive authority of rule allocation. In addition of analysis, Lima *et al.*[7] proposed personality traits prediction in text groups and extended the problem of personality prediction into a multi label classification problem. This is a novel approach as an individual may possess more than one personality traits. The authors used the Naïve Bayes Algorithm to analyze tweets and named their model as Bayesian Personality Predictor .This approach was divided into three

steps i.e preprocessing, transformation and classification, where in preprocessing certain attributes were extracted from the tweets and then in the second phase multi label sets were mapped training sets of five single labels. Finally, third phase semi supervised classification takes place with the help of these training sets and meta attributes as input to the classifier. Similarly, Motamedi *et al.*[8] conducted user behavior using three metrics such as user connectivity, user activity and user reactions. They basically used data mining and statistical approach to measure group level user behavior. Moreover, Roja *et al.*[9] in his research applied graph theory, language dispensation, and information speculation give a top-down plan of grouping dynamics to study the behavior of users on social news websites and also Edoardo *et al.*[13] provided complete review of quantitative method foretell the activities of terrorist group. Then applied game theoretic method to verify the rules which predict the user's behavior. In addition Matsutani *et al.*[15] addresses the trouble of predicting faith relations. These links created between the active users of article analysis sites. They adopted NMF (non negative matrix) approach for solving these problems. Users in item review sites can post a review along with ranking on different items inside particular sets can be done by using NMF for trust links prediction and the collaboration approach to suggest places to customer was used by Hea *et al.*[16] and projected the recommendation systems to collect information regarding users and study users behaviors based on achieved information. The user's interest was depending on variety of travelling and co-travelling. Similarly, Yu *et al.*[17] analyze the users data and adopted Apriori algorithm through which the frequent travel pattern of users is obtained. Through travel prototype it becomes easy to understand the user behavior and interest which in turn is used for suggesting travelling plans completely depend upon user behavior and interest. Travel sight composed from travel blogs and trendy travel websites. They gather them mutually based on similarity and besides this Gupta *et al.*[19] described that the regular summarization played essential role in information recovery method and document processing method. The feature of information extraction saves the time of text readers. Through this the useful information from text extracted and rest are removed. Therefore, these methods offer only valuable information following testing. Similarly, Amolik *et al.* [20] proposed other accurate representation for analysis. Twitter information is regarding evaluation of upcoming Bollywood and Hollywood shows using features vector, classifiers for example Naive- Bayes along with SVM(simple vector machine), the classification of tweets will be correct. Naive- Bayes have improved exactness rather than SVM, but have a little inferior exactness and recall. If the training data is increase then precision of classifier will also increase. A new approach for sentiment analysis was described by Sawhney *et al.*[21] and examined the government strategy like the demonetization with the help of normal person opinion using Twitter data. Tweets are collect with assured symbols. Analysis is done base upon geo-location (collection of tweets is done according state wise). Sentiment analysis API is use from significance cloud moreover classify the state in six categories. In a survey, Shigenaka *et al.*[23] focuses on the novel system that leverage user contribution image contents for predicting the location category that satisfied vendor would stopover regularly and frequently. Visit behavior would help the customer relation management and market survey. This is due to check -in locations on the posts of user profile. This will be helpful in the business and marketing A hybrid and novel approach was introduced by Ajantha *et al.* [24] to identify and represent the relation between customer and there point of interest in travelling destination using customer spot vector. Using this vector generate 65%-100% efficiency rather than accessible method in ranking and prediction. Moreover, Mortanch *et al.*[25] said that content summarization has a variety of approaches and classified abstractive approach and extractive approach. The authors offered inclusive evaluation of drawing out which is depend on content summarization. It was classified in two approaches i.e. supervised learning and unsupervised learning and Gulati *et al.* [26] expressed that there is requirement of text summarization when individual need real meaning of particular matter slightly from one source of information available on net. The classifications carry out standard accuracy of 73% in excess of various Hindi papers. Likewise, Mungen *et al.*[27] method focuses on motif based post. This post is used to calculate and analyze the users effective posts. Calculating influence of a user is more correct than the post and it depends upon the user behavior during posting. Additionally, Bhargava *et al.* [28] presented the usage of sentiment analysis. The author supported the English language which acts important role for the investigation in this field and planned an approach through which different languages break down and explore sentiments. This approach was inclined by various techniques of machine learning.

IV. Sentiment Analysis on Twitter Data

Sentiment analysis workflow figure 1 as mentioned below. The structure collects user information from twitter, and point of interest. Information like user id, tweets, likes. This method contains four key modules: data collection of data, processing of data, classification and output analysis.

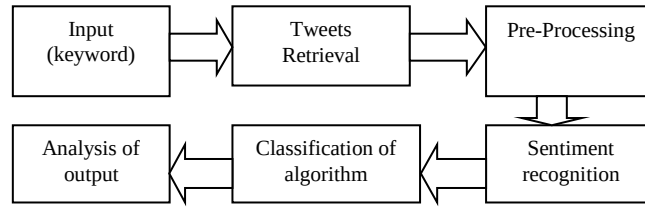


Fig.1 Sentiment Analysis Workflow

a) Input (keywords): This is the first step done by choosing a particular subject collecting tweets with keywords and then analysis on tweets is performed. The information such as user id, posts, likes, comments, emotions.
 b) Tweets Retrieval. Tweets can be collect by using programming languages such as python and many more.
 c) Data Preprocessing: In this step, filtration of required user data is done for removing inconsistent and incomplete noisy data. Here, following steps are involved:

- Removing special characters, URL's
- Removing stop words
- Removing re-tweets
- Tokenization
- Normalization

d) Sentiment detection: this step classifying the polarity of given tweets feature. The polarity may be optimistic, pessimistic and neutral. The recognition of polarity is completed by various lexicons like SentiWordNet, Sentiment Lexicon etc. It will help in calculating sentiment score, strength and many more. [18]

e) Classification algorithm: The classification of Twitter data is through by supervised learning. The effectiveness of the classifier built upon which dataset used for which categorization method. Supervised learning approach can be used to guide categorization method. [18]

f) Analysis output: Crucial opinion of sentiment analysis is revolutionizing unstructured data into important data. After completion of analysis, outcomes are display on graphs, bar charts, line charts, pie charts etc.

Algorithm: To build a sentiment analysis of Twitter data to show how might integrate an algorithm like this into applications. First start by choosing a topic, then gathering tweets with that keyword and perform sentiment analysis on those tweets.

Step 1: Gather Tweets First, choose a topic want to analyze. Inside sentiment-analysis, input can be defined Whatever phrase you like and using a word expect to return positive, negative or neutral results.

Step 2: Provide the name of a public profile from whom tweets to be extracted

Step 3: Perform Sentiment Analysis on Tweets. After gathering and cleaning data set, it is ready to execute the sentiment analysis algorithm on each tweet. Then, calculate an average score for all tweets combined.

Step 4: Calculating the percentage of particular tweets

Step 5: Visualizing the percentage of tweets through graph or pie chart

V. Comparative Case Study of Sentiment Analysis Techniques

Authors & Year	Datasets	Techniques	Accuracy
Apoorv aggarwal (2011)[3][29]	11,875 manually annotated tweets	Kernel Unigram Senti-features	73.93% 71.35% 71.27%
Sayed-Ali Bahrainian (2013)[10][29]	Twitter data on smartphones	Unigram features, Naive Bayes, SVM	89.78%
Neetu, M.S., Rajasree (2013)[11][29]	Twitter posts about electronic items	SVM Naive Bayes MaximumEntropy	90% 89.5% 90%
Geetika, Gautam, Divakar (2014)[12][28]	Customer review Twitter dataset	SVM Maximum Entropy	85.5% 83.8%
Gurkhe, Dhiraj, Rishab (2014)[14][29]	Twitter data	Bigram Unigram	15% 81.2%

VI. Conclusion and Future Work

Twitter, micro-blogging site through which million peoples cooperate with everyone and share their feelings, emotions and express their opinions or views through posts, tweets. The analysis is prepared in special summit of views for extracting sentiments and opinions and defines various levels of sentimental and opinion mining, different methodologies and techniques of sentiment analysis are discussed, it is very important to be familiar with about Twitter, how to extract tweets, composition, its consequence and enhance the analysis of Twitter data for

specific topics to measure the views or opinions and behavior of users towards different categories. The essential required information of sentiment analysis using Twitter data is discussed and situates indication regarding origin of tweets, preprocessing and sentiment analysis. This paper shows accurateness be enhanced after semantic analysis is follow via machine learning techniques and predictive model has been suggested which can be utilized in understanding the behavior of users and a process which incorporate the recognition result of words on tweet sentiment classification and also prominence on how to preprocess the Twitter data for maximum information taking out from the small content.

References

- [1] <http://en.wikipedia.org/wiki/Twitter>.
- [2] Golbeck, Jennifer. "Predicting personality from twitter", Privacy Security, Risk and Trust (PASSAT) Third International Conference on Social Computing (Social Com), IEEE 2011.
- [3] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in socialmedia. Association for Computational Linguistics, 2011.
- [4] B.Lee and B.Hwang, "A study of the correlation between the spatial attributes on twitter". In eight international conference on ICDEW conference, 2012.
- [5] J. Lehmann, B. Gonc alves, J. J.Ramascio, and C. Cattuto, "Dynamical classes of collective attention in twitter", in proceeding of the 21st International World Wide Web Conference (WWW), 2012.
- [6] E. T. R. Rosenman, "Retweets —but not just re-tweets: Quantifying and predicting influence on twitter" bachelor's thesis, applied mathematics, Harvard College, Cambridge. 2012
- [7] Lima, Ana CES, and Leandro N. De Castro. "Multi-label Semi-supervised Classification Applied to Personality Prediction in Tweets." Computational Intelligence and 11th BrazilianConferences on Computational Intelligence (BRICS-CCI & CBIC), 2013.
- [8] Reza Motamedi, Roberto Gonzalez, Reza Farah bakhsh, Reza Rejaie, Angel Cuevas, and Ruben Cuevas"Characterizing group-level user behavior in major online social networks, Technical report available at ,<http://mirage.cs.uoregon.edu/pub/> CISTR-2013
- [9] Roja Bandari, Gestalt Computing and Study of Content-Oriented User behavior Doctoral Thesis University of California, USA, 2013
- [10] Bahrainian, Seyed-Ali, and Andreas Dengel."Sentiment analysis and summarization of twitter data."Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE, 2013.
- [11] Neethu, M.S, and R. Rajasree."Sentiment analysis in twitter using machine learning techniques."Computing,Communication and Networking Technologies (ICCCNT), Fourth International Conference, IEEE, 2013.
- [12] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh International Conference , IEEE, 2014.
- [13] Edoardo Serra and V. S. Subrahmanian, A Survey of Quantitative Models of Terror Group Behavior and an Analysis of Strategic Disclosure of Behavioral Models, IEEE Transactions on Computational Social Systems, vol.-1, pp.66-88, 2014
- [14] Gurkhe, Dhiraj, and Rishit Bhatia. "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification." (2014).
- [15] K. Matsutani, M.Kumano, M.Kimura, K.Saito, k.Ohara, H.Motoda, combining activity Evaluation information with NMF for Trust-link Prediction in Social Media, IEEE International Conference on Big Data(ICBD) pp.2263-2272, 2015.
- [16] Jiangning Hea, Hongyan Liua, "SocoTraveler: Travel package recommendations leveraging social influence of different relationship types.", Journal of Information and Management, 2016.
- [17] Zhiwen Yu, Huang Xu, Zhe Yang, Bin Guo, "Personalized Travel Package With Multi Point- of- Interest recommendation Based on Crowd Sourced User Footprints." IEEE Transactions On Human-machine Systems, Vol. 46, No.1,pp. 151 - 158, 2016.
- [18] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment Analysis of Events from Twitter using Open Source Tool." (2016).
- [19] Manisha Gupta, Dr. Naresh Kumar Garg, Text Summarization of Hindi Documents using Rule Based Approach, International Conference on Micro- Electronics and Telecommunication Engineering, vol. 8, pp. 1- 4, 2016
- [20] Amolik, Akshay,"Twitter sentiment Analysis of movie review using machine Learning Techniques." International Journal of Engineering and Technology, vol.7.6, pp. 1-7, 2016.
- [21] Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon."Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government."ICT Express, 2017.
- [22] Bhumika Gupta, Monika Negi, Kanika, Vishwakarma, Goldi Rawat, Priyanka Badhani, "Study of twitter sentiment analysis using machine learning algorithms on python", International journal l of computer application vol.165-no.9, 2017.
- [23] R.Shigenaka, Y.Chen, F.Chen, D.Joshi, Y.Tsuboshita, "Image-Based Profiling Of Frequent And Regular Venue Categories".(ICME),pp.541-546, IEEE, 2017
- [24] Ajantha.D, J.Vijay, R.Sridhar, "A User Location Vector Based Approach for Personalized Tourism and Travel Recommendation." vol. 17, pp.440-446, IEEE, 2017.
- [25] N. Moratanch, S. Chitrakala, A Survey on Extractive Text Summarization. IEEE International Conference on Computer, Communication and signal Processing (ICCCSP), vol.8, pp.1-4, 2017.
- [26] Archana N.Gulati, Dr. S.D.Sawarkar, A novel Technique for Multi-document Hindi Text Summarization. International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017), vol. 8, pp. 1-4, 2017.
- [27] A.Mungen, M.Kaya," Influence Analysis of Posts In Social Networks By Using Quad- Motifs." pp.1-5, IEEE, 2017.
- [28] Shah Nawaz, Parmanand Astya, "Sentiment Analysis: Approaches and Open Issues." International Conference On Computing, Communication and Automation, vol. 9, pp. 1-5, 2017.
- [29] Rasika Wagh, Payal Punde," Survey on sentiment Analysis using twitter dataset." 2nd International Conference on Electronics, Communication and Aerospace Technology (ICEA) 2018.