Attitude Analysis through Twitter using Data Mining Techniques

Rishal Shah[*] Kruti Mody[†] Apoorva Kuckian[‡] Dr Radha Shankarmani[§]
[*]shahrishal@gmail.com [†]krutimody1995@gmail.com [‡]apurvakuckian@yahoo.co.in [§]radha_shankarmani@spit.ac.in
Sardar Patel Institute of Technology
Andheri(W), Mumbai

***Abstract*—Social Media is a place where users express themselves ranging from political opinion to fashion style. It makes one feel connected to a larger community and fosters conversations. Nowadays there are scads of social media websites like Twitter, Facebook, Instagram, etc and with thorough analysis of a user's social media accounts it is possible to gauge attitude of the user. In any organization the correct attitude of the staff members can prove to be effective in great progress. It is thus necessary for an organization to inculcate correct attitude within the staff members. There is certain level of attitude required for a particular designation, which we call the 'Ideal Attitude'. Social Media data pertaining to a specific social media website is obtained and is categorized into Big Five Personality Traits - Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (OCEAN). The attitude thus observed is termed as 'Observed Attitude'. Attitude analysis will prove to be very helpful for recruiters for hiring deserving candidates by comparing the ideal and observed attitude. This paper focuses on use of Attitude Analysis for recruitment.**

***Index Terms*—Attitude Analysis, Big Five Personality Traits (OCEAN, CANOE), Prediction, Twitter.**

I. Introduction

The growth of inter-net gave an endless variety of ways to its users to express their opinions, feelings and expression to a large heterogeneous audience. Hence it has provided an opportunity to study and harvest enormous data which is being produced rapidly at a continuous rate. An abundance of profiles are created by on-line users to connect with the universe and use the services of social media platforms. Twitter is one such platform wherein recent issues and subjects are discussed. At Twitter, 288 million active users per month express themselves with short informal text messages called tweets which amounts to 500 million tweets in a single day.[8]

In the professional world where recruiters find it difficult to choose the perfectly deserving candidate, the aspects they judge are: Knowledge, Skills and Attitude. Although it is possible to judge knowledge and skills through interviews, no method has been defined to analyse attitude of a person. Hence building a tool that analyses attitude was a requirement. Attitude has been previously analysed by questionnaires. But it is possible for a candidate to answer them in a certain way to falsely portray themselves so as to appear perfect for the particular position.

Attitude influences how people interact on-line. An individual tends to expresses his true attitude towards varied issues and discussions be it social reforms, fashion styles, politics, sports or entertainment. The relationship between real world social networks and personality has been usually studied using a personality test called ”The Big Five”. Different implementations have been proposed to predict personality of an individual based on twitter [1][2][3]. Also sentiment analysis and opinion mining techniques were used to predict personality [5]. Classifying the text by analysing sentiment of each word as positive, negative or neutral has already been researched and studied by researchers but the topic of recognition of detailed attitudes by analysing text has been ignored. Personality prediction is possible using sentiment analysis and social media. Similarly, the five factor model along with social media networks can also be used to study and analyse the attitude of an individual. Figure 1. summarizes the five factor model wherein the upward arrow indicates high score in that trait and downward arrow indicates lower score in that trait.

Consider a case where an organization wants to hire an employee for a specific position. The position requires the position holder to have a specific attitude. It requires the employee to be open, extrovert, agreeable, organized and lower emotional instability. So, the goal is to provide OCEAN value to an individual who interviews for the position. The values in each trait are assigned using his social media account with minimal error. On comparing the observed value with the ideal value (standards set by the organization) deciding whether to accept or reject the candidate is much simpler.

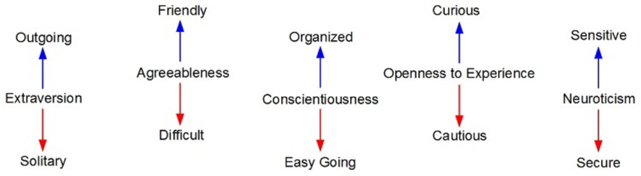


Fig. 1. Summary of Big Five Personality Traits[11]

II. Background and Related Work

A. The Big Five Traits

The ”Big Five” model has emerged as one of the most established and well-researched measure for defining human personality in recent years. It can be extended to define an individuals attitude and preferences. The Big Five Traits can be explained as:

- Openness: People having high degree of Openness like change, new reforms, appreciate new and unusual ideas, and have a good sense of aesthetics. Openness refers to degree of new thoughts, ability to imagine, amount of creativity, political liberalism, curiosity, tolerance and appreciation for culture.
- Conscientiousness: People with high on Conscientiousness level tends to be more organized and managed. Qualities include being organized, managed, reliable, and consistent. They have a planned approach towards life and thrive to pursue long-term goals.
- Extraversion: People with high extraversion enjoy other's company, and express positive emotions. They tend to be socially active, friendly, energetic, talkative and outgoing most of the times. They have a tendency to expand their friend circle and interact with more people. They like to be at the center of attention.
- Agreeableness: People with high Agreeableness maintain positive social relations, are friendly, agreeable, cooperative, and compassionate. People having high Agreeableness people tend to trust others easily. They are adaptive and supportive to others needs.
- Neuroticism: Neuroticism is a category which involves all the negative emotions. People with high Neuroticism experience emotions such as guilt, depression, anger, stress and anxiety. They experience mood swings and tend to have a spontaneous behavior. It is desirable to score low in this category, they tend to be more confident are calm by nature.

An individual is associated with five scores that correspond to the five main personality traits and that form the acronym of OCEAN [2].

B. *Attitude and Twitter*

Our personality can help define our attitudes toward specific things, situations, or people. According to the book *"Attitudes, Personality and Behaviour"* by *Icek Ajzen* just like personality, attitude is a hypothetical construct which is inaccessible by direct observations and must be inferred from measurable responses.There is no limit to the kinds of attitude and hence it is better if these attitudes are categorized under few heads. For the most parts the book emphasizes on similarities between personality and attitude hence it is possible to analyse attitude using twitter and big five traits just as it is possible to predict personality [1].

The relationship between the social media networks and personality has already been established. There are several online web portals, software, tools, etc which uses the Big Five Inventory to predict the personality of the subject under observation with minimal error rate. However, analysis of attitude using the textual information available on social media is yet to be done. To the best of our knowledge this is the first study looking at the relationship between Twitter and Attitude. However, there have been previous studies that relates personality to social media by using classification and regression methodologies for processing the textual data. The utility of Big Five Model can be enhanced to analyse the attitude along with predicting personality. According to [12] people with high extroversion and conscientiousness shows ease in use of social media websites [1]. Also, individuals with high agreeableness scores were selected more often as friends by people with similar agreeableness score [13].

III. Data Collection

A. *Web User Interface*

Python is used as a code base for all the processing. The server is switched on and the url *https://localhost:5000* is entered into the web browser. The interactive Web page created using Flask is displayed. Here, the subject enters his credentials. These credentials include their name, email id, Twitter handle. The credentials entered are stored into a database for future use.

B. *Extracting Tweets*

Tweepy is open-sourced API, hosted on GitHub. Communicating with Twitter platform using a python code base is available using tweepy. Along with tweepy, Twython is also used to extract twitter data which is a python library. In order to use tweepy, it is necessary to log in and obtain four unique credentials namely, *consumer key, consumer secret, access key and access secret.* These credentials are then used in python code base wherever interaction with live tweets is required. However, there is limitation of 3200 tweets which can be retrieved per user, at once. The tweets of the user under study thus obtained are stored in an excel sheet which can be further processed.

C. *Formation Of Dataset*

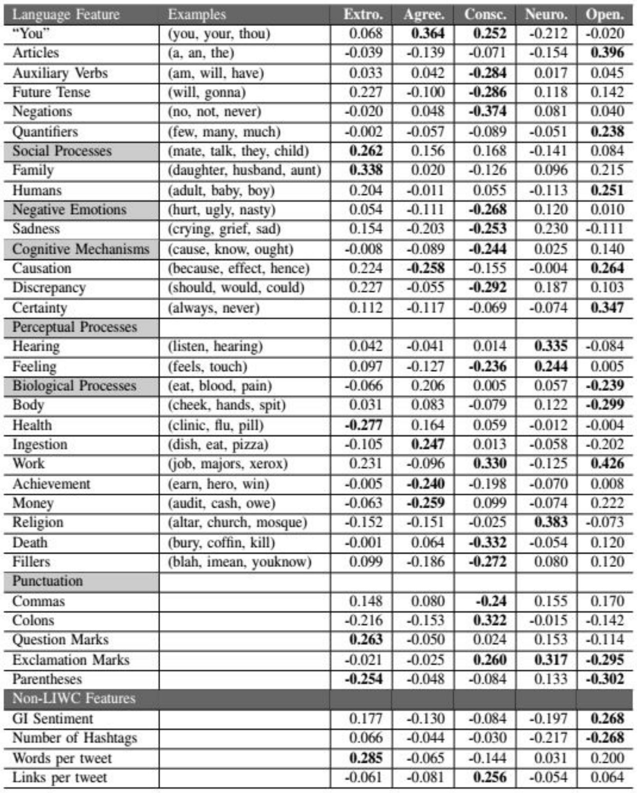
As publicly available data set proved to be inefficient and incomplete there was a need to create a new data set that could relate words obtained from twitter profile and OCEAN traits. Figure 2. shows the categories in which all the words in a dictionary fall and it's respective OCEAN value [1]. Each word is categorized among these categories and assigned these OCEAN values. Initially the number of words taken in data set are limited but with continuously running the algorithm the data set is trained and new words are added to the data set.

IV. Attitude Analysis

After collecting the tweets of the individual under study, the tweets are preprocessed. The data thus obtained is then subjected to the attitude analysis algorithm. the result thus obtained is represented in the form of kiviat or radar or spider chart.

A. *Preprocessing Tweets*

Once the user enters his credentials on the web page his tweets are stored in an Excel sheet *Twitter_Timeline.xlsx.* These tweets won't be long but will contain pictures, URL, twitter user handles, emoticons, etc. These tweets are filtered such that stop words like a, the, in, of, etc are removed. Stop words are those words which cannot have any sentiment and are basically just connectives. Also, the URL is replaced with a word *”URL”* and twitter user references are replaced with *”AT_USER”*. Excessive punctuation marks at the start and at the end of the tweets are removed.



Language Feature	Examples	Extro.	Agree.	Consc.	Neuro.	Open.
“You”	(you, your, thou)	0.068	**0.364**	**0.252**	-0.212	-0.020
Articles	(a, an, the)	-0.039	-0.139	-0.071	-0.154	**0.396**
Auxiliary Verbs	(am, will, have)	0.033	0.042	**-0.284**	0.017	0.045
Future Tense	(will, gonna)	0.227	-0.100	**-0.286**	0.118	0.142
Negations	(no, not, never)	-0.020	0.048	**-0.374**	0.081	0.040
Quantifiers	(few, many, much)	-0.002	-0.057	-0.089	-0.051	**0.238**
Social Processes	(mate, talk, they, child)	**0.262**	0.156	0.168	-0.141	0.084
Family	(daughter, husband, aunt)	**0.338**	0.020	-0.126	0.096	0.215
Humans	(adult, baby, boy)	0.204	-0.011	0.055	-0.113	**0.251**
Negative Emotions	(hurt, ugly, nasty)	0.054	-0.111	**-0.268**	0.120	0.010
Sadness	(crying, grief, sad)	0.154	-0.203	**-0.253**	0.230	-0.111
Cognitive Mechanisms	(cause, know, ought)	-0.008	-0.089	**-0.244**	0.025	0.140
Causation	(because, effect, hence)	0.224	**-0.258**	-0.155	-0.004	**0.264**
Discrepancy	(should, would, could)	0.227	-0.055	**-0.292**	0.187	0.103
Certainty	(always, never)	0.112	-0.117	-0.069	-0.074	**0.347**
Perceptual Processes						
Hearing	(listen, hearing)	0.042	-0.041	0.014	**0.335**	-0.084
Feeling	(feels, touch)	0.097	-0.127	**-0.236**	**0.244**	0.005
Biological Processes	(eat, blood, pain)	-0.066	0.206	0.005	0.057	**-0.239**
Body	(cheek, hands, spit)	0.031	0.083	-0.079	0.122	**-0.299**
Health	(clinic, flu, pill)	**-0.277**	0.164	0.059	-0.012	-0.004
Ingestion	(dish, eat, pizza)	-0.105	**0.247**	0.013	-0.058	-0.202
Work	(job, majors, xerox)	0.231	-0.096	**0.330**	-0.125	**0.426**
Achievement	(earn, hero, win)	-0.005	**-0.240**	-0.198	-0.070	0.008
Money	(audit, cash, owe)	-0.063	**-0.259**	0.099	-0.074	0.222
Religion	(altar, church, mosque)	-0.152	-0.151	-0.025	**0.383**	-0.073
Death	(bury, coffin, kill)	-0.001	0.064	**-0.332**	-0.054	0.120
Fillers	(blah, imean, youknow)	0.099	-0.186	**-0.272**	0.080	0.120
Punctuation						
Commas		0.148	0.080	**-0.24**	0.155	0.170
Colons		-0.216	-0.153	**0.322**	-0.015	-0.142
Question Marks		**0.263**	-0.050	0.024	0.153	-0.114
Exclamation Marks		-0.021	-0.025	**0.260**	**0.317**	**-0.295**
Parentheses		**-0.254**	-0.048	-0.084	0.133	**-0.302**
Non-LIWC Features						
GI Sentiment		0.177	-0.130	-0.084	-0.197	**0.268**
Number of Hashtags		0.066	-0.044	-0.030	-0.217	**-0.268**
Words per tweet		**0.285**	-0.065	-0.144	0.031	0.200
Links per tweet		-0.061	-0.081	**0.256**	-0.054	0.064

Fig. 2. Categories along with OCEAN values[1]

Now, after the tweets are filtered for stop words, URL and punctuation marks the resulting words obtained are stored to a vector array called feature vector. The words in feature vector account for some value in each OCEAN trait. These words are termed as Feature Words (FW).

B. OCEAN Values for Tweets

After formation of Feature Vector, each word is analysed. Each tweet has a separate feature vector. Let us consider a case where we have 5 tweets. So, we will have 5 feature vector each containing FW. Now, each tweet is processed as per the flow chart shown in Figure 3. FW is checked whether it is present in the data set or not. If it is present then the corresponding OCEAN values are inserted into the sentiment_list. If not, then FW is searched to have a synonym within the dictionary API. If synonym is found then the algorithm checks whether the synonym is present in the data set, if present in data set then the corresponding OCEAN value is added to sentiment_list. After adding the value of synonym the FW along with other remaining synonyms are inserted into the data set so that next time that word is encountered as FW it can be directly found from the data set and the need to invoke the dictionary API is eliminated. Now, if the synonym is also not found then the algorithm displays the error message of ’synonym not found’ and the FW is added to *buffer.txt* file. The words in *buffer.txt* are checked whether they are capable of having sentiment or not and are then manually inserted to the data set. For next tweet the process continues and the OCEAN values are appended in the sentiment_list such that length of the sentiment list will be five times the number of tweets evaluated. Here, since we considered 5 tweets the length of sentiment list will be 25 where 1st, 6th, 11th, 16th, 21st will correspond to the openness of tweet 1, 2, 3, 4 and 5; 2nd, 7th, 12th, 17th, 22nd will correspond to the conscientiousness of tweet 1, 2, 3, 4 and 5 and so on. In order to find the cumulative openness value for each tweet the weighted average is calculated. The openness value of each FW is multiplied by the weights assigned to the word.

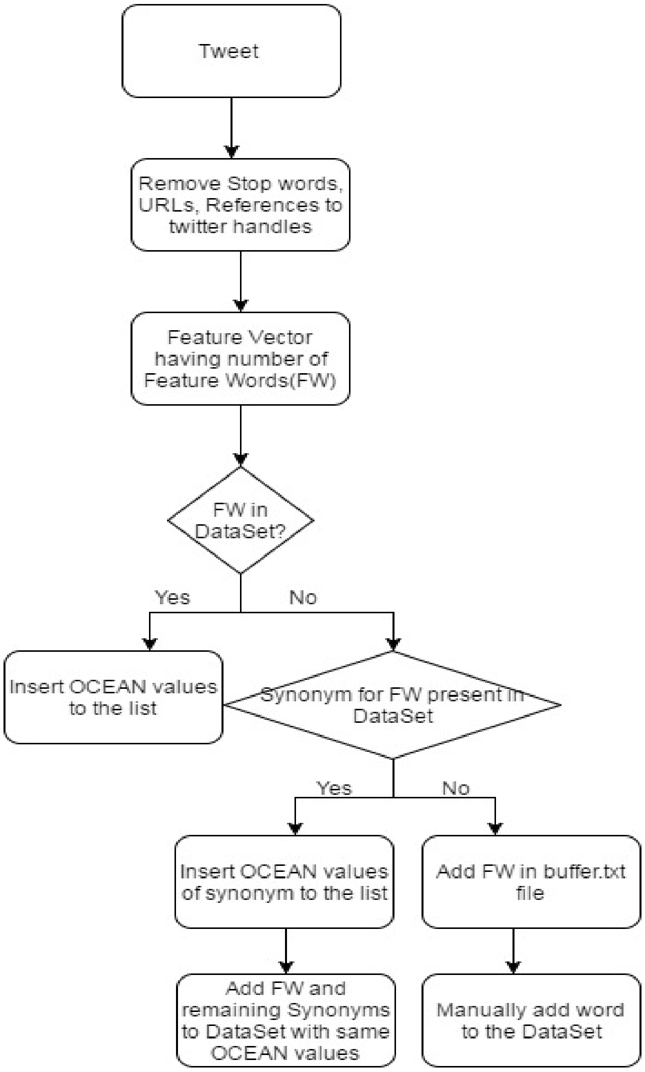

Tweet
Remove Stop words, URLs, References to twitter handles
Feature Vector having number of Feature Words(FW)
FW in DataSet?
Yes
No
Insert OCEAN values to the list
Synonym for FW present in DataSet
Yes
No
Insert OCEAN values of synonym to the list
Add FW in buffer.txt file
Add FW and remaining Synonyms to DataSet with same OCEAN values
Manually add word to the DataSet

Fig. 3. Flow chart to process each tweet

Weights are assigned to the words based on degree of comparision and level of significance. The value thus obtained is added and is divided by the sum of weights of each FW (equation (1)). Hence the 1st, 6th, 11th, 16th and 21st values are the weighted average openness value of each tweet.

$$Openness_{tweet} = \frac{\sum_{i=1}^{no.ofFW} openness_i * weight_i}{\sum_{i=1}^{no.ofFW} weight_i} \quad (1)$$

Just like openness the value for remaining traits is obtained similarly. These values are then averaged to find the final trait value of the individual under study.

C. Output Visualization

The kiviat or radar chart is obtained on the use interface just after he/she enters his credentials on the web page. The radar chart formed is made by using d3.js (Data Driven Documents, a java script library). The OCEAN values calculated are represented on the radar chart as shown in figure 4. The individual's observed attitude value is shown along with the ideally required attitude value in each of the Big Five Trait.

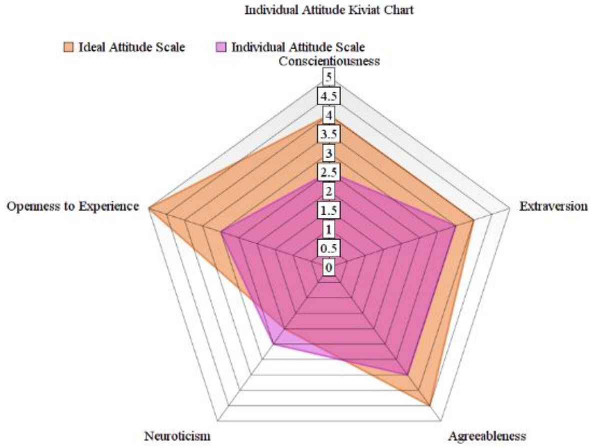


Fig. 4. Individual's Observed and Ideal Attitude Kiviat Chart

V. Result And Discussion

Users of social media reveal a lot about themselves, the question that arises from this research is how the results can be used. The kiviat chart can be studied to find out how close the *observed value* of the subject is to the *ideal value* defined by the organization which utilizes the tool for analysing the attitude. The difference between the observed and ideal value can be calculated and lesser the obtained value, greater is the closeness to the desired attitude. Considering the case previously listed, lesser the difference greater is the probability of acquiring the position offered by the organization. The person is more likely to be recruited if he/she fits perfectly to the expectations in terms of attitude. Organizations recruit employees based on skill set and by personal interview which lasts for maximum 1 hr. It is not possible to correctly determine the interviewee's attitude within few minutes or hours, that is when this research comes into picture. People easily acquire the skill set require for a position in an organization, what lacks is the attitude. What is the use of the skill set if you do not know where to apply it and how to apply it? Thus, with this research the person can work effectively so as to reach the ideal attitude standards along with the skill set thus helping him/her to excel. By knowing his attitude he/she will also know in which position does his attitude suits the best and discover his true potential.

VI. Conclusion

In this paper we have shown how big five traits - openness, conscientiousness, extraversion, agreeableness and neuroticism and twitter data of an individual can be used to analyse attitude of that individual. Organizations set a standardized measure required for a position and based on the result obtained after analysing the candidate's twitter account the recruiter will know how close the candidate's attitude is to ideal attitude. The recruiter will thus be as ease while making a decision to accept or reject the candidate. The way it connects to the bigger picture is that it saves the overall time and effort spent in the existing recruitment system. The researched can be further extended by categorizing based on gender. Big expanded organizations are always resourceful, so this research can be used for people who are less resourceful such as recruiting for startups where the recruitment of employee with correct attitude decides the progress of the startup. It can be used for finding a marriage partner, for hiring domestic help. Thus, this research has varied uses ranging from common man to mega organizations.

References

[1] Jennifer Golbeck, Cristina Robles, Michon Edmondson, Karen Turner, Predicting Personality from Twitter IEEE International Conference on Social Computing, 2011

[2] Daniele Querica, Michal Kosinski, David Stillwell, Jon Crowcroft, Our Twitter Profiles, Our Selves: Predicting Personality with Twitter 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011

[3] Ravikiran Janardhana, Twitter Sentiment Analysis and Opinion Mining, Department of Computer Science,university of North Carolina at Chapel Hill

[4] Margaret M. Bradely, Peter J. Lang, Affective Norms for English Words: Instruction Manual and Affective Ratings, Intelligent Sensors, Technical report C-1,The center for research in Psychophysiology, university of Florida, 1999

[5] Deepali Virmani, Vikrant Malhotra, Ridhi Tyagi, Sentiment Analysis Using Collaborated opinion Mining, NASA ADS , January 2014.

[6] Finn Arup Nielson, A new ANEW: Evaluation of a word list sentiment analysis in microblogs, Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings : 93-98. 2011 May

[7] Francois Mairesse, Marilyn A Walker, Matthias R. Mehl, and roger K Moore, Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text, in Journal of Artificial Intelligence Research 30 (2007) 457-500,November 2007.

[8] Kanupriya Sharma, Amanpreet Kaur, A review of the existing state of Personality prediction of Twitter users with Machine Learning Algorithms, IOSR Journal of Computer Engineering (IOSR-JCE), July Aug. 2015.

[9] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of Things for Smart Cities, IEEE Internet of Things Journal, Vol. 1, No.1, 2014, pp. 22-32.

[10] Jennifer Golbeck, Predicting Personality from Social Media Text, Transactions on Replication Research, 2016.

[11] Big Five Personality Traits,

[12] P. Rosen and D. Kluemper, The Impact of the Big Five Personality Traits on the Acceptance of Social Networking Website, AMCIS 2008 Proceedings, page 274, 2008.

[13] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. van Aken, and W. Meeus, Emerging Late Adolescent Friendship Networks and Big Five Personality Traits:A Social Network Approach, Journal of personality,78(2):509538, 2010.