Personality Analysing on Watson Cloud by tracking the digital footprints of the user

SARATH CHIRAYIL SUBHASH

Submitted as part of the requirements for the degree
of MSc in Cloud Computing
at the School of Computing,
National College of Ireland
Dublin, Ireland.

December 2015

Supervisor Dr. Anu Sahni

Abstract

Artificial intelligence has become prime defacto in offering solution to complex and unstructured data. The intelligence architecture consists of a network of computers interlinked with each other that uses the data mining technology to extract information and custer them to give a valid answer to the existing complex query. Use of machine language in predicting new algorithm for the complex solution is done based on the unstructured data that is fed in. The process that we undergo is to repeat and analyze the input fed into the machine and validating the output for all the use cases. Social networks has become an integral part in the our society. They are build and deployed in cloud using the available infrastructure. Social networks has served us in many ways to become a basic necessity in our day to day activities. Social networks has grown their business and has had an impact on our current generation. Today one can judge others character based on the posts or tweets made in social networking platforms. It is evident to characterize the person and determine the skills acquired by the user to build up his career. In this research we propose that the personality of a human can be determined by the way someone uses his/her social networking account. The lean measure of personality is assessed based on Big 5 personality traits model. We have used personality insight API in Watson cloud to analyze and predict the personality of the user. In this research we have constantly analyzed the users personality scores and studied the pattern on the personality quality attained by each politician. The application is tested in cloud and the personality of the Twitter user is extracted and analyzed. The novelty lies in ideally predicting the character of the Twitter user based on statistical analytics we will determine weather the politician will turn out to be a good politician with the personality that he possess. Our sample dataset is a cluster of politicians who are active Twitter users. This application can be used to predict the outcome based on personality insight using Twitter feeds. Here in this research paper, we start by analyzing the Watson Cloud, Twitter lexical pattern, and the Big 5 personality trait model.

Acknowledgements

I would like to give my special thanks to Dr. Anu Sahni of the School of Computing, National College of Ireland for her invaluable support, technical advice and feedback before submission. I have never met a lecturer with such work ethic and care for students to do well. I could not have completed this research paper without the unwavering love and support from my family and friends who supported and encouraged me throughout the process.

Submission of Thesis and Dissertation

National College of Ireland

Research Students Declaration Form

(Thesis/Author Declaration Form)

Name: SARATH CHIRAYIL SUBHASH

Student Number: 14111241

Degree for which thesis is submitted: MSc CLOUD COMPUTING

Material submitted for award

(a) I declare that the work has been composed by myself.

(b) I declare that all verbatim extracts contained in the thesis have been distinguished by quotation marks and the sources of information specifically acknowledged.

(c) My thesis will be included in electronic format in the College Institutional Repository TRAP (thesis reports and projects)

(d) ***Either*** *I declare that no material contained in the thesis has been used in any other submission for an academic award.

Or *I declare that the following material contained in the thesis formed part of a submission for the award of

_____

(State the award and the awarding body and list the material below)

Signature of research student: _____

Date: _____

Submission of Thesis to Norma Smurfit Library, National College of Ireland

Student name: Sarath Chirayil Subhash Student number: 14111241

School: Computing Course: MSc in Cloud Computing

Degree to be awarded: MSc in Cloud Computing

Title of Thesis: Personality analysing on Watson cloud by tracking the digital footprints of the user

One hard bound copy of your thesis will be lodged in the Norma Smurfit Library and will be available for consultation. The electronic copy will be accessible in TRAP (http://trap.ncirl.ie/), the National College of Ireland's Institutional Repository. In accordance with normal academic library practice all theses lodged in the National College of Ireland Institutional Repository (TRAP) are made available on open access.

I agree to a hard bound copy of my thesis being available for consultation in the library. I also agree to an electronic copy of my thesis being made publicly available on the National College of Ireland's Institutional Repository TRAP.

Signature of Candidate: _____

For completion by the School:
The aforementioned thesis was received by_____ Date:_____

This signed form must be appended to all hard bound and electronic copies of your thesis submitted to your school

Contents

List of Figures

Chapter 1

Introduction

Internet usage among general users has increased high time. As a medium for communication, internet serves millions of people. The users of social networking needs internet enabled devices for communication and these users have separate language among the networking community. The main social networks that are used by the general public are Twitter, Facebook and Instagram. These users vary in their character depending on factors like age, gender and even the demography. The social networking users share each and every moment in the internet without even considering the privacy factor. The information being posted can be accessed by a third party without even having to notify the user. These social networking platforms have become a stage for entertainment who are entertained by another set of users forming a community. Assuming that there are millions of users who have created social networking accounts for communication and stating their views, the volume of data information that these people share will be humongous. The need for data management plays an important role and these data which are shared, has to be analyzed to address the traits of the user. The research paper address Twitter users and predicting the personalities based on their tweets. In this research paper we are trying to prove that the set of politician who uses twitter can serve to be a good politician who is likely to win based on tweets and the pattern of tweets and the usage of twitter. Politicians are assumed as a role model or the voice of millions of general public who are elected based on their character trait. The character of a politician is analyzed based on the pattern of tweets as depicted by the Big 5 personality traits model. The application use the predictive analysis to determine the character trait of the twitter user. Usage of this application will determine if the user character is good enough to run the government. In this research paper we will be considering the politicians across the world as our sample data. The application will be saved in the cloud so that it will be easier to scale up and scale down with the

humongous data. The availability of the application is also analyzed. The application is run and tested in cloud. This has resulted in great resource reliability and even the the management of resources is efficient.

In this research paper, using the twitter API we will be analyzing the twitter user profile, the user tweets, tweeting behaviour and even the social networking language. It should be known that there are humongous data transfer across network. To predict the sentiments of the user based on the above context is a handy process. Considering the tweets by a particular user, one have to assume that the length of tweet that the user share is generally short and are unstructured, or without any logic as its the language pattern for tweeting. The use of hashtag has to be taken into account when we go ahead with the sentiment analysis of the user. We also have to analyze the abbreviations or short messaging pattern as observed in the mobile networking users. There are various real world application that uses the Twitter analytics to find the nature of the user and then customize the web page for the user to display advertisements that the user prefer most. The e-commerce is a mega business among internet community as there is always a quality to understand what the user needs. The value and needs of the user is also considered for the same. The research utilizes Twitter API for retrieval of user information including tweets and the timestamps and the frequency of the usage. This is used along with the pattern of the tweets shared and then we analyze the personality of the user based on analytics using IBM Watson cloud. In this journal we advocate more on characterization of the political leaders and the way in which it will affect their political future. The tweets hold an essential role in this aspect as it determines the sentiments of the user and our API predicts the personality trait based on the tweets and the Twitter usage. Hence again the novelty of this work is the ability to predict the candidate who is likely to win the political campaign. A study is done based on the personality of politicians who have successfully been elected and comparing the values obtained using the for new user. The limitation sighted out is that we will not be able to determine the traits possessed by the politicians if the tweet size is less or under par words which is set as 3000 words at least. Again it also depends on the demography and even those associated with the linguistic ability. Our sample is restricted to the politicians who have tweeted in English and not any other language.

Chapter 2

Background

The use of electronic gadgets both offline and online have served the purpose of the data being accessed by the third party. In this literature review we focus on determining the human personality based on the tweets from Twitter. The chapter starts with the study on IBM Watson Cloud and how it varies from the rest of the Cloud Service providers. In section 2.1, we will be investigating on Watson Cloud and section 2.2 addresses the twitter analytics services that will yield better outcome for the proposed paper. The evaluation of personality traits of human based on the Big 5 psychological traits [27](Tupes and Christal, 1992) is carried in section 2.3.

2.1 Watson Cloud

IBM have always come up with innovations. IBM Watson being the latest, demonstrates how cognitive computing could make use of unstructured information into straightforward answers. IBM has made shift in bringing the services for the developers to make changes in internet applications in general. Watson is utilizing a new kind of technology wherein it is capable of learning new kind of information that is passed onto it and is unlike classical computing which is built on logic and designed with rules and handles structured data [2]. IBM Watson can hence be assumed to be a technology platform that uses natural language and machine learning to sight out relevant things articulated at present from a set of unstructured data. The intelligence of Watson is purely dependent on its Corpus, where the dataset is fed manually. It then parses across its corpus to analyze the data that is fed in and gives insight into the subject. Hence, the AI employed in Watson will yield desired output based on the the information we fed manually. Perera, 2014[21] forecasts that the use of machine

learning and artificial intelligence on many use cases can be useful in managing areas where it requires batch processing and those that requires real world use-cases. The areas can be town planning and even stock markets. In our context this helps us in identifying the traits of the user to determine the personality of the politician.

Feldman,2012[9] analyzes the search evolved from an expert-only context to an end user tool. The book dwells deep into the Watson and analyses the concepts behind the cognitive computing. IBM Watson is utilizing predictive analytics as discussed above. The predictive analytics use the information from the past to determine the future outcome based on statistical algorithm and machine learning techniques. Considering the outcome of the result to be genuine, one has to keep in mind about the quality of the data being handled, and even the level of data being analyzed repeatedly. The framework as discussed by Babu and Sastry,2014[23] integrates big data analytics and predictive analytics for automated decision making process in an organizational level. Here the predictive analysis is carried out in such a way that the resultant output is used as an input and the given input is analyzed each time to ensure the quality of the service. In the predictive models algorithm the input is evaluated for each case to provide a known output if the instruction set is already present else the outcome is served and the rules are generated for the unknown outcomes. This serves the essential framework for an AI. The studies in the area of big data and predictive analytics can be used to judging knowledge-intensive tasks. Earley,2014[7] summarizes that predictive analytics can help mine online reference applications and systems known for actionable knowledge. Watson is known for its intelligence in the AI. The machine has outwitted many persons in a Television show. This super-powerful predictive computing could be brought into medical field. Considering the term of cancer, Watson can efficiently categorize different kinds of cancer and its relative side effects on different age groups after understanding more about the patient records from its corpus (Strickland and Guy,2013)[25]. IBM implements statistical AI for the operational purposes of Watson. To gain more intelligent performance, IBM implements the methodologies of AI and higher levels of statistical data (Apte, Morgenstern and Se June Hong, 2015)[5]. In our journal we will be implementing the knowledge harvestation based on temporal, multi-lingual, visual and even common sense. We will be implementing, a knowledge base for the time-spans and time-stamps as and when the event occurs (Suchanek and Weikum, 2013)[26]. The extraction of knowledge is a time consuming process and again the need for mining the data is also considered. The mining speed is dependent on the need for relevant data and storing it in a summary table (Wen-Chi Hou,1999)[12].

2.2 Twitter Text Analysis

The words play a significant role in a community and the psychological character of the user. Fast and Funder,2008[8] identifies a category of word use which clearly identifies the personality using a broad range of personality data provided. The work demonstrates that the more words that a user outputs, the more easier it is to characterize his personality (Yarkoni,2010)[30]. This is carried out by word based analysis and linguistic inquiry and word count (LIWC). The aforementioned LIWC uses different set of categories with different range of word count ranging the sets as articles, pronouns, work related words and even social words.

The journal proposed by (Ikeda et al.,2013)[13] have an insight on the demography and the linguistic pattern of Twitter user by analyzing the content of the information they share. A deeper insight into the pattern the way user tweets, determines what age he is in and and he requires others to understand. The journal studied the pattern of market analysis for business. An overview of the journal published by (Jiantao and Ning,2014)[14] proposes user interest based on latent dirichlet allocation and singular value decomposition method. Abascal-Mena, Lema and Sedes,2014)[3] extracts Twitter data carefully selected around hashtag pattern so as to have an insight into overall subject. It determines the hashtag and the related information may be extracted to study the overall nature of the subject. The study reveals that tweets has become a medium for communication among millions of users and have a specific culture for interactions within each other. The article presented the constitution of graphs from the tweets that the user shares and how it evolve over the time and on the the behavioural attitude towards the subject. The overall studies on twitter analytics serves us to determine the tweeting pattern and the time stamps associated with the tweets. This is basically help us in dwelling the user characteristics in more depth.

2.3 Personality Traits

The widely accepted model is the Five Factor Model (FFM) also known as the Big 5 personality traits for measuring the human personality. FFM is accredited for finding the human personality or psyche[1] in recent years. The FFM is listed down as below and it delineates the the broad human traits, which binds most of the variations in personality seen across humans. Apart from this we account for the needs and values of the human in this research paper, as it helps us effectively determine the human personality . Tupes and Christal,1992[27] in their journal, takes the sample data based on the lexical approach whereas the FFM is known for its questionnaire pattern (McCrae

and Costa, 2003)[17].

2.3.1 Openness to Experience

The acceptance to new techniques and methodologies is a quality for openness to experience. It is assumed to be in two subsets namely inventive/curious vs consistent/-cautious . The high score in openness implies that the person is open to new ideas and techniques. They tend to be curious, intelligent and even imaginative. The high score also implies that the category falls into having artistic ability trait and sophisticated in taste. They have very high aesthetic values. Lambiotte and Kosinski,2014[16] On the other hand we can consider the high score in openness to be perceived as usually unpredictable and have lack of focus towards the things that the human will be doing. Conversely a human with low points in the openness are likely to be close minded and will be doing their activities with a set piece of instructions. They generally possess down-to-earth character and are conventional in nature. A human who is intellect is said to have high scores in openness to experience (Pervin and John,1999)[15].

2.3.2 Conscientiousness

(McCrae and Costa,2003)[17] in their book sights conscientiousness as a dimension of an individual differences in any organization and achievement. The people falling in this category with high conscientiousness tend to be self-disciplined, responsible, organized, and devoted to their duties. They are high achievers, hard workers and planners often prioritizing the tasks assigned to them.They being ambitious tend to be liberal in views and values. The people are most likely to be loyal and faithful in nature. The people scoring low values in the conscientiousness possess flexible nature and are spontaneous. We can subsume them to be easy going, unreliable and careless in other words. Psychometric-success.com[1] forecasts that conscientiousness is the way in which we control and directs our impulsive actions which may end up in trouble or even fun. The impulsive actions can cause a human to think in a way which he assumes to be best at that point of time but eliminates the fact of taking a wise and appropriate decision. The result is that a person who scores high points in conscientiousness avoids all kind of troubles and attains success in the desired areas of success through their purposeful planning and sheer persistence.

2.3.3 Extraversion

Extraversion sometimes known as surgency (McCrae and Costa,2003)[17] is the intensity in being active, talkative, assertive over things. The studies suggest that the person who values more in extraversion is often perceived as dominant and even into socializing. They possess high energy and will be an integral part in giving lot of positive emotions in a community. The basic traits exhibited by the extroverts are shown below in figure 2.1 (McCrae and Costa,2003)[17]. It should be also noted that the introverts arent opposite to extroverts in their character as because the person cant lack energy level always. Its because the person is not into socializing and neednt require any external simulations to do a set of work. The nature of introverts are assumed to be arrogant and unfriendly as a result. In reality, an introvert who have high level of agreeableness, is likely to be a team player.

2.3.4 Agreeableness

Agreeableness reflects a person to get along with others discarding the individual differences which is more towards a social cause and harmony among a community. They are characterized by the optimistic behaviour towards the subject. The people who tend to achieve high values in accordance with agreeing towards the concerned subjects are generally peace keeps and attain high popularity. On the other side, the disagreeable people tend to be stubborn in their decision and are likely to take rough decision which is considered good at times depending on the circumstances for the welfare of the mass.

2.3.5 Neuroticism

Neuroticism is subjected on the basis of sensitive or nervous human trait against the socially secure and confident trait. In simple words a person who is having less control over his thoughts acts impulsively. Humans possessing high scores in neuroticism tends to be anxious and depressed in nature and are usually unapproachable by others or even uninspiring in the thoughts . The humans falling in this category are assumed that they do not think clearly over the matter of concern and the humans with less neuroticism are considered to be free from negative thoughts while we cannot perceive them to posses positive values.

Tuten and Bosnjak,2001[28] in their journal forecasts that the human personality trait is determined on the basis of openness to experience and neuroticism. The personality

Neuroticism
Calm—Worrying
Even-tempered—Temperamental
Self-satisfied—Self-pitying
Comfortable—Self-conscious
Unemotional—Emotional
Hardy—Vulnerable

Extraversion
Reserved—Affectionate
Loner—Joiner
Quiet—Talkative
Passive—Active
Sober—Fun-loving
Unfeeling—Passionate

Openness to Experience
Down-to-earth—Imaginative
Uncreative—Creative
Conventional—Original
Prefer routine—Prefer variety
Uncurious—Curious
Conservative—Liberal

Agreeableness
Ruthless—Softhearted
Suspicious—Trusting
Stingy—Generous
Antagonistic—Acquiescent
Critical—Lenient
Irritable—Good-natured

Conscientiousness
Negligent—Conscientious
Lazy—Hardworking
Disorganized—Well-organized
Late—Punctual
Aimless—Ambitious
Quitting—Persevering

Figure 2.1: The Five Factor Model of Personality

traits explained in (Tupes and Christal,1992)[27] is analyzed and we successfully determine which trait is exposed more to internet usage. The study forecasts that when an individual is exposed to new technology, he is eager enough to know more about the subject which is the trait associated with openness to experience. This is related to positive side of the internet usage while on the other side, neuroticism is considered to be negative. Golbeck et al.,2011[10] cites the journal by using the information of human published online to determine the personality traits they possess. This is even used for the suggestion of friends or engage the friend circle to a distinct community. Selfhout et al.,2010[24] studied on the development of emerging friendships in social networks. The journal reveals that the process of selecting friends and selected as friends is dependent on extraversion and agreeableness respectively. Over the time, openness to experience plays a vital role in enhancing friendship in social networks.

2.4 Social Networking and Personality

Social networking is becoming popular day by day and the information shared on the profiles are humongous. The data varies from different social networks. In this research paper we assume social networking websites to be used for tweeting, messaging and even used for the job profiles. The popular social networking platforms are Facebook, Twitter and Linkedin. The personality research on social networking platforms is a complex process. We have to analyze the lexical pattern applied for different networking platforms. In Twitter, we have the hash-tag pattern so as in Instagram. The formal way of writing and posting an article is seen in Linkedin as the networking platforms houses the basic need for job and related contents in the area. In general, people follow a specific behaviour towards the social networking platforms. We observe the personality analyzed in (Nie et al.,2014)[19] through a predictive analysis using local linear semi-supervised regression algorithm for Microblog users. The study also demonstrated that the accuracy of prediction is improved by the usage of unlabelled data. Adali and Golbeck,2012[4] analyzed the behavioural pattern in terms of overall activity in internet, message pattern sighting if it was manually created or forwarded or if it had any URLs attached, pair behaviour determined their attitude towards their friends and followers. They also analyzed the pattern within a group of friends online and if they generally stick towards that attribute mostly. Mostly the activities on social networking sites tend to capitalize on posts photos, posting opinions, views on subjects and personal information and posts (Pagani, Hofacker and Goldsmith,2011)[20]. The advantage here for the network content providers are that the data that is being consumed by other users are generated by the users of interest. This itself shows the openness to experience in the social networking community. Research shows that personality is

intertwined with job-success, attractiveness, drug use, marital satisfaction and happiness (Lambiotte and Kosinski,2014)[16]. Moreover, the analysis on the research shows the Facebook profile to portray the human character than the idealized role (Back et al.,2010)[6]. Once the trait is know we will be able to predict the personality trait of online users or community in future (iberna and Vehovar,2009)[31]. The journal articulates about an algorithm where the user tends to respond to the subsequent followers who the user has interacted. Unlike (Pagani, Hofacker and Goldsmith,2011)[20] (iberna and Vehovar,2009)[31] didnt went into check the messages of the users. If that was to be considered, the journal published would have yield better results. The journal presented by (Guo, Lin and Chen,2009)[11] suggest that the interest of the user tend to decline with the time lapse. Henceforth we can assume that the human possessing openness to experience will be more interested in making his profile active. The journal article presented by (Plummer, Hiltz and Plotnick,2011)[22] is carried out to find the success rate of job applicants using the social networks as these social networks have a large pool of members associated to them and have access to relevant profiles and characters within that networking platform. Social networking users also show attention towards same age, nationality and even gender. This is delved in further to check the selective behaviour of social networking user in the journal (Xiao et al.,2012)[29]. Mogadala and Varma,2012 [18] computes the mood swings of the twitter user using regression testing. Experiments from their work have attained less root-mean-square error as when compared to other regression approaches for predicting mood transition.

Chapter 3

Design

In this chapter we will be specifying the overall design and specification of the techniques that we have implemented in our work. The application is built using Node.js and MongoDb.

At present we start discussing about node.js. It is one of the powerful server-side scripting language. On analyzing we come to a point that the node.js is a runtime javascript that uses Googles V8 engine. Node.js contains event loop which is a single thread used for managing the I/O operations. The I/O operations are asynchronous in nature. This makes an effective language for our application development as we have to read and write to network connections and even to the databases. It also allows us to build a scalable applications which is fast. In this project we use npm as the package manager as it is fast, robust and consistent. npm ensures that the project are isolated and the version control is managed efficiently.

MongoDB is a cross platform document- oriented database which makes it easier to integrate with applications. The information that MongoDB stores are in BSON format which is likely to be JSON objects having dynamic schema. The application requires a higher volume of data to be saved and written to analyzing them is a tedious job. So MongoDB is the best choice as it sets its replica as like Master-Slaves which make it available and fast in real world scenarios. As our application is dealing with the Twitter API to access the tweets of the user from the start and determining the character trait, we require a humongous database to store the information. As for the case we have selected MongoDB which will serve the basic purpose in our application.

The application is implemented using Twitter API to retrieve the information of the user. The retrieved data is saved in the Target database. Using the MongoDb application we mine the data and access the required fields for the application structure.

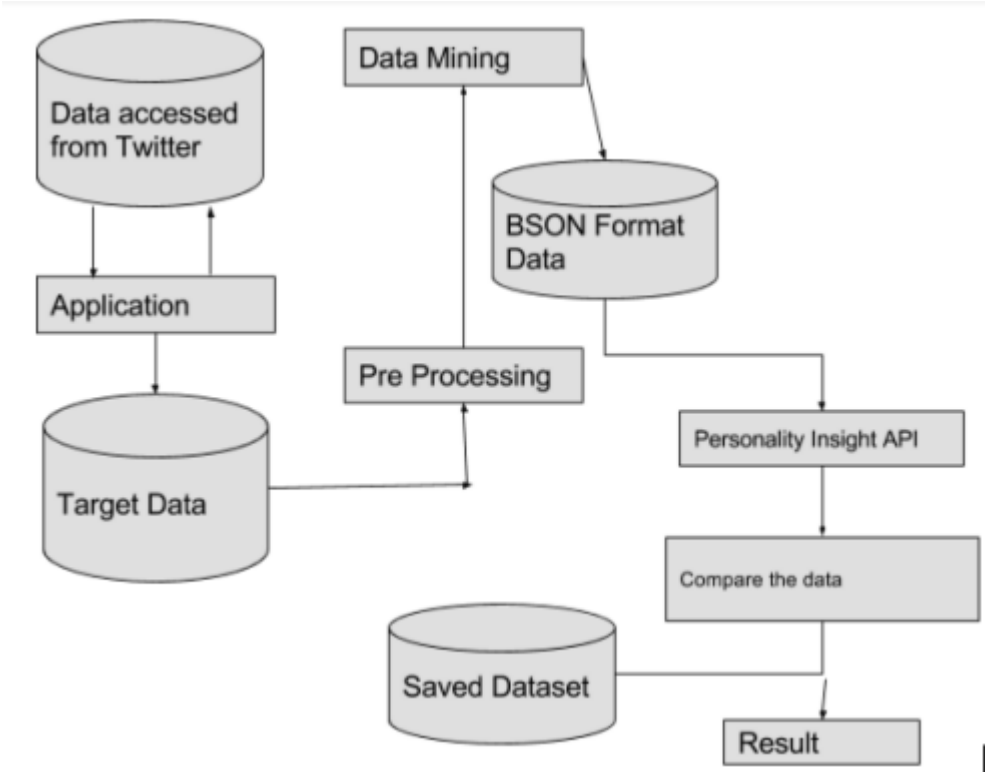

Figure 3.1: Framework for the application.

The retrieved application is then saved as BSON format which is the JSON format for MongoDB. The information is then passed on to the personality insight API for further analysis to find the trait of Twitter user. In our context the Twitter user will be the politician.

The design can be summarized into data collection, data preprocessing and data evaluation after consuming the API. The data collection done through Twitter API is passed on to corpus of textual analytics. After that as per the application requirement we extract the features from the file using MongoDB. In the data evaluation model, we pass the information onto the personality insight API where the user traits are extracted and are checked against the saved dataset to see if they become successful candidate by examining the dataset which was saved prior.

The main page of the application handles the username to be used as an input. The username provided will be handled by the twitter api which through npm twitter package manager. This will access the twitter consumer key, consumer secret, access token key and access token secret. Once the twitter api has been configured with the application we access the twitter feeds using the twitter id.

The twitter user id analyzed is used to retrieve the information about the user. The

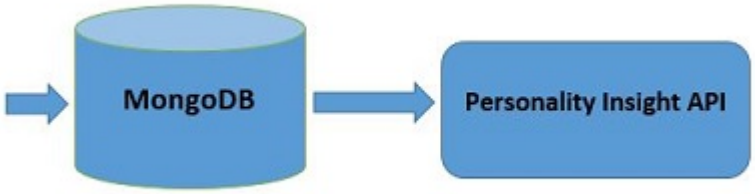

Figure 3.2: MongoDB and API interaction.

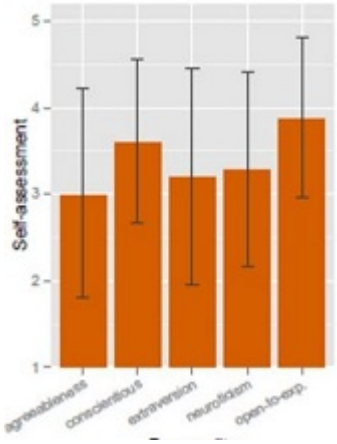

Figure 3.3: Personality

tweets are checked and the time-stamps are verified and are used by the data collection agent. The data is saved in the user profiles handled by Mongodb. The extraction of information is done by the personality insight api and the extracted data is used for the analysis as depicted in the following diagram.

The extraction of personality is done by the API. The needs and values are eliminated in the application developed. The personality scores are attained through the API. Once attained the values are extracted and saved as a collection in MongoDB. This ensures that the user profiles are extracted from the JSON format. The extracted formats are then saved into different politician profiles which will be another collection in NodeJS. The data that is saved in the collection will be used for comparing the results that we have extracted from the twitter user account. Time stamps and user's nature is analyzed through the application by the use of api. This gives the quantitative results of the user in terms of the personality score. The twitter api uses the information based on the timestamps that the user logs in and is using the twitter which is extracted from cookies and sessions. The further manipulation can be done on this api to get the

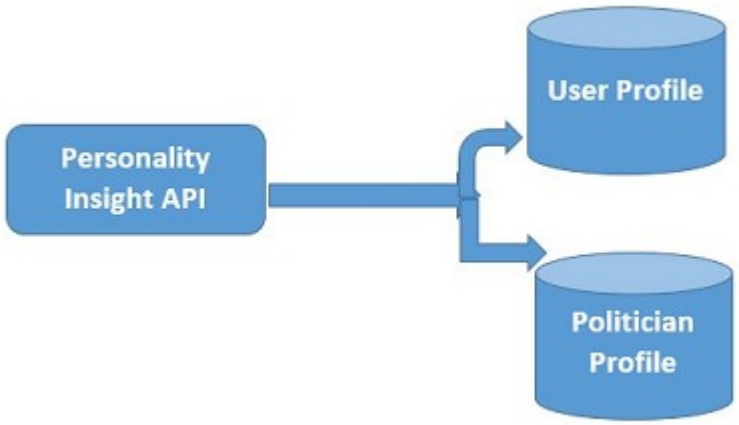

Figure 3.4: Collections in MongoDb and Personality Insight

details of the overall interest of the user outside twitter. This is not analyzed in our current context.

Chapter 4

Implementation

The application is built using Node.js and MongoDb. Node.js being light weight we have built and efficient program that runs in bluemix account. Further detail about implementation is listed below.

Software details: Operating System: Linux Ubuntu Trusty Tahr 14.04 LTS Nodejs: 0.10.31 MongoDb: 3.2.0

The application deployed in watson cloud can be accessed from using the below mentioned url. URL: http://your-celebrity-match.eu-gb.mybluemix.net/

The below shown figure 4.1 shows the running instance in cloud.

The application contains Twitter API. The Twitter API is responsible for the call made on specific username. The Twitter username starts with @ sign and for this program we have omitted the symbol so that it will be easy for the users to key in the username and get to see the personality trait using Watson API. The input of the Watson service API will be handled by the Twitter and we will be interested in the output provided by the API in JSON format. As since the personality is calculated based on Big 5 personality traits model, Needs and values of the user based on the tweets and time and behaviour of the tweets.

Figure 4.1: Instance running on bluemix account

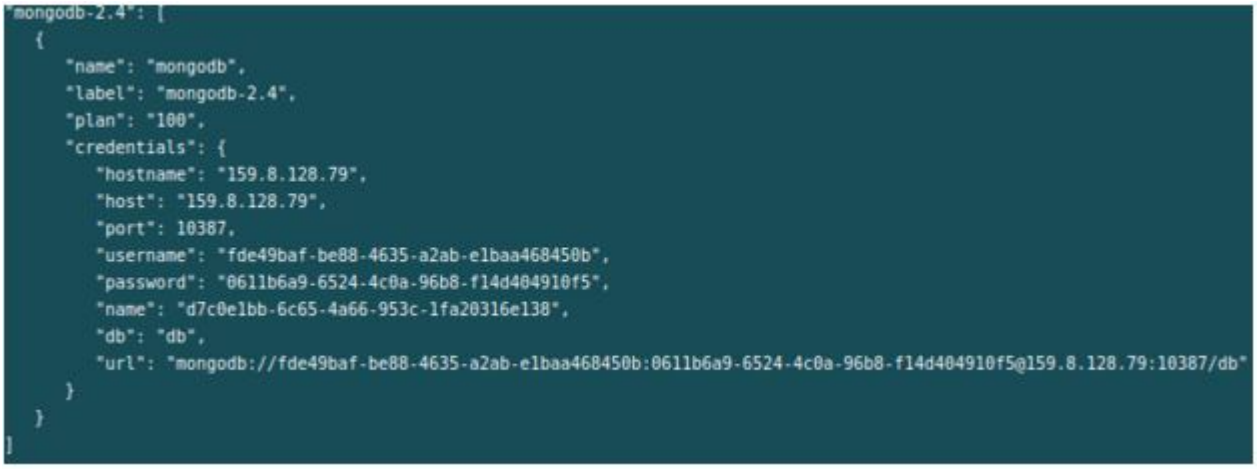

Figure 4.2: Service for MongoDb

Figure 4.3: Dependencies in package.json

While implementing the application in node.js we have included the npm modules, as the framework and the use of kerberos module was handled. We have also implemented mongoose thin client that connects between the node.js application and the server mongodb.

The environment variables for mongodb is shown in figure 4.2 below.

We have selected MongoDb as a mean that we can efficiently scale and be able to design the application. The querying is also is considered to be efficient in for operational and administrative procedures. The dependencies required for our application is shown in the below figure 4.3.

Once the Twitter API is set up, the twitter username are validated in the code below.

```
var showUser    = Q.denodeify(req.twit.showUser.bind(req.twit)),
  getTweets  = Q.denodeify(req.twit.getTweets.bind(req.twit)),
  getProfile = Q.denodeify(req.personality_insights.profile.bind(req.personality_insights)),
  getUserFromDB = Q.denodeify(User.findOne.bind(User)),
  saveUserInDB = Q.denodeify(User.createOrUpdate.bind(User));
```

Figure 4.4: Instance running on Bluemix account

Figure 4.5: Database saved locally within the application.

We use the denodeify of the Q module to convert the POST callback function to a promise function.

The application is implemented in such a way that the user details are saved into a database. There will be a local database that is attached to the application where we will be saving the details of the politicians who have great attitude and perspective towards the political aspects in the current world. Once the database have been implemented, we will be comparing the new users/ politicians against the current database that is attached to the application to determine the personality of the newly added member. We will also be able to update the data set by adding many members as it will work efficiently in determining the efficient candidate. The figure 4.5 below shows the database of the twitter users that we have saved in the application. As a matter of fact we have added the number of followers and even the number of tweets that these users have made.

The need for adding more records could be handled by exporting the personality traits of the user and updating the database file manually. The retrieved result is stored and accessed in JSON format. For any The application hosted in cloud is up and running and is used to abstract the character trait of the user. In our application we focus on predicting the political candidates result. The application is so useful as individual we will be mapping the details of the user based on the personality traits and even his twitter id. The sample figure 4.7 shows the details of the pattern in which a user record is saved in our database.

Figure 4.6: Operating in Cloud

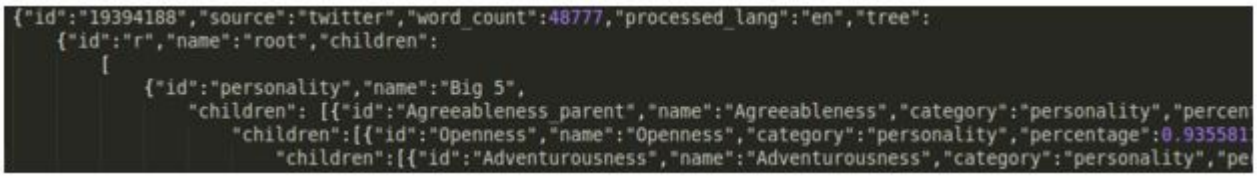

Figure 4.7: Personality trait based on Big 5

The application houses two database, one for handling the username and the other for comparing the results against the set of politicians against whom we will be comparing the results. The evaluation of results is discussed below.

Chapter 5

Evaluation

In this section we evaluate the application and see if it is working as desired. In the research paper we have selected the politicians dataset. Here we are determining the candidates win or lose probability based on the character trait analyzed through Twitter API and Personality insight API. We take a sample data of 50 politicians across the world who have won and who have lost. The analysis can be used to determines the candidateship for the politicians in the coming election. Using the Twitter API we can delve into the tweets of the politicians to determine the winning probability.

On evaluation we conclude that our model is extracting traits of the user. It is taken into consideration that the Big 5 personality traits, needs and values. We will be listing the traits such as cautious, outgoing, stability, practicality, self-discipline and being curious. We will be analyzing it on a small dataset of 100 politicians.

Based on the predictions we plot the politicians as per the graph below based on some specific traits such as being cautious, outgoing, stability, practicality, self-discipline and being curious. The x-axis plots the username of politician and the y-axis plots the score the user have received using our personal insight API based on the tweets.

Studies shows that the user value is dependent on the traits they possess. Our research has revealed that Barack Obama has similar traits with Al Gore. Unfortunately Al Gore didnt make it to president whereas Barack Obama did even though they had similar traits. Our studies on other politicians reveal that, Bill Clinton has similar traits with Mitt Romney, Mitt Romney with Barack Obama, Enda Kenny with Mitt Romney, Micheal Martin with Al Gore and finally George W Bush with Bill Clinton.

The below attached is the graphical representations for the same.

The dataset of 50 politicians is charted below represented in terms of percentage. The

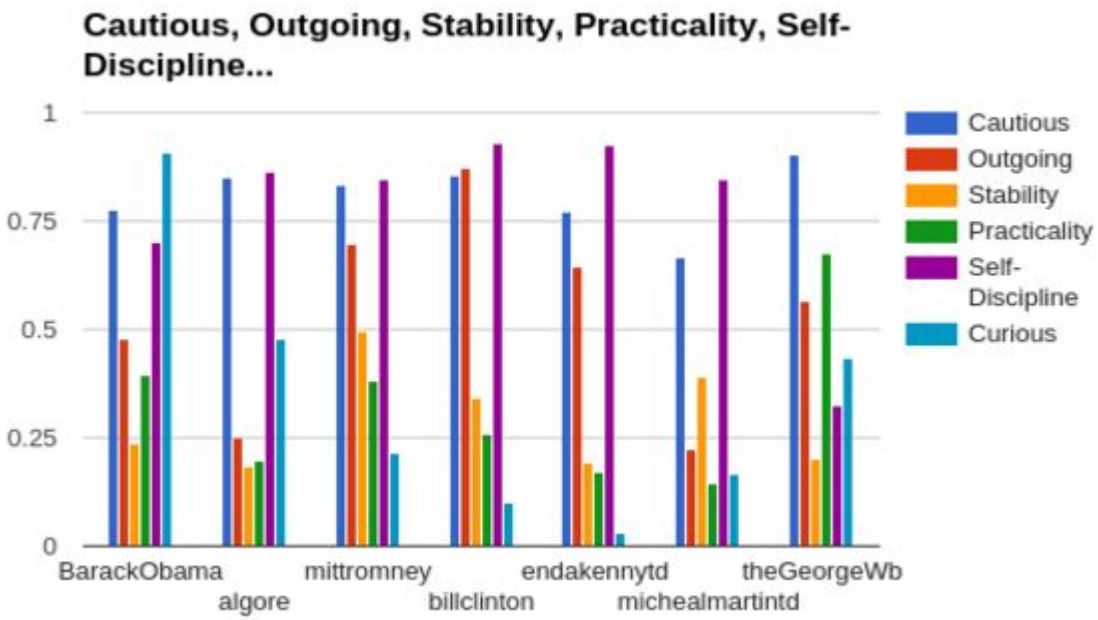

Figure 5.1: Graph plotting the personality trait and the politicians scores

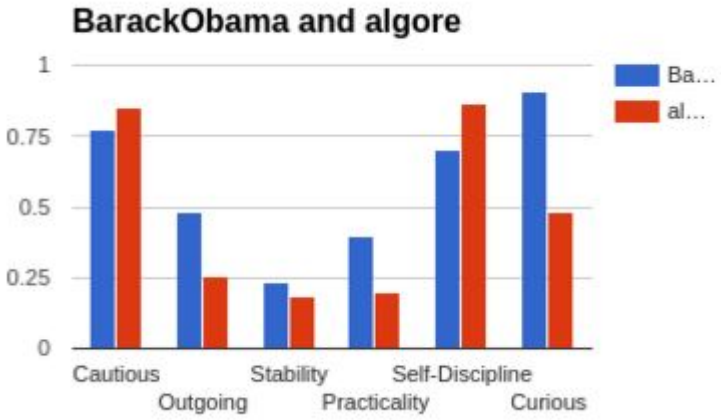

Figure 5.2: Statistical measure of user traits- Obama and Algore

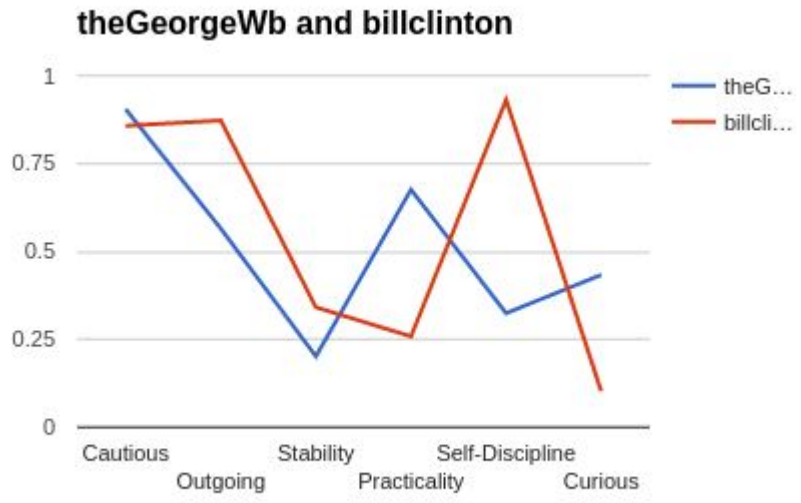

Figure 5.3: Statistical measure of user traits- George W Bush and Bill Clinton

sampling error is also set so as that we will be able to determine the mean deviation from the character trait.

The dataset is shown below.

Politician	BarackObama	theGeorgeWb	Algore	billclinton	mittromney
Cautious	77.56+-5.94	90.46+-9.67	85.04+-7.37	85.73+-8.23	83.32+-7.20
Outgoing	47.96+-6.06	56.52+-7.97	25.21+-5.69	87.24+-6.68	69.68+-5.69
Stability	23.53+-6.98	20.18+-11.13	18.51+-7.61	34.14+-9.22	49.69+-7.60
Practicality	39.56+-5.64	67.58+-9.16	19.69+-6.56	25.86+-7.55	38.14+-6.51
Self-Discipline	70.25+-4.44	32.45+-4.99	86.36+-4.06	93.09+-4.17	84.73+-4.15
Curious	90.97+-7.48	43.35+-12.49	47.96+-6.17	10.26+-5.86	21.50+-8.19

Politician	gipper76	lobekarnish	ebbush	llthingshill	enSanders
Cautious	91.44+-8.72	47.56+-6.76	71.89+-5.94	50.85+-7.89	57.36+-5.94
Outgoing	8.79+-7.11	4.47+-5.89	61.36+-6.06	79.93+-6.27	32.75+-6.06
Stability	7.34+- 10.03	22.12+-7.73	59.22+-6.98	15.03+-8.70	18.20+-6.98
Practicality	4.39+-8.20	3.79+-6.73	29.06+-5.64	3.82+-7.18	21.4+-5.64
Self-Discipline	94.41+-4.44	76.86+-4.49	81.35+-4.41	91.22+-4.00	76.25+-4.41
Curious	99.58+-11.28	17.59+-7.96	26.96+-7.48	11.10+-9.62	12.76+-7.48

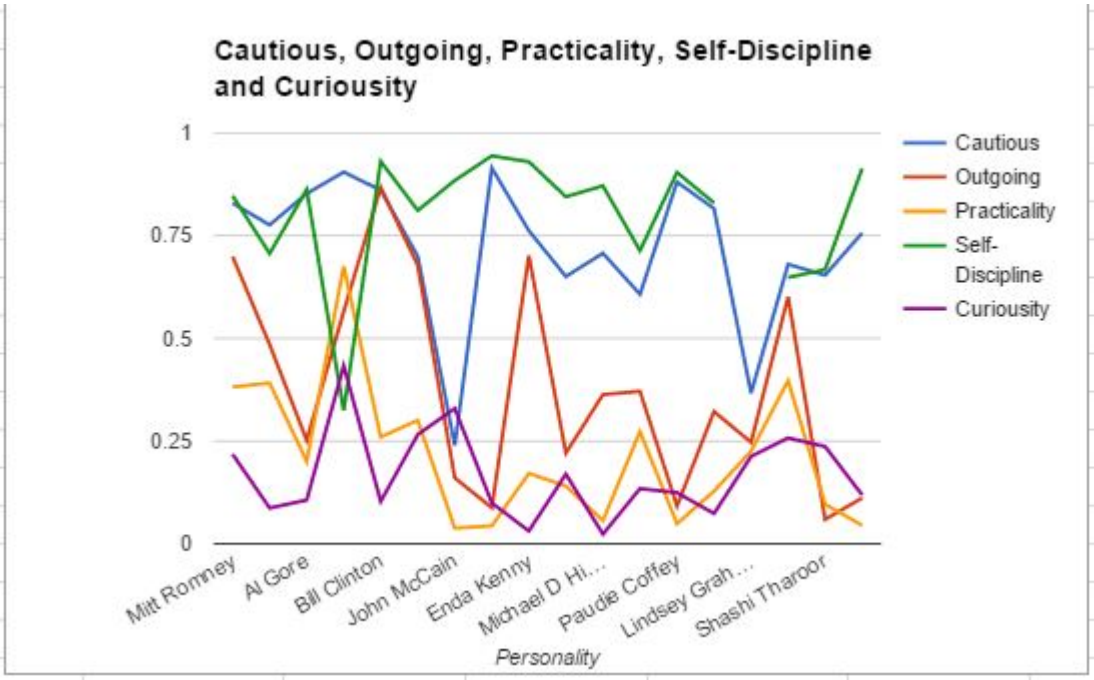

Figure 5.4: Variations of the desired qualities

Politician	lincolnchafee	tedcruz	grahamblog	carlyfiorina
Cautious	77.56 +- 5.94	90.46 +- 9.67	85.04 +- 7.37	85.73 +- 8.23
Outgoing	11.16+-6.06	76.76+-6.06	24.74+-6.06	63.08+-6.06
Stability	20.71+-6.98	30.42+-6.98	55.95+-6.98	16.2+-6.98
Practicality	45.19+-5.64	13.46+-5.64	22.59+-5.64	14.14+-5.64
Self-Discipline	91.34+-4.41	78.97+-4.41	71.34+-4.41	79.49+-4.41
Curious	11.88+-7.48	99.70+-7.48	21.17+-7.48	5.95+-7.48

On analysis of the results we can say that George W Bush had similar vision as Bill Clinton though they were from different political parties. This work makes an insight into the success of the candidate in the presidential elections that happened in US. We can also determine the future outcomes if the tweets are genuine and are tweeted by the concerned politician. Another area of application is that we can use this application in various roles in the industry.

The data retrieved is analyzed based on the personality scores attained and a brief

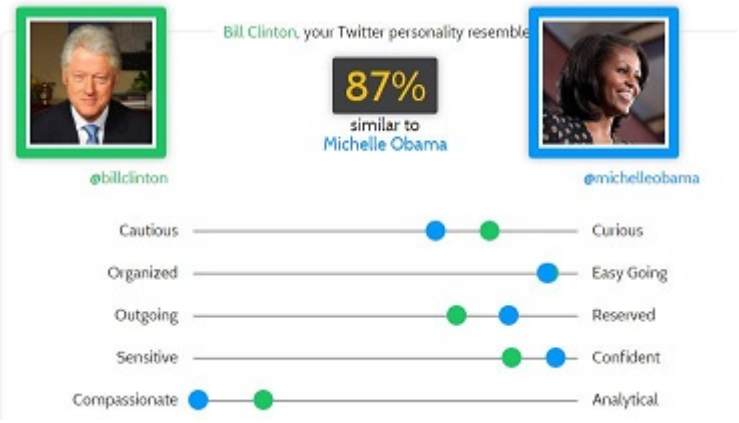

Figure 5.5: Comparison of user trait

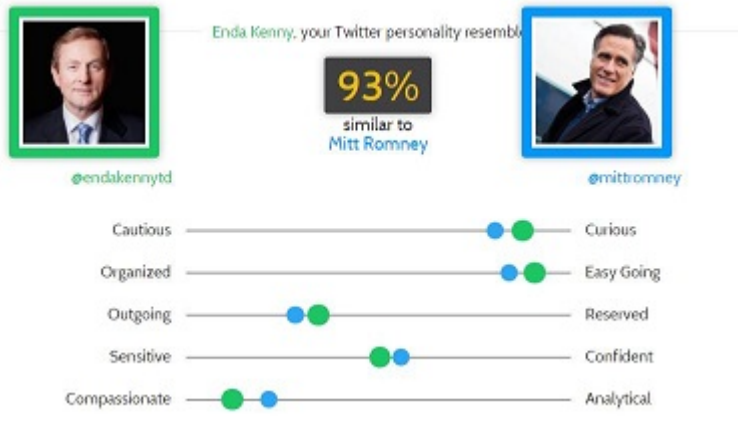

Figure 5.6: Comparison of user trait

Personality trait	Barack Obama
Cautious	0.7759526731
Outgoing	0.4851799343
Stability	0.2325158221
Practicality	0.3913554229
Self-Discipline	0.7064001155
Curiousity	0.08718710679

Table 5.1: Personality scores of Barack Obama

study is made on them to see if the candidate is a winning candidate or not. A dataset of winning candidates is analyzed and further it is analyzed against a losing candidates. We determine the key qualities of the candidates who are winning and analyze their score. Once the scores and key factors have been analyzed we plot a graph to see the success rate. The new user is compared against a set of winning candidates or the successful candidates who have high reputation. Based on the comparison we analyze the values of the values of the new user and determine if the candidate is likely to win or lose. Fig 5.4 determines the variations of the desired qualities based on personality of politicians.

The work based on individual user is explained below. Assuming Barack Obama as the president of US, the traits of politicians are examined to determine their success.

The evaluation is examined based on the facts described below.

Analysis: To analyse the personality of Barack Obama and check if he is likely to win.

Fact: The personality score of Barack Obama is depicted in table 5.1. Based on the personality scores of candidates who've contested in election and won we will analyze the results and check if Obama is likely to win. The cautious rate of Obama is 0.7759526731 which is similar to Bill Clinton and George Bush. The Outgoing scores determine that Obama is moderate compared to others and considering the factor of stability Obama is more stable that two of the previous presidents of United States. The practicality score is seen more in George W Bush unlike the past presidents of United States. George W Bush also sores more in curiosity and self-discipline.

Prediction: The results as retrieved by the data portrays that Barack Obama has won election.

The analysis is done on 100 samples and we were able to determine that the api yielded results which was 72.4278 percent accurate. More data being added into the database will make us understand if the api can be used successfully in determining the traits. As of now we have tested the api which is good enough in determining the user trait.

The sample data is analyzed and the originality of personality trait is dependent on the number of tweets performed. The limitation of the application is that we need to have a bigger dataset consisting of politicians who have performed more number of tweets or even explained about them in not less than 3000 words. We did not analyze the politicians from rest of the world who tweeted in deifferent language outside English.

Chapter 6

Conclusions

In this research paper, we have used the Twitter API to retrieve the tweets of the user and broadly analyze the personality score in the Big Five personality traits model. The area that need more attention in our study is that when the sample data set is improved, we will be able to determine the human character in terms of social networking behaviour and the actual behaviour.

The research on personality determination using the Twitter on Watson cloud have affirmed that the result is quite unique on the way they use their tweets and lexical typography. It is now evident that the traces that we leave behind the social network can determine ones personality. Determining the personality have affirmed that the human is grounded to the concerned subject by keeping him in the comfort zone ensuring that there is nothing happening more creative outside the comfort zone that the user is in. Studies reveal that the people now are more addictive to social networking platforms. The change for the human trait with age is persistent which we will witness in near future, and here humans will be divided based on the trait they possess by the cyber world. The analysis on politicians is considered to be quite accurate in certain cases where the sample size of tweets tend to be more rather that the politician with less tweets.

The research paper has provided the desired result. We have analyzed the the traits and have predicted the outcomes of the success rate of the candidate who have contested and this application can be utilized to successfully determine the outcome of the Presidential elections. The genuinity of the application depends on whether the politician himself is using the application or else some of his colleague.

The limitation of such a system is that we cannot determine the personality of the user if the sample tweet size is less. It is advised that the user should be active in

Twitter so that we obtain the relevant information based on his tweets. Our work is based on politicians who uses English as a language for tweeting. Outside English we have not determined the personality trait. The results obtained through our analysis can be either used for useful purposes or for negative purposes. We will be able to get most of the information about the Twitter user without even bothering about legal consequences. The optimistic thing to be considered is we will be able to know our traits and work on it or even can be used in medical field. As said if one knows about himself he refrains from troubles. So our sample data set on the politicians can be used for either the change in the country or simply choosing the right candidate that is needed for the development process and welfare of the nation. The future work that the research area propose is a machine learning technique that will handle the precision personality on large data sets. The application can be used in medical field to determine the mindset of the surgeons before major surgery. This will eliminate the cause of human error in major surgery. For this we will have to know the mindset of the doctors before and after operations.

Bibliography

[1] The 'big 5' aspects of personality, howpublished = http://www.psychometric-success.com/personality-tests/personality-tests-big-5-aspects.htm. Accessed: 2015-12-14.

[2] Developer cloud documentation — watson developer cloud, howpublished = https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/getting_started/, note = Accessed: 2015-12-14.

[3] R. Abascal-Mena, R. Lema, and F. Sedes. From tweet to graph: Social network analysis for semantic information extraction. In *Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on*, pages 1–10, May 2014.

[4] S. Adali and J. Golbeck. Predicting personality with social behavior. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 302–309, Aug 2012.

[5] C. Apte, L. Morgenstern, and Se June Hong. Ai at ibm research. *Intelligent Systems and their Applications, IEEE*, 15(6):51–57, Nov 2000.

[6] Mitja D. Back, Juliane M. Stopfer, Simine Vazire, Sam Gaddis, Stefan C. Schmukle, Boris Egloff, and Samuel D. Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 2010.

[7] Seth Earley. Big data and predictive analytics: What's new? *IT Professional*, 16(1):13–15, 2014.

[8] Lisa A Fast and David C Funder. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, 94(2):334, 2008.

[9] Susan E. Feldman. The answer machine. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 4(3):1–137, 2012.

[10] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156, Oct 2011.

[11] Dan Guo, Fuhong Lin, and Changjia Chen. User behaviors in an online social network. In *Network Infrastructure and Digital Content, 2009. IC-NIDC 2009. IEEE International Conference on*, pages 430–434, Nov 2009.

[12] W.-C. Hou. A framework for statistical data mining with summary tables. In *Scientific and Statistical Database Management, 1999. Eleventh International Conference on*, pages 14–23, Aug 1999.

[13] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35 – 47, 2013.

[14] Zhao Jiantao and Shi Ning. User interest prediction in microblog using recommendation method. In *Information Technology and Artificial Intelligence Conference (ITAIC), 2014 IEEE 7th Joint International*, pages 367–370, Dec 2014.

[15] O.P. John, R.W. Robins, and L.A. Pervin. *Handbook of Personality, Third Edition: Theory and Research*. Guilford Publications, 2008.

[16] R. Lambiotte and M. Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, Dec 2014.

[17] Robert R McCrae and Paul T Costa. *Personality in adulthood: A five-factor theory perspective*. Guilford Press, 2003.

[18] Aditya Mogadala and Vasudeva Varma. Twitter user behavior understanding with mood transition prediction. In *Proceedings of the 2012 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media*, DUBMMSM '12, pages 31–34, New York, NY, USA, 2012. ACM.

[19] Dong Nie, Zengda Guan, Bibo Hao, Shuotian Bai, and Tingshao Zhu. Predicting personality on social media with semi-supervised learning. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 02*, WI-IAT '14, pages 158–165, Washington, DC, USA, 2014. IEEE Computer Society.

[20] Margherita Pagani, Charles F. Hofacker, and Ronald E. Goldsmith. The influence of personality on active and passive use of social networking sites. *Psychology and Marketing*, 28(5):441–456, 2011.

[21] S. Perera. Large scale data processing in real world: From analytics to predictions. In *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*, pages 8–8, Dec 2014.

[22] Maria Plummer, Starr Hiltz, and Linda Plotnick. Predicting intentions to apply for jobs using social networking sites: An exploratory study. In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences*, HICSS '11, pages 1–10, Washington, DC, USA, 2011. IEEE Computer Society.

[23] M.S. Prasad Babu and S.H. Sastry. Big data and predictive analytics in erp systems for automating decision making process. In *Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on*, pages 259–262, June 2014.

[24] Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel Van Aken, and Wim Meeus. Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of Personality*, 78(2):509–538, 2010.

[25] E. Strickland and E. Guy. Watson goes to med school [2013 tech to watch]. *Spectrum, IEEE*, 50(1):42–45, Jan 2013.

[26] F. Suchanek and G. Weikum. Knowledge harvesting from text and web sources. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1250–1253, April 2013.

[27] Ernest C. Tupes and Raymond E. Christal. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2):225–251, 1992.

[28] Tracy L. Tuten and Michael Bosnjak. Understanding differences in web usage: The role of need for cognition and the five factor model of personality. *Social Behavior and Personality: an international journal*, 29(4):391–398, 2001-01-01T00:00:00.

[29] Chunjing Xiao, Ling Su, Juan Bi, Yuxia Xue, and Aleksandar Kuzmanovic. Selective behavior in online social networks. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 206–213, Washington, DC, USA, 2012. IEEE Computer Society.

[30] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363 – 373, 2010.

[31] A. Ziberna and V. Vehovar. Using social network to predict the behavior of active members of online communities. In *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, pages 119–124, July 2009.