Recognising Personality Traits using Social Media

Vanshika Varshney[1], Aman Varshney[2], Tameem Ahmad[3], Asad M. Khan[4]
Department of Computer Engineering
Z. H. College of Engineering & Technology,
Aligarh Muslim University, Aligarh, India
[1] 607vanshika@gmail.com, [2] amanvars@gmail.com, [3] tameemahmad@gmail.com, [4] masadiitr@gmail.com

***Abstract*— With the fast-emerging technology, the use of social media has also increased to a great extent. People use various social media sites like Facebook, Twitter, Google+, etc. The users are allowed to share various kinds of personal information on their social media accounts. People write what they feel, they also share posts, re-tweets, etc. Therefore, social media can be used to predict personality of a person. Personality is a fundamental aspect of human behavior and can help a lot in various fields of business and marketing. In this paper, we explore the techniques of various machine learning algorithms in order to deduce about personality of a user from his activities on social media. We use three algorithms namely SVM, KNN and MNB and then compare the results of these three algorithms. Finally, in the end we provide a combined result of all the three algorithms.**

Keywords*— Machine learning, Social media, Personality Recognition, SVM, KNN, NB

I. INTRODUCTION

Social media has become an integral part of our life today. People today share a large amount of information about themselves in the forms of status updates, photos, check-ins, and by updating their descriptions on various platforms such as Facebook, Instagram, Twitter, etc. This sharing of information reveals very important personality traits about the users. This is quite an important aspect, as people use this information to judge others, and this can impact an individual's career, romantic and life prospects. Furthermore, this information can be beneficial for the corporations as this allows them to perform targeted advertising and marketing. These features can also be used in employee recruitment, career and health counseling.

"Personality is a set of individual differences that are affected by the development of an individual: values, attitudes, personal memories, social relationships, habits, and skills" [1]. The most accepted model for defining personality traits is the Big Five model. The five traits are extraversion, agreeableness, openness, conscientiousness and neuroticism. Each user can have more than one personality trait at once and therefore we use binary classifiers which separate users as having a certain personality trait or not.

In this paper, firstly we have discussed about the analysis of social networks and then we have described the various steps involved in predicting personality traits. We have described how data will be collected and then pre-processed in order to prepare it for the purpose of classifying. Finally, we have discussed about the three algorithms used and inferred which algorithm is the best among them.

II. ORGANISATION

The aim of this work is to provide an introduction to the concept of social networks and to explain how this concept can be used for better understand the ecology and evolution of personalities and behavioral strategies in general[24]. We first discuss the Big Five personality model and then use various classification algorithms to come out with the results. We will get to know about the various personality traits of a user from his social media activities. We then compare the results of different algorithms. Finally, we provide final output of personality traits in the form of a combined result of all the three algorithms.

III. SOCIAL NETWORK ANALYSIS

"Social network analysis is the process of investigating social structures through the use of network and graph theories" [2].

Analysis task of social networks includes getting information about the structure of network, finding various parameters such as radius, diameter, etc. of the network, finding relationships and communities in the social network[3].

There are two kinds of network analysis:

A. Ego network analysis

It mainly deals with the analysis of individual nodes. "ego" represents an individual node. They can be persons, organizations, etc. Behavior of an individual node is mined in this kind of analysis.

B. Complete network analysis

Rather than just dealing with individual nodes, it deals with the relationships between certain number of nodes. A lot of important tasks like equivalence analysis, measures such as closeness, between-ness and centrality all depend upon complete network analysis.

978-1-5386-0814-2/17/$31.00 ©2017 IEEE

IV. METHODOLOGY

The overview diagram of the methodology has been shown in the *Fig. 1*. Firstly, we need to have a dataset for our algorithms to work upon. For this purpose, we collected various tweets, status, etc. (Selected, for testing) from different users and then pre-processed this data to represent it in vector space model. Classification process applied further to give us labelled dataset. The final results will then predict the personality of the users based upon the Big five personality traits. The result will also predict primary and secondary personality characteristics which are obtained from the combination of different traits. A better knowledge based engineering approach would be required [25].

General procedure is shown in *Fig. 2*.

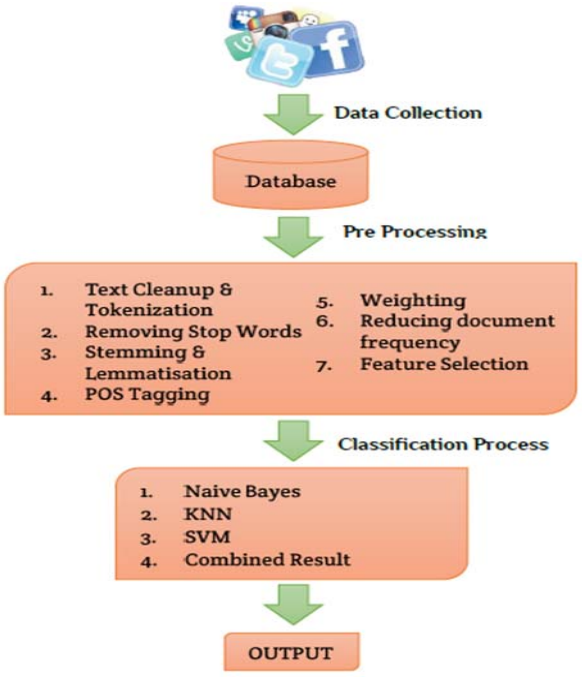

Fig. 1. Overview Diagram

A. Data Collection or Text Mining

User data in form of text is obtained from the collection of tweets and re-tweets of a Twitter user, status from a Facebook user, etc. User tweets are obtained using the Twitter API [4] which provides access to the twitter data. For data from Facebook, we use the MyPersonality dataset which is used for the purpose of various research tasks and activities. This dataset is acquired from a user's interaction with Facebook application [5][23]. A small test dataset is prepared from multiple sources.

All posts from a single user ID are then appended to from a long string which is then considered as a single document.

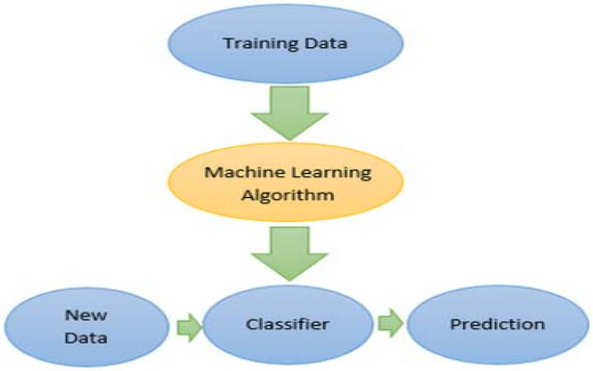

Fig. 2. General procedure

1) MyPersonality Data: The study of a user's personality reflects a lot about him. Statistical features such as number of likes, groups, tags, events, etc. [22] can prove to be of great significance. Demographic activities predicted such as age and gender also play an important role. An additional insight into a user's social network behavior can be accounted by Egocentric network parameters which represent number of friends, and measures such as density, brokerage, and betweenness.

2) Training the classifiers: There are two basic steps of using a classifier i.e. training and classification. It is an iterative process which helps us to build the best classifier possible. We train the classifiers on a corpus of various essays which are labelled with different personality traits. These essays are usually much longer than the tweets and status updates of the users. This process helps in creating a classifier which generalize across all domains. For the purpose of training data, we can make use of digital dictionary such as WordNet. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept [6].

B. Cleansing and Pre-processing of data

We first collect the data and remove all the inconsistencies and redundancies. For the classification purposes, the data has to represented in a vector space model. "Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in

information filtering, information retrieval, indexing and relevancy rankings" [7].

There are various steps involved in the pre-processing of data as shown in the *Fig. 1*.

1) Text Cleanup and Tokenization: We first perform text cleanup. By text cleanup we mean removing unnecessary information such as ads, etc., dealing with tables, formulas and various figures.

We then perform the step of tokenizing. "Tokenization is the process of converting a sequence of characters into a sequence of tokens" [8]. These steps are shown in the *Fig. 3*.

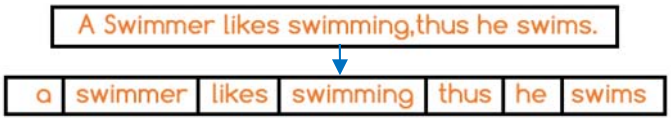

Fig. 3. Tokenization

2) Removing stop words: We perform this step by means of filtering. Stop words are the words which have very little or no significance at all but are necessary for the structural and grammatical purposes. This step is shown in *Fig. 4*.

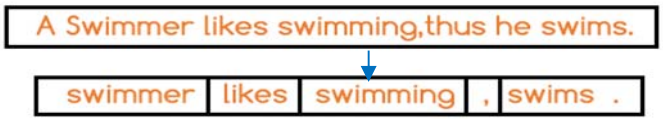

Fig. 4. Removing Stop Words

3) Stemming and Lemmatization: We then perform the step of stemming on our data. "Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form"[9]. Stemming algorithms are commonly known as stemmers. The most commonly used stemmer is the Porter stemmer [10]. An example of Porter stemming is shown in the *Fig. 5*.

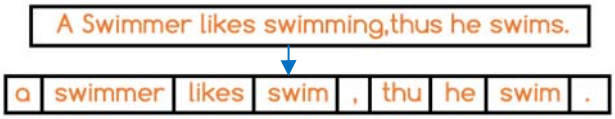

Fig. 5. Stemming

After stemming, lemmatization is performed. "Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form" [11]. Lemmatization is required to know the context of a word because the entire process is based upon the fact whether the word is a pronoun, noun, adjective, etc. An example of lemmatization is shown in the *Fig. 6*.

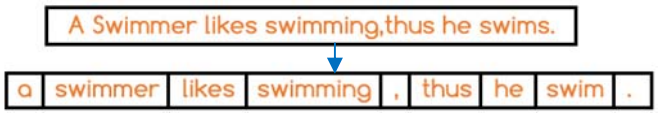

Fig. 6. Lemmatization

4) Part of Speech Tagging (POS Tagging): Each and every word involved in the dataset is a representative of a different human emotion and has a different meaning in a given context. Therefore, we need to associate each word in our document with a unique tag. "In corpus linguistics, part-of-speech tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a sentence"[12]. This step is shown in *Fig. 7*.

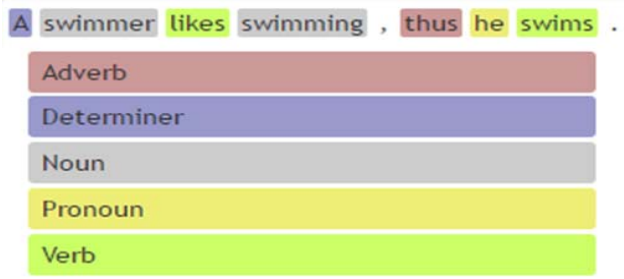

Fig. 7. POS Tagging

5) Calculating Tf-Idf (Weighting): It is a weighting scheme for vector space model in text mining. "Tf–Idf is a numerical statistic that is intended to reflect how important a word is to a document. It is used as a weighting factor in information retrieval" [13].

Tf stands for term frequency which represents the frequency of each word in a document. Df stands for document frequency i.e. the words that are used more in a document have more weight.

Collectively, Tf-Idf stands for the combination of sublinear Tf and inverse document frequency. The formula for calculating Tf-Idf is given by-

$$tfidf_t = f_{t,d} \times log \frac{N}{df_t}$$

$tf \cdot idf_t$ = weight of term *t*
$f_{t,d}$ = number of occurrences term *t* in document *d*
N = total document
df_t = number document contains term *t*

6) Reducing document frequency: The number of words in the dataset is limited to the words that appear more frequently in a document. By doing so, we can increase the efficiency and accuracy of the classification process and also reduce load on the system.

7) Feature Selection: "It is the process of selecting a subset of relevant features (variables, predictors) for use in model construction" [14].

This process takes place before the training of the classifier. It serves two main purposes. Firstly, it helps to make the training process simpler by reducing the effective vocabulary size. It proves to be of great help as some classifiers like NB are very expensive to train. Secondly, it also removes noisy features which increases the overall accuracy of the system. It can also help to reduce the total memory usage.

C. Applying the Classifiers

Personality classification from the text requires multi-label classification. "Multi-label classification and the strongly related problem of multi-output classification are variants of the classification problem where multiple target labels must be assigned to each instance" [15].

It means that a person can have more than one personality trait at the same time or it may also be possible that a person does not have any one dominant personality characteristic. We will then use binary classifiers on the dataset.

1) Multinomial Naïve Bayes (MNB): Naïve Bayes is a classifier which is based upon the probability. This algorithm performs very well under complex conditions also. MNB is a slight variation of the Naive Bayes algorithm. It proves to be very useful when we have limited resources in terms of memory.

We use MNB because multiple occurrences of a word in our text plays a very important role. NB assumes that the various features used in the classification are independent. During the text classification, we use the tokens of a document in order to classify it on appropriate class. MNB estimates the conditional probability of a particular token given a class as the relative frequency of a token in a document belonging to a certain class. It takes into account the number of occurrences of a token in a training document from a certain class also including multiple occurrences. MNB equation is as under-
Naïve Bayes can be used to predict a user's personality from his self-written text [16].

$$P(X|c) = \log\frac{N_c}{N} + \sum_{i=1}^{n} \log\frac{t_i + \alpha}{\sum_{i=1}^{n} t + \alpha}$$

$P(X|c)$ = probability of the document X in any class *c*
N_c = total number of documents in class *c*
N = total number of documents
t_i = weight of the term *t*
α = total weight of the term *t* in class *c*
$\sum_{i=1}^{n} t$ = represents smoothing parameter

Below are the definitions of some terms used in pseudocode:

Prior Probability: It is the probability of a document being present in a certain category from some set of documents.

Likelihood: Provided that a document belongs to a certain category, it is the conditional probability of some word occurring in the document.

Below is the pseudocode of this algorithm:

Naïve Bayes Pseudocode

1. Given training data which consists of documents belonging to different classes.
2. Calculate the prior probability.
3. Find the word frequency of each class.
4. Calculate likelihood
5. Classify a new document X based on the probability P(X/W)
   i) Find P(A/W) = P(A)* P(word1/class A) ..... *P(wordn /class A)
   ii) Find P(B/W) = P(B)* P(word1/class B) ..... *P(wordn /class B)
   And so on.
6. Assign document to class that has higher probability.

Fig. 8. Naïve Bayes Algorithm

2) K-Nearest Neibors (KNN): It is a non-parametric and also a lazy learning algorithm. The input here consists of the k closest training examples in the feature space. Any object will be classified on the basis of majority of votes of its neighbors. The value of k is a positive integer. This algorithm uses distance function to calculate results i.e. the nearer neighbors contribute more towards the output than the distant ones [17]. The distance function that we use is Cosine similarity. It finds similarity between various documents.

Mainly it is used for the purpose of detecting emotions from texts. Basic emotions are happiness, anger, love, hatred, etc. Below is the pseudocode of this algorithm:

KNN Psuedocode

1. Determine parameter K=number of nearest neighbours.
2. Calculate the distance between the query instance and all the training samples.
3. Sort the distance and determine nearest neighbours based on the K-th minimum distance.
4. Gather the category of the nearest neighbours.
5. Use simple majority of the category of nearest neighbours as the prediction value of the query instance.

Fig. 9. KNN Algorithm

3) SVM: It is a supervised learning algorithm. It analyses the data and marks it as a part of one of the two classes and thus it is a non-probabilistic binary classifier. SVM constructs a hyperplane which is used for classification purposes [18].

4) Combined Result: Different algorithms give different results. For such a purpose, we use the concept of combined result. Majority vote of all the three methods mentioned above. If an algorithm gives wrong output, then the result of other two algorithms can be used to come at the conclusion. This approach helped to deduce the final results in an state where there were different outputs.

V. LIMITATIONS AND FUTURE ASPECTS

Some of the limitations of personality deductions are that some users often misrepresent information[26] about themselves on social media, and analysis of that information will lead to inaccurate results. To avoid this, it is important to analyze different social media accounts of the same user, and pick up information from those accounts and compile it in a single document. Along with that we can also combine a user's self- written information in that document in order to get better results and to improve accuracy. Another challenge is the usage of special characters and slang language by the user.

Since social media breaks through all barriers of regions and language, and has a user base comprising of people who speak thousands of languages, it is important to have multilingual support. We therefore need mechanisms to convert these languages into English. We can make use of the various already existing mechanisms for this purpose.

VI. CONCLUSION

A number of machine learning techniques have been utilized to predict the personality traits of users from social media platforms, based on the Big 5 model of personality traits. The various algorithms explored were Multinomial Naïve Bayes, Support Vector Machine, K-Nearest neighbors.

In this paper [19], it is determined that the best results were obtained by using MNB with an average accuracy of about 60%. SVM and KNN had worse performances than MNB. Owing to difficulties in separating the data among two classes, SVM gave inferior results. KNN also couldn't match up to MNB, as there was a difficulty in choosing the accurate value of k for the required purpose. As the value of k plays an important part in calculating the probability, this is a major disadvantage.

Combined result gave the best performance. In the combined result, we use the majority vote of the three aforementioned algorithms. In case an algorithm predicts incorrectly, this combined result helps to filter out the wrong result as it takes into account the correct result obtained by the other two algorithms.

References

[1] Wikipedia contributors. "Personality." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 Apr. 2017. Web. 30 Apr. 2017.

[2] Wikipedia contributors. "Social network analysis." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 18 Apr. 2017. Web. 1 May. 2017.

[3] Akhtar, Nadeem. "Social network analysis tools." In Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, pp. 388-392. IEEE, 2014.

[4] Twitter API available: https://apps.Twitter.com/ [18- Apr- 2017].

[5] Kosinski, Michal, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." American Psychologist 70, no. 6 (2015): 543.

[6] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu.

[7] Wikipedia contributors. "Vector space model." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 17 Oct. 2016. Web. 28 Apr. 2017.

[8] Wikipedia contributors. "Tokenization (lexical analysis)." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 10 Jan. 2017. Web. 29 Apr. 2017.

[9] Wikipedia contributors. "Stemming." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 1 Mar. 2017. Web. 29 Apr. 2017.

[10] Porter, Martin F. "An algorithm for suffix stripping." Program 14, no. 3 (1980): 130-137.

[11] Wikipedia contributors. "Lemmatisation." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 6 Nov. 2016. Web. 29 Apr. 2017.

[12] Wikipedia contributors. "Part-of-speech tagging." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 5 Mar. 2017. Web. 30 Apr. 2017.

[13] Wikipedia contributors. "Tf–idf." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 19 Apr. 2017. Web. 30 Apr. 2017.

[14] Wikipedia contributors. "Feature selection." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 24 Mar. 2017. Web. 30 Apr. 2017.

[15] Wikipedia contributors. "Multi-label classification." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 30 Jan. 2017. Web. 30 Apr. 2017.

[16] Kibriya, Ashraf M., Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. "Multinomial Naive Bayes for Text Categorization Revisited." In Australian Conference on Artificial Intelligence, vol. 3339, pp. 488-499. 2004.

[17] Wikipedia contributors. "K-nearest neighbors algorithm." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 12 Mar. 2017. Web. 30 Apr. 2017.

[18] Wikipedia contributors. "Support vector machine." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 27 Apr. 2017. Web. 30 Apr. 2017.

[19] Pratama, Bayu Yudha, and Riyanarto Sarno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM." In Data and Software Engineering (ICoDSE), 2015 International Conference on, pp. 170-174. IEEE, 2015.

[20] Farnadi, Golnoosh, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. "Recognising personality traits using Facebook status updates." In Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13). AAAI, 2013.

[21] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "others, Introduction to information retrieval, vol. 1." (2008).

[22] Khan, Ayesha Tooba, Ibra Husain, and Yusuf Uzzaman Khan. "Seizure onset patterns in EEG and their detection using statistical measures." In India Conference (INDICON), 2015 Annual IEEE, pp. 1-5. IEEE, 2015.

[23] Ahmad, Shamim, and Tameem Ahmad. "Offering a Guided Learning of Quranic Knowledge to the Seekers through the Internet." In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, pp. 72-76. IEEE, 2013.

[24] Ahmad, Tauseef, and Tameem Ahmad. "Using the concept of Multi-Threaded Programming Preparing the Object Oriented Design Model." International Journal of Advanced Computer Research 2, no. 4 (2012).

[25] Ahmad, Tameem, Shamim Ahmad, and Mohammed Jamshed. "A knowledge based Indian agriculture: With cloud ERP arrangement." In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on, pp. 333-340. IEEE, 2015.

[26] Bansal, Palak, and Tameem Ahmad. "Methods and Techniques of Intrusion Detection: A Review." In International Conference on Smart Trends for Information Technology and Computer Communications, pp. 518-529. Springer, Singapore, 2016.