Machine Learning for Personality Analysis Based on Big Five Model

Joel Philip, Dhvani Shah, Shashank Nayak, Saumik Patel and Yagnesh Devashrayee

Abstract The proposed research attempts to emulate a statistical report making system, which takes into considerations, the activities of user and their behavior online by means of their interactions on varied array of social media platforms. It is possible that youngsters may come across incidents on Internet, which probably may be inappropriate for their age group or may push them towards certain erratic psychological behaviors. This study caters to such arising needs, for various individuals—young or old, alike, so as to keep a tab upon their own online activities through browsing history which may be directly/indirectly blend into their human characteristics. On social media, people express their likes, dislikes, thoughts, opinions, and feelings which sum up to be their own personality. This data (thoughts and opinions on social platform and browsing history) can be exponentially aggregated to identify user's personality traits. It can then be used for self-monitoring, parental monitoring, or for businesses who wish to hire employees based on their personality criteria, if approved by concerned users. For this study, we have used supervised machine learning algorithms like Naïve Bayes and Support Vector Machines. We have evaluated their performance through the combinations of different feature extraction process like BOW, TF, and TF-IDF with each classifier. In conclusion, we have found that TF-IDF with SVM has the best performance.

Keywords Personality · Social media · Machine learning · Psychology
Deep learning · Data analytics · Data mining · Cognitive science
Big Five personality traits · Neural network

J. Philip · S. Nayak · S. Patel · Y. Devashrayee
Universal College of Engineering, Thane, India
e-mail: tjoelphilip@gmail.com

D. Shah (✉)
St. John College of Engineering and Management, Palghar, India
e-mail: shahdhvani08@gmail.com

V. E. Balas et al. (eds.), *Data Management, Analytics and Innovation*,
Advances in Intelligent Systems and Computing 839,
https://doi.org/10.1007/978-981-13-1274-8_27

1 Introduction

A social platform is a place where a person expresses or presents themselves the way they are. The data which comes under social media is nothing but series of intercommunications among humans and post in any format. This can be in tremendous volume and is ever-increasing in size. The person opens about their likes, dislikes, thoughts, opinions, and feelings which sum up to be their own personality. Thus, this data can be collected to identify user's personality. According to the statistics [1], there were a total of 2.01 billion active users on Facebook in June itself which meant a rise of 17% since the last year and these numbers are to increase over the coming years. Also, 29.7% of users are of age 25–34, and it is the most common age demographic. Also, it comes with the responsibility of handling huge amount of data, for every 60 s: around 500,000 comments 290,000 statuses are updated, and 136,000 photos are uploaded on Facebook. All these comments, status, and photos roughly help generating a user's profile thus revealing a lot about their personality [1].

A person's personality is their characteristics and aspects of other's perception. The capability of finding connections between behavioral aspects derived from the data collected from social media can help us in bifurcating users into various groups based on their personality, this is the main objective of this research. Personality refers to individual characteristics pattern of thinking, behaving, feeling which makes them different from other individual. Personality of person is dependent on the type of situation and mood. When it comes to analyzing personality the one thing that comes into the picture is the "Big Five Model" or "Five Factor Model" which comprises of five major components such as Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism [2].

The most famous application of machine learning which everyone is aware of is the sentimental analysis. In sentiment analysis, the machine learning model attempts to predict the emotional value based upon the input supplied to it. This research aims at doing far more than just the analysis of emotions. The aim is to label a person's personality by analyzing the textual information generated by him on social media to create a statistical report depicting the resources on which the user is investing maximum time and efforts by analyzing his browser history.

2 Related Work

In [3], the authors made use of naive Bayes and SVM models in order to classify a Tweet message into positive or negative category where positive and negative indicate the emotion based on the message. The SVM model when used correctly does work as expected but the ill-effects of it results in overfitting. In [4], the researchers have presented a method to accurately predict the user's personality by making use of data obtained from smartphones, cell phone application, Bluetooth,

and SMS usage along with the use of supervised learning algorithm help constitute a data for predicting. This process of collection of data and then running it through a prediction model helps generate characteristic for the users. In [5], the personality traits of microblog users are generated. It proposes multi-task regression and incremental regression algorithms to predict the Big Five personality from online behaviors. This indicates that even online microblog can be used to accurately predict personality of users. However, no NLP method has been used as well as data set is very low, i.e., only blog of topic. In this work [6, 7], the authors have proposed a unique methodology to extract human emotions from text. It applies various machine learning algorithms to predict readers reaction.

This study aims to collect data from various resources like Twitter, Facebook, Quora, and the browsing history of a user. This study comes with two aspects: the first task is to analyze his opinions to predict his personality and the second task is to generate a report on his daily online activities to know his interest and area of research. In conclusion, people express opinions/feelings in complex ways, which makes understanding the subject of human opinions a difficult problem to solve. One of the biggest challenges of this research is data collection from various resources and categorizing the text according to the Big Five model for training purpose.

3 Proposed Idea

The basic idea here is to generate subcategories to the five factors which can help classify the personality with much better meaning. One problem faced here is that the personalities of people change according to the social networking site they use. This would mean the personalities obtained from the data provided would not be accurate, but it changes according to the social site used, e.g., A person's behavior on any social networking websites like Twitter or Facebook would be different as compared to their behavior on professional networking websites like Stack Overflow or Quora. People tend to behave differently according to the social website on which they have their online presence. If we are to follow the conventional method and try to get the personality for a person, then the output obtained would not defy correctness in their value. Our framework will be designed keeping the exact problem in mind assuming that one single user has account on social media sites as well as he is active on educational/informational sites. The framework will analyze different data differently according to the platform used by the person to check their personality. This way we can find traits related to behavioral aspects from social networking websites whereas technical qualities from professional networking websites. After we have the basic structure defined, we can start the process of determining the personality of the user.

3.1 Data Collection

The process begins with collection of information or data from the respective social networking websites.

Twitter: Twitter API provides tweets, likes, and number of followers. This helps us perform analysis on the tweets of the users to generate their personality chart.
Facebook: Graph API can help us get all the publicly available information like liked pages, liked movies, etc., can help us generate personality by going through the likes of the user.
Quora: The Quora API helps us get all the questions or answers provided by the users.

3.2 Data Cleaning

After the data collection process, comes the data cleaning process. Texts written by humans usually contain noise and we can extract features out of the texts only after the removal of the noise. Natural Language Processing helps the machines to understand the human language [8]. Techniques used for text cleaning include tokenization, POS tagging, stemming, removal of stopping words, etc. (Fig. 1).

On inspecting the dataset, we noticed that the raw data contained a lot spelling mistakes along with redundancy. To solve this problem, we ran the whole of the dataset through a spellchecker and deleted all the redundant data. Sometimes, the user writes about his feeling for a situation on his blog or writes a technical article. This sort of user data contains lots of words as compared to a simple tweet. To deal with this situation, we can use the concept of text summarization and then pass the sentences into the data cleaning process. For browsing history, all what we have is a csv file containing the URLs of the web pages visited so far. To find that where the user is investing his time, we should extract information from the URLs. To do so, we have a python library named, 'Extraction' used for extracting titles, descriptions, images and canonical URLs from web pages. Once we get the topic, we can classify it into an educational, informative, or entertainment-based content. As stated earlier, the difficult task is getting the classified historical data for the training purposes.

3.3 Features

We made use of two features in our model: a bag of words (BOW) and WordNet synset.

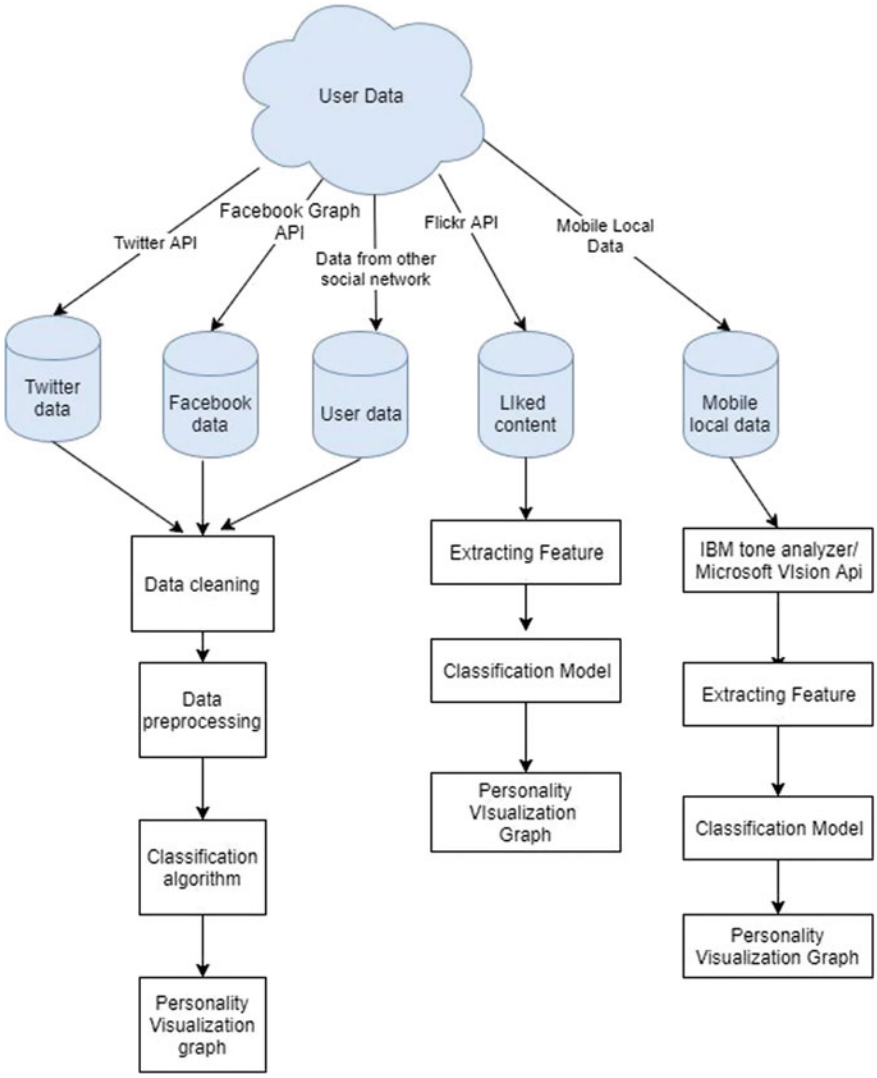

Fig. 1 System architecture

3.3.1 Bag of Words

Bag of words approach (BOW) is the most common methodology for extracting feature from sentences and documents. Here, each word counts as a feature and the histograms of words in the text are examined. We use Bag of words concept in methods of document classification where the repetition frequency of each is utilized as a feature for training a classifier.

3.3.2 WordNet

Now, for improving the quality of the feature set and reduce overfitting, the concept of WordNet comes handy, as we use it to map the words in the tweets/messages onto their synonym set (synset). By mapping words, we conclude that words having same meaning prompt same personality criteria. The quantity of features is lessened but improves the coverage of particular feature. This proves beneficial as it covers the words outside training set but are in similar synset.

3.4 TF-IDF

Observing the patterns certain function words as 'the', 'and', 'he', 'she', 'it' repeat commonly across most descriptions. As a result, it would not be a smart decision to emphasize on such words while implementing Bag of Words to classify the documents. Clipping off the words in a list of high-frequency stop words can be one method which can be used. Considering Term Frequency-Inverse Document Frequency (TF-IDF) weight of each word can be an alternative [9]. Here, more is less and less is more meaning that sometimes the less occurring words weight heavy in terms of information content whereas more occurring words can be not much useful considering information point of view. Following equation helps us to produce weights for each word:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\mathrm{idf}_i = \log \frac{|D|}{|dt_i|}$$

$$\mathrm{tfidf}_i = t\mathrm{f}_{i,j}\mathrm{idf}_i$$

$\mathrm{tf}_{i,j}$: importance of term i in document j

$n_{i,j}$: number of times term i occured in document j

$\sum k\, n_{k,j}$: total number of words in document j

idf_i : general importance of term i

$|D|$: total number of documents in corpus

$|dt_i|$: number of documents where the term t_i appears

3.5 Classifier

For this study, we used classification algorithms like Naïve Bayes and SVM and evaluated their performance. We need to classify the text into five main categories (Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism). Before classification, the feature extraction process takes place using the NLP techniques mentioned above. Then, the features along with the label are fed into supervised machine learning models as shown in Fig. 2. Now, the system is ready to classify the given unknown input after the training and validation process. We have used the scikit-learn machine learning library in python. Also, the feature extraction process is done using sklearn library. For this study, we have trained the data only to output only two classes, i.e., Conscientiousness (Label 1) and Neuroticism (Label 2). Label 1 contains 570 sentences whereas Label 2 contains 621 sentences for the training purpose and 20% of the data was used for the development tests. We need to provide very less amount of training data which acts as a benefit in terms of N.B and SVM classification and help decide criterion necessary for categorization [10].

4 Results/Expected Output

Till now, we have collected data for three classes, trained the data for two classes and carried out three experiments for each classifier. For each experiment, we use a "feature vector", a "classifier" and a train-test splitting strategy. For experiment 1, we used a Bag of Words (BOW) representation of each document, for second experiment we used Term Frequency and for the final experiment, we used TF-IDF along with NB classifier and then with SVM classifier.

Table 1 shows the classification report for MultinomialNB classifier whereas Table 2 shows the same for SVM model.

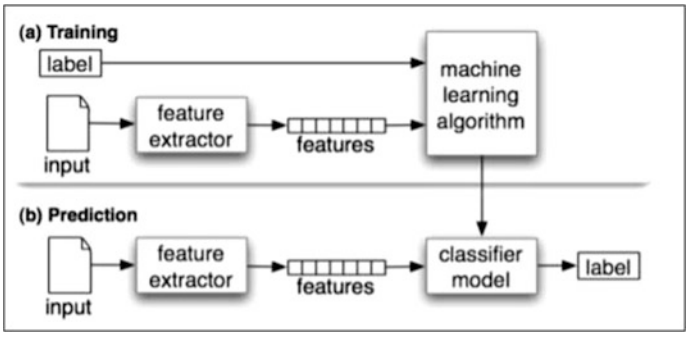

Fig. 2 System flow

Table 1 Classification report NB classifier

	Precision	Recall	F1-Score	Support
Conscientiousness	0.95	0.97	0.96	577
Neuroticism	0.97	0.96	0.96	619
Avg./Total	0.96	0.96	0.96	1196

Table 2 Classification report SVM classifier

	Precision	Recall	F1-Score	Support
Conscientiousness	0.98	0.97	0.98	616
Neuroticism	0.97	0.98	0.96	585
Avg./Total	0.98	0.97	0.97	1196

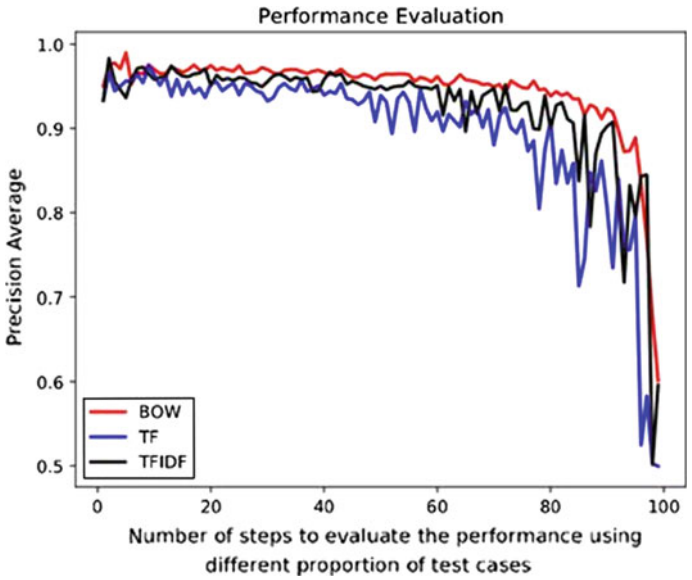

Fig. 3 Naïve Bayes classifier

Figures 3 and 4 show the average precision value for random train and test subsets for cross-validation and performance evaluation process for NB and SVM classifier models, respectively [11]. It compares the different feature extraction process like BOW, TF, and TF-IDF with BOW for each classifier. From the below results, SVM with TF-IDF outperforms Naïve Bayes Algorithm for this research study [10].

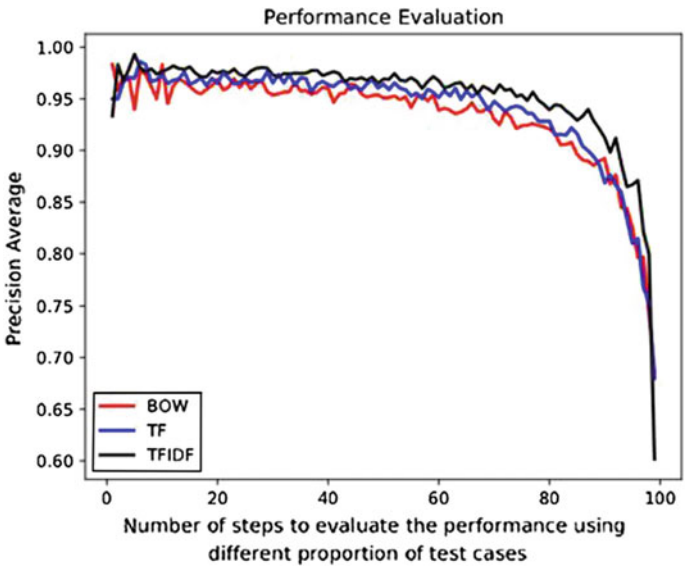

Fig. 4 Support vector machine

5 Limitation and Future Work

People with extensive use of social media platforms tend to post or upload a lot about themselves through timely updates, caption-description, images, videos, and other interests. Also, at times a user share status on ongoing topics currently happening in the real world. At such circumstances, it is required that the system should have information about that topic so that system can relate the user's status with topic and can analyze the actual meaning of status.

Figure 5 shows the expected output in the future. It will show Big Five model and the sub traits of every factor to help us know more about user. So, the system will generate multiple personality graph as system will take data from different sites. Every graph will have different values of personality traits for one user as user tends to act different on different social site. At the end, system will merge all value of personality traits and will generate one final personality analysis graph.

In Future, we can consider images, videos and voice to better understand a human's personality trait from all aspects and not only text incorporating deep learning algorithms with far better results.

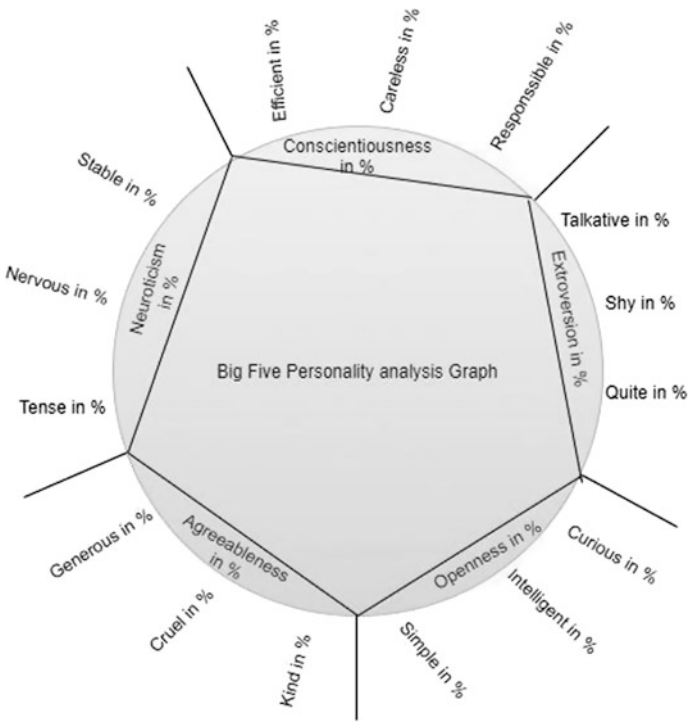

Fig. 5 Personality graph

6 Conclusion

This study comes with an efficient approach of summarizing a human's personality which is based on Big Five model using natural language processing and machine learning techniques to train the model with the personality traits and thus generating a graphical report of user's character. We found that SVM with TF-IDF feature extraction process produced better results. With more datasets and better feature extraction process like using bigrams or n-grams instead of BOW would produce better outcome. Such system will be an aid in understanding a person's characteristics which would help during recruitment process, employee satisfaction, for parental monitoring, etc.

References

1. The top 20 valuable facebook statistics. Updated September 2017. Available https://zephoria.com/top-15-valuable-facebook-statistics/. Accessed August 16, 2017.
2. McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215.
3. Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011, October). Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)* (pp. 149–156). IEEE.
4. Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011, June). Who's who with big-five: Analyzing and classifying personality traits with smartphone. In *2011 15th Annual International Symposium on Wearable Computers (ISWC)* (pp. 29–36). IEEE.
5. Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., & Zhu, T. (2013, November). Predicting big five personality traits of microblog users. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 1, pp. 501–508). IEEE.
6. Kalghatgi, M. P., Ramannavar, M., & Sidnal, N. S. (2015). A neural network approach to personality prediction based on the big-five model. *International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2*(8), 56–63.
7. Zhang, W., Zhao, G., & Zhu, C. C. (2014). Mood detection with tweets.
8. Natural language processing. Available https://en.wikipedia.org/wiki/Natural_language_processing. Accessed August 16, 2017.
9. Available http://www.markhneedham.com/blog/2015/02/15/pythonscikit-learn-calculating-tfidf-on-how-i-met-your-mother-transcripts/. Accessed November 2, 2017.
10. Available http://scikit-learn.org/stable/modules/feature_extraction.html. Accessed November 8, 2017.
11. Available https://www.datasciencecentral.com/profilesblogs/7-important-model-evaluation-error-metrics-everyone-should-know. Accessed November 25, 2017.