Personality Analysis using Social Media

Sagar Patel, Mansi Nimje, Akshay Shetty, Prof. Sagar Kulkarni
Department of Information technology,
Pillai College of Engineering, Navi Mumbai, India – 410206

***Abstract*— Social media has become a platform for users to present themselves to the world openly by revealing their personal views and insights on their lives. Hence, extracting information from social media and yielding insightful results about the person has become easier. We are beginning to understand that this information can be efficiently utilized to analyze the personality of the concerned person. In this paper, we aim to gain knowledge of the personality of a user by using the social media platform of the concerned user. These social media platforms could be Facebook or Twitter. Personality analysis can help to reveal many types of interactions: it can be used to predict a suitable job for a person and also know about his efficiency in the same; professional, romantic, his nature's traits can also be studied. Personality analysis may even be able to detect the roots of any kind of suspicious, immoral or wrongful trait in a person.**

Keywords—Personality, social media, MBTI, Big Five Personality, analysis

1. INTRODUCTION

Personality is defined as the characteristic set of behaviors, cognitions and emotional patterns that evolve from biological and environmental factors. While there is no generally agreed-upon definition of personality, most theories focus on motivation and psychological interactions with one's environment. Trait-based personality theories, such as those defined by Raymond Cattell define personality as "The traits that predict a person's behavior". On the other hand, more behavioral-based approaches define personality through learning and habits. Nevertheless, most theories view personality as relatively stable.

Since the inception of social media, a prodigious amount of status updates, tweets and comments have been posted online. The language people use to express themselves can provide clues about the kind of people they are, online and off the digital media. Some personality psychologists study publicly available social media data in addition to solicited surveys. However, they still start with predefined traits like extroversion, neuroticism or narcissism and correlate them with the writing. In other research, linguists have used algorithms to identify topics of conversation, but they do not have much to say about the personalities of the conversationalists. Hence research in this sector is important.

2. BACKGROUND AND RELATED WORK

2.1 The Big Five Personality Model

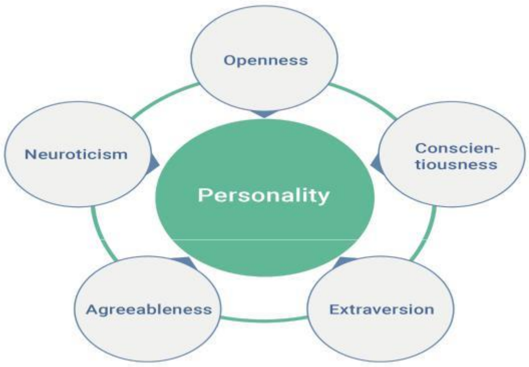
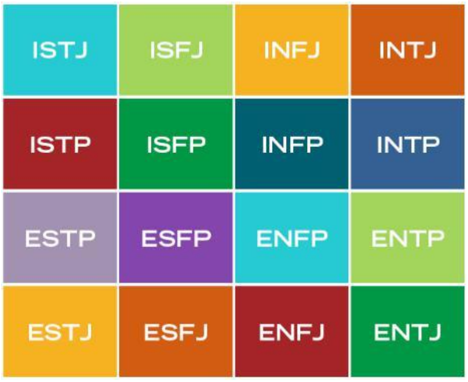


Figure 1: Big Five Personality Model

The "Big Five" model of personality [12] dimensions is one of the most well-researched and well-regarded measures of personality structure in recent years. The models five domains of personality are Openness to experiences, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.

2.2 The Myers Briggs Personality Model

The Myers Briggs Type Indicator Personality model [13] is also called the MBTI model. The Myers–Briggs Type Indicator (MBTI) is an introspective self-report questionnaire indicating differing psychological preferences in how people perceive the world and make decisions."The underlying assumption of the MBTI is that we all have specific preferences in the way we construe our experiences, and these preferences underlie our interests, needs, values, and motivation."



ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTP	ESFP	ENFP	ENTP
ESTJ	ESFJ	ENFJ	ENTJ

Figure 2: The Myers Briggs Personality Combinations

2.3 Related Work

Before we actually began to start working on building our own system for Personality Analysis, we referred multiple research and papers from around the world. We studied and tried to identify all the drawbacks and flaws of the existing system architectures proposed. We realized that the predictive models must be scalable and dynamic to meet the requirements of ever-growing data and vast possibilities. Also, Very little work is done on semantic analysis for personality detection from tweets. It was identified that different approaches can be followed to improve recognition of linguistic constraints such as slang usage, communal bias, abbreviations and sentiment of tweets. Most of the work done is limited to the English language and hence the involvement of different languages is required.

The authors of [1] have used the supervised learning approach to compute personality using SVM, [2] authors have explored the importance of twitter profiles in predicting the personality using Logistics Regression and neural network. Also the authors [3] have used SVM, NaiveBayes for personality prediction from written text. For the personality prediction of people in the USA use of images was made [5]. Also the use of NaiveBayes, LogisticsRegression, KNN and SVM is made for classification of different personality traits from social media [6][7], the authors [8] have worked on dutch, Spanish and Italian language for personality prediction, some authors [9][10] have used the deep learning approach using RNN and LSTM for getting semantic analysis for personality classification.

3. DATASET

The dataset used is from Kaggle[14]. This dataset is publicly available. It consists of tweets tagged with one of the 16 MBTI types. These tags are combinations of four characters. The dataset consists of a total 8660 rows and 2 columns. The distribution of MBTI traits (number of rows out of 8600) in each class is as follows:

- Introversion(I) : 6664; Extraversion(E) : 1996
- Sensing(S) : 7466; Intuition(N) : 1194
- Thinking(T) : 4685; Feeling(F) : 3975
- Judging(J) : 5231; Perceiving(P) : 3429

There are fifty posts included for every user. This data comes from uses of personalitycafe.com, an online forum where users first take a questionnaire that sorts them into their MBTI types.

4. METHODOLOGY

This section describes the methodology(figure 3) followed to obtain the personality scores on the MBTI scale:

4.1 Pre-processing

In this step we preprocess the dataset by removing the unwanted characters and words from our dataset.

- **Hyperlink removal:** As we are only dealing with the simple textual data, hyperlinks are of no use and are removed using a regular expression.
- **Emoticons handing:** The emoticons are converted to text which adds to the quality of our training dataset.
- **Unwanted character removal:** Unwanted characters like punctuation marks, numbers, multiple spaces and symbols which do not provide meaningful information are removed.
- **Stopword removal:** In the dataset there are many stopwords like a, for, etc which does not provide meaningful information while training so we remove it by using the nltk stopword library.
- **Lemmatization:** Grouping of different inflected forms of a word, called lemma is done (gone, going, went to go) since inflected forms of the same words carry one same meaning.
- **Stemming:** Stemming is a crude heuristic process that chops off the ends of words or removes derivational affixes. To achieve stemming, a snowball stemmer was used.
- Finally we have used the one-hot encoding technique to encode the set of four MBTI personality traits into 0’s and 1’s for classification tasks.

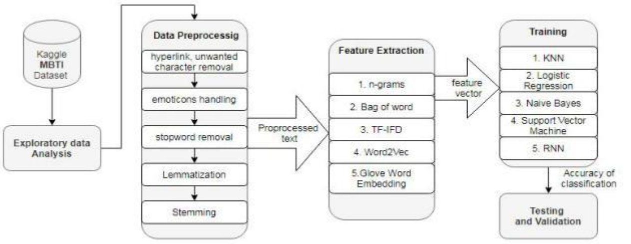


Figure 3: Proposed system

4.2 Sampling

Training the model with unbalanced data will yield a biased result. Hence sampling by using the Upsampling method is done where the sample of minority class is matched to the majority class.

4.2 Feature Extraction

A core step in NLP is to convert raw or annotated text into features, which will give the machine learning model a simpler, more focused view of the text. The models applied for this step are -

- **N-grams:** N-grams are contiguous sequences on n-items in a sentence. It helps to give more features of different grams and gain semantic analysis from the text.
- **Bag-of-words (BOW):** Using the BOW technique different features related to each personality type are extracted based on words for classifying the personality traits at the word level.
- **Term Frequency and Inverse Document Frequency (TF-IDF):** Term Frequency and Inverse Document Frequency or TF-IDF, produces the product of term-frequency and inverse-document-frequency. As our dataset is unbalanced for a few personality types, it results in having fewer words related to a few personality types. So to extract the features from low-level TF-IDF is useful.

- **Word2Vector (W2V):** Word2vec takes as its input, a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. For getting the semantic analysis the W2V features extraction is used.
- **Glove Word Embedding:** In the RNN model for our embedding layer, we use an embedding matrix to the 50-dimensional Glove representation of that word.

4.3 Training

Text classification is an important task in Supervised Machine Learning. The dataset has been split into 80:20 for training and testing purposes. The training dataset initially when trained on K Nearest Neighbour (KNN) with Bag of Words (BOW) and Term Frequency and Inverse Document Frequency (TF-IDF) gave in low-performance results. Later the Logistics Regression model was trained using Word2Vec of 200 dimensions and Word2Vec of 200 dimensions with Upsampling which gave excellent results with semantic analysis. Naive Bayers was also trained using BOW with upsampling, simple BOW and TF-IDF with Upsample and the resulting performance was similar to that of Logistics Regression, but the only drawback was that the analysis was done at word-level. An SVM model was also used and its results were slightly lesser accurate than the ones gained using Naive Bayes and Logistic Regression.

We have also trained our model on neural networks using Recurrent Neural Network (RNN) with Long term short memory (LSTM), where we have used the Glove embedding vector for the Embedding layer. Although due to a short dataset the RNN performance was recorded as low. Yet it was useful to capture some of the information in text data which tends to get ignored by the above-specified models.

4.4 Testing and Validation

For analyzing the performance of the above-trained model we used the testing dataset with the models. The process was used to gain details of the performance of the above-trained models, the accuracy of the models, confusion matrix, f1-score, recall and precision of models. Based on the analysis of the results the best model was selected. Using the Validation the hyperparameter tuning is done and the best-tuned parameter for our training model is obtained. We emphasized the fact that the best model selection does not depend only on the accuracy factor but also on the confusion matrix which gives the matrix of correctly classified classes which is very helpful to see that our model is under fitted or overfitted on testing dataset. Hence while the process of testing and validation each and every crucial factor were taking into consideration so that the resulting model will come out to be the most efficient one.

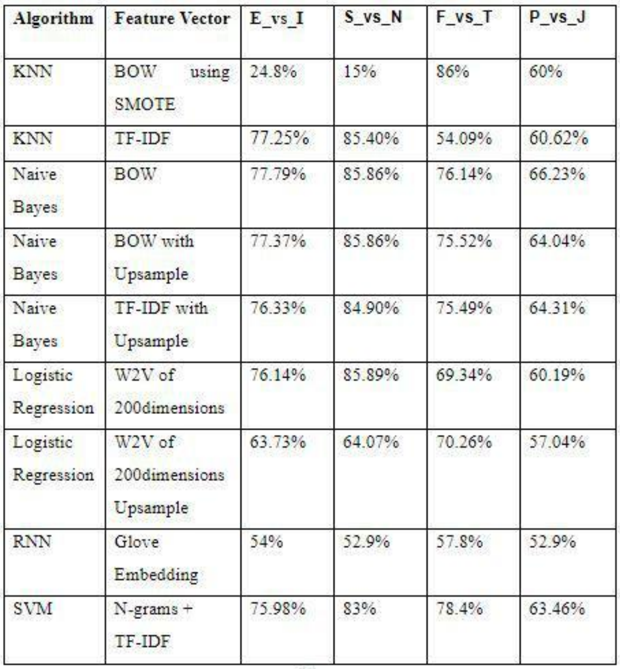
4.5 Web Development

After performing all the steps involving the data collection and its processing along with the critical training, evaluation and validation of the models mentioned above, the best model was selected. This final model having the highest performance was used in web deployment. A Web Application was developed to provide a simple user interface to take the personality test. Web application was developed with the Flask framework which helps well to serve the web and can be easily integrated with python which is used in the backend working. For passing the tweets of any user for personality prediction the TwitterAPI is used which gives the list of tweets of the specified user and then it is sent to the backend for preprocessing and also converting it to a similar format that of features extraction technique used during the training phase and then to saved model. Finally the results are displayed in a graphical format to the user at the user interface.

5. RESULTS

The goal of the proposed approach was to analyze and predict the personality traits of a person. A comparative study(Table 1) of various classifiers and feature vectors was done to obtain the most accurate prediction of the personality trait.

KNN performs very badly and also it is computational very slow. Logistic Regression shows a better result compared to KNN. Naive Bayes classifier has greater performance than KNN and Logistic Regression for E_vs_I, S_vs_N, F_vs_T, and P_vs_J, where accuracy for P_vs_J is low as compared to others. RNN and SVM classifiers also reported with low performances for classifying the user personality types, as SVM and RNN require the larger dataset for better accuracy (Table 1).



Algorithm	Feature Vector	E_vs_I	S_vs_N	F_vs_T	P_vs_J
KNN	BOW using SMOTE	24.8%	15%	86%	60%
KNN	TF-IDF	77.25%	85.40%	54.09%	60.62%
Naive Bayes	BOW	77.79%	85.86%	76.14%	66.23%
Naive Bayes	BOW with Upsample	77.37%	85.86%	75.52%	64.04%
Naive Bayes	TF-IDF with Upsample	76.33%	84.90%	75.49%	64.31%
Logistic Regression	W2V of 200dimensions	76.14%	85.89%	69.34%	60.19%
Logistic Regression	W2V of 200dimensions Upsample	63.73%	64.07%	70.26%	57.04%
RNN	Glove Embedding	54%	52.9%	57.8%	52.9%
SVM	N-grams + TF-IDF	75.98%	83%	78.4%	63.46%

Table 1: Comparative study of algorithms

From the above Table 1, the two models Naive Bayes and Logistic Regression classifier perform better than others. For selecting the one best model we consider the confusion matrix(Table 2)[15] and other classification metrics.

Algorithms	Feature Vector	E_vs_I	S_vs_N	F_vs_T	P_vs_J
Naive Bayes	TF-IDF with Upsample	[[1760 256] [360 227]]	[[2124 105] [288 86]]	[[807 398] [240 1158]]	[[481 544] [385 1193]]
Logistic Regression	W2V 200dimension with Upsample	[[1328 675] [269 331]]	[[1476 768] [167 192]]	[[834 361] [413 995]]	[[590 441] [677 895]]

Table 2: Confusion matrix of algorithms

Figure 4 shows the graphical representation of the accuracy of Naive Bayes and Logistics Regression with different Features Vector.

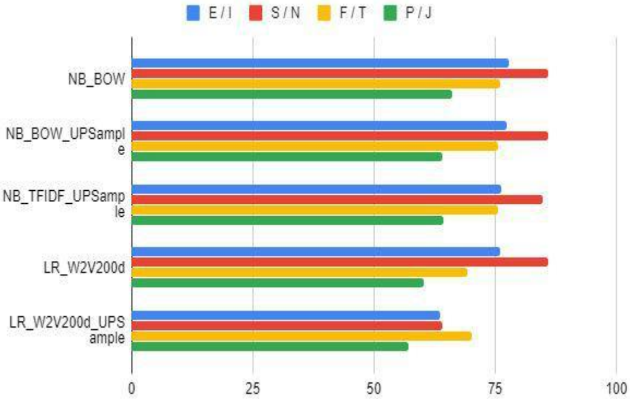
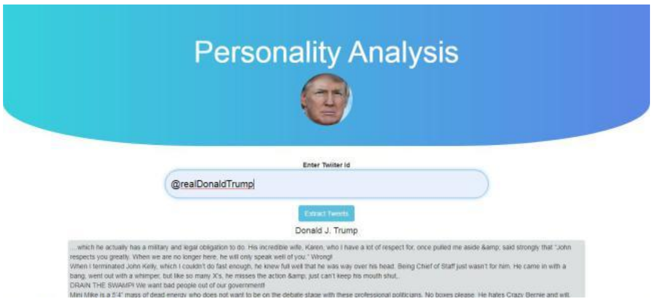


Figure 4: Comparison of model

From Figure 4 it is clear that the NaiveBayes with TF-IDF UpSampling perform better than the other classifiers.

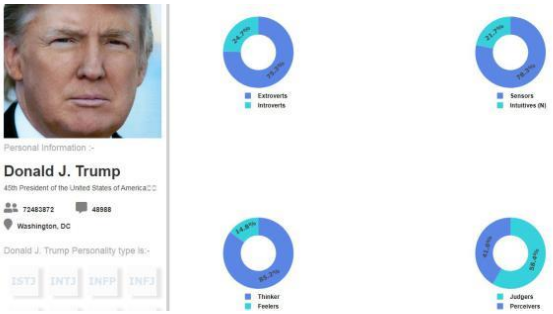


@realDonaldTrump

Donald J. Trump

....which he actually has a military and legal obligation to do. His incredible wife, Karen, who I have a lot of respect for, once pulled me aside &amp; said strongly that "John respects you greatly. When we are no longer here, he will only speak well of you." Wrong!
When I terminated John Kelly, which I couldn't do fast enough, he knew full well that he was way over his head. Being Chief of Staff just wasn't for him. He came in with a bang, went out with a whimper, but like so many X's, he misses the action &amp; just can't keep his mouth shut,.
DRAIN THE SWAMP! We want bad people out of our government!
Mini Mike is a 5'4" mass of dead energy who does not want to be on the debate stage with these professional politicians. No boxes please. He hates Crazy Bernie and will

Figure 5: GUI of Web



Personal Information :-

Donald J. Trump

45th President of the United States of America

72483872 48988

Washington, DC

Donald J. Trump Personality type is:-

Figure 6: GUI of Web

6. CONCLUSION

The system was able to gather the tweets of the user when given his twitter handle. The algorithms used were able to process the tweets and generate the required results. Finally the system was able to provide an efficiently accurate result by giving the personality type of the user.
An interactive GUI was created to use the system.

7. FUTURE WORK

The established system can be made more efficient by conducting more testing and feeding the system with a much accurate dataset. Once the system is highly accurate it can be used in corporations or even by the government to analyze the personalities of concerned individuals. The system can also be used in the crime sector. The currently implemented system of personality analysis can be extended and features like gender detection, age detection etc can be added to it.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Sagar Kulkarni for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Satishkumar Varma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, Maurizio Morisco," TwitPersonality: Computing Personality Traits From Tweets Using Word Embeddings and Supervised Learning, 2018"

[2] Mehul Smriti Raje(B) and Aakarsh Singh," Personality Detection by Analysis of Twitter Profiles, 2018"

[3] Srilakshmi Bharadwa j, Srinidhi Sridhar, Rahul Choudhary, Ramamoorthy Srinath," Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification, 2018"

[4] Chaowei Li, Jiale Wan, Bo Wang," Personality Prediction of Social Network Users 2018"

[5] Shafaan Khaliq Bhatti, Asia Muneer, M Ikram Lali, Muqaddas Gull," Personality Analysis of the USA Public Using Twitter Profile Pictures"

[6] Bayu Yudha Pratama, Riyanarto Sarno," Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM"

[7] Kanupriya Sharma, Amanpreet Kaur, "Personality prediction of Twitter users with Logistic Regression Classifier learned using Stochastic Gradient Descent 2015, IOSR Journal of Computer Engineering (IOSR-JCE)"

[8] Mounica Arroju, Aftab Hassan, Golnoosh Farnadi," Predicting Myers-Briggs Type Indicator with Text Classification"

[9] Hernandez, Rayne and Knight, Ian Scott, " Predicting Myers-Briggs Type Indicator with Text Classification"

[10] Anthony Ma and Gus Liu, "Neural Networks in Predicting Myers Brigg Personality Type from Writing Style "

[11] FIRE, Forum for Information Retrieval Evaluation.

[12] https://en.wikipedia.org/wiki/Big_Five_personali ty_traits, Five Big Personality traits.

[13] https://www.myersbriggs.org/my-mbti-personalit y-type/mbti-basics/home.htm?bhcp=1, Myers-Briggs Type Indicator.

[14] https://www.kaggle.com/datasnaek/mbti-type, Kaggle MBTI Dataset.

[15] https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html,Confusi on matrix.