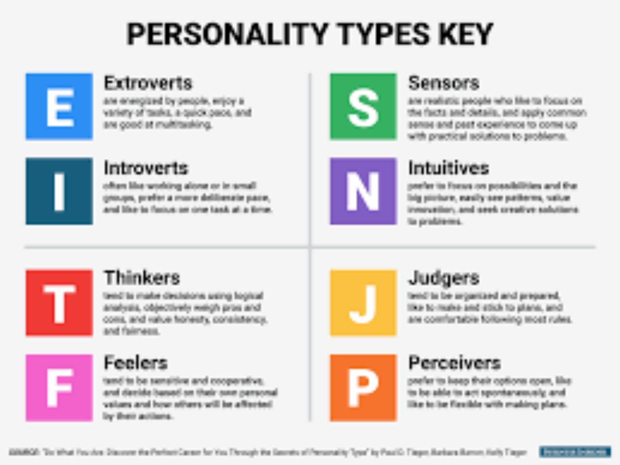
PERSONALITY PREDICTION AND GROUP DETECTION USING SOCIAL MEDIA POSTS

Muskan Kuchhal
2K18/CO/217
Delhi Technological University
Computer Science Department
muskankuchhal_2k18co217@dtu.ac.in

Muskan Saini
2K18/CO/218
Delhi Technological University
Computer Science Department
muskansaini_2k18co218@dtu.ac.in

I. INTRODUCTION

Personality, in broad term, refers to the characteristic patterns of thoughts, feelings and behavior exhibited by a person. Although earlier this concept of importance of personality of people was considered trivial but in the present times, as population bombarded, personality types of people reveal a lot more use such as mental health and psychology of people, establishing friendships and social groups of similar interests and even guiding people to pursue careers that might seem to be most suitable to them. Understanding personality type can be applied to the workplace. It can help with our leadership style, to resolve conflicts more effectively, to communicate more effectively, to understand how others make decisions, to coach others, to improve sales skills and to retain key staff.



PERSONALITY TYPES KEY

E Extroverts
are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.

I Introverts
often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.

S Sensors
are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.

N Intuitives
prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.

T Thinkers
tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.

F Feelers
tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

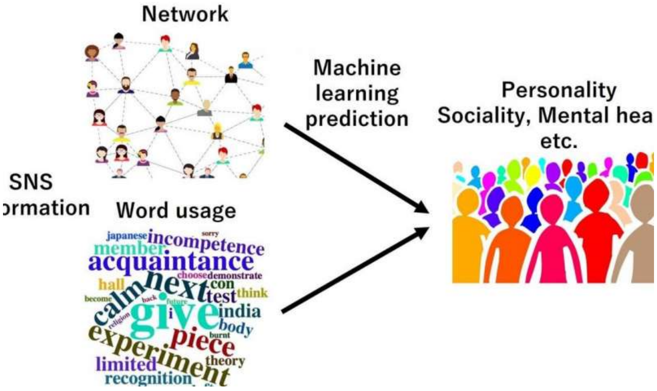
J Judgers
tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.

P Perceivers
prefer to keep their options open, like to be able to act spontaneously and like to be flexible with making plans.

SOURCE: "Do What You Are: Discover the Perfect Career for You Through the Secrets of Personality Type" by Paul D. Tieger, Barbara Barron, Kelly Tieger

II. PROBLEM STATEMENT

The project aims on predicting the personality of a person by different broad dimensions which can be used to describe the human personality and psyche. Different machine learning methods can be used for the prediction including K-Means clustering, Multi-layer perceptron, Random Forest Classifier, K nearest neighbours. Also, an analysis and comparison could be done on the personality predictions of different models. Elbow method can be applied for selecting the optimal number of clusters for K-means clustering. A personality score would be predicted for each personality trait(such as introversion, judging etc), for each person in test data.



III. MOTIVATION

Personality testing and assessment refer to techniques designed to measure the characteristic patterns of traits that people exhibit across various situations. Personality tests can be used to help clarify a clinical diagnosis, guide therapeutic interventions, and help predict how people may respond in different situations. Previously a person had to explicitly report to psychologists and get done with tests and psychometric analysis. But now using automatic personality predictor, one's personality can be predicted on the basis of various factors and thus can be further used in organisations and academics.

IV. DATA COLLECTION (link)

The dataset that we will be using for the project is (MBTI) Myers Briggs Personality Type Dataset from Kaggle.

This dataset contains over 8600 rows of data, on each row is a person's:

- Type (This persons 4 letter MBTI code/type)
- A section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters)).

V. DATA CLEANING

The dataset acquired contains textual posts by individuals on social media and may contain separators, links and URLS and various other symbols and punctuation marks.

The data thus, was cleaned using the following techniques:-
1. Replacing all '|||' (post separators) with ' ' in the posts
2. Removing all the stop words from the posts
3. Applying post stemming
4. Conversion of words to lowercase
5. Removal of the hyperlinks with ‘URL’
6. Removal of digits and punctuations

VI. DATA PREPROCESSING

In order to transform the textual posts into columns of input data, TF-IDF vectorizer is used that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

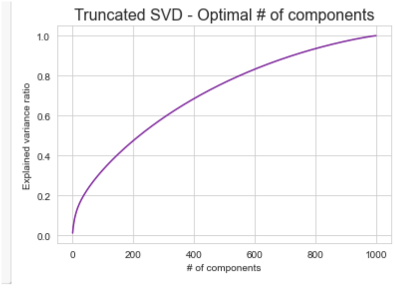
For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

TF-IDF believes that high frequency may not able to provide much information gain. In another word, rare words contribute more weights to the model.

Further, due to a large variety of words used in the posts, the feature vector seemed to be quite large, therefore, truncated SVD was used for dimensionality reduction.



$$A_{nxp} = U_{nxn}\, S_{nxp}\, V^{T}_{pxp}.$$

VII. MULTI CLASS CLASSIFICATION

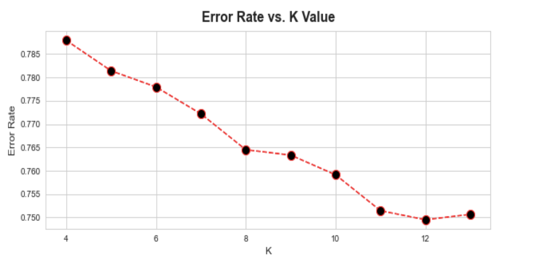
The four models of Random forest classifier, K nearest neighbours, One vs Rest classifier and multi-layer perceptron models were used to classify the data into their respective personality types.

Random Forest classifier

- *It* fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
- Using n_estimators=30, min_samples_leaf=50, oob_score=True, n_jobs= -1 gives a training Accuracy of 41.99% and testing Accuracy of 27.96%, a little greater than the baseline of 21.1%.
- All predictions limited to only within four of the sixteen classes: INFJ, INFP, INTJ, and INTP, which are the most abundant classes in the dataset.

K nearest Neighbours

- *It* finds the observations in its training data that are "nearest" to the observation to predict; it then averages or takes a vote of those training observations' target values to estimate the value for the new data point.
- The optimal number of neighbours were found to be equal to 15 using GridSearchCV.



- The accuracy on training set was found to be 100% and that of test set was 27.19%.
- The model was clearly overfit.

One vs Rest Classifier

- *OneVsRestClassifier* is a multiclass classifier in which a class is fitted against all the other classes; since each class is represented by one and only one classifier, it is possible to gain knowledge about the class by inspecting its corresponding classifier.
- The accuracy on training set was 44.38% and test set was 36.84%.
- The predictions were scaled out to all the classes and not just the dominant classes.

Multi-layer perceptron model

- A multi-layer perceptron model with two dropout layers and two dense with activation function of relu with a the last dense layer of activation function sigmoid was used for the classification of posts. The loss of sparse categorical crossentropy was used.

- 500 epochs with a batch_size of 128 was used to train the data.

- The model provided the training accuracy of 72.46% and a test accuracy of 31%.

VIII. BINARY CLASSIFICATION

We have seen in the above classification that the models could not achieve a high enough accuracy and therefore, binary classification on individual axes of I-E, N-S,F-T and J-P was done using the same multi-layer perceptron model.

Further, some performance was improved by handling the imbalance of data using SMOTE technique of oversampling using synthetic data creation.

Introversion vs Extroversion

On this axes, the training accuracy of 98% and a testing accuracy of 76.3% was achieved. However, even after handling imbalance, the results are quite skewed.
The f1 score for introversion and extroversion was found to be 0.86 and 0.21 respectively.

INTUITION VS SENSING

On this axes, the training accuracy of 98.8% and a testing accuracy of 74.9% was achieved. However, even after handling imbalance, the results are highly skewed.
The f1 score for intuition and sensing was found to be 0.92 and 0.07 respectively.

Feeling vs Thinking

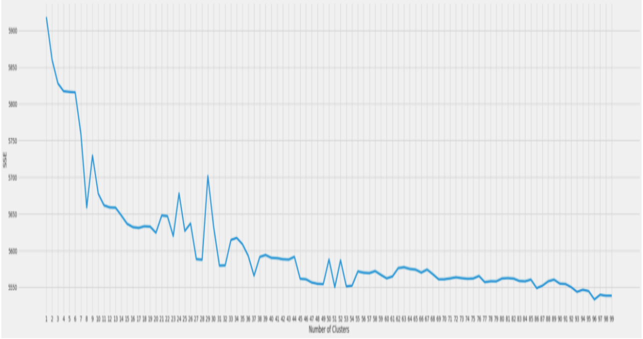
On this axes, the training accuracy of 91.11% and a testing accuracy of 68.92% was achieved. The results along this axes was much balanced when compared with others. The f1 score for Feeling and Thinking was found to be 0.75 and 0.59 respectively.

Judging vs perceiving

On this axes, the training accuracy of 85.9% and A testing accuracy of 53.28% was achieved. The results along this axes was the most balanced when compared with others. The f1 score for Judging and Perceiving was found to be 0.51 and 0.56 respectively.

IX. CLUSTERING AND GROUP FORMATION

On social media platforms, people often post to express themselves and thus, their posts can very well be used to find appropriate friend groups and community for them. Thus, we used the unsupervised K means clustering algorithm for finding the appropriate number of clusters using the elbow method and thus, can be used to suggest them relevant group options.



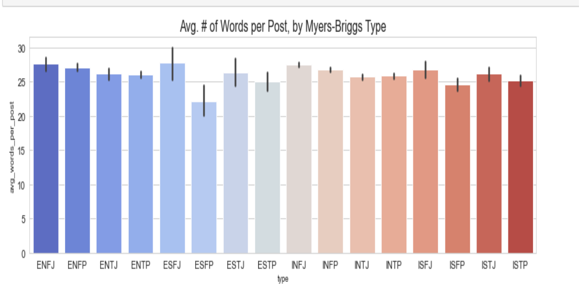
The best K found by studying the above graph was 37.

X. SNETIMENT ANALYSIS

Length of posts

- The average post length is 26.4 words per post.

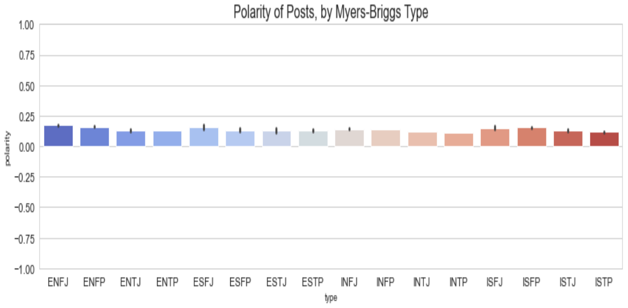
- From the graph, it is clear the ESFP tend to use fewer words in their posts.



Polarity of posts

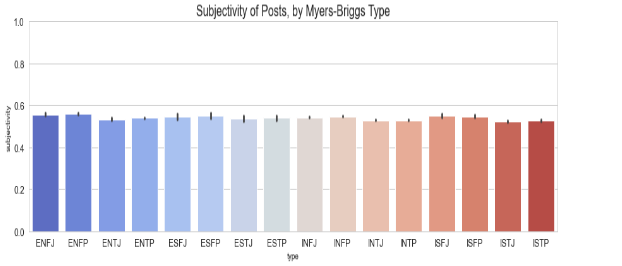
- The average polarity is 0.13.

- The overall polarity is quite neutral although ENFJs depicted the highest polarity.



Subjectivity of posts

The average subjectivity was 0.54. Thus, the posts have a moderate tone of neither objective or completely subjective.



XI. OTHER EXPERIMENTS

Due to the imbalance in the dataset, we tried using several techniques such oversampling, SMOTE and ensemble with oversampling, however, the application of these even reduced the accuracy of the models.

Pre-processing using count vectorizer was also tried out but TF_IDF since being an improved version was incorporated.

Convolution networks were tried instead of the multi-perceptron model since they perform well with textual data however, in this case the model performed poorly.

XII. FUTURE WORK

Further work can be done in future to improve the efficiency and try out other techniques to improve the study on this data.
1) Improving the imbalances in the data and incorporating data from other personality types with more of sensing quality
2) Using word2vec and RNN models for classification
3) Doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.).
4) Use the cross validation set to understand overfitting.
5) Using other algorithms such as collaborative filtering for community detection

XIII. REFRENCES

- https://medium.com/@bian0628/data-science-final-project-myers-briggs-prediction-ecfa203cef8
- https://thesai.org/Downloads/Volume11No3/Paper_58-Personality_Classification_from_Online_Text.pdf
- https://www.ijert.org/personality-prediction-from-social-media-text-an-overview -:~:text=Sentiment analysis is also done,level based on their posts.
- https://link.springer.com/article/10.1007/s11276-018-01913-4
- https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/