I.J. Information Technology and Computer Science, **2021, 6, 48-60**
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.06.04

Myers-briggs Personality Prediction and Sentiment Analysis of Twitter using Machine Learning Classifiers and BERT

Prajwal Kaushal[1], Nithin Bharadwaj B P[2], Pranav M S[3], Koushik S[4] and Anjan K Koundinya[5]
Department of Computer Science and Engineering, BMS Institute of Technology and Management, Bengaluru, India
E-mail: prajwalkkp@gmail.com[1], bharadwajnithin99@gmail.com[2], annjank2@gmail.com[5]

Received: 14 July 2021; Revised: 16 August 2021; Accepted: 09 October 2021; Published: 08 December 2021

Abstract: Twitter being one of the most sophisticated social networking platforms whose users base is growing exponentially, terabytes of data are being generated every day. Technology Giants invest billions of dollars in drawing insights from these tweets. The huge amount of data is still going underutilized. The main of this paper is to solve two tasks. Firstly, to build a sentiment analysis model using BERT (Bidirectional Encoder Representations from Transformers) which analyses the tweets and predicts the sentiments of the users. Secondly to build a personality prediction model using various machine learning classifiers under the umbrella of Myers-Briggs Personality Type Indicator. MBTI is one of the most widely used psychological instruments in the world. Using this we intend to predict the traits and qualities of people based on their posts and interactions in Twitter. The model succeeds to predict the personality traits and qualities on twitter users. We intend to use the analyzed results in various applications like market research, recruitment, psychological tests, consulting, etc, in future.

Index Terms: BERT, MBTI, Machine Learning, Personality Prediction, Sentiment Analysis.

1. Introduction

Social media is generating a colossal amount of data that is rich in sentiment. This data can be in the form of news, posts, articles, tweets, messages, etc. Nowadays, people from every corner of the globe use social media platforms to share and expresses their opinion. Twitter is one of the leading social media platforms where users interact and share their views in the form of tweets with other users. Till a few years ago, analyzing public sentiments was not something that was actually taken seriously. Today, thanks to advancement in technology and the growing social media users in a productive way, sentiment analysis is emerging as a viable tool. What makes it interesting and rather different from the other forms of data analytics is that this one deals with emotions and helps in getting closer to the public. Therefore, it is necessary and becomes the need of the hour to draw insights and investigate the sentiments of the public.

There exists wide range of Machine Learning algorithms and methods for performing sentiment analysis on Twitter data. SVM, Naïve Bayes, Recurrent Neural Networks, Convolutional Neural Networks are some of the machine learning algorithms that are existing to perform sentiment analysis on twitter. But until recently, no model was capable of understanding the context of people's statements. Bidirectional Encoder Representation from Transformers (BERT) is able to understand context like difference between 'this painting is pretty ugly' and 'this doll is pretty'. Both the words have pretty but the emotion is completely different between those words. This was not possible in previous popular NLP models used for sentiment analysis. Hence, sentiment analysis using BERT is a fine-grained task that aims to identify the sentiment polarity of the social media users, in particular Twitter.

Moving to the concept of personality prediction, it helps in analysing and understanding the views, characteristics and behaviour of an individual. It helps in understanding one's one behaviour as well as others. People of all walks of life are unique and different in their own ways and at times this information can prove to be vital and revolutionary. There are various techniques available to mine the social media data and predict the personality of users including Naïve Bayes, classification trees, and association rules. Various algorithms and methods do exist to extract useful information on the behaviour of users based on their tweets, posts and interactions.

The aim of this paper is to predict the personality of twitter users under the umbrella of Myers-Brigg Type indicator concept. MBTI is a widely used tool for testing the personality of a person. It is a personality framework that helps us in getting a basic idea or an overview of the traits and characteristics of an individual. MBTI acts as a powerful tool in many scenarios like – recruitment, psychometric tests, counselling, market research, etc. MBTI unlike many other types of psychological evaluations, your results are not compared against any standard norms but to simply offer further information about your own unique personality. Also, no personality type is better or best than another. MBTI

characterizes people in four basic pairs of qualities which further extends to 16 different types. The four basic pairs are sensing-intuition, thinking-feeling and judging-perceiving.

Combining the advantage of Sentiment Analysis and Profiling the users on social media using MBTI as an instrument can be blessing in disguise in many applications. For instance, consider a software company that is hiring for people with good communication skills, a team worker, adaptable and creative. The recruiter can crosscheck for these desired qualities in an individual based on his/her social media posts and match with the results generated from the system, thereby selecting the most appropriate candidate and also builds confidence. At the same time, it comes as a useful tool in deciding the appropriate domain for the fresher. We intend to achieve the personality prediction using ML classifiers like KNN, Logistic Regression, Naïve Bayes, Random Forest, SVM and Stochastic Gradient Descent. And find the classifier producing the best accuracy.

The structure of this study as follows: the next section is about the literature survey. Section three provides details about the research methodology and the data analysis and implementation can be found in section four. The results and conclusions are section five and six.

2. Literature Survey

The basic concept of Sentiment Analysis starts with finding the polarity of the sentence. The polarity of a sentence is found using various ML classifiers as in [1] and it is found that Neural Network Classifier algorithm provides the best accuracy. Also, researchers in [2] have explained that the polarity of the sentence is found using the concept of term frequency. The performance of Logistic Regression, SVM and Multinomial Naïve Bayes algorithms are compared and is observed that Logistic regression outperforms the SVM and Multinomial naïve bayes with an accuracy of 86.23%. A study and comparative analysis of existing techniques used for opinion mining through machine learning approach is provided in [3] which also uses SVM and Naïve Bayes. But there are challenges faced here when using other languages and handling of negation expressions.

A BERT-CNN sentiment analysis model has better performance than the original BERT Model by adding a CNN representative layer, is proposed in [4] so as to improve the accuracy. Using the sentiment classification layer, the sentiments of online commodity reviews is predicted. From the experiments it is observed that the F1 value of the new BERT-CNN model is increased by 14.4% and 17.4% from the original BERT and CNN models respectively. Representative power in target-dependent sentiment classification with BERT [5], which is a subtask of ABSA is explored. And in [6], a general classifier model is proposed, which uses the pre-trained language model BERT as the base for the contextual word representations. It makes use of the sentence pair classification model to find semantic similarities between a text and an aspect.

With respect to basic sentiment polarity classification is discussed in [7] Input is given i.e., either the username or a hashtag. Then the tweet is retrieved from twitter information that undergoes feature extraction and subsequently undergoes classification. The accuracy of feature vector is tested victimisation Naive Thomas Bayes classifier. In [8], we further come to know that the accuracy of classifying tweets as positive, negative and neutral can be increased. The presented methodology combined the use of unsupervised machine learning algorithm and lexicon-based algorithm. The work in [9] evaluates the people's sentiment about a person, trend, product, etc. The outcome is again polarity of the analysis but visualization and depiction are done using histogram and pie chart. The author follows a different dataset in [10] by determining the polarity of news articles (textual data) with the help of statistical methods, based on frequencies of positive and negative words.

There are various other techniques and different levels at which sentiment analysis performed as discussed in [11]. In particular, A detailed survey of various deep learning architectures used in sentiment analysis at both sentence level and aspect level are discussed. The advantages and the drawbacks of the different state-of-art methodologies are also discussed in [12].

In view of analytical profiling, users are profiled based on having highest number of keywords in respective clusters. To profile the user discussed in social media, "Profiling Social Media Users (PSMU)" algorithm is proposed in [13]. Simulation results prove that the proposed algorithm forms unique clusters and profiles the Twitter Users accurately (97.53% accuracy). Standards in personality models such as the Big Five model, DISC and the Myers-Briggs Type Indicator have been the basis for all such personality prediction. A user's social media data can thus be used to predict his/her personality. The main aim in [14] is to understand the concept of personality prediction using the data that is available on social media. MBTI is a viable tool for personality prediction as discussed in [15] is to propose a prediction analyzer for a Project Manager to allocate members in a given project by testing their ability using MBTI testing method and hence justifying that the person selected is most suitable for the project. There are many ML Classifiers like Random Forest, Naïve Bayes, SVM, etc. used in predicting Kaggle Users personality which is evident in [16]. It has been observed that Logistic Regression outperforms the other algorithms yielding the best accuracy and F-measure. The work in [17] talks about the analysis of social media data generated from Facebook using the Big Five method of personality prediction.

The accuracy is eventually increased by using the Random Forest Regression algorithm on the large datasets of Facebook generated data. A ML Based classifier is used in [18] that will take in input text, process it and predict the MBTI based personality type of the author of said text. As an example, Donald Trump's social media handle is

analyzed and the personality type is found. On the other hand, researchers in [19] talk about another personality prediction indicator, Big- 5 using SVM, Naïve Bayes and decision tree on facebook posts. It is observed that SVM produces the best accuracy. Similarly, as in [20], SVM is used as the classifier for sentiment analysis on twitter and IMDB reviews as it is the most accurate algorithm.

Finally, in our work we have used the concept of BERT as described in [21] for sentiment analysis. This model training is done based on the dataset of IMDB reviews [22] using BERT. On the other hand, the dataset for personality prediction is chosen from [23] for the purpose of training. The papers above uses various algorithms and various levels at which sentiment analysis is done and different models for personality prediction. We intend to perform sentiment analysis using BERT and apply six different machine learning classifiers for personality prediction using MBTI as a tool on twitter user.

3. Research Methodology

This paper work can be divided into two parts - sentiment analysis and personality prediction. For sentiment analysis we have used BERT and for personality prediction we did a comparative study to see which algorithm performs best for the task. It is shown in the figure 1 below.

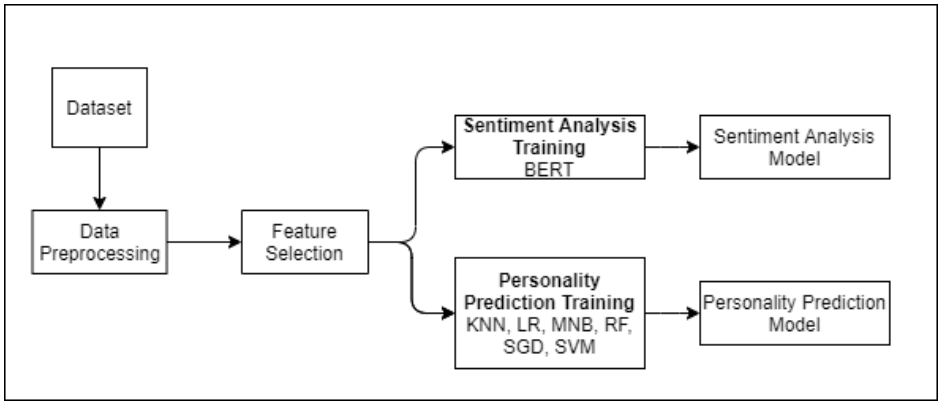

Fig.1. Shows the training process of the two models

As shown in the above figure 1, the collected dataset has to go through data pre-processing in which data cleaning, tokenization and stemming is done. After the pre-processing and feature selection is done, the clean data is used to train models for sentiment analysis and personality prediction. As explained earlier, the sentiment analysis model training is done using BERT and for personality prediction we have trained the models using 6 machine learning classification algorithms namely - K nearest neighbors, logistic regression, multinomial Naive Bayes, random forest, stochastic gradient descent and support vector machine. After this, the best performing algorithm was selected for the final personality prediction.

The sentiment analysis model outputs the sentiment of the user tweets that is, it classifies each tweet as positive or negative and gives the percentage of positive and negative tweets. The personality prediction model predicts the personality of the twitter user using MBTI scale which is Introvert(I) / Extrovert(E), Intuition(N) / Sensing(S), Thinking(T) / Feeling(F), Judging(J) / Perceiving(P).

3.1. Sentiment Analysis using BERT

Here BERT comes under semi supervised learning model, model is trained for a specific task that enables it to understand the patterns of the language. It enables better predictions based on context. Bert is specific large transformer masked language [21] model.

Language model refers class of statistical model that predict the probability of a sentence Eg: GitHub is open source is more likely than open is github source. There are many different statistical language types. Coming to BERT it comes under Bidirectional [21] type, where text is analysed in both forward and backward directions. Masked BERT is trained, with some percentage of input tokens (words) being masked randomly and the model tries to predict what the hidden token could be, thus improving the contextual learning. Transformers is an approach that aims to solve sequence-to-sequence tasks and also handles long-range dependencies with ease which means transformers make BERT faster and improves contextual prediction accuracy. There are 2 versions of BERT base and Large. BERT Large has around 345[21] million parameters which is the largest model of its kind, while even the smaller version of BERT has 110[21] million parameters. BERT can be used for Neural Machine Translation, Question Answering, Sentiment Analysis and Text summarization. BERT is training occurs in 2 phases, Pretraining [21] phase and Fine tuning [21] phase. Pretraining is used to learn the language, whereas fine tuning is used to perform a specific task like sentiment analysis. Here the fully connected output layers of the network can be replaced with our own output layers.

3.2. Personality Prediction using ML Classifiers

This section deals with predicting the personality using the concept of MBTI. In order to achieve this, we have used six different ML Classifiers – KNN, Logistic Regression, Naïve Bayes, Random Forest, SVM and Stochastic Gradient Descent.

A. K Nearest Neighbors (KNN)

KNN classifier is supervised ML algorithm which is easy, simple to implement which is used for both classification and regression predictive problems. But most of times KNN is used in classification problems in the industry level. It assumes that familiar objects or things will be in close proximity. In figure 2, the classifier identifies the k nearest neighbors for the new data point (blue in color) and based on those the neighbors it classifies the new data point as category A.

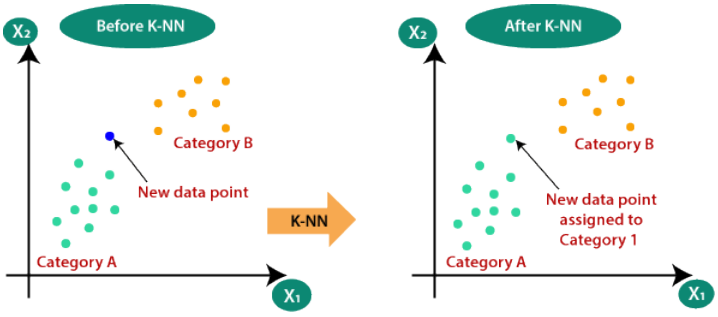

Fig.2. Shows the classification process using KNN

B. Logistic Regression

Logistic regression is used to model the probability of a certain class or event existing. For example, yes/no, pass/fail etc. This can be used to classify several classes of events such as determining whether an image contains a dog, elephant, fish, tiger etc. Each object that are detected will be assigned a probability between 0 and 1 and their sum will be 1. Consider a model with two predictors x1 and x2 and one binary output variable Y and p = Probability(Y=1) then

$$l = \log_b\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

where l is the log-odds, b is the base of logarithm and coefficients of x1, x2 are the parameters.

C. Naïve Bayes

This algorithm is the supervised ML algorithm that learns the likelihood of an article with specific components having a place with a particular gathering or class. The algorithm is called “naive” because in light of the fact that it expects that the event of a one component is free of the event of another element given the class variable.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{P(B)} \tag{2}$$

where, P(A|B) is the posterior probability and P(B|A) is the likelihood probability of event B occurring given event A has already occurred. P(A) is the priori of A and P(B) is the evidence.

D. Random Forest

It is a supervised machine learning algorithm which is simple, elaborately and easy to produce even without hyper-parameter tuning. It is a standout amongst the most utilized calculations as a result of its straightforwardness and it tends to be utilized for both relapse and arrangement assignments. More often than not arbitrary Random Forest information is prepared with the "BoW (Bag of Words)" strategy and the "forest" is a gathering of Decision Trees. The words bagging strategy though is a blend of many learning models which builds the general outcome. One bit of leeway of RF is that it very well may be utilized for both classification and regression issues, which structure most of current AI frameworks. How a RF would look like can be seen below in figure 3.

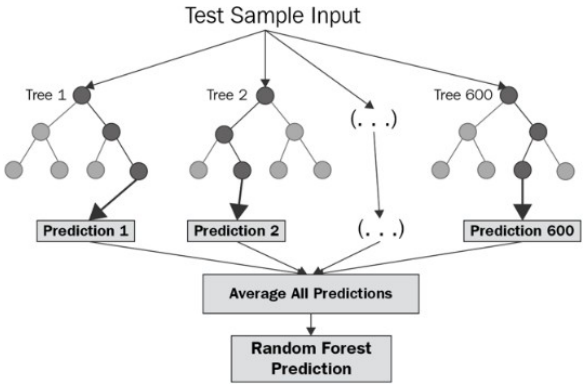

Fig.3. Classification using Random Forest

E. Support Vector Machine (SVM)

It is a machine learning approach which is used for classification and regression problems but usually it is used for classification problems. Each data item of a data set is plotted in m-dimensional space as a point where the value of particular coordinate is the value of each feature. Later, classification is calculated by identifying the hyperplane where two of the classes is been differentiated. A SVM creates parallel segments by producing two parallel lines. For every class of information in a high-dimensional space and uses practically all attributes. It isolates the space in a solitary go to produce level and straight parcels. Two classifications are partitioned by a reasonable hole that ought to be as wide as possible. They do this by dividing a plane called hyperplane. The process of SVM is indicated in figure 4 below.

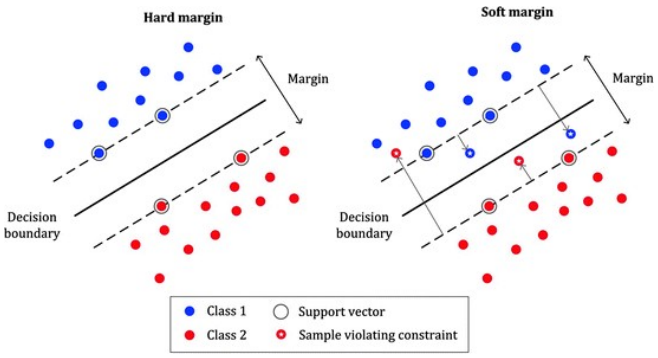

Fig.4. Classification using SVM

F. Stochastic Gradient Descent

Stochastic Gradient is one of the widely used machine learning algorithm. It is the starting point of neural network. It is an iterative method for optimizing any given function. The main aim is to reach the lowest part of the slope by using a small learning rate. The figure 5 below shows the gradient descent for the equation with two dimensions (x,y).

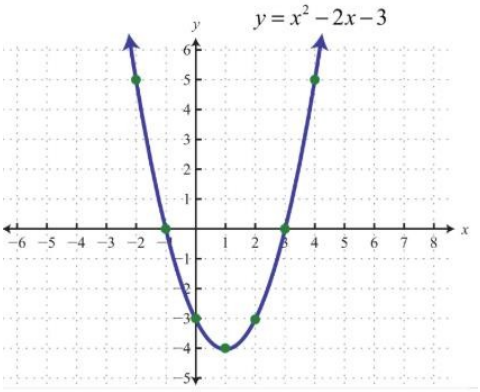

Fig.5. Stochastic Gradient Descent

4. Data Analysis and Experiments

This section talks about the dataset analysis and implementation of the model both in terms of sentiment analysis and personality prediction.

4.1. Sentiment Analysis

Data used is IMDB dataset [22] having 50,000 movie reviews for NLP. It contains binary sentiments and has more data than previous benchmark datasets. This data provides a set of 25K highly polar movie reviews for training and 25K for testing. As the data is fairly balanced and widely used in the field of sentiment analysis irrespective of platforms, in this paper it is decided to use them during the training phase for creating the model and testing the accuracy of the model.

To implement BERT, we have used ktrain which is a light weight library to implement machine learning models which is influenced by fastai for pytorch. Once the data is collected, it is sent for pre-processing. Pre-processing of the data involves tokenizing, embedding and transforming the data with respect to BERT model, max length is restricted to 400 words, which means if the sentence length exceeds 400 words will be truncated. The batch size is set to 6, to get optimal performance, the optimal learning rate is 2e-5 for sentiment analysis. The hyperparameter epochs is set to 1, which means the internal parameters are updated during each cycle. Fitting based on 1 cycle policy makes training faster and gives better model through super convergence.

4.2. Personality Prediction

This sub section is all about the data analysis, data pre-processing, training and testing of the model for personality prediction using the six aforesaid machine learning classifiers.

A. Data Analysis

The dataset that we are using is Kaggle [23] dataset. It contains Tweets of a twitter user and tis corresponding MBTI Type. MBTI is a personality type system that divides everyone into 16 distinct personality types across 4 axes: Introversion (I) – Extroversion (E), Intuition (N) – Sensing (S), Thinking (T) – Feeling (F), Judging (J) – Perceiving (P). This dataset contains 8675 rows and there are 2 columns. On each row is a person's: Tweet i.e, a column of the last 50 things they have posted. There are no missing values or null values in this dataset. One major problem is that all the values are in the form of texts so we had converted them to numerical form to train our model. INFP has the highest frequency so it will have a lot more data and the least is ESTJ so it will have the least amount of data as seen in the figure 6. There are also no repeating posts in the dataset. We can see that the dataset is not balanced throughout the different classes. There is unbalance in Introversion/Extraversion and Intuition/Sensing whereas Feeling/Thinking and Perception/Judgment are properly balanced. Although we have created trained models for each pair. Only the last 2 pairs are somewhat reliable in predicting MBTI type. So, it is not advised to depend on first 2 pairs i.e., IE and NS pairs.

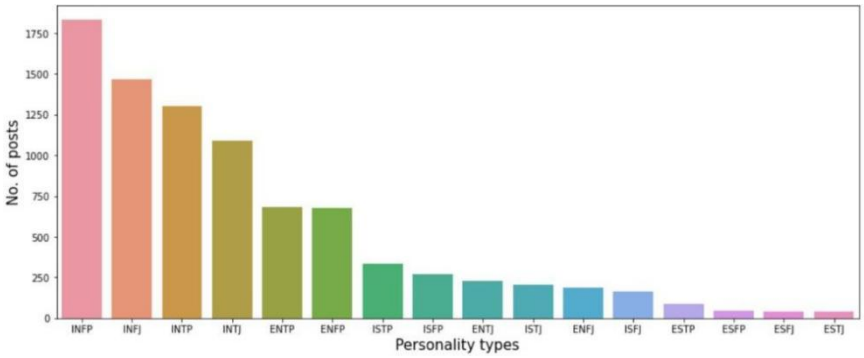

Fig.6. Bar graph of the dataset

Figure 7 shows the swarm plot of the dataset. They are used to plot all the points in a dataset. We can say that people who mostly use social media are of the introversion type. INFP has the most cluttered which shows that most people who tweeted are INFP which further shows the imbalance.

And as shown in figure 8, the joint plot, there is a bar graph on top of x axis and a bar graph on top of y axis and a scatter plot in between which shows the distribution of data for both the columns. The down area helps in calculating the probability density function. The bar graph shows Gaussian distribution of a sample space, that is the number of words per tweet and corresponding variance of word counts from the dataset. In the middle part the posts with a greater number of words gets darker colour. Here more number of tweets have words near 100. There is a relationship when there are 25-30 words per tweet & the variance of word counts is 100-150. This is also clear from the histogram plots on the axes.

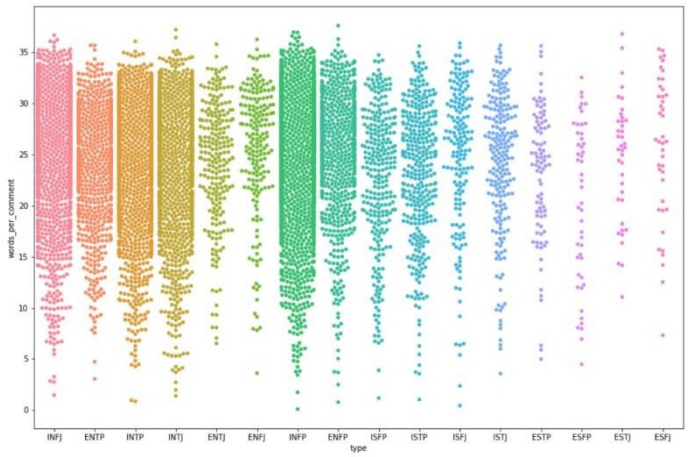

Fig.7. Swarm Plot of the dataset

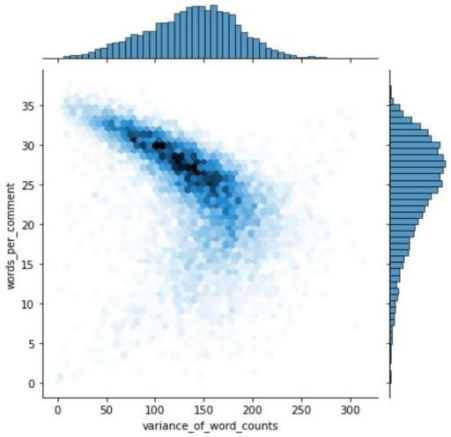

Fig.8. Joint Plot of the dataset.

The list of most common words like “I, to, the, and, of, is, this, that, etc” make no contribution. Also, many irrelevant words, a lot of url links, mentions are removed using data pre-processing techniques as these things have no impact and gives no insights on the personality. Next, we will see the word cloud of the dataset. Here also we can see many stop words which are apparently removed. In the word cloud as shown in figure 9, size is proportional to the frequency. This can give us insight as to how each personality type chose their words.

Fig.9. Word Cloud for all the MBTI types.

B. Pre-processing

We have added one column for each MBTI characteristic pair as we will be training independent classifier model for each pair separately. The reason for this is because of imbalance present in our dataset. Name of the columns are - ie, ns, ft and pj. So, for a person who is INTJ the newly created columns will have I, N, T, J respectively. These will be converted to numbers in preprocessing step. As a part of data pre-processing, we have performed tokenization, lemmatization and stemming. Finally applied the tf-idf algorithm to get the weighting plan which contains weights.

C. Training

After all the pre-processing steps are done, we move towards training our models. We have to create 4 different models namely - IE_Model, NS_Model, FT_Model, PJ_Model and IE_Model is responsible for predicting whether a person is of the type - Introversion or Extraversion. NS_Model is responsible for predicting whether a person is of the type - Intuition or Sensing. FT_Model is responsible for predicting whether a person is of the type - Feeling or Thinking. PJ_Model is responsible for predicting whether a person is of the type - Perceiving or Judging. These models are trained using the six algorithms mentioned earlier. At the end of training, we are left with 24 models (6X4) from which we have selected the best IE_Model, NS_Model, FT_Model, PJ_Model.

4.3. Model Deployment

The figure 10, explains the sequence of events that occurs when the model is deployed. Here, tweepy is a twitter API which is used to retrieve twitter data like tweets, likes, retweets etc, in our case we will be using it only for retrieving the tweets. First the User makes an API call to tweepy for tweets from a particular twitter handle. The User Credentials are authenticated. After authentication the user's requests are served. The tweets are given as input to the models, where BERT model predicts the sentiments of the tweets and personality prediction model predicts the MBTI characteristics of the individual twitter user based on the tweets.

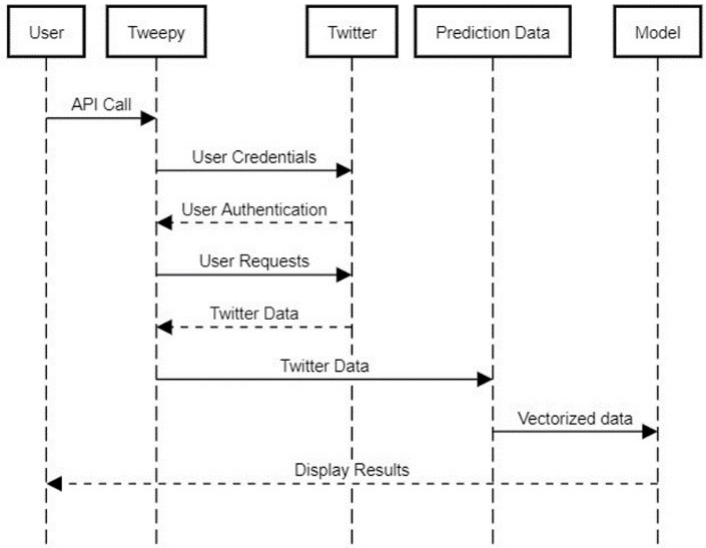

Fig.10. Model Deployment and Real Time Tweet Analysis

5. Results and Discussions

This section deals with the results and outputs obtained from the models and its validation. Here also we can divide it into phases where the first sub section talks about the results and its discussions with respect to the sentiment analysis model based on BERT and the second sub section deals with the personality prediction results and analysis.

5.1. Senitment Analysis

The model takes around 90 mins to be trained, when it is trained achieves an accuracy 82.58% on test data. From output it can be observed that training loss is 0.3744, validation loss is 0.1387, as training loss is not lesser than validation loss it indicates that the model is not overfitted. This is the case as training loss is measured during each epoch and validation loss is measured after each epoch, the model should ideally improve after each epoch and since validation loss is measured after epochs it should be less. Model's performance can truly be tested when it is deployed in real world environment, we proceed in this direction in this project, to retrieve most recent tweets from twitter handle twint library is used it is done so by retrieving the tweets which are in English (as BERT is finetuned in English) and then give only tweets as input to the model, then we measure the sentiments of each tweet and the aggregate of the sentiments is displayed visually through a pie chart. To check the data in real scenario we are showing select tweets of celebrities' twitter handle

and analyzing the reason for classification.

Figure 11 below, shows a recent tweet of actor Sudeep and it is being classified by the model as positive or negative and pie chart shows the aggregate sentiments of Sudeep starting from the most recent tweets and to older tweets. The pie chart keeps on changing as model keeps on classifying older tweets. The pie chart is showing the percentage of tweets classified as positive and negative. The actor has tweeted about his disheartenment over the early demise of actor Sanchari Vijay, this is being correctly classified by the model as negative. We can also notice that the percentage of positive tweets for Sudeep's twitter handle comes about to be 84.2% and negative to be about 15.8%.

Fig.11. Aggregate sentiment of Sudeep's tweets

The figure 12, shows sentiments of a particular tweet of cricketer Virat Kohli and also the aggregate sentiments of Virat Kohli's tweets starting from the most recent through pie chart. The cricketer asks people who are tested covid negative to stay together and be positive and this is being correctly classified as positive by the model.

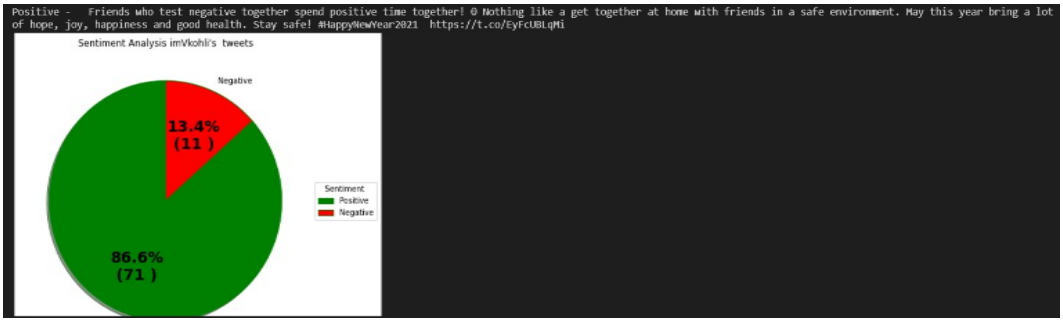

Fig.12. Aggregate sentiment of Virat Kohli's tweets

5.2. Personality Prediction

The performance of all the algorithms with their accuracy on the four different models IE, NS, PJ and FT is shown in the tables 1-4. From the above tables 1-4, we can see that Support vector machine, SVM performs best in all the models among all the six machine learning classifiers. Therefore, we have chosen SVM as the ML Classifier for personality prediction.

Table 1. IE Model Results

Algorithm	***Accuracy***
K Nearest Neighbours	0.771784232365145
Logistic Regression	0.822498847395113
Multinomial Naive Bayes	0.789764868603043
Random Forest	0.770862148455510
Stochastic Gradient Descent	0.814661134163209
Support Vector Machine	0.842784693407100

Table 2. NS Model Results

Algorithm	***Accuracy***
K Nearest Neighbours	0.843706777316736
Logistic Regression	0.880129091747349
Multinomial Naive Bayes	0.855232826187183
Random Forest	0.861226371599816
Stochastic Gradient Descent	0.874135546334717
Support Vector Machine	0.881973259566621

Table 3. FT Model Results

Algorithm	***Accuracy***
K Nearest Neighbours	0.639004149377593
Logistic Regression	0.793453204241586
Multinomial Naive Bayes	0.810511756569848
Random Forest	0.763485477178423
Stochastic Gradient Descent	0.781927155371139
Support Vector Machine	0.834485938220378

Table 4. PJ Model Results

Algorithm	***Accuracy***
K Nearest Neighbours	0.638543107422775
Logistic Regression	0.736745043798986
Multinomial Naive Bayes	0.733978792070078
Random Forest	0.654679575841402
Stochastic Gradient Descent	0.725680036883356
Support Vector Machine	0.785615491009682

To check the real test cases, we took some of the famous celebrities – Facebook owner Mark Zuckerberg and Bollywood actor Akshay Kumar and following is the results of their personality prediction.

Fig.13. The word cloud of Mark Zuckerberg's tweets.

The figure 13 shows the word cloud of Mark Zuckerberg's tweets. The word cloud represents the types of words used. The size of the word in the word cloud depends on the number of times that word is used. And we can notice that find, following, facebook are some of the most used words in his tweets whereas words like good, agree, yes are less used words.

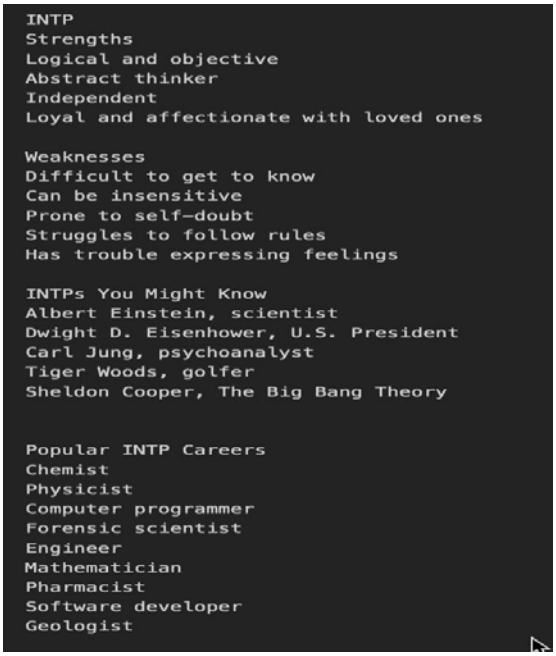

Fig.14. Personality Prediction Result for Mark Zuckerberg.

This figure 14 shows the final result of Mark Zuckerberg's tweets. As shown, Mark Zuckerberg is classified as INTP (Introvert, Intuition, Thinking, Perceiving) type. The strengths and weaknesses of the INTP type are also displayed. Famous people who were of the INTP type are displayed next. At last, the popular career of INTP type is displayed. As we can see, Popular INTP Careers has careers like - computer programmer, engineer, software developer, etc. This is exactly what Mark Zuckerberg is now. So we can say that the prediction is good.

Coming to the second result, the figure 14 shows the word cloud of Akshay Kumar's tweets. The word cloud represents the types of words used. The size of the word in the word cloud depends on the number of times that word is used. And we can observe that capeofgoodfilms, laxmii, year are some of the most used words and best, coming, part are less used words.

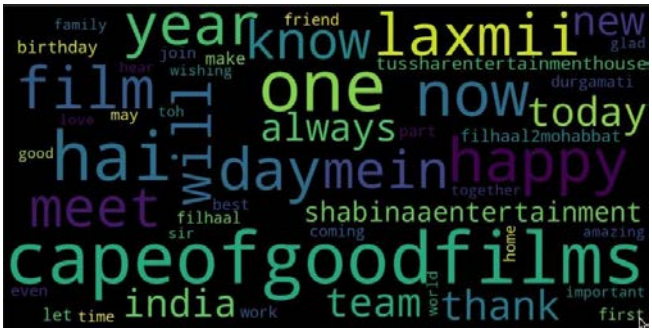

Fig.15. The word cloud of Akshay Kumar's tweets.

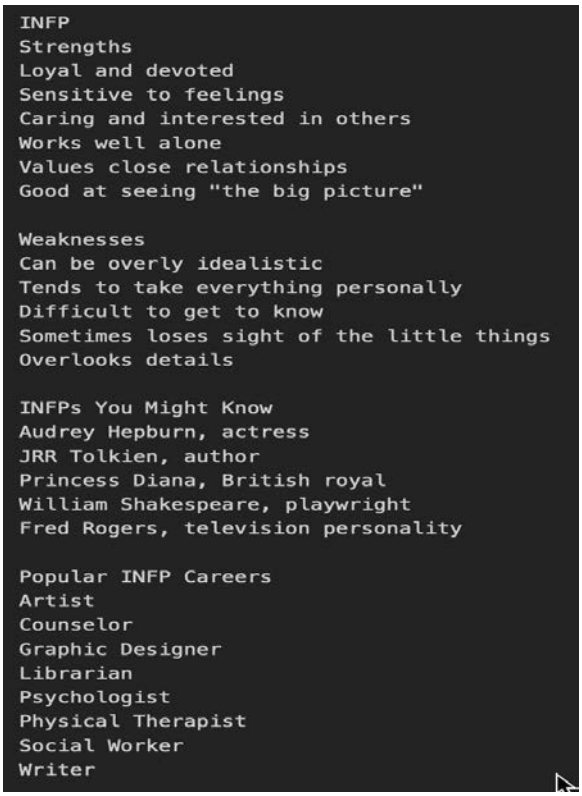

Fig.16. Personality Prediction Result for Akshay Kumar

The figure 16 shows the final result of Akshay Kumar's tweets. As shown, Akshay Kumar is classified as INFP (Introvert, Intuition, Feeling, Perceiving) type. The strengths and weaknesses of the INFP type are also displayed. Famous people who were of the INFP type are displayed next. At last, the popular career of the INFP type is displayed. As we can see, Popular INFP Careers has careers like - artist, etc. This is exactly what Akshay Kumar is now. So we can say that the prediction is good.

6. Conclusion

Personality traits are the behaviour that is exhibited by individuals, sentiments are the emotions that the person feels at the moment. Personality traits are likely to be exhibited during a lifetime, sentiments are transitory in nature. Each

person responds to a particular sentiment differently based on their personality traits. So, in our research work, we have successfully deployed two models, first model analyses the sentiments of the Twitter user using BERT, whereas for predicting personality we did a comparative study of six prominent machine learning classifiers and it is experimentally found that SVM performs better than other algorithms. In future work, we intend to use these models for the recruitment process where the recruiter can get to know about the personality of the prospective employees and based on that, place them in a suitable position in the organization. This research work can also be extended in the field of market research where based on the personality and sentiments of the people living in an area, appropriate products and services can be provided.

References

[1] Golam Mostafa, Ikhtiar Ahmed, Masum Shah Junayed, "Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data", International Journal of Information Technology and Computer Science, Vol.13, No.2, pp.38-48, 2021.

[2] Poornima A, K Sathiya Priya, “A Comparative Sentiment Analysis of Sentence Embedding Using Machine Learning Techniques”, 6th International Conference on Advanced Computing & Communication Systems (ICACCS), March 2020.

[3] Deep Kaneria, Brijesh Patel, “Sentiment Analysis for Twitter Data”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-7S, May 2020.

[4] Junchao Dong, Feijuan He, Yunchuan Guo, Huibing Zhang, “A Commodity Review Sentiment Analysis Based on BERT-CNN Model”, 5th International conference on Computer and communication Systems, ISBN:978-1-7281-6137-2, May 2020,

[5] ZHENGJIE GAO, AO FENG, XINYU SONG, AND XI WU “Target-Dependent Sentiment Classification With BERT”, accepted October 5, 2019, date of publication October 11, 2019, date of current version November 4, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2946594

[6] Mickel Hoang, Oskar Alija Bihorac “Aspect-Based Sentiment Analysis Using BERT”, Issue 167, Article – 20, ISSN: 1650-3740, 2019.

[7] A Brahmananda Reddy, D.N. Vasundhara, P. Subhash “Sentiment Research on Twitter Data” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

[8] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi, Wejdan Abdullah AlShehri “Sentiment Analysis of Twitter Data”, Pulished in International Conference on Computer and Information Sciences (ICCIS) April 2019.

[9] Prakruthi V, Sindhu D, Dr S Anupama Kumar, “Real Time Sentiment Analysis of Twitter Posts”, 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, December 2018, IEEE.

[10] Vishal S. Shirsat, Rajkumar S. Jagdale, S. N. Deshmukh, “Document Level Sentiment Analysis for News Articles”, International Conference on Computing, Communication, Control and Automation (ICCUBEA), August 2017, IEEE.

[11] Abdullah Asaeedi, Mohammad Zubair Khan, “A Study on Sentiment Analysis Techniques of Twitter Data” International Journal of Advanced Computer Science and Applications, Vol.10, No.2, 2019.

[12] Indhra om Prabha M, G. Umarani Srikanth “Survey of Sentiment Analysis Using Deep Learning Techniques”, Published in 1st International Conference on Innovations in Information and Communication Technology (ICIICT),2019

[13] Vasanthakumar G U, Shashikumar D R, Suresh L, “Profiling Social Media Users, a Content-Based Data Mining Technique for Twitter Users” 1st International Conference on Advances in Information Technology (ICAIT), 2019.

[14] P. S. Dandannavar, S. R. Mangalwede and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018.

[15] P. B. Kollipara, L. Regalla, G. Ghosh and N. Kasturi, "Selecting Project Team Members through MBTI Method: An Investigation with Homophily and Behavioural Analysis," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2019, pp. 1-9, doi: 10.1109/ICACCP.2019.8883022

[16] Shristi Chaudhary, Ritu Singh, Syed Tausif Hasan, Ms. Inderpreet Kaur, “A Comparative Study of Different Classifiers for Myers Brigg Personality Prediction Model”, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 05, May-2018

[17] Medhavini Rao, Pooja Jayant Kanchugar, Pooja R, Prakshitha M N, Anitha R, “Personality Recognition using Social Media Data”, International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue:04, April-2019

[18] Hernandez Rayne, Knight Ian Scott, “Predicting Myers-Briggs Type Indicator with Text Classification”, 31st Conference on Neural Information Processing Systems (NIPS), 2017

[19] Azhar Imran, Muhammad Faiyaz, Faheem Akhtar,"An Enhanced Approach for Quantitative Prediction of Personality in Facebook Posts", International Journal of Education and Management Engineering, Vol.8, No.2, pp.8-19, 2018.

[20] Munir Ahmad, Shabib Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets", International Journal of Modern Education and Computer Science, Vol.9, No.10, pp. 29-36, 2017.

[21] “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova Google AI Language

[22] Large Movie Review Dataset for Sentiment Classification, https://ai.stanford.edu/~amaas/data/sentiment/

[23] Mitchell J, “MBTI, Myers-Briggs Personality Type Dataset”, 2017.

Authors' Profiles

Prajwal Kaushal, is a student pursuing his B.E. (CSE) at BMS Institute of Technology and Management affiliated to Visveswariah Technological University (VTU), Belagavi, India. His research interests include Statistical Machine Learning, Probabilistic Graphical Models, Deep Learning, Neural Networks and Generative Adversarial Networks.

Nithin Bharadwaj B P, is a student pursuing his B.E. (CSE) at BMS Institute of Technology and Management affiliated to Visveswariah Technological University (VTU), Belagavi, India. He aims to work and research in the field of Data Science, Artificial Intelligence, and Machine Learning. He has recently worked as an intern in Nokia Networks and Solutions as a Software Developer.

Pranav M S, is a student pursuing his B.E. (CSE) at BMS Institute of Technology and Management affiliated to Visveswariah Technological University (VTU), Belagavi, India. He is much fond of research and his interests lie in Natural Language Processing, Machine Learning, Algorithms, Game theory.

Koushik S, is a student pursuing his B.E. (CSE) at BMS Institute of Technology and Management affiliated to Visveswariah Technological University (VTU), Belagavi, India. He is much fond of research and his interests lie Natural language processing, Neural Networks, Algorithms, Machine Learning.

Dr. Anjan Koundinya has received his B.E (CSE), M. Tech (CSE), and Ph.D. degree from Visveswariah Technological University (VTU), Belagavi, India. He has been awarded the Best Performer PG 2010, First Rank Holder (M. Tech CSE 2010) and recipient of Best Ph.D Thesis Award by BITES, Karnataka for the academic year 2016-17. He has served in industry and academia in various capacities for more than a decade. He is currently working as Associate Professor and PG Coordinator in Dept. of CSE, BMSIT&M, Bengaluru.

How to cite this paper: Prajwal Kaushal, Nithin Bharadwaj B P, Pranav M S, Koushik S, Anjan K Koundinya, "Myers-briggs Personality Prediction and Sentiment Analysis of Twitter using Machine Learning Classifiers and BERT", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.6, pp.48-60, 2021. DOI: 10.5815/ijitcs.2021.06.04