Personality Prediction of People Based on Social Media Posts

Tirumala Manukonda(1001662386)
Department of Computer Science
University of Texas Arlington
Texas, TX 76019
tirumalareddy.manukonda@mavs.uta.edu

Abstract

With the increase in the usage of social media, people often express their feelings using social media posts. The main aim of this project is used to come up with a model that predict individuals personality based on their social media posts. In this project, Machine Learning classification techniques are used to predict four personality traits based on Myers- Briggs Type Indicator (MBTI) model, namely Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P) from input text. This work provides the basis for developing a personality prediction system that aids organizations in recruiting and selecting appropriate personnel.

1 Introduction

Personality of a person encircles every aspect of life. It describes the pattern of thinking, feeling and characteristics that predict and describe an individuals behaviour and also influences daily life activities including emotions, preference, motives and health [1]

The increasing use of Social Networking Sites, such as Facebook and Twitter have propelled the online community to share ideas, sentiments, opinions, and emotions with each other; reflecting their attitude, behaviour and personality. Obviously, a solid connection exists between individuals temperament and the behaviour they show on social networks in the form of comments or tweets [2].

The relation between the social media posts and the individuals personality or behavior is very high, using this, want to build a machine learning model that could assist across multiple domains like Health sectors - to verify patients mental health, Organizations - to know the personality of the individual they are going to recruit, also in Social Media Platforms - to understand the depression levels of a person from their posts and notify his family members, when someone is highly depressed.

In this proposed work, machine learning classification techniques like LogisticRegression, DecisionTree, MultinomialNB, RandomForestClassifier and XGBoostClassifier are applied on the benchmark personality recognition MBTI dataset to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S),Feeling-Thinking(F-T) and Judging-Perceiving(J-P). The major issue of this prediction is that the dataset is highly imbalanced, to address this issue data-re-sampling techniques like Under-Sampling and Over-Sampling are applied.

2 Method

The working procedure of this proposed system are as follows: (i) Data acquisition, (ii) Preprocessing and feature selection, (iii) Text-based personality classification using MBTI model, (iv)Applying different classification techniques, (v) Data re-sampling and evaluate performance

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2.1 Data acquisition

The MBTI dataset is downloaded from kaggle and it has two columns, (i) type and (ii) posts. By type it has 16 MBTI personality types, such as INTP, ENTJ and INFJ, etc. As we are interested in MBTI traits rather than types, therefore added four new columns to the original dataset for the purpose of traits determination. As a result, the new modified dataset Table 1 will have additional four new Binary columns namely Introversion-Extroversion(IE),Ntuition-Sensing(NS), Feeling-Thinking(FT), Judging-Perceiving(JP).

New modified dataset					
type	posts	IE	NS	TF	JP
INFJ	PTypeToken PTypeToken moments sport...	0	0	1	0
ENTP	finding lack these posts very alarming eo...	1	0	0	1
INTP	good course which know thats bles...	0	0	0	1
INTJ	dear PTypeToken enjoyed conversation ot...	0	0	0	0
ENTJ	youre fired eostokendot thats another silly...	1	0	0	0

Table 1: After adding personality traits indicator columns

2.2 Preprocessing and feature selection

Different pre-processing techniques and feature selection are exploited. These techniques include tokenization, removal of links, Punctuation, Remove Personality Types Words,and convert text to lower and feature selection using Countvectorizer.

2.2.1 Preprocessing

Data preprocessing steps

a).Removing Personality Types Words :- Words like 'INTP', 'INTJ', 'ENTP'.. are removed from the text of posts so that model predicts with the unseen data.

b).Word stemming:- Stem words are produced by eliminating the pre-fix or suffix used with root words.

c).Tokenization:- Sentences are divided into small words.

2.2.2 Feature selection

CountVectorizer:-

Using machine learning algorithms, it cannot execute text or document directly, rather it may firs be converted into matrix of numbers. This conversion of text document into numbers vector is called tokens. The count vector is a well-known encoding technique to make word vector for a given document. CountVectorizer takes what's known as the Bag of Words approach. Each message or document is divided into tokens and the number of times every token happens in a message is counted.

CountVectorizer perform the following tasks:

- It tokenizes the whole text document.
- It constitutes a dictionary of defined words.

2.3 Text-based Personality Classification Using MBTI Model

MayersBriggs Type Indicator is used for classification and prediction. This model categorize an individual into 16 different personality types based on four dimensions, namely

(i) Attitude- Extroversion vs Introversion: this dimension defines that how an individual focuses their energy and attention, whether get motivated externally from other people's judgement and perception, or motivated by their inner thoughts

(ii) Information- Sensing vs iNtuition (S/N): this aspect illustrates that how people perceive information and observant(S), relying on their five senses and solid observation, while intuitive type individuals prefer creativity over constancy and believe in their guts

(iii) Decision- Thinking vs Feeling (T/F): a person with Thinking aspect, always exhibit logical behaviour in their decisions, while feeling individuals are empathic and give priority to emotions over logic

(iv) Tactics- Judging vs Perceiving (J/P): this dichotomy describes an individual approach towards work, decision-making and planning. Judging ones are highly organized in their thoughts. They prefer planning over spontaneity. Perceiving individuals have spontaneous and instinctive nature. They keep all their options open and good at improvising opportunities [3].

2.4 Applying different classification techniques

ML classifiers like LogisticRegression, DecisionTree, MultinomialNB, RandomForestClassifier and XGBoostClassifier are imported from scikit-learn library and used to classify the dataset. MBTI type indicators were trained individually and then the data was split into training and testing datasets. The model was fit onto the training data and the predictions were made for the testing data.

The evaluation metrics, such as accuracy, precision, recall and f-measure, describe the performance of a model. Therefore, different evaluation metrics has been used to check the overall efficiency of predictive model.

2.5 Data re-sampling and evaluate performance

2.5.1 Data re-sampling

Data manipulation sampling approaches focus on rescaling the training datasets for balancing all class instances. Two popular techniques of class resizing are over-sampling and under-sampling.

a).Over-sampling :- Over-sampling is the way toward expanding the number of classes into the minority class. SMOTE() is used for Over-sampling data.This process is continued till the majority and minority class occurrences are balanced out.

b).Under-sampling :- Under-sampling approach is used to level class distribution by indiscriminately removing or deleting majority class instances. This process is continued till the majority and minority class occurrences are balanced out.

2.5.2 Evaluate performance

After data re-sampling, the best performing classifier from the above model performance is used and then model was fit onto the re-sampled training data and the predictions were made for the testing data to find the evaluation metrics.

3 Result

The Classification Report of each indicator is considered and model performance is evaluated. All the accuracies of each inidicator and classifier are tabulated below in table 2

XGBoostClassifier performed better when compared to the other classifiers.

Random Forest gives highest for all traits. However, TF, JP Indictaor accuracy is less compared to XG Classifier is obtained.

MultinomialNB classifier gives overall low performance, however its TF is a little bit high.

DecisionTree has low accuries of TF and JP indicators.

Comparison of accuracy of classifiers				
Classifier	IE	NS	TF	JP
LogisticRegression	76	84	73	63
DecisionTree	76	85	62	59
MultinomialNB	70	79	74	61
RandomForestClassifier	78	86	70	62
XGBoostClassifier	78	86	72	64

Table 2: Table with accuracies of indicators of each classifier

Using XGBoost as baseline method and with the re-sampled data, model was fit on the resampled data and predicted using testing data. The model has overall accuracy of around 50 with over sampling and around 40 with under sampling. Model performance hasn't improved at all even with the re-sampling.

XGBoostClassifier is better performing model of all the classifiers with the accuracy of 78, 86, 72, 64(Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P)) Indicators.

4 Conclusion

Overall aim of this project is to come up with a model that performs better in predicting the individual posts on the social media, even though dataset is highly imbalanced, XGBoost model did better to give that accuracy. More work has to be done to get proper data and need to comeup with better preprocessing techniques to analyze posts and do more feature engineering to achieve high performance model.Then to build an application that will be very useful where it predicts the individual personality from there posts, Mainly which gives a flag or alerts to the family of individuals when someone puts highly depressed posts on thier social media accounts

References

[1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep LearningBased Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr.2017. [2] D. Xue et al., "Personality Recognition on Social Media With Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017

[3] M. C. Komisin and C. I. Guinn, "Identifying personality types using document classification methods," In Twenty-Fifth International FLAIRS Conference, 2012.

[4] A. Khan, H. Ahmad,"Personality Classification from Online Text using Machine Learning Approach" vol. 11, no. 3, 2020

[5] M. Hossein, H. Kazemian,"Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®",: 14, 2020

[6] L M. Tadesse, Bo Xu, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform"

[7]S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Personal Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.

[8]G. Carducci, G. Rizzo, 1 "TwitPersonality: Computing Personality Traitsfrom Tweets Using Word Embeddings and Supervised Learning ", : MDPI, 18, 2018