Personality Prediction

Ayush Gangwar (B20CS008)

Abstract

The growing number of people using social media has resulted in a considerable increase in the amount of information available online. The contents that these users post on social media can often provide valuable insights into their personalities (for example, in terms of predicting job satisfaction, specific preferences, and the success of professional and romantic relationships) without the hassle of taking a formal personality test. The approach, known as personality prediction, entails breaking down digital input into components and mapping them to a personality model. The motivations for building such a classifier are the pervasiveness of social media means that such a classifier would have ample data on which to run personality assessments, allowing more people to gain access to their personality type, and perhaps far more reliably and more quickly. Our initial impressions of the proposed work were positive, since it surpassed similar current studies in the literature. On the dataset, our results achieve a maximum accuracy of 60 percent by LGBM classifier.

1.Introduction

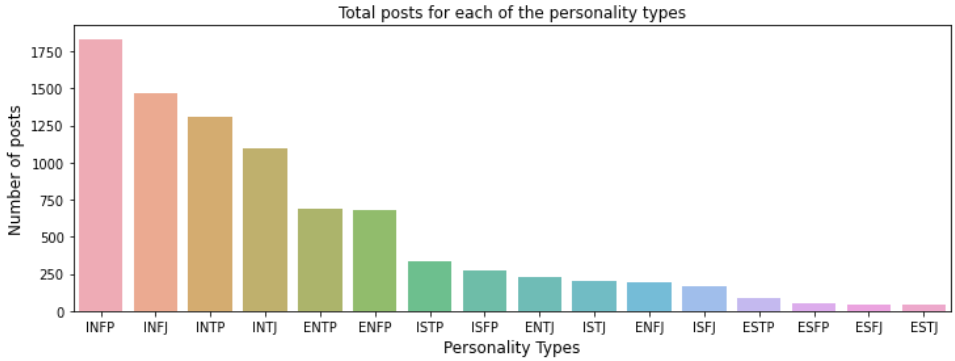
The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axis. The following are the measurements. Extraversion (E) vs. Introversion (I): a measure of whether a person favours the outside or the inside world. Sensing (S) vs. Intuition (N): a measure of how much information is processed through the five senses vs. impressions based on patterns. Thinking (T) vs. Feeling (F): a preference for objective principles and facts over weighing other people's emotive viewpoints. Finally, Judging (J) vs Perceiving (P): a measure of how much an individual likes a planned and structured life over one that is more flexible and spontaneous.

Dataset type:

The dataset contains 8675 rows where each row represents has posts and type of person

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Distribution of personality type in dataset:



2.Approach

We did exploratory data analysis, data preprocessing , followed by training the models and evaluation.

- **Preprocessing :**

  The text processing is broken up into two operations, cleaning the post data, lemmatization, tokenization and vectorizing the cleaned post data by making use of the Tf-idf Vectorizer.

  The cleaning process is defined as follows:

  - Removing the url links
  - Striping punctuations
  - Convert text to lowercase
  - Removing numbers from the dataframe

  Tokenizer divides a text into smaller bits to aid comprehension and the development of NLP models. By analysing the sequence of words, tokenization aids in interpreting

the meaning of text. Lemmatization reduces different forms of a word to the root word.

3. Methodologies

We implemented the following classification algorithms

- Logistic Regression
- Random Forest Classifier
- LGBM Classifier
- Multinomial Naïve Bayes

- **Evaluation**

Model	Train			Test		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
Logistic Regression	0.62	0.68	0.59	0.42	0.41	0.40
Multinomial NB	0.15	0.17	0.26	0.11	0.23	0.11
Random Forest	0.6	0.79	0.52	0.10	0.09	0.05
LGBM Classifier	0.98	0.98	0.98	0.59	0.58	0.60

- ***Accuracies:***

Model	***Accuracy on test data***
Logistic Regression	57.08797541298501
Random Forest	28.19823280829812
LGBM Classifier	60.16135228582405
Multinomial NB	22.973492124471764

4. Results and observation

Multinomial Naïve Bayes performed worst because of it's poor assumption.

Further , we can see ensemble model like LGBM perform the best.

Accuracies of models can be increased by required hyperparameter tuning.