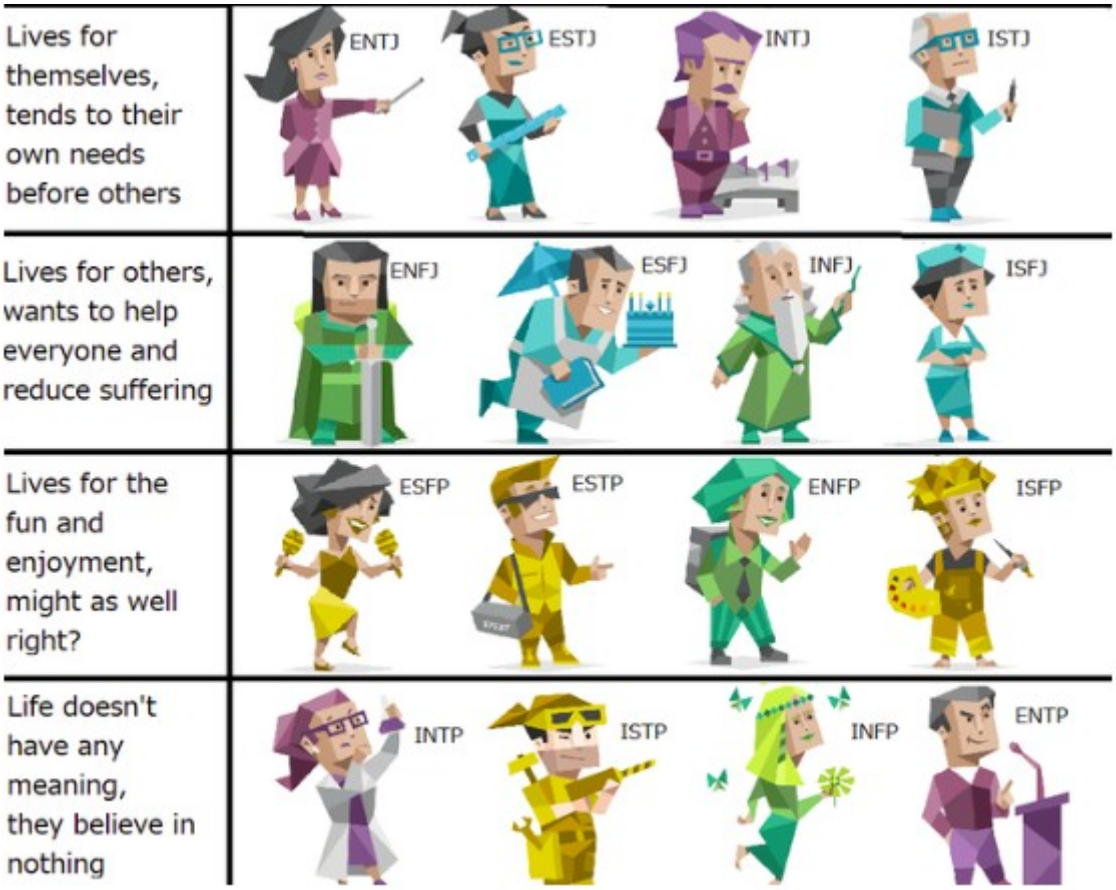
MBTI Personality Classification

Abu Shahid
B20CS003

Abstract- This paper reports author's experience with building a MBTI Personality Predictor. In data to day life, people share personal data and broadcast it between the user all over the internet. This data can be useful for organizations, marketing and sentiment analysis. We create a classification model using text-data features and meta features from user comments, messages and posts to predict their personality and analyze it.



I. 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 MBTI lays out a binary classification based on four distinct functions, and draws the typology of the person according to the combination of those four values (e.g. INFP, ESTJ):

- **Extraversion/Introversion (E/I)** - preference for how people direct and receive their energy, based on the outer or inner world
- **Sensing/INtuition (S/N)** - preference for how people take information in, by five senses or by interpretation and meanings
- **Thinking/Feeling (T/F)** - preference for how people make decisions, by relying on logic or emotions towards people and special circumstances

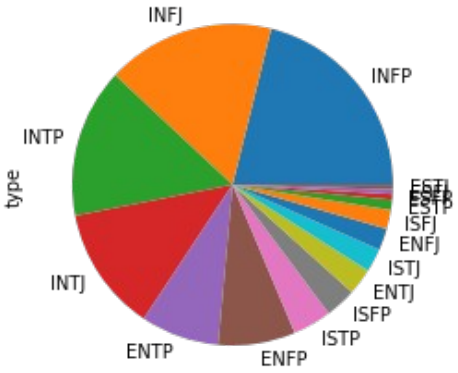
- **Judgment /Perception (J/P)** - how people deal with the world, by organizing it or staying open for new information

So for example, someone who prefers introversion, intuition, thinking and judging would be labelled an INTJ in the MBTI system, and there are lots of personality based components that would model or describe this person's preferences or behaviour based on the label.

Dataset

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))



II. Methadologies

Overview

MBTI personality prediction is a classification task requiring supervized machine learning on the data given. For our task, we did exploratory data analysis, data preprocessing and preparation, followed by training and evaluation.

We implemented the following classification algorithms

- Gaussian Naive Bayes
- Logistic Regression
- Kneighbours Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- MLP Classifier

1. Exploratory Data Analysis

No NA values were encountered in the dataset. The personality types did not have equal representation in the dataset. Two more features: words_per_comment and variance_in_word_count were added to the dataset.

2. Data-preprocessing and preparation

To better appretiate the relationship between text and personality types, we tokenize text to form Bag of Words. Tokenizer breaks a text into smaller chunks to help under stand the context and develop model for NLP. Tokenization helps in interpreting the meaning of text by analysing the sequence of words.

Further, posts were cleaned. URLs, vague punctuation, non-unicode text and dots between the words was removed. Tokenization and Lemmatization was performed. Lemmatization reduces different forms of a word to the root word.

TFIDF vectorizer was used for the same. TF-IDF is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words.

The TFIDF matrix was fitted into a TruncatedSVD model to get our trainable dataset.

3. Training and Evaluation

A function was defined which fits our data in the aforementioned models and gives a baseline report with accuracy, precesion, recall, F1 score and specificity. Model was first trained on small resampled data to check the functioning of the implementation and then later was trained on complete dataset.

Resampled Dataset

model	accuracy	precision	recall	f1score	specificity
randomforest	0.500	0.489	0.474	0.480	0.964
GNB	0.415	0.474	0.396	0.403	0.962
xgboost	0.389	0.414	0.386	0.386	0.961
DT	0.258	0.310	0.262	0.282	0.954
logit	0.178	0.260	0.192	0.175	0.950
MLPC	0.106	0.134	0.120	0.076	0.940
KNN	0.074	0.071	0.068	0.055	0.937

Complete Dataset

model	accuracy	precision	recall	f1score	specificity
xgboost	0.628	0.621	0.624	0.621	0.974
randomforest	0.610	0.630	0.609	0.586	0.971
MLPC	0.588	0.580	0.584	0.548	0.973
GNB	0.542	0.586	0.539	0.547	0.967
DT	0.414	0.413	0.414	0.418	0.959
KNN	0.147	0.139	0.140	0.142	0.938
logit	0.210	0.097	0.210	0.101	0.937

In our classification, precision is not more relevant than exhaustivity neither the opposite, plus F1 is much less prompt to overfitting or underfitting issues compared to accuracy.

The tables show that we get a maximmum accuracy of 62.8% with XGBoost trained on complete data. This is satisfactory and any attempt to hyperparameter tuning was not performed, considering the small size of our dataset. The best English models were trained on over 1M Twitter instances using logistic regression classifier and combining linguistic and count-based meta-features, nevertheless, they outperformed the majority-class baseline only on the IE dimensions, achieving the accuracy of 72.5% on those binary tasks(Sanja and Seren).

Two hypothesis can posed for such an outcome.

1. Social media commentary is not representative of one's personality.
2. Textual data does not resonate with MBTI personality.

References

- Sanja and Seren- Why Is MBTI Personality Detection from Texts a Difficult Task?
- Using TF-IDF to Determine Word Relevance in Document Queries
- How to build a Lemmatizer and why?