Immediate MBTI Personality Diagnosis Through Online Posts

Group 7: Yihang Hu, Jiacheng Yang, Aijia Zhang

1. Summary

This project explores the concept of classification of personality types using the MBTI test, which is usually an introspective self-report questionnaire that shows people's different psychological preferences when perceiving the world and making decisions. It arises from four dimensions that consist of two categories each. Those dimensions are Extraversion (E) versus Introversion (I), Sensing (S) versus Intuition (N), Thinking (T) versus Feeling (F), and Judging (J) versus Perceiving (P).

The main focus of the project is constructing an immediate MBTI personality diagnosis through individuals' responses to certain topics. To begin with, two precleaned datasets are adopted from Kaggle for model construction. Both datasets originate from dataset MBTI1, but with different cleaning standards. The first dataset extracts several features from the original posts such as average content length, and frequency of certain punctuations, while the other dataset is content only.

After doing literature reviews on papers under similar settings, three commonly used classification methods are selected: Naive Base, Linear SVM, and Random forest. The first dataset with features is selected in this section. Instead of applying direct multi-classification on sixteen personality types, this study carefully carries out four separate binary classifications according to four major dichotomies in the original MBTI test. Regardless of the four classifications, Random Forest seems to outstand others with the highest accuracy rate of 76.8%(E or I), 86.4% (N or S), 65.2% (T or F), and 59.1 % (J or P).

Later, Long short-term memory (LSTM) is used to accomplish a content-only classification, which is a typical machine learning model based on a bag of words model that considers the words in isolation but is a type of neural network that has hidden states and allows past outputs to be used as inputs. To make the comparison, the study also applied three previous methods to the second content-only dataset. However, LSTM fails to show superior performance compared to others in both binary and multi-classification scenarios.

In later explorations with a larger dataset mbti500, the result shows LSTM personality predictive power increase and surpass the other three as data size increases. LSTM obtains an overall 72% accuracy rate in multi-classification while others only obtain around 20% accuracy.

2. Introduction

This section introduces the basic concept, existing applications, and historical progress of MBTI, which contributes to the motivation of this study.

2.1 Definition

Mayes bridges type indicator, denoted as Mbti, is one of the most popular personality tests in the country. MBTI is created based on the theory of famous Swedish psychoanalyst Carl Jung, who believed that humans experienced the world using four different functions; these functions were the dichotomies that we associate with the test today. So what are the four dichotomies?

- Extraversion or Introversion
  Differ by the focus of attention: An extrovert is energized by people and things in the external world; while an introvert is energized by ideas or impressions in the inner world.
- Sensing or Intuition
  Differ by information input: A sensing person gathers details and facts that can be confirmed by experience; while an intuition person gathers ideas and sees future possibilities
- Thinking or Feeling
  Differ by decision making: A thinking person makes decisions by logic and analysis; while a feeling person makes decisions based on personal values
- Judging or Perceiving
  Differ by lifestyle: A judging person enjoys planning and deciding; while a perceiving person enjoys remaining open to new options.

Four dichotomies together form up to sixteen different personality types, each denoted with four capitalized letters as shown below(Figure 1),

Figure 1: 16 MBTI personality types diagram [1]

2.2 Application of MBTI

The MBTI results help individuals to explore and understand their strengths and weaknesses, what their likes and dislikes are, and their compatibility with other people. It can help people to understand how others view them and what career paths might be most suited to their talents.

Besides, MBTI is relevant to psychological counseling in three main ways: (1) by making the core qualities, e.g. empathy and acceptance, more tangible; (2) as a technique and framework; and (3) as a perspective on counseling practice and on other counseling theories.

Furthermore, MBTI is gradually being employed by more and more companies as an alternative or supportive tool for recruitment and employee assessments. A shocking 89 of the fortune 100 companies actually use this test for hiring and development purposes.

2.3 The Motivation of The Study

Throughout history, the MBTI test manual was modified several times since its first publication, and became much shorter and easier to interpret than the original version; Nevertheless, the current US version still contains over 90 questions. Test-taking is still a time-consuming process, which largely lowers the accessibility and willingness of individuals who are interested. The goal of the project is to provide an immediate and convenient tool for individuals to get their MBTI personality test results directly through their online posts.

3. Dataset

Two datasets are used in the project: mbti1 and mbti500. They are both derived from Kaggle, an online forum that has many professional and pre-processed datasets. According to the description, the data in the two datasets are originally collected from the PersonalityCafe forum on which individuals with different personalities share their daily posts.

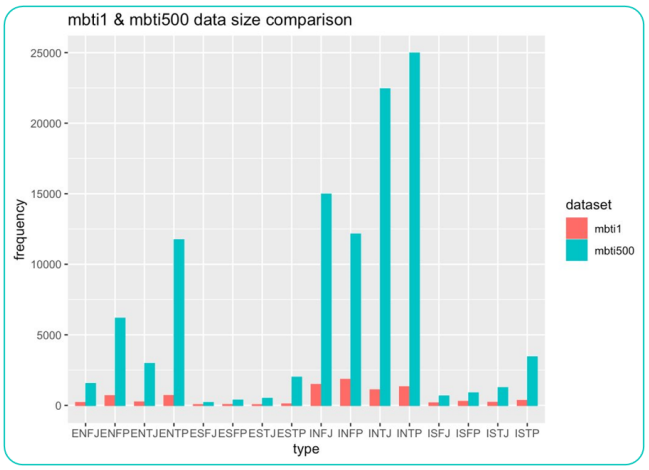

Figure 2: Data size comparison between MBTI1 & MBTI500 datasets

Mbti500 is the original dataset, containing 106067 individual observations (each observation includes 50 posts of that individual) and their corresponding MBTI personality types. Mbti1 is a small subset of Mbti500 with only 8675 observations but it has been preprocessed and has more features, such as the number of emojis in each observation and the calculated Sentiment Score. [2]

The sentiment Score is used to understand the emotion behind comments and text. Here Sentiment Score is calculated through the VADER Sentiment algorithm. VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities called sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

This project mainly uses mbti1 for traditional machine learning classification algorithms to avoid too much computation and only switches to mbti500 when doing experiments to see if data size influences the effectiveness of the LSTM(Long short-term memory) algorithm. Below are the features derived from the original posts in dataset 1 (Figure 3).

```
 [1] "X"              "type"           "posts"             "EorI"              "NorS"
 [6] "TorF"           "JorP"           "avg_comment_length" "comment_length_var" "Sentiment"
[11] "Ellipses"       "Exclamation"    "Question"          "Links"             "Picture"
[16] "Emojies"        "Upper"
```

Figure 3: List of extracted features from dataset 1.

4. EDA

In this section, all the exploratory data analysis is based on mbti1. This is because mbti1 is pre-processed and has many interesting features that really worth exploring.

4.1 Data Distribution

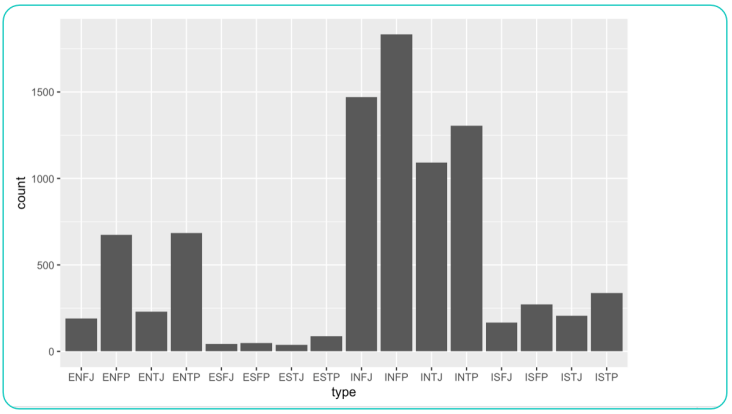

Figure 4: Distribution diagram of 16 MBTI personalities

There are in total 8675 observations in mbti1. As shown in the plot(Figure 4), there is a huge imbalance between the number of people in different personality groups. In Particular, there is almost no data from the ES group(ESFJ, ESFP, ESTP, ESTJ) but a huge amount of data from the IN group(INFJ, INFP, INTJ, INTP).

4.2 Exploratory Questions

4.2.1 Question 1: Do extroverted people overall speak more than introverted people?

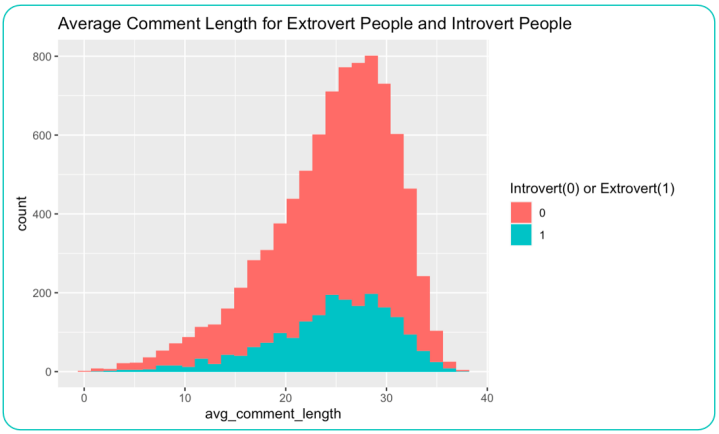

Figure 5: Average comment length comparison between extroverted & introverted groups

In order to answer the question, the feature avg_comment_length, which stands for the average length of comments for each individual, is used and analyzed. From this plot(Figure 5), it is obvious that the number of introverts and extroverts varies a lot: most individuals are introverts in this dataset. However, both extroverts and introverts do show a similar distribution for average comment length. The mean value of avg_comment_length for introverts is 24.5057, a little bit smaller than that of extroverts(24.5887).

4.2.2 Question 2: Do extroverted people have more positive feelings(measured by Sentiment Score) than introverted people?

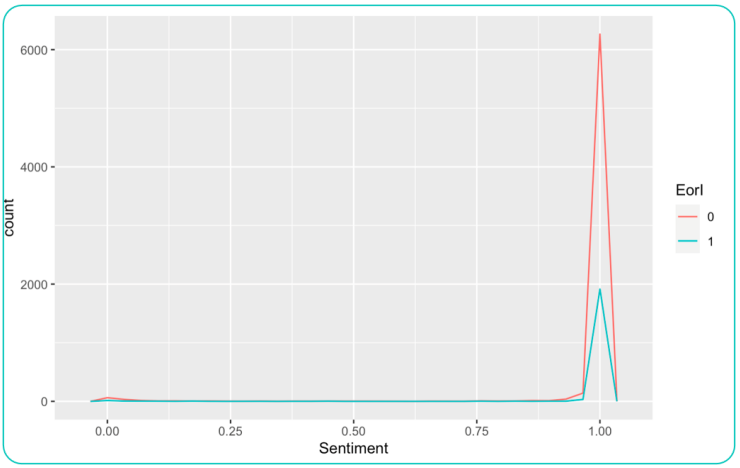

Figure 6: Sentiment scores comparison between extroverted & introverted groups

In order to answer this question, the Sentiment Score needs to be used for analysis. From the plot above(Figure 6), despite the imbalance of introverts and extroverts, the mean value of the Sentiment

Score for the two groups is similar: introverts' average sentiment score is 0.973, which is a little bit lower than that of extroverts(0.98).

5. Methodology

This section includes the basic concepts and implementations of the four distinct classification methods involved in the study. Apart from LSTM, which is conducted through python, the other three are implemented using R markdown.

5.1 Machine Learning Models

1. **Naive Base:** It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors, which fits our data situation of independent extracted features.

2. **Linear SVM**: The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. In our case n is two for binary classification. The selection of linear SVM is under the assumption that our data is linearly separable by using straight lines.

3. **Random Forest**: It is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result.

5.2 Deep Learning Model

Long short-term memory(LSTM)

The deep learning model the project implemented is the Long short-term memory, which is a special kind of Recurrent Neural Network (RNN).

A typical machine learning model based on the bag of words model considers the words in isolation, which means that the bag of words model will count the frequency of each word and based on that do the modeling, treating each word independently. However, Recurrent Neural Networks(RNN), a type of neural network that has hidden states and allows past outputs to be used as inputs, provide the previous information to the current cell and treat it as part of the training, considering not only every word in a sentence but also the sequence of each word.[3] For instance,

"The concert was boring for the first 15 minutes while the band warmed up but then was terribly exciting."

Here, when doing sentiment analysis, a typical machine learning model based on the bag of words model may consider this sentence as a negative one since more negative words are

covered in this sentence like “boring” and “terribly.” However, if RNN is implemented here, the sentence will be classified as positive which is the true meaning, since the sequence of words is taken as a count of the modeling.

However, there are some downsides to RNN. It is not able to capture the long-term dependencies in practice though theoretically speaking it can achieve that. Besides, RNN may suffer from vanishing gradients and be caused by long series of multiplications of small values. Hence, LSTM is introduced here to prevent these two problems from happening. Another example here:

"The cat, which already ate a fish, still tries to find something to eat."

Here, if the purpose is to find the subject of *“**still tries to find something to eat,**”* the RNN will provide the answer “fish,” since it is the closest subject next to this part of the sentence. However, it is not the truth. The right one should be “cat,” which RNN cannot capture but LSTM can find.

In Experiment 2 and Experiment 3, the project implements the improved RNN model LSTM to do the training and do the comparison analysis with other machine learning models.

6. Experiment

In this section, the models introduced above are implemented to do the comparative analysis. By comparing the test accuracies among these models, it will provide the best model under different conditions.

6.1 Experiment Design

Experiment 1—-The project focuses on the mbti_1 dataset at the beginning, which includes every element in the posts. The first thing to do is split the dataset into training and testing. The project puts 80% of the dataset into training and 20% of the dataset into testing. It first tries to do the multi-class classification on 16 types of personalities but the result is not very ideal. Then since all personality dichotomies are independent of each other, the project tries to do the binary classification four times: Introversion vs Extroversion, Sensing vs Intuition, Thinking vs Feeling and Judging vs Perceiving.

Experiment 2—-After the first experiment, the project provides a precondition that the classification only concentrates on the contents of every post. Therefore, using the same dataset, the project implements LSTM to do the modeling only forcing on the content itself.

Experiment 3—-At the end of the second experiment, the LSTM’s accuracies are not the best among all the models the project uses. Hence, the project uses a new dataset, mbti_500, to improve the backsides of the original dataset.

6.2 Results and Comparisons

In the following section, this paper used tables and plots to show how well every model performs under different circumstances in two datasets.

Experiment 1

The project tries to do the multi-class classification of 16 personality types on the mbti_1 dataset.

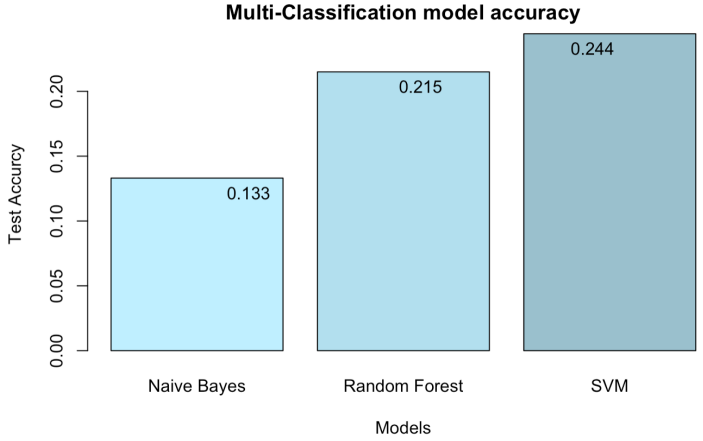

Figure 7: Histogram of three multi-classification models' accuracy rates (Naive Bayes, Linear SVM, Random Forest)

From the histogram above(Figure 7), it shows that the Linear SVM does the best job among these three models, but the overall accuracy is only around 24.4%, which is not very ideal. Therefore, due to the dependence of all personality dichotomies, the project chooses to do the binary classification four times as an alternative method.

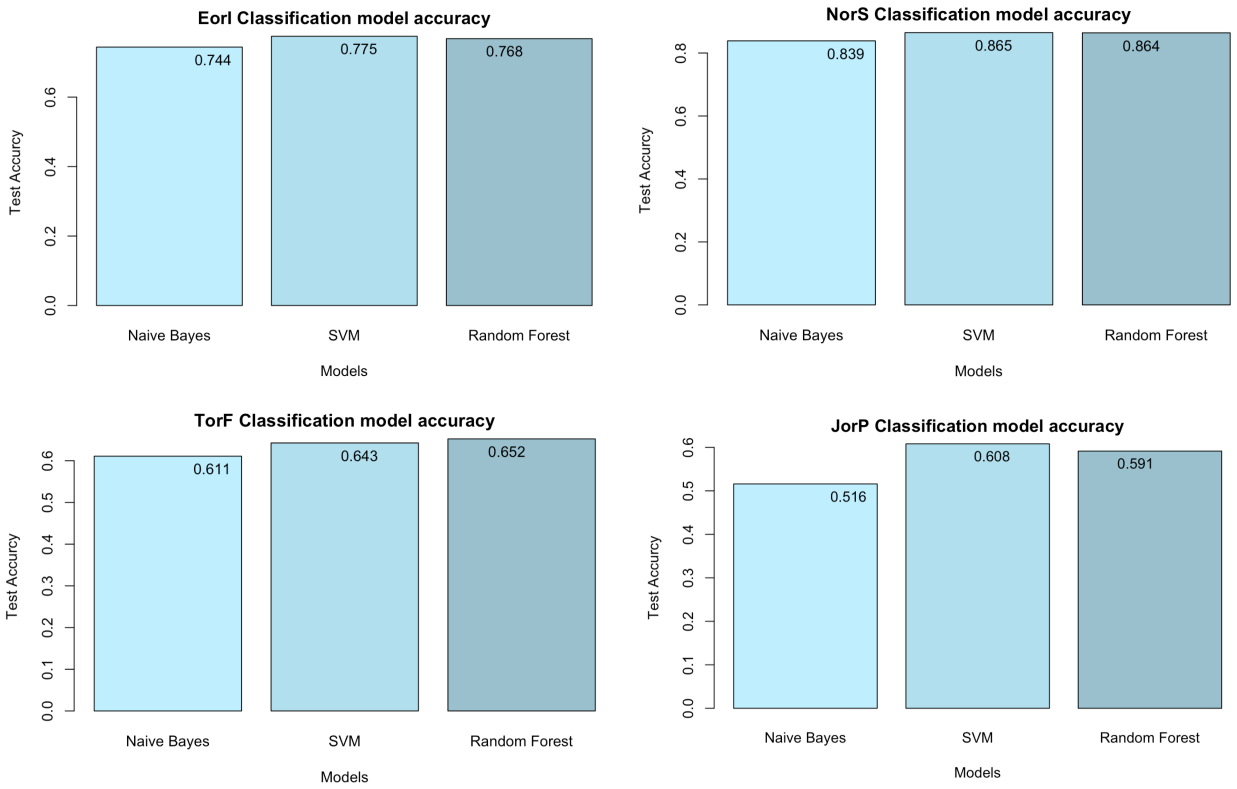

Figure 8: Histograms of three Binary classification models' accuracy rate under four dichotomies(EorI, NorS, TorF, JorP).

From the plots(Figure 8), it can be easily concluded that Random Forest and Linear SVM do a quite better job than the Naive Bayes. However, by observing the confusion matrices of Linear SVM, it seems that there are some problems going on here:

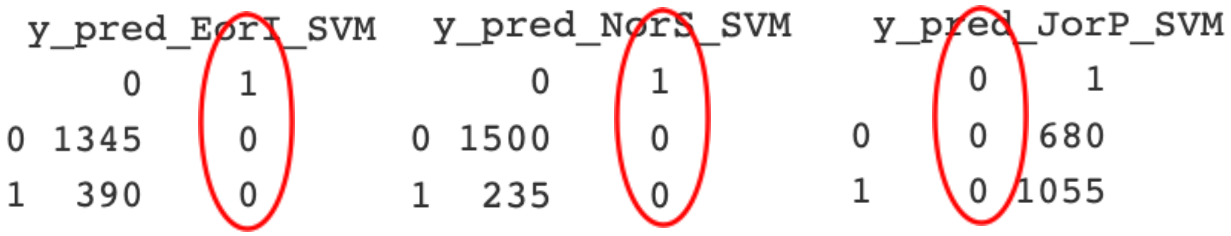

Figure 9: Confusion Matrices of Linear SVM model

Hence, from the tables shown above(Figure 9), it seems that there is an overfitting issue happening here. It is very much possible that due to the imbalanced dataset, the model fails to learn the pattern of one class and gets biased towards another class. Thus, till now, Random Forest provides the best performance.

Experiment 2

In this experiment, the project only forces the posts themselves, concentrating on the content. After applying LSTM to the same dataset, the project compares its results with previous models not only in multi-class classification but also in four binary classifications.

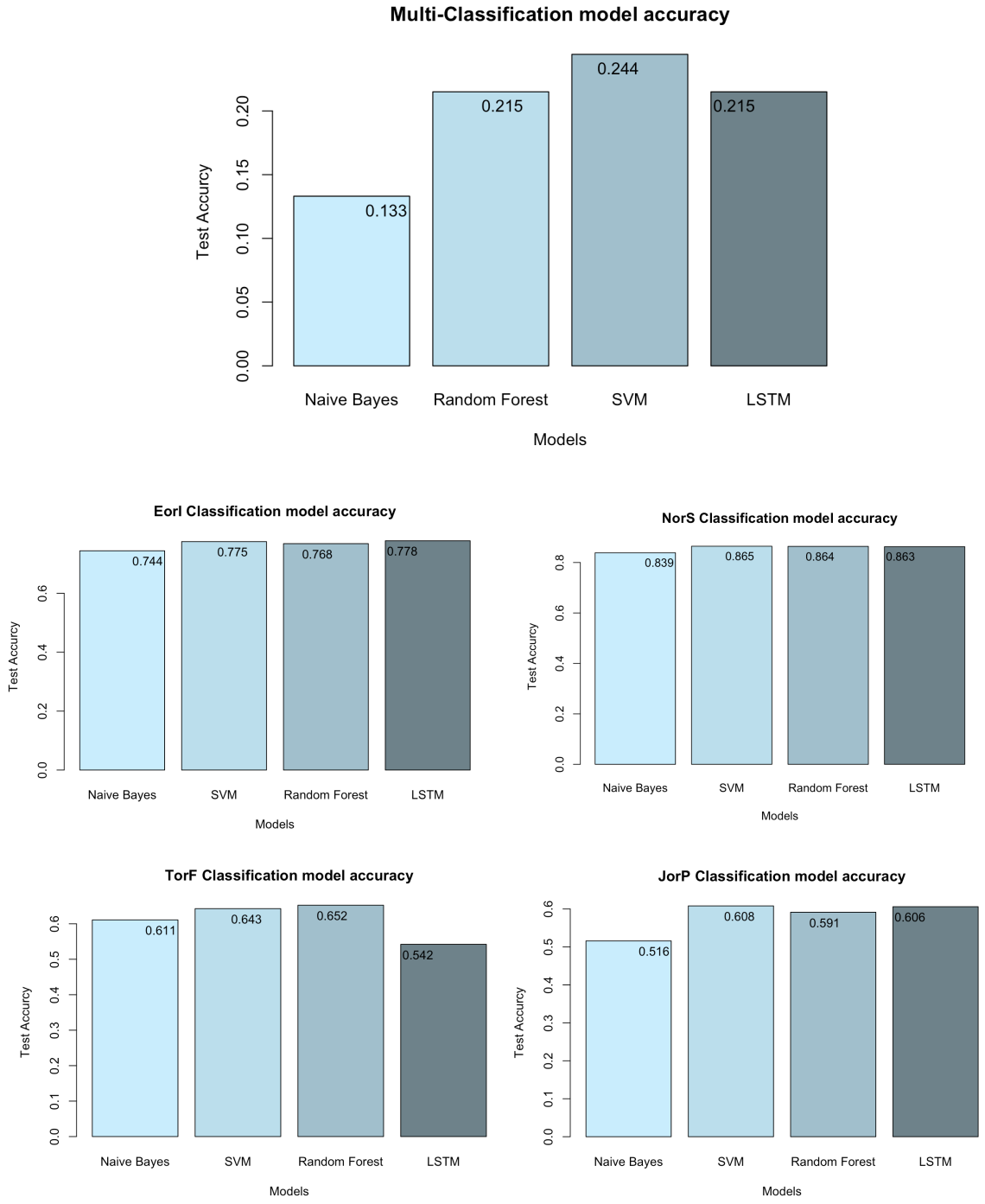

Figure 10: Histogram of four multi-classification models' accuracy rates (upper). Histogram of four binary-classification models' accuracy rates under four dichotomies(lower). Both with the mbti1 dataset.

Hence, from the histograms shown above(Figure 10), it seems that LSTM does not perform better than the Random Forest model and is even worse from time to time during the binary classifications.

The reason this situation happens may be information loss or the small size of the dataset. When the project uses LSTM to do the classification, it ignores many features such as emojis, links, etc., since, before the process, it needs to do the data cleaning. However, these elements are covered when it uses Naïve Bayes, Linear SVM, and Random Forest to do the training. Furthermore, the content in mbti_1 may not be enough for LSTM to learn the pattern during training, which leads to low accuracy.

Experiment 3

Based on the potential problems the project finds in the last experiment, this experiment forces on the new dataset–mbti_500, which has a larger data size and more content, excluding other features like links, but only content.

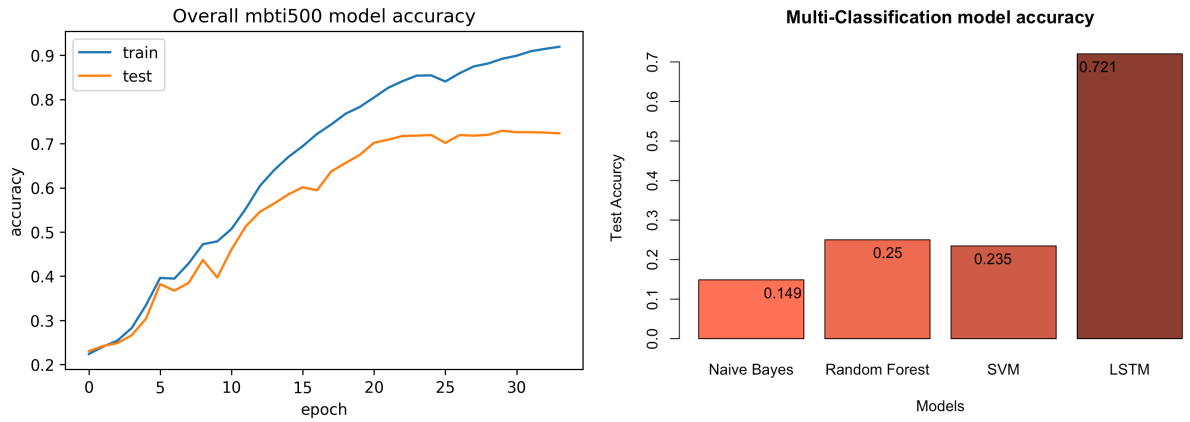

Figure 11: *Histogram of four multi-classification models' accuracy rates (with mbti500 dataset)*

Hence, from the plots shown above(Figure 11), LSTM shows a great performance of 72.1% accuracy, which is much higher than the other three models.

6.3 Conclusion and Evaluation

From these three experiments, it shows that if people want to predict a person's personality type based on his posts and along with all kinds of elements his posts have, like number of links and emojis, Random Forest can do a great job. However, if we want to do the prediction only based on the poster's writing pattern and content, LSTM seems to be the better choice, as long as enough data is provided.

7. Challenges & Future Direction

7.1 Challenges

7.1.1 Unrepresentativeness

One of the major challenges for this project is the unrepresentativeness of certain groups within the whole sample data. For example, the ES group(ESTJ, ESTP, ESFJ, ESFP) takes up around 2.5% of sample data but actually takes up around 30% of the real population. This could lead to a huge potential error when conducting personality classification in real life. In order to solve this, it is important to get a larger dataset that is more representative of the whole population.

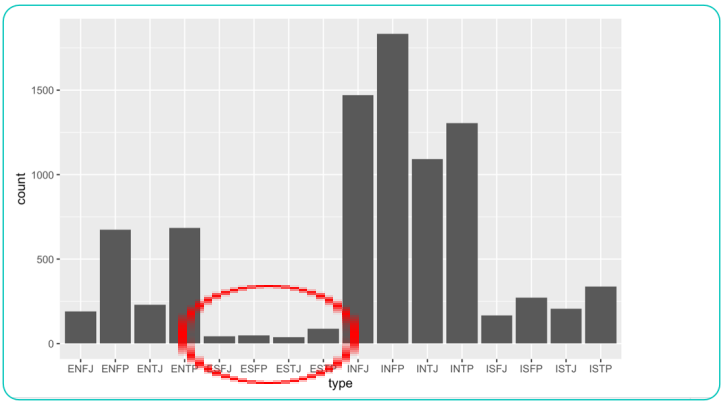

Figure 12: Unbalance of original data.

7.1.2 Inconvenience in the application

The classification models in this project require collecting 50 posts from individuals' social media, which is much better compared to the traditional time-consuming questionnaires that normally require individuals to spend an hour or even more. However, it is still not very practical to conduct this information retrieval process because not everyone has the habit of sharing posts on social media.

7.2 Future Direction

In order to solve the problem mentioned in 7.1.2, accurate voice recognition technology is needed. By transforming voices into text, information about individuals no longer needs to be retrieved from their social media but can be directly derived from conversations. With the help of voice recognition technology, the classification models of this project could be very helpful in scenarios like teaching, job interviews, and psychological counseling.

Acknowledge

The code implementation can be found in the attached GitHub link:
https://github.com/Psyduck572/STOR565_PROJECT_MBTI_CLASSIFICATION

Reference:

[1] Wikimedia Foundation. (2022, November 27). *Myers–Briggs Type indicator*. Wikipedia. Retrieved November 30, 2022, from https://en.wikipedia.org/wiki/Myers%E2%80%93Briggs_Type_Indicator

[2] Khalid, Z. (2021, December 30). *MBTI personality types 500 dataset*. Kaggle. Retrieved November 30, 2022, from https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset

[3] Olah, C. (2015, August 27). *Understanding LSTM networks*. Understanding LSTM Networks -- colah's blog. Retrieved November 30, 2022, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/