Classification of Myers-Briggs Type Indicator (MBTI) Type Based on Past Posts on Internet Forums

December 5, 2021

1 Introduction

The rise of the Myers-Briggs Type Indicator (MBTI) as one of the most popular personality test in the past decade as part of the greater mainstream usage of personality types in popular culture is not surprising, at the very least. It functions as a framework in understanding one's psychological tendencies through profiling four different dichotomies: Extraversion/Introversion (E/I), Sensing/Intuition (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P).[1] Part of its widespread appeal can be attributed to the ease of use and interpretation of the model.[2] Self-assessments—oftentimes through the internet—allow for a democratisation of the personality test.[3] The four-letter MBTI types are also then interpretable through tangible and understandable names (e.g. "The Architect", "The Mediator", etc.) as an extension thereof (Keirsey Temperament Sorter) and descriptions that highlight the personalities' assumed best traits.[4]

Within more established organisations, there is also a desire for a personality framework that can be easily used in order to optimise working relationships and to create more efficient groupings of people of varying types.[5] In this case, there is more emphasis on justifying the use of the MBTI in these places. Oft-cited justifications include the Jungian basis as mentioned by its founders, Katherince Briggs and Isabel Myers. While Carl Jung's contributions to the domain of psychology and psychoanalysis cannot be disputed, the nature of his work on psychological types cannot be considered scientifically rigorous.[6] As such, the Jungian basis cannot be used as a justification for the scientific nature of the MBTI, neither can the various studies endorsing it (many studies are either written by beneficiaries of the sales of the MBTI, methodologically weak, or both).[7]

As a part of the popular internet culture, observations of social media biographies indicate that the MBTI has become a way to communicate one's identity and preferences along with other indicators, heralding from astrology (e.g. Pisces, Virgo, etc.) or even fiction (e.g. Hogwarts' Slytherin House from Harry Potter). The appearances of the MBTI alongside these indicators form the basis of one of the main criticisms on the personality test. A phenomenon described as

1. Isabel Briggs Myers and Peter B. Myers, *Gifts Differing: Understanding Personality Type* (Mobius, March 5, 1995), ISBN: 978-0-89106-074-1, Google Books: KJ_kBfphgQgC.

2. En Jun Choong and Kasturi Dewi Varathan, "Predicting Judging-Perceiving of Myers-Briggs Type Indicator (MBTI) in Online Social Forum," *PeerJ* 9 (June 23, 2021): e11382, ISSN: 2167-8359, accessed December 5, 2021, https://doi.org/10.7717/peerj.11382, https://peerj.com/articles/11382.

3. Choong and Varathan.

4. Scott O. Lilienfeld, Steven Jay Lynn, and Jeffrey M. Lohr, *Science and Pseudoscience in Clinical Psychology, Second Edition* (Guilford Publications, October 12, 2014), ISBN: 978-1-4625-1751-0, Google Books: q50gBQAAQ BAJ.

5. Christopher J Lake et al., "Trust in Name Brand Assessments: The Case of the Myers-Briggs Type Indicator.," *The Psychologist-Manager Journal* 22, no. 2 (2019): 91, ISSN: 1550-3461.

6. Robert T. Carroll, "Myers-Briggs Type Indicator - The Skeptic's Dictionary - Skepdic.Com," accessed December 5, 2021, http://skepdic.com/myersb.html.

7. Frank Coffield et al., "Learning Styles and Pedagogy in Post-16 Learning: A Systematic and Critical Review," 2004,

the Barnum effect (or the Forer effect) can be used as a framework in describing the MBTI's rise in popularity.[8] In this case, the descriptors of the personality can be said to be vague enough to be applicable to almost everyone, while being somewhat specific enough to allow for the feeling that it was specifically tailored for the particular person.

While there are various criticisms on the usage of the MBTI, it is possible to observe that the widespread and democratising effect of the usage of the MBTI may lead to more scientifically rigorous methods. One such personality type construct is the Big Five Inventory, of which some correlations between its factors and MBTI dichotomies were found.[9] Although its creation and the studies on it that follow may have scientific gaps, the inference of the MBTI type seemed to have been influential in the field of social media advertisement personalisation.[10] The use of this technique has proven to positively impact consumer brand engagement, brand attachment, and subsequently revenue.[11]

Within this context, therefore, the usage of open social media posts by various users can be used to either predict the users' MBTI types or to validate their self-reported assessments and types. The current research on the matter have had successes in predicting the four binary dichotomies separately.[12] However, the predictive power of the MBTI type will be penalised quite severely due to the nature of having four independent probabilities combined together. Knowing that all four dichotomies are predictable at around 90% accuracy each in various researches, it averages out to around 66% accuracy when combined together.[13] This is significant as even a single letter difference in the prediction makes a noticeable difference in the types of recommendations and assumptions of the person assessed.[14]

As such, the focus of this research will be on the holistic prediction of MBTI based on textual contents written by users of a social media platform. The methods of processing the textual data will be discussed, both in terms of its shape and transformations. Afterwards, a discussion on the model best suited for this classification task will be shown, followed by the choice for good baseline models. In the end, there will be an exploration on the models chosen in the attempt to improve on them or to combine them.

8. Jerome Tobacyk et al., "Paranormal Beliefs and the Barnum Effect," *Journal of Personality Assessment* 52, no. 4 (December 1, 1988): 737–739, ISSN: 0022-3891, accessed December 5, 2021, https://doi.org/10.1207/s15327752jpa5204_13, https://doi.org/10.1207/s15327752jpa5204_13.

9. Paul T Costa Jr and Robert R McCrae, *The Revised Neo Personality Inventory (Neo-Pi-r).* (Sage Publications, Inc, 2008), ISBN: 1-4129-4652-2.

10. Myeong-Yeon Yi, O-Joun Lee, and Jason J. Jung, "MBTI-Based Collaborative Recommendation System: A Case Study of Webtoon Contents," in *Context-Aware Systems and Applications*, ed. Phan Cong Vinh and Vangalur Alagar, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (Cham: Springer International Publishing, 2016), 101–110, ISBN: 978-3-319-29236-6, https://doi.org/10.1007/978-3-319-29236-6_11.

11. Cheol-Ho Yoon and Dong-Sub Lim, "The Effect of the Big Five and the MBTI on Impulsive and Compulsive Buying Behaviors: An Integrated Analysis in Online Shopping," *Journal of International Trade & Commerce* 14, no. 3 (2018): 101–117.

12. Kosuke Yamada, Ryohei Sasano, and Koichi Takeda, "Incorporating Textual Information on User Behavior for Personality Prediction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (Florence, Italy: Association for Computational Linguistics, July 2019), 177–182, accessed December 5, 2021, https://doi.org/10.18653/v1/P19-2024, https://aclanthology.org/P19-2024.

13. Choong and Varathan, "Predicting Judging-Perceiving of Myers-Briggs Type Indicator (MBTI) in Online Social Forum."

14. Choong and Varathan.

2 Dataset

2.1 Source

The dataset being used will be from Kaggle, named '(MBTI) Myers-Briggs Personality Type Dataset'.[15] It contains two columns: a section of the last fifty posts by each user, separated by three pipe characters ("|||") and the users' MBTI type. Each row is represented by a user from the PersonalityCafe forum, a social media site mainly focused on discussions on the self and personality types. In total, there are about 8600 rows of data.

2.2 Exploratory analysis

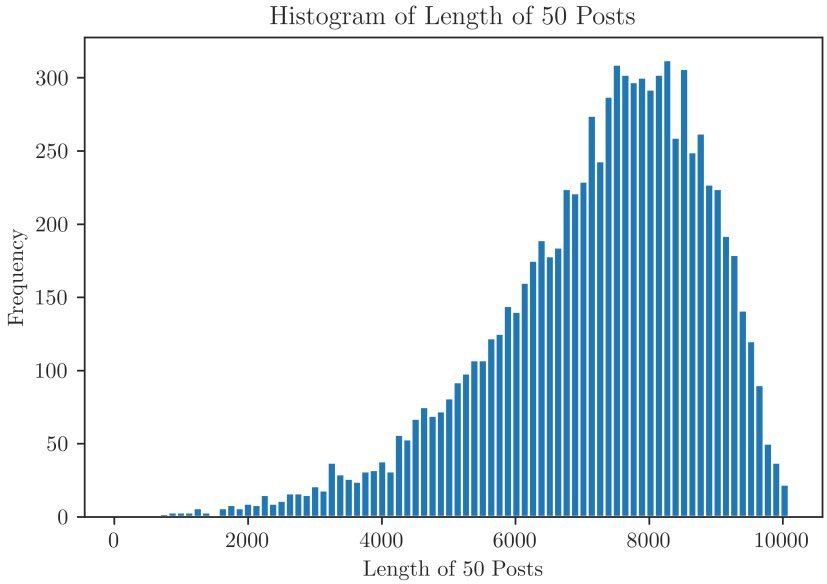

Figure 1: Histogram of the length of posts in the dataset

The lengths of the posts seem to be negatively skewed with a somewhat positive kurtosis. This indicates a consistency for the aggregate of fifty posts to be further from zero, which is advantageous as there are, in general, more data to work with per user. When considering the average length of posts when divided into the different types, there does not seem to be any appreciable difference in the distributions as seen in Figure 2.

However, the population distribution based on the MBTI types seem to be extremely uneven. In order to find out if this is the case for most of the population, a cross reference into a different dataset is done. This time, the dataset is from Reddit, with one post per row.[16] The observation from that dataset indicates that it is likely that, at least on the internet, the labelled MBTI type distribution is very uneven as seen from Figures 3 and 4.

This presents a problem in that since the difference between the type with the highest population and the lowest population is extremely large, classifying these minority types might be much more difficult and with more errors than the majority types. The problem of uneven types can also be seen even when dividing them into their respective dichotomies and trichotomies as seen in Figures 5, 6 and 7.

15. "(MBTI) Myers-Briggs Personality Type Dataset," accessed December 5, 2021, https://kaggle.com/datasnaek/mbti-type.

16. Dylan Storey (Myers Briggs Personality Tags on Reddit Data; accessed December 5, 2021), July 30, 2018, https://doi.org/10.5281/zenodo.1482951, https://zenodo.org/record/1482951.

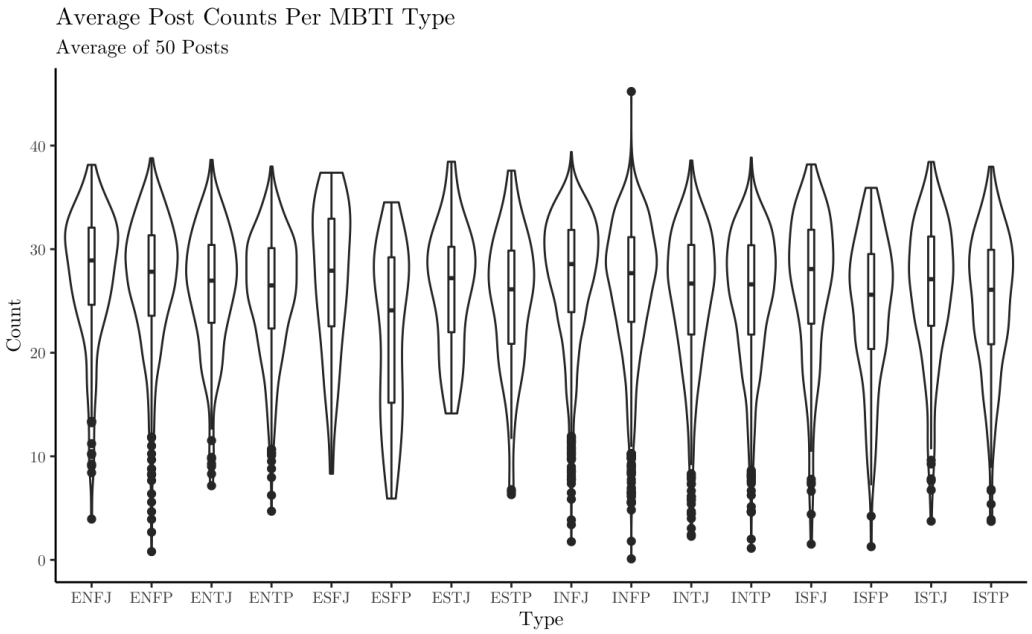

Figure 2: Average length of posts within each type

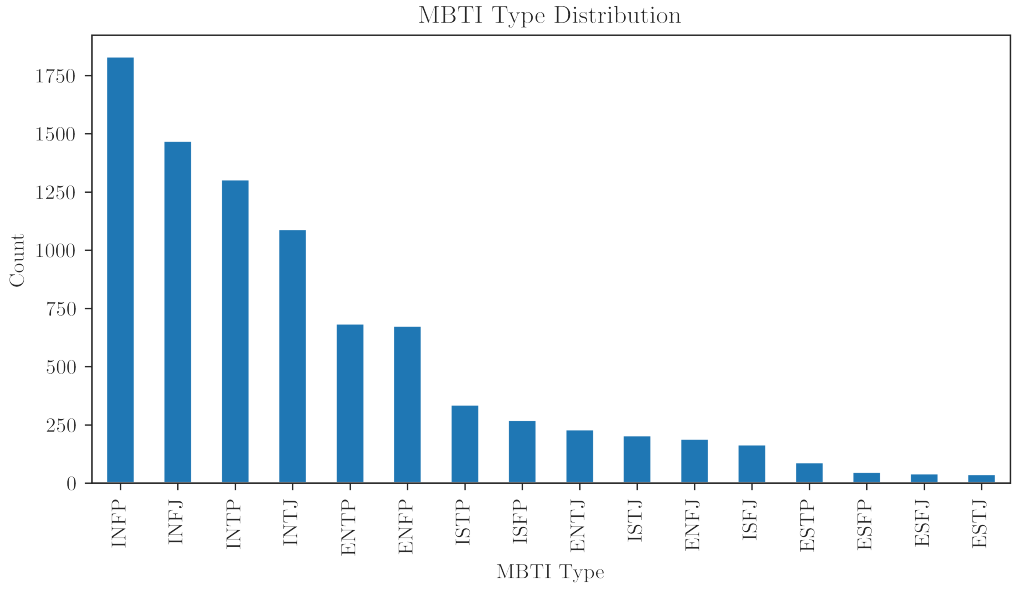

Figure 3: Population distribution of the MBTI types

When observing the word clouds in Figure 8, there seems to be an indication that the corpus themselves contain the correct classification. This is to be expected as the data came from a forum based on conversations on personality. Therefore, the removal of these words is one of the considerations that is going to be made when preprocessing the text data. Knowing that the types of words used are different in each class means that there is some confirmation that classification based on text data might be feasible.

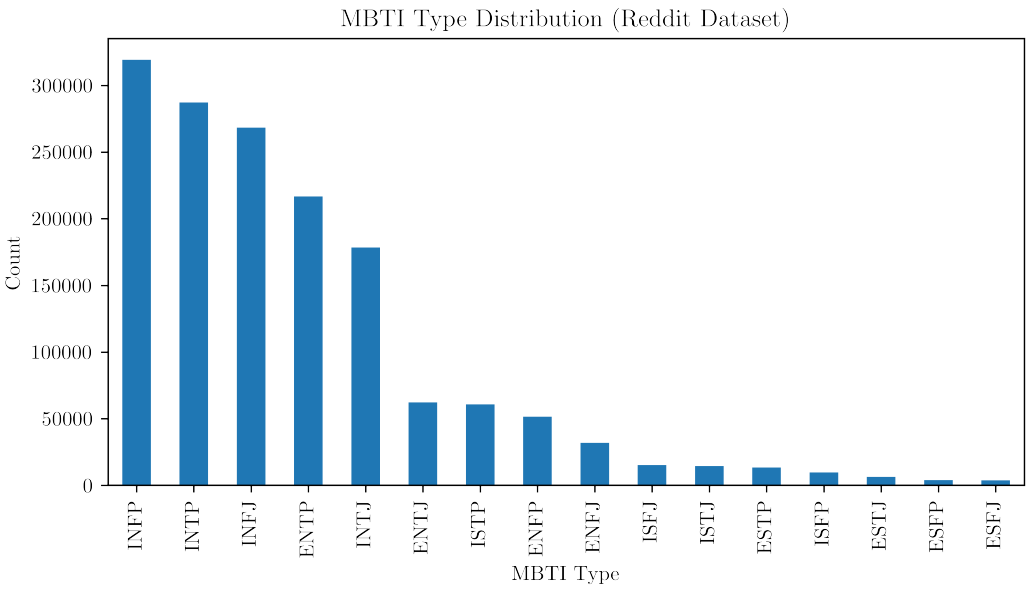

Figure 4: Population distribution of the MBTI types from Reddit

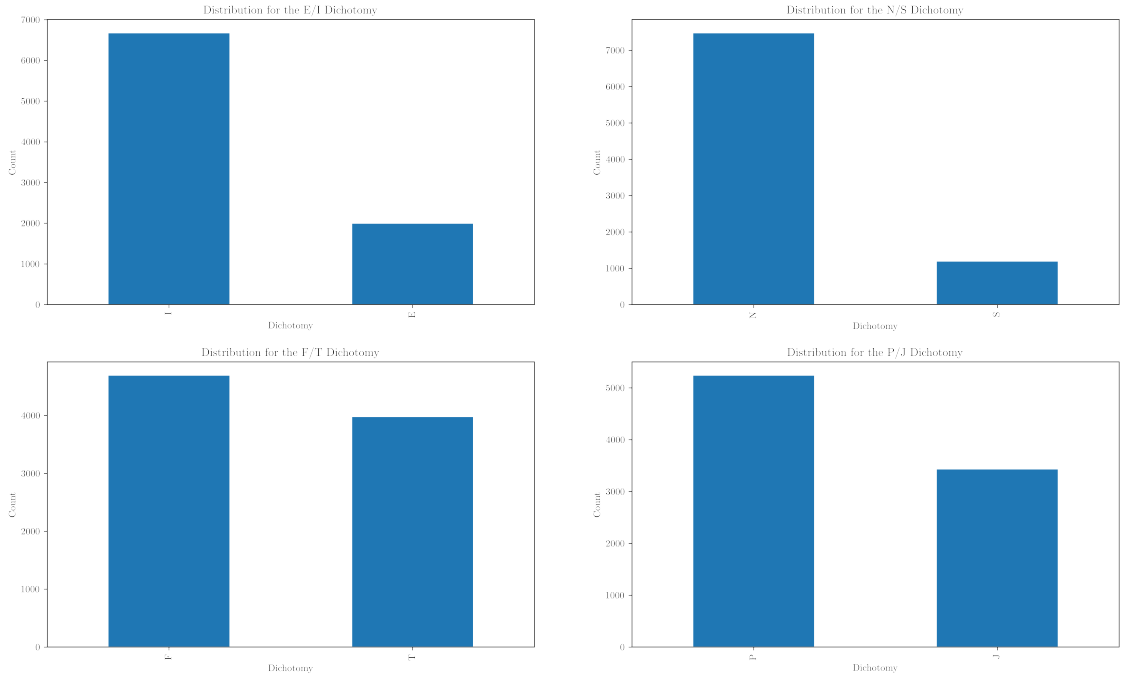

Figure 5: Population distribution of the MBTI types when divided into different dichotomies

3 Modelling

On a macro scale, the process of modelling the data will involve the input of the dataset as mentioned above with the objective of obtaining a model that best predicts the MBTI type of the user based on their posts.

Observing the process a little bit deeper, there are roughly four main steps, happening somewhat sequentially, though there will be many iterations based on the model accuracy. The data goes through text preprocessing first, involving removals of irrelevant data and transformations required for the models being tested. Here, processing time is a concern as it will also impact not only the modelling, but also the prediction time using the model. The processed text is then

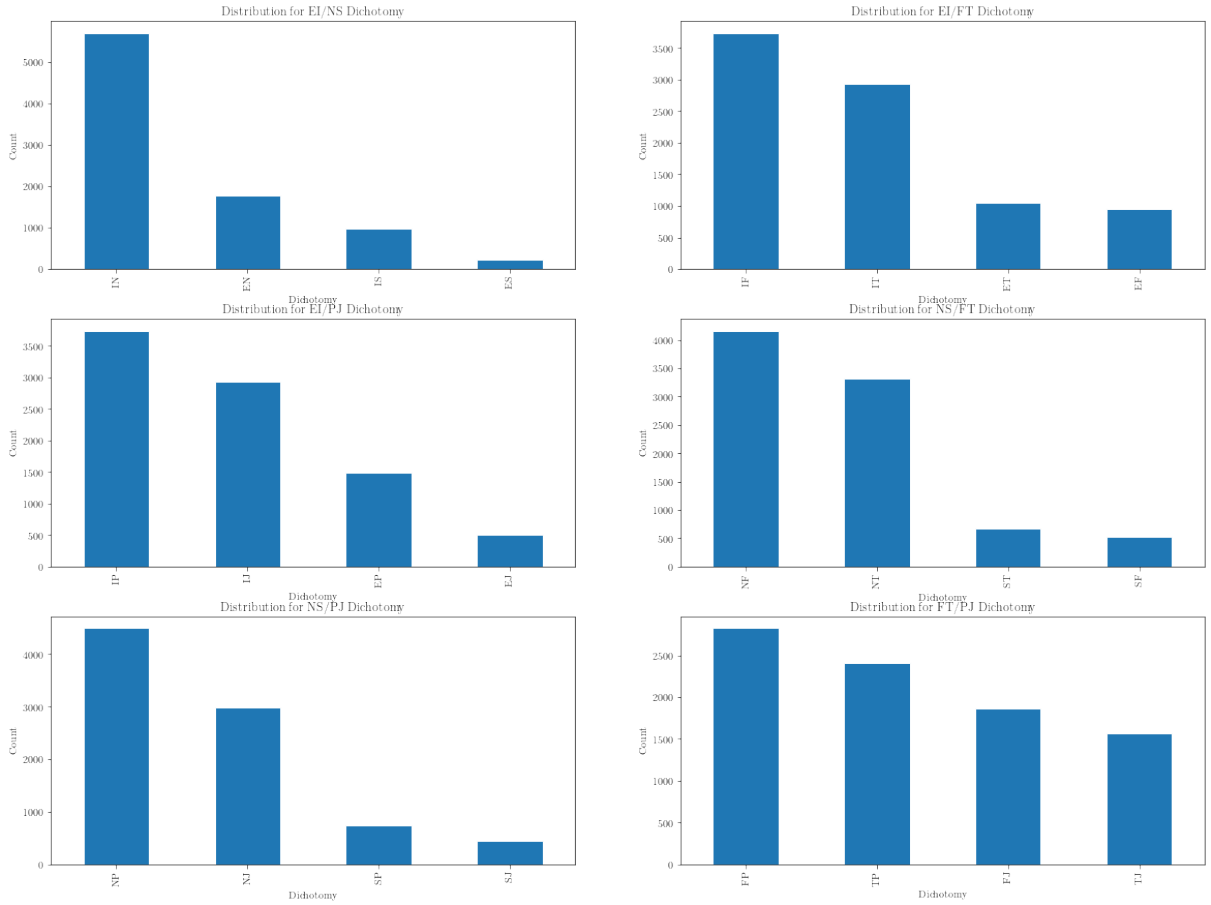

Figure 6: Population distribution of the MBTI types when divided into different two-dimentional dichotomies

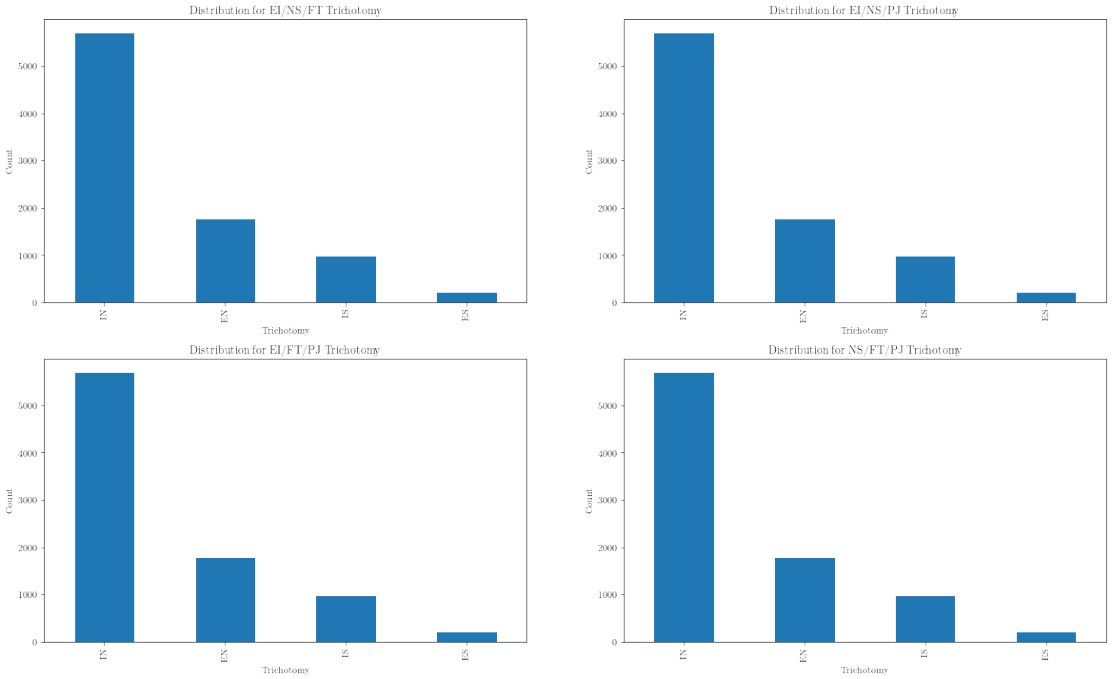

Figure 7: Population distribution of the MBTI types when divided into different trichotomies

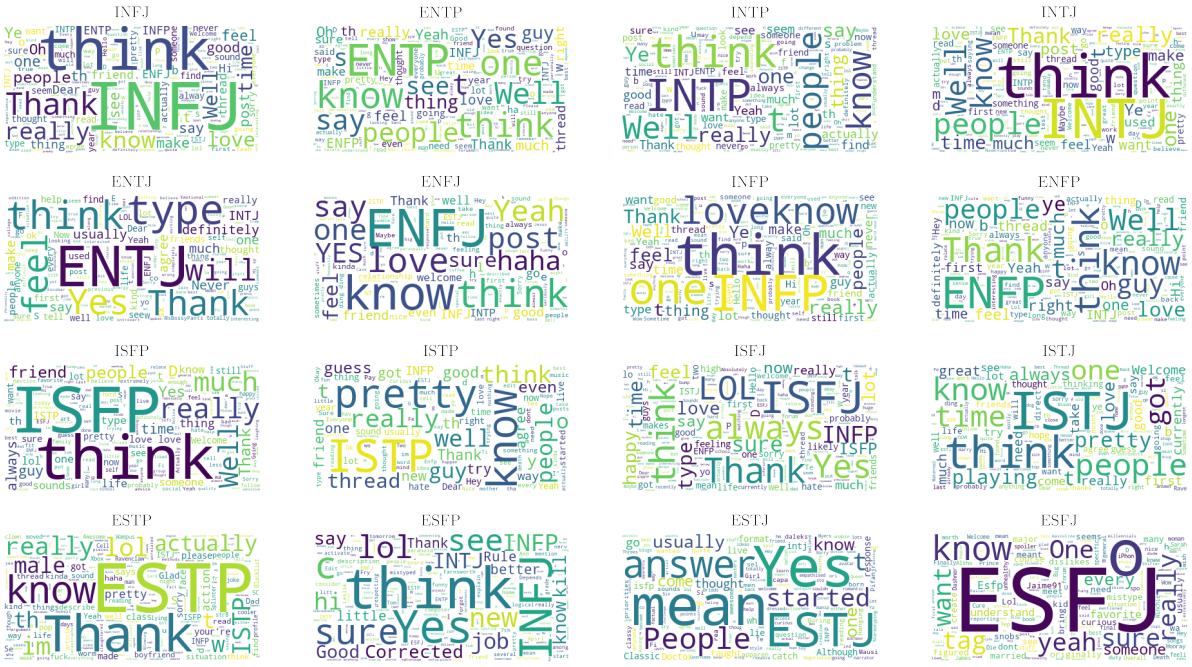

Figure 8: Word clouds from each MBTI type

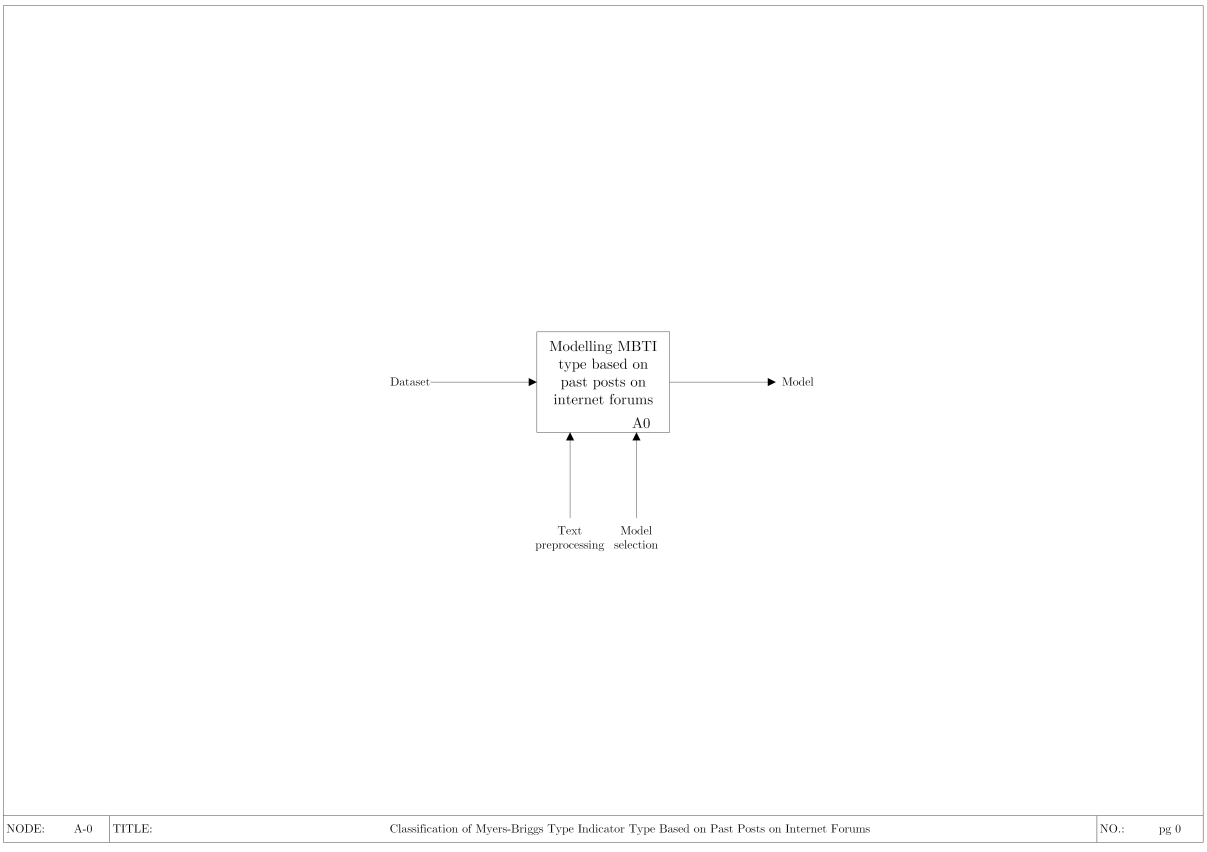

Figure 9: Macro view of the modelling process

fed into the model selection process where several candidates from machine learning and neural network models are chosen and tested based on their accuracies. Again, there are constraints based on the resources had and the complexity of the model also affects the time taken in training, knowing the relatively limited timespan had in going through the various models. After the

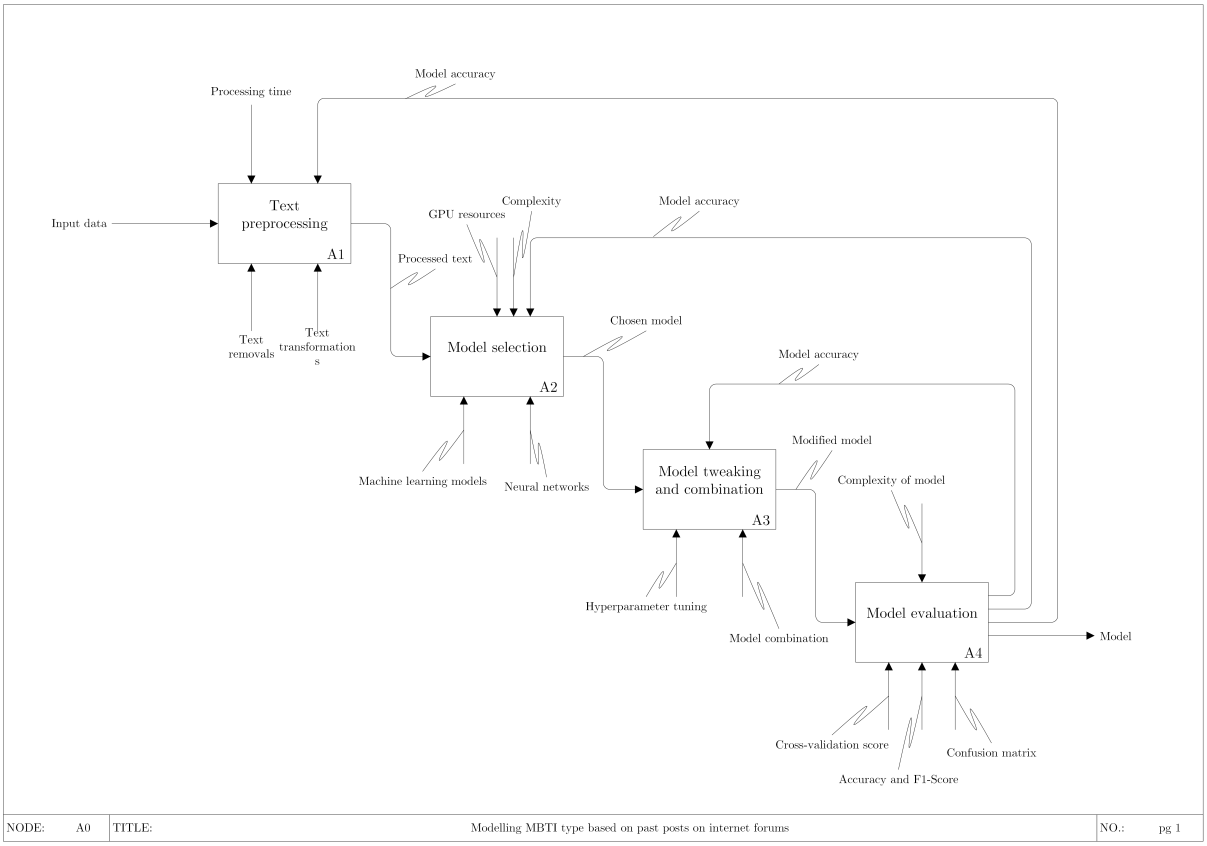

Figure 10: More in-depth view of the modelling process

top few models are chosen, they will be tweaked and even be combined together, and the ones with the highest accuracy will be chosen as the final model. The evaluation process will depend on the complexity of the model, potentially including not only accuracy and confusion matrices, but also cross-validation scores. The ROC/AUC curve was also considered, but knowing the unevenness of the MBTI types in the dataset, it was decided that it will not be as reliable as the F1-score.[17]

4 Text preprocessing

4.1 Cleaning

As with any kind of text corpus, the cleaning thereof is required before converting them into vector representations. The main considerations in the model will be as follows:

1. Changing all the text into lower case. This is in order to prevent fragmentations of identical words in different cases.

2. Removal of URL links, mentions, and hashtags. Those sets of words usually do not add to the meaning of the sentence. Sometimes they are accompanied by symbols, which makes the overall message more indistinct.

3. Expansion of contractions. This is to preserve more meaning and to prevent further fragmentations of words with identical meaning in different forms.

17. "Machine Learning - F1 Score vs ROC AUC," Stack Overflow, accessed December 5, 2021, https://stackoverflow.com/questions/44172162/f1-score-vs-roc-auc.

4. Removal of accented characters. This is, again, to prevent fragmentations (e.g. Naïve vs Naive).

5. Removal of non-alphabetic characters. The assumption is that numeric and punctuation characters do not add much meaning to the corpus.

6. Removal of English stop words. This is to reduce the noise associated with common English words that do not specifically point to a meaning.

7. Lemmatising the words. This is to reduce fragmentations even further by combining words of different nature into the same lemma.

8. Removal of single characters and extra whitespace. This is to reduce the information noise in the corpus.

This will form the baseline for the preprocessing framework. There will be modifications to this during the model tweaking and combination stage. The decision to keep the MBTI type words in the corpus is done as it keeps the context of the dataset intact—that it is the natural conversation topic in a forum centred around the self and personality types.

4.2 Word embeddings

There are three different choices that were compiled: GloVe, FastText, and Paragram.[18]EnglishWordVectorsJoh
These three will be tested during the model tweaking and combination stage. The baseline word embedding chosen will be FastText due to its advantage of being able to work well with rare and unrepresented words.[19] Moreover, its vector file is the smallest of the three.

5 Model selection

5.1 Classical machine learning models

5.1.1 Support Vector Machine

The model was implemented via GridSearchCV with 3 cross validation folds to facilitate hyperparameter tuning to enhance model performance. Support vector machine was used as it is a powerful classifier which divides data points according to a decision boundary. The highest accuracy score that SVM managed to generate was 0.63.

5.1.2 Random Classifier

The model was implemented via GridSearchCV with 3 cross validation folds to facilitate hyperparameter tuning to enhance model performance. The dataset is fed into the model and is divided into smaller components and are constructed into decision trees. Afterwards, they are made to predict and the tree with the most number of votes will have its prediction result set as the final random forest's prediction result. The accuracy result obtained is 0.38.

5.1.3 Logistic Regression

The model was implemented via GridSearchCV with 3 cross validation folds to facilitate hyperparameter tuning to enhance model performance. The model follows a sigmoid function whereby if the output is more than 0.5 it will result in a y-value of 1 and otherwise 0. The accuracy result recorded is 0.61.

18. "GloVe: Global Vectors for Word Representation," accessed December 5, 2021, https://nlp.stanford.edu/projects/glove/.

19. "GloVe and fastText — Two Popular Word Vector Models in NLP - DZone AI," dzone.com, accessed December 5, 2021, https://dzone.com/articles/glove-and-fasttext-two-popular-word-vector-models.

5.1.4 Decision Tree

The most suitable Attribute Selection Measures will be selected to divide the data and it will be further sectioned off to build trees until the condition of a tree's child node is met. The accuracy for this model is 0.21.

5.1.5 XGBoost Classifier

Unlike traditional ML models this classifier was faster than its counterparts but however in terms of performance it was poorer. Even though it was regarded to be superior. It works via an optimised form of gradient boosting algorithms to enhance the overall performance. The accuracy for this model is 0.15.

5.1.6 CatBoost Classifier

Similar to XGBoost, CatBoost is another gradient boosting algorithm which is designed to lessen overfitting and maintaining a loss rate across training, validation and testing. It is very robust to the point that XGBoost adapts some of its features. The accuracy result of the model is 0.67. Hence, it is very evident that unlike other classical and gradient boosting models the CatBoost algorithm is far superior and high-performing.

5.1.7 Gaussian Naive Bayes

The accuracy obtained from this model is 0.38.

5.1.8 Multinomial Naive Bayes

This model specifically caters to discrete variable classification, which would mean entities like word counts. As a result, the usage of token count matrix is apt for this model due to it being compatible with integers rather than fractional values. Hence, the accuracy obtained is relatively higher than the most of the models with a value of 0.58.

5.1.9 Complement Naive Bayes

This model serves to improvise on the shortcomings of the previous model by eliminating prior assumptions and aiding dataset which are very disproportionate across the classes. The accuracy obtained is 0.63.

5.1.10 Categorical Naive Bayes

This models performs well for discrete features which are categorical distributed, similar to Multinomial Naive Bayes. The accuracy scored for this model was relatively higher with a value of 0.63 again.

5.2 Neural Networks

Both the neural network models below are used with the Adam optimiser with the Cross Entropy Loss function, and the MBTI types are encoded into integers using the LabelEncoder function from Scikit-learn.

5.2.1 Long-Short Term Memory (LSTM)

By building a model with an LSTM single direction layer with hidden dimension size of 256 and a SimpleAttention layer, the accuracy of about 0.64 is attained in the validation set. To implement self attention, there are several dimensional changes have been done. Apart from that, LogSoftmax was used in the end of the forward function to convert a vector of numbers into a vector of probabilities.

5.2.2 Convolutional Neural Network (CNN)

At first, the two-dimentional CNN layer with five filters did not work as well, giving an accuracy of around 0.67. However, when the text preprocessing was modified to exclude the removal of single characters, lemmatisation, and the removal of stop words, the accuracy jumped to around 0.68. This is a first indication that these processes can be detrimental to the model's performance.

5.3 Conclusion

While CatBoost seemed to be very promising, the ultimate choice here is to attempt to use a CNN model due to its greater flexibility. Also, LSTM will also be considered as an extension to the CNN model with the hopes of improving the accuracy. Therefore, the baseline model used to tweak and combine will be the CNN model that excluded the removal of single characters, lemmatisation, and the removal of stop words.

6 Model tweaking and combination

There were a few comparisons made in order to find out ways to improve the model further.

6.1 Text preprocessing

As outlined earlier, the accuracy seemed to be better when the single characters and stop words are not removed and the words are not lemmatised. However, there is also the question of which word embedding model to use. When using GloVe, the accuracy reached 0.70, while using Paragram yielded 0.71. Therefore, the usage of Paragram seems to be better.

6.2 Tweaks and combinations

Using the baseline CNN model with FastText, the tweaks included the changing of dropout rates and the number of filters. However, this did not improve the accuracy much. Therefore, an LSTM layer is inserted after the convolutional layer in the hopes of improving the accuracy. With that, the accuracy went up to around 0.78. However, even though the Paragram word embedding seemed better in the CNN-only model, when using it with this new model, the accuracy was worse at around 0.71.

6.3 Conclusion

Therefore, the best model in this case is the CNN-LSTM model with the FastText word embedding. The preprocessing of the text does not include the removal of single characters, lemmatisation, and the removal of stop words, which improves the time taken to both model and predict significantly.

```
class CNN_LSTM(nn.Module):

    def __init__(self):
        super(CNN_Text, self).__init__()
        filter_sizes = [1,2,3,4,5]
        num_filters = 100
        n_classes = len(le.classes_)
        self.embedding = nn.Embedding(max_features, embed_size)
        self.embedding.weight = nn.Parameter(torch.tensor(embedding_matrix, dtype=torch.float32))
        self.embedding.weight.requires_grad = False
        self.convs1 = nn.ModuleList([nn.Conv2d(1, num_filters, (K, embed_size)) for K in filter_sizes])
        self.dropout = nn.Dropout(0.1)
        self.lstm = nn.LSTM(len(filter_sizes)*num_filters, 256, bidirectional=True, batch_first=True)
        self.fc1 = nn.Linear(256*2, n_classes)


    def forward(self, x):
        x = self.embedding(x)
        x = x.unsqueeze(1)
        x = [F.relu(conv(x)).squeeze(3) for conv in self.convs1]
        x = [F.max_pool1d(i, i.size(2)).squeeze(2) for i in x]
        x = torch.cat(x, 1)
        x = self.dropout(x)
        x = x.view(x.size(0), 1, x.size(1))
        x,_ = self.lstm(x)
        x = x.view(x.size(0), x.size(2))
        logit = self.fc1(x)
        # print(logit.size())
        return torch.log_softmax(logit, 1)
```

Figure 11: Population distribution of the MBTI types

7 Model evaluation

Due to the model chosen being a neural network, it was decided that a cross-validation technique will be impractical. As such, the evaluation will be in the form of a confusion matrix.

From this confusion matrix, it can be concluded that the model works relatively well for the majority data, but falls short when considering the minority data. Some of the confusions shown are from MBTI types that differ only by one letter, particularly within the J/P dichotomy.[20] Other works have shown that this particular dichotomy is quite challenging to predict.

8 Conclusion

In conclusion, it seems like it will be very difficult to push the accuracy beyond 0.80. This makes sense when considering the nature and weakness of the MBTI. As this is the aggregation of the past fifty posts by users in a forum, it stands to reason that, knowing that people's MBTI types tend to change over a period of time, the MBTI of a user can change slightly over the course of the fifty posts, affecting the aggregated data. Moreover, the inability for the model to predict the minority data well can be attributed to the lack of data. As such, an improvement to this model will include this in mind.

Another future improvement that can be considered is in exploring the usage of CatBoost a bit further. This is as it has shown a remarkable accuracy. Also, the usage of transfer learning on transformer models such as BERT, ULMFiT, and the likes, can allow for an increase in the accuracy.

In the end, the model is unlikely to have the potential to be generalised to a more public dataset. This is due to the conversation that revolves around the forum being more focused and

20. Choong and Varathan, "Predicting Judging-Perceiving of Myers-Briggs Type Indicator (MBTI) in Online Social Forum."

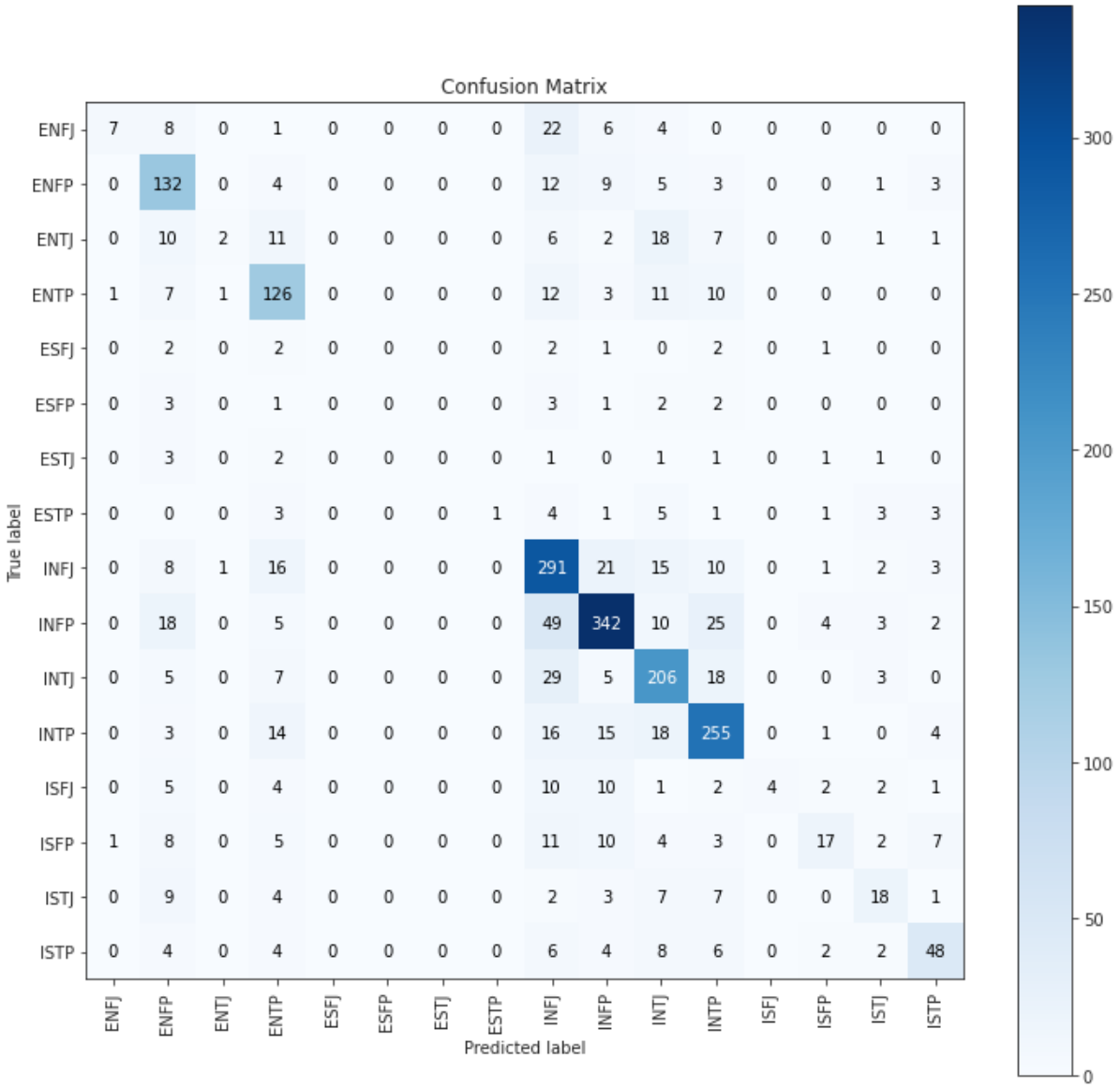

Figure 12: Population distribution of the MBTI types

niche, thus potentially creating bias in the model when exposed to datasets from other sources.[21] The issue of uneven numbers in the distribution of the MBTI types is also one challenge in need of solving. Thus, a larger, more even labelled dataset is needed to improve the model.

21. Choong and Varathan, "Predicting Judging-Perceiving of Myers-Briggs Type Indicator (MBTI) in Online Social Forum."