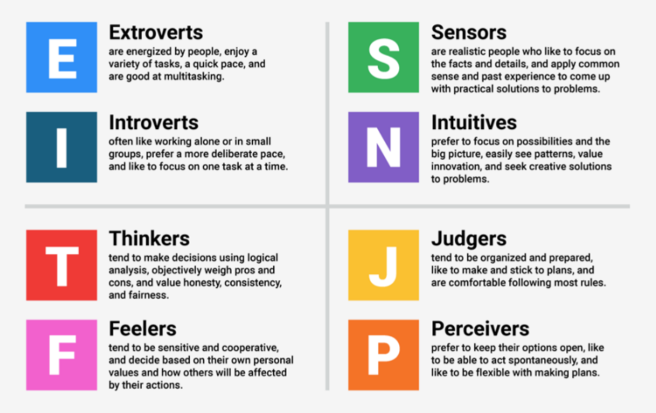
MBTI Text Classification

The Myers-Briggs Type Indicator (MBTI) is a well-known and widely used personality inventory based on the psychological theories of Carl Gustav Jung. It is often used as a profiling tool for discovering and understanding our personalities and behaviors.

It should be noted that for the MBTI there are no right or wrong answers as it is based subjectively on the perception each person has about themselves. Nevertheless, it has been a source of personal and professional growth for many people since it can shine a light into hidden aspects of our self.

This tool provides insights into our preference for each of four dichotomies:

- *Extraversion (E) vs. Introversion (I)*
- *Sensing (S) vs. Intuition (N)*
- *Thinking (T) vs. Feeling. (F)*
- *Judging (J) vs. Perceiving (P).*



E **Extroverts** are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.

I **Introverts** often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.

S **Sensors** are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.

N **Intuitives** prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.

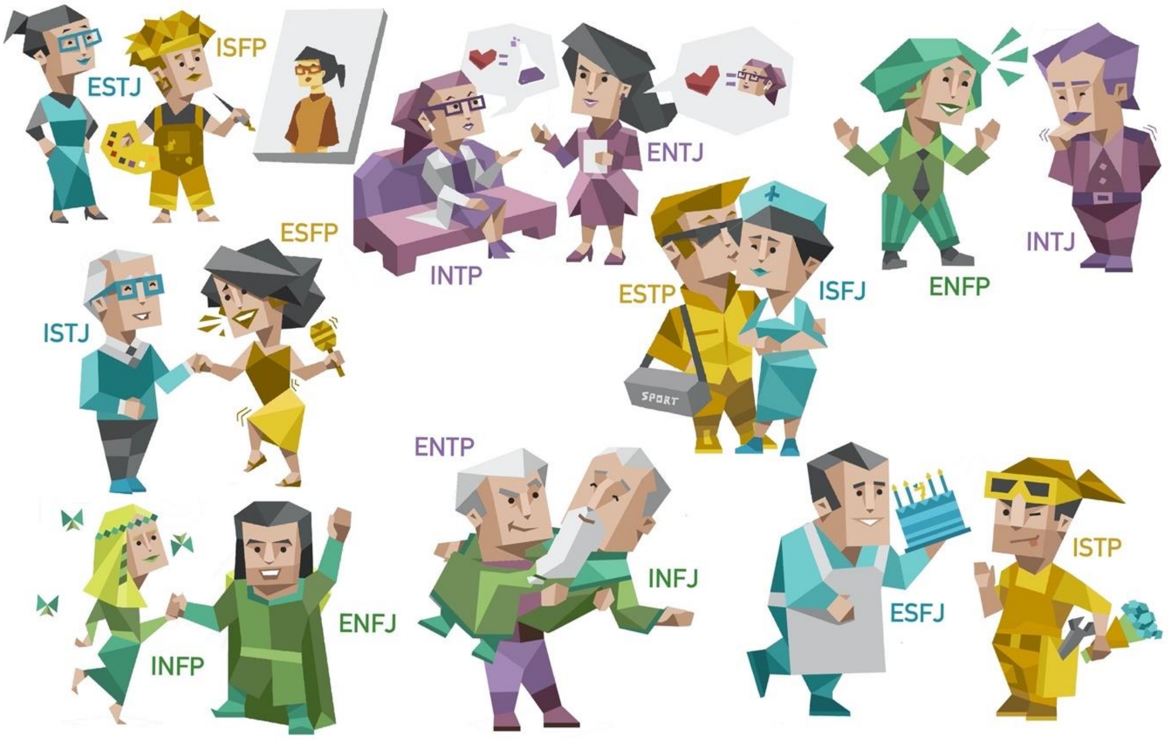
T **Thinkers** tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.

F **Feelers** tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

J **Judgers** tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.

P **Perceivers** prefer to keep their options open, like to be able to act spontaneously, and like to be flexible with making plans.

These preferences will result in 16 possible combinations of 4 different "letters" that will determine a person's MBTI type.



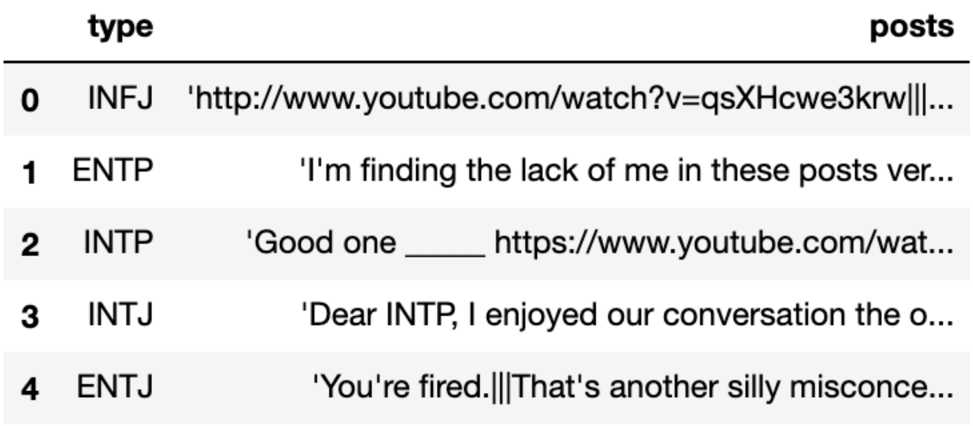
1. Problem Identification

Is it possible to determine a person’s MBTI personality type through how they express themselves in written communication?

Complete Problem Identification Report

2. Data

With over 8600 entries, this Kaggle dataset contains two columns, a **predictive column** with a compilation of many posts made by a certain individual in the Personality Cafe forum and a **target column** with their MBTI type. Here is an example of five individuals:



	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...

3. Methodology

I followed a method that consists of 7 steps.

- **Problem Identification:** this step involves identifying the correct problem to solve. Mentioned above.
- **Data Wrangling:** data collection, organization, and definition.
- **Exploratory Data Analysis:** creating plots and charts to understand the relationship between data.
- **Pre-processing and Training Data Development**: standardizing our data for future modeling.
- **Model Creation:** selecting, training and deploying a model to make predictive insights.
- **Prediction:** using our model to predict unseen data.
- **Presenting results:** creating a final report and presenting the results.

4. Data Wrangling

Data Wrangling Notebook

Luckily, this step was very simple for this project. The whole dataset had 0 null values and for our modeling section I did not have to make any adjustments to our predictive column. Nonetheless, it was a good opportunity to create new features from our predictive column (the posts). With little effort I was able to extract information like:

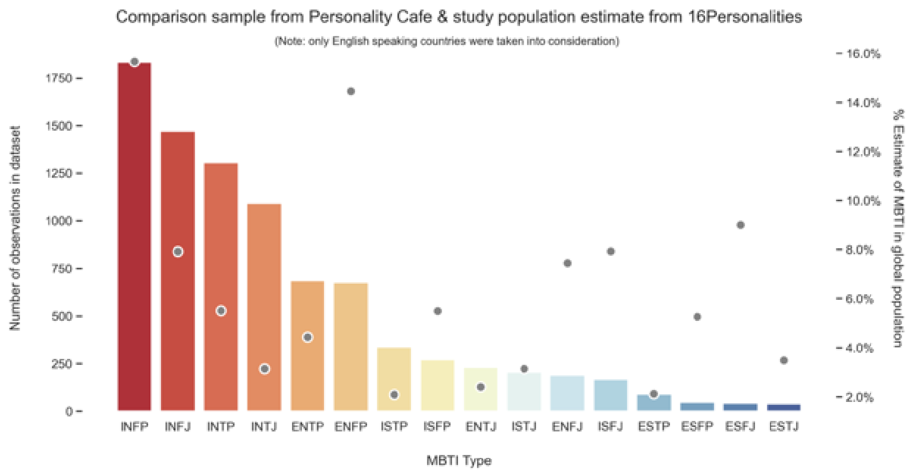
- The individual dichotomies (IE, NS, TF, JP)
- The Keirsey Temperaments (NF, NT, SP, SJ)
- The average length of the words used by each individual
- The number of emoticons they use
- The number of times they reference other MBTI types
- The use of positive and negative words
- The use of first-person pronouns vs. second and third person pronouns

5. EDA

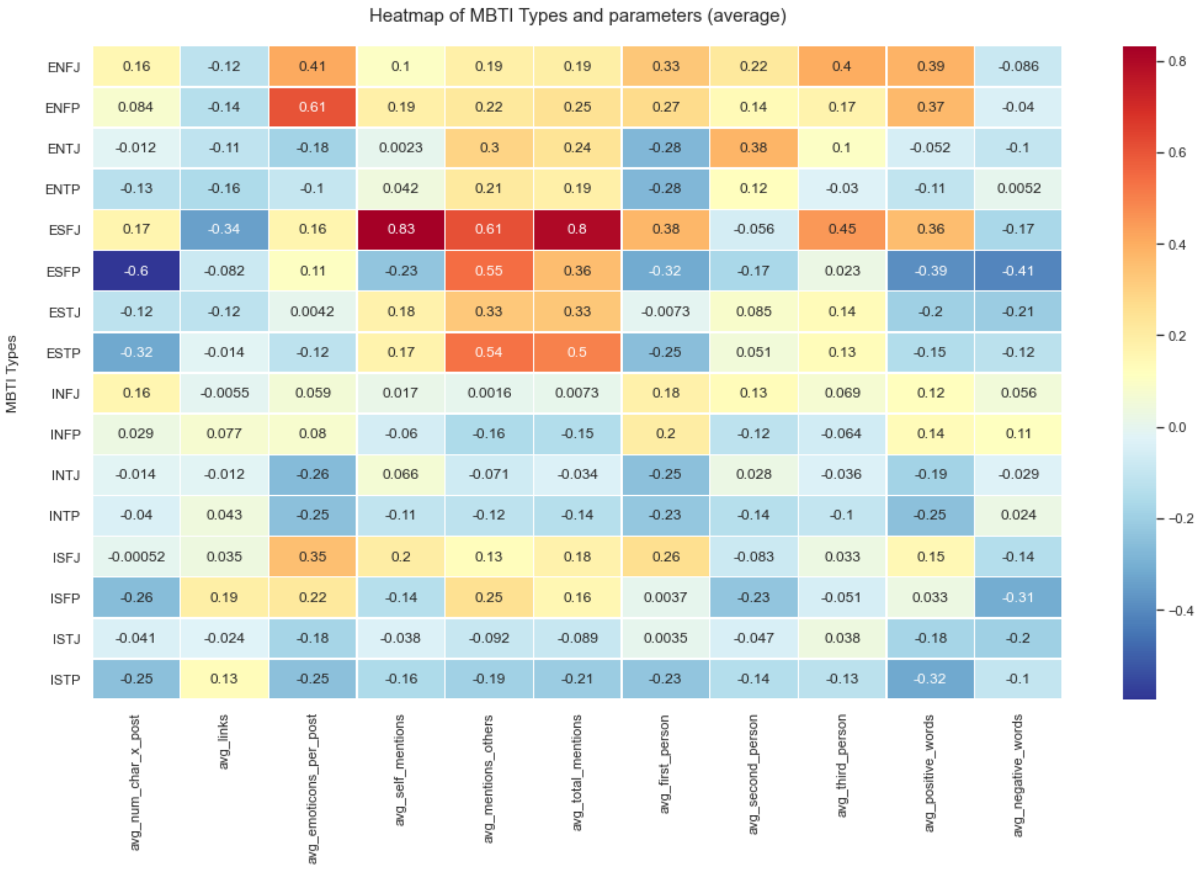
EDA Notebook

When looking at our dataset, the first relevant piece of information I come across is that there is a strong imbalance towards “intuition” (N) and “introvert” (I) types. While the “sensing” (S) and “extraverts” (E) are a bit more underrepresented. This makes sense since the former types are abundant in the Personality Café forum.

Additionally, to determine if our sample was representative of the whole population, I scrapped the website 16personalities to get a glimpse of how many people fall into each MBTI type (the % are from English speaking countries). Web scraper.



When analyzing the 16 personality types separately, I found interesting differences in some of the features I created in the data wrangling step. A first glimpse into these differences can be seen in the following heatmap.



I carried out the appropriate statistical analyses for each of the features to see if there was a significant difference between each of the 16 MBTI types. Since our process included comparing the means between more than two groups, I used the Kruskal-Wallis Test (the non-parametric equivalent of the ANOVA). Furthermore, since I was doing multiple testing, I could easily end up observing at least one significant result due to random chance. If I wanted to keep our desired significance level, I had to do some type of post-hoc test. We carried out a Dunn's test adjusting the p-values through the Bonferroni method. Finally, it is important to note that statistical significance does not mean practical significance therefore, I also looked at the size of the effect to get more insights on the observed differences. There are several ways to calculate effect size, a very common one is Hedges' g. The formula for Hedges' g is:

$$\text{Hedges' } g = \frac{M_1 - M_2}{SD^*_{pooled}}$$

The magnitude of Hedges' g may be interpreted using Cohen's (1988) convention as small > 0.2, medium > 0.5, and large > 0.8. With all this in mind I found that:

- ESFPs show a statistically significant difference in the number of characters used in their posts. On average they use less characters than the other types. This difference has a medium effect size. On the other side of the spectrum, we find that INFJs and ENFJs use slightly more characters per post than other types.

- I did not find any relevant difference in the number of links used by the different MBTI types.

- In relation to the use of emoticons, ENFPs show a significant difference with a medium positive effect size (they use more). ENFJs, ISFJ's, ISFPs, also show a positive but small effect while ENTJs, INTPs, and ISTPs show a small negative effect. All this seems to indicate that "Thinking" types use less emoticons than "Feeling" types.

- If we look at how frequently a certain MBTI type mentions other MBTI types, we find that ESFJs talk about a lot more than other types. This result can be biased due to the small number of instances of ESFJs in our dataset. What seems to be a recurrent trend, however, is the fact that "Extraverts" tend to talk much more about other MBTI types than themselves in comparison to "Introverts".

- In terms of the use of 1st, 2nd, and 3rd person pronouns, we only find small effect size differences amongst MBTI types. Our data shows that in general "Thinking" types use less 1st person pronouns than "Feeling" types. In terms of 2nd and 3rd person pronouns, the data does not show any relevant differences.

- Finally, in relation to the use of positive and negative words. We see that ENFJs, ESFJs, and ESFPs use more positive words on average and ESFPs, INTPs, and ISTPs use less.

Another way to analyze the data is to look into the 4 dichotomies directly instead of the 16 types separately. This way we end up having a bigger number of observations for each group. As we will see in the next section, this was actually what we did in the modeling phase.

6. Feature Engineering & Preprocessing

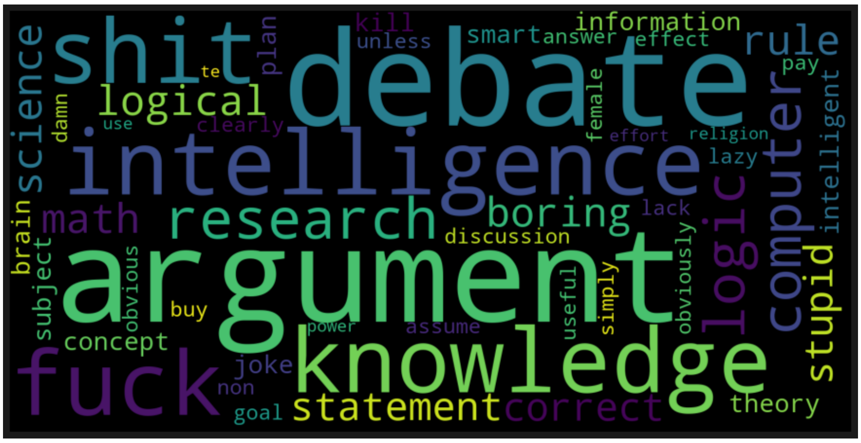
Feature Engineering and Preprocessing Notebook

For this section, I did three things:

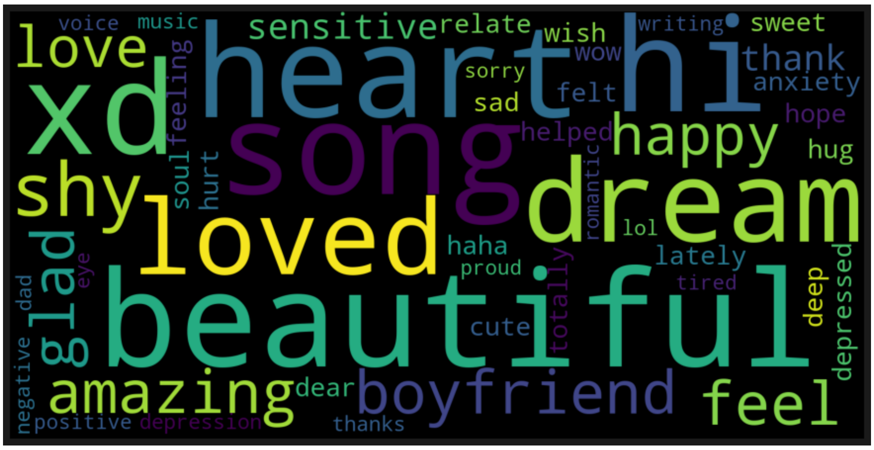
- I used the NLTK and TextStat libraries to further analyze the dataset.
- I used CountVectorizer and TF-IDF to convert the text data into numbers understandable by the computer
- I used Multinomial Naïve Bayes to see if I could find the most predictive words for each preference pair.

From these three parts, the most interesting one for the sake of this report is the third step. When I calculated the most predictive words for every dichotomy, I was very surprised to see how well the model categorized words for each group. I provide the examples of "Thinking" and "Feeling" which I believe are easier to understand by a lay audience but the other word clouds can be found in this section of the preprocessing notebook.

Thinking Word Cloud



Feeling Word Cloud



7. Algorithms & Machine Learning

Modeling Notebook

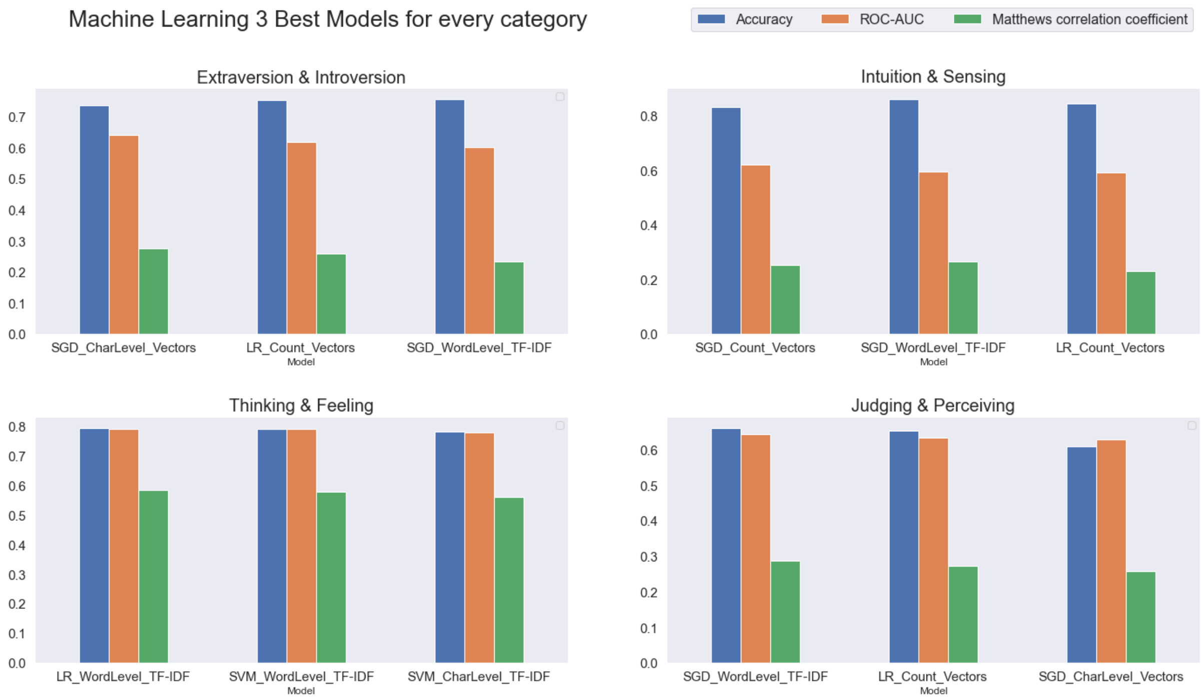
The first thing I did in the modeling part was generate 4 different datasets according to how words were transformed into vectors:

- CountVectorizer Word Level
- TF-IDF Word Level
- TF-IDF N-Grams Level (2 and 3 words)
- TF-IDF Character Level

Then I used these 4 datasets as inputs to generate several models. A summary of the different **models** I have tried follows:

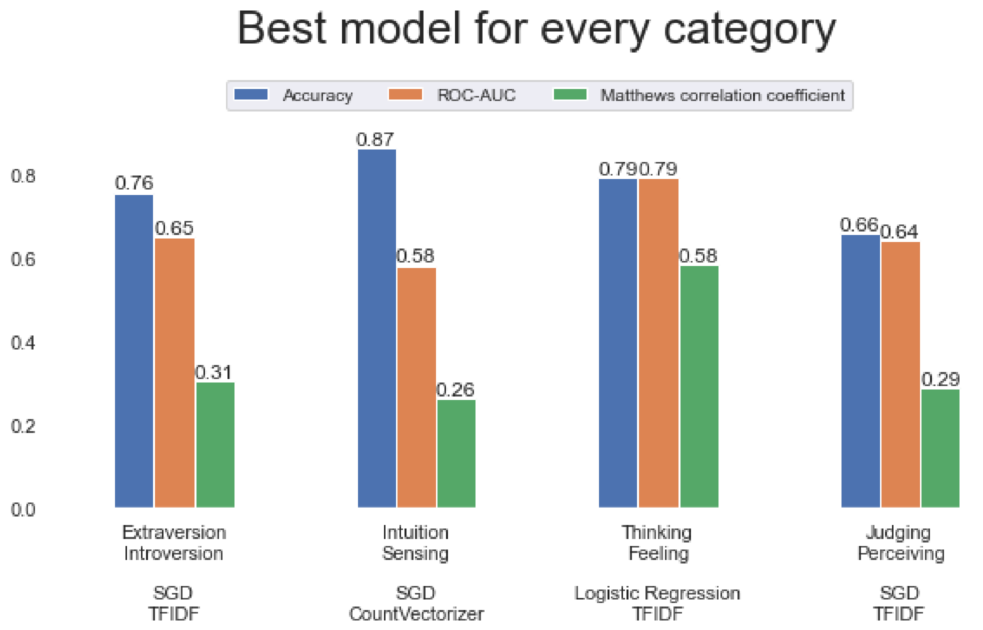
- **Models tried:** Naive-Bayes Classifier (NB), Logistic Regression (LR), Support Vector Machines (SVM),k-Nearest Neighbours (kNN), Stocastic Gradient Descent (SGD), Gradient Boosting (GB), XGBoost (XGB), Random Forest (RF)
- **Pending models:** Some models to be tried include Catboost, Adaboost, LightGBM, and deep learning models and transformers.

The top 3 models for each category were:



Metrics: *After looking around I saw several studies that used "accuracy" to measure the effectiveness of the machine learning models for text classification (some examples include this paper, and this paper). I added accuracy as a way to compare against this type of studies, however, according to Stephan Kolassa accuracy is not the best metric we can use since it is sensitive to class imbalance. This article by Boaz Schmueli says: "For binary classification, there is another (and arguably more elegant) solution: treat the true class and the predicted class as two (binary) variables, and compute their correlation coefficient (in a similar way to computing correlation coefficient between any two variables). The higher the correlation between true and predicted values, the better the prediction. Schmueli is referring to the Matthews Correlation Coefficient (MCC). Taking all of this into account, I also added ROC-AUC and Matthews Correlation Coefficient to analyze the results of the models.*

After doing a quick run through of the many different models, I took the best performing one and did used GridSearchCV to see if I could improve them a little bit by finding the best parameters. I was able to improve the E/I, N/S, and T/F models, although only a little bit. For the JP model I did not find a better model than the one I originally computed.



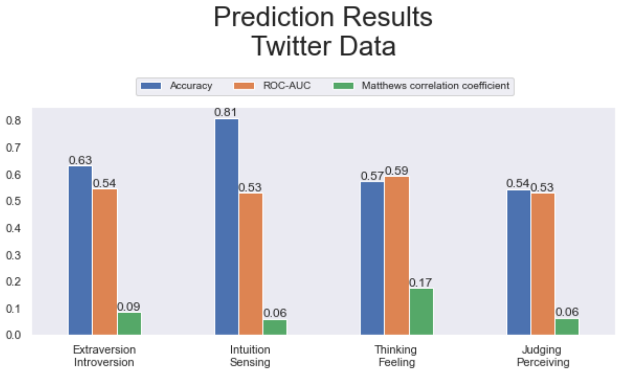
8. Predictions

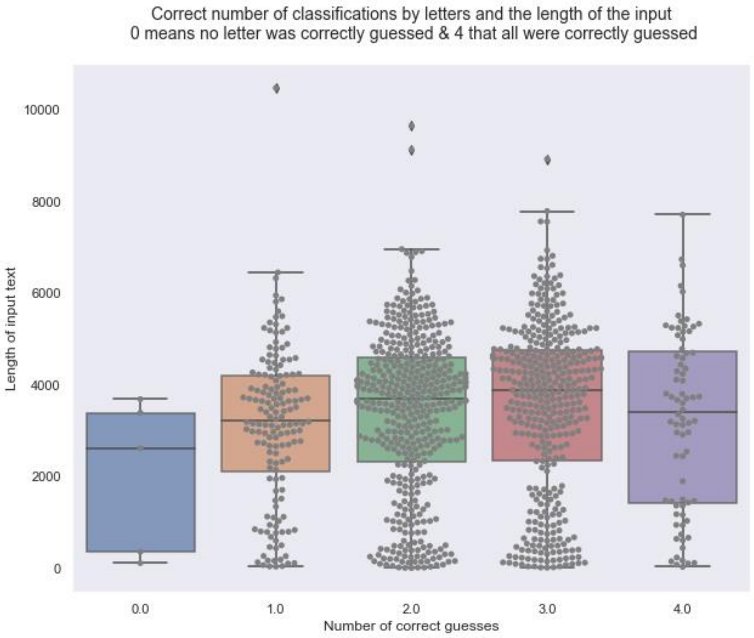
Predictions Notebook

Twitter Web Scraper Notebook

I ended up webscraping Twitter for users who include their MBTI types in their bios. When I found these individuals, I then proceeded to extract 100 of their tweets. I then used these tweets to predict their MBTI types to see if the machine learning models created in the previous steps can correctly classify people according to their tweets.

The results are a little bit underwhelming compared to our results in the training and test dataset. We need to take into account that these results are being computed with models trained on data from Personality Cafe and not from Twitter, this could be one of the main reasons this did not work out as well as we would have liked. Nevertheless, it is interesting to see the results and this can lead us to further improve our models (see section *"9. Future improvements and Next Steps"*)





9. Future Improvements & Next Steps

There are several improvements to this project:

- **Using other features**: to create the models I only used the text data from the original dataset. Throughout the project I extracted new features that I did not use for the modeling part. These could be added to our existing models to check if they improve.

- **Additional Data**: As aforementioned, there are some imbalances in two of the dichotomies (E/I and N/S). Gathering more data can help correct this imbalance. To access new data three of our options are:
  - Use a Reddit MBTI web scraper
  - Get tweets from people with their MBTI on their twitter bio
  - Go to Personality Cafe Forum to extract more data
  - Note: there are probably other places where we can get text data and a target column with the MBTI types.

- **Deep learning Models and Transformers:** In this project I did not use deep learning models and transformers. These have been found to be good at NLP and text classification in particular. Libraries like pytorch, tensorflow, or transformers can help us improve these models.

Regarding the **next steps**:

- A good next step is to convert the best models into a **web application** that people can use where they put their Twitter screen name and the application returns the probabilities for each of the dichotomies.

- **Cross-profiling tools comparison**: one big area of research in psychology and the study of personality is how different profiling tools relate to each other. A further analysis on how MBTI and other profiling tools like the Big 5 model or the Enneagram relate to MBTI could be done.

10. Credits

Thanks to Ben Bell for his amazing support and recommendations as a Springboard mentor and to Inés Guix for all her recommendations on statistical analyses.