Milestone Report

Detection of Myers-Briggs Personality Type via Twitter posts

The work described in this report is contained in the mbti.ipynb notebook of the same repository.

The Myers-Briggs type indicator (MBTI) represents a type of psychological assessment that helps people to understand themselves better and determine where they stand across 4 axes: Introversion (I) – Extroversion (E), Intuition (I) – Sensing (S), Thinking (T) – Feeling (F), Judging (J) – Perceiving (P). The MBTI assessments usually consists of 93 multiple-choice questions. However, psychology research shows that personality traits can be highly correlated with linguistic behavior. Therefore, the aim of this project is to develop a web application that could assign MBTI using a client's 50 last Twitter posts based on a corpus of 1.2M English tweets annotated with Myers-Briggs personality types, gender, and user statistics.

The data set is obtained through a reference from Plank B., Hovy D., "Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week" (2015) which represents a novel corpus of 1.2M English tweets available at https://bitbucket.org/bplank/wassa2015.

Import packages, .all files, and inspect the data

After the raw data set was imported from the Bitbucket repository, the initial inspection of .all files established that there were multiple files with varying number of tweets available with the same number of rows (1,499). The highest number of tweets was obtained from "2000g.all" which contains the most number of tweets (up to 2000 tweets) for 1,499 twitter users. The .all file was converted to a data frame with 4 columns for the MBTI type, gender, number of tweets, and tweets for each user. It was also discovered that additional meta-data file was available for each user which contained follower count, number of statuses, favorites, list participation, and profile background in a hexadecimal form. This file was parsed and added to the existing data frame.

Organize and clean columns

The resulting data frame consists of 9 columns which are given the names with respect to the type of information contained as original .all files did not contain any identification for column names. The first four columns which represent MBTI type, gender, number of tweets, and tweets were formatted appropriately and did not require any additional cleaning. The other five columns which contained additional user information were not formatted and had unnecessary string information with the column name for each observation (for example, "profile_background_color=…"). Therefore, each observation for these columns was changed using del_col_name function. The resulting string columns were converted to integers for further numerical manipulations. The only column that was left in the string form was the profile_background_color which was manipulated to add “#” for further analysis of color using data visualization. Each column was checked for null values. It was shown that there were no null values present in the data set.

Initial data exploration and statistics of MBTI

It was found that the existing corpus of twitter data from 1,499 users is shifted towards introverts (64%) and females (63%) as was expected from the description of the data set. The most frequently mentioned MBTI type was INFJ (257 instances). The least popular type was represented by ESTP (15 instances). The resulting distribution of different MBTI types is shown in Figure 1 sorted by personality trait. It seems like the most rare types of personality were reported more frequently which is probably not surprising as people tend to share mostly unique information with their followers.

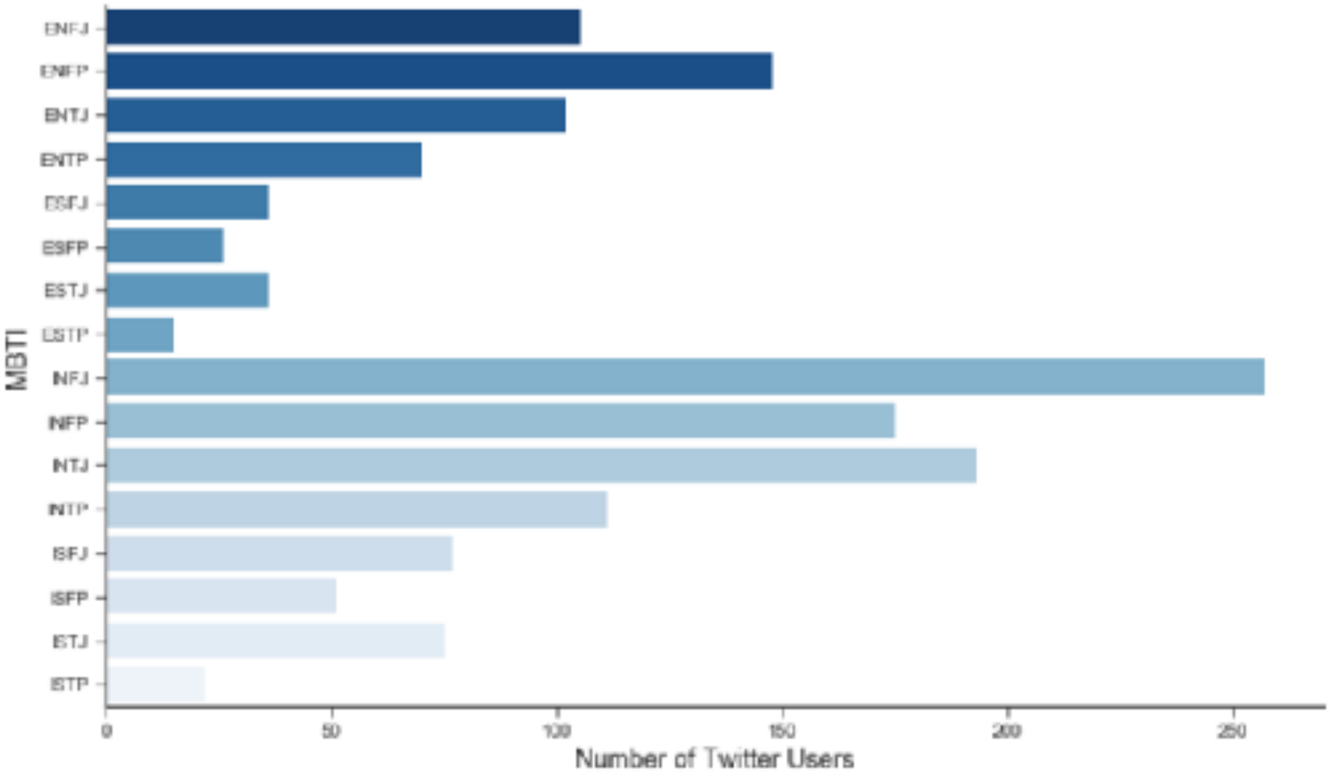

Figure 1: MBTI types by frequency.

The investigation of separate personality traits which include Introversion vs. Extroversion, Intuition vs. Sensing, Feeling vs. Thinking, and Judging vs. Prospecting showed that majority of the traits are skewed towards a certain type, but there are some gender-specific differences (see Table 1) . For example, introversion is present approximately at the same level for female and male participants. Males showed more propensity towards Intuitive (N) behavior than women (5% more prominent). One of the most obvious differences was in the distribution of Thinking vs. Feeling trait which was unequal for women as the majority of females are more feeling (64%) while the majority of males are more thinking (51%). Judging vs perceiving demonstrates approximately the same results for both males and females. However, it was shown that males are less judging than females (less by 4%).

Table 1: Distribution of personality traits for all users, females, and males.

Personality Trait	Total	Female	Male
Introversion-Extroversion	**64%** / 36%	**65%** / 35%	**61%** / 39%
Intuitive-Sensing	**77%** / 23%	**75%** / 25%	**80%** / 20%
Thinking-Feeling	42% / **58%**	36% / **64%**	**51%** / 49%
Judging-Perceiving	**58%** / 42%	**60%** / 40%	**56%** / 44%

The comparison of meta-data results for different gender shown in Figure 2 demonstrated that females have more followers, more favorites, and are more frequently added to the public lists on Twitter. They also prefer darker color schemes for their profile background as shown in Figure 3. It was determined that the average number of statuses published by the user was not influenced by gender.

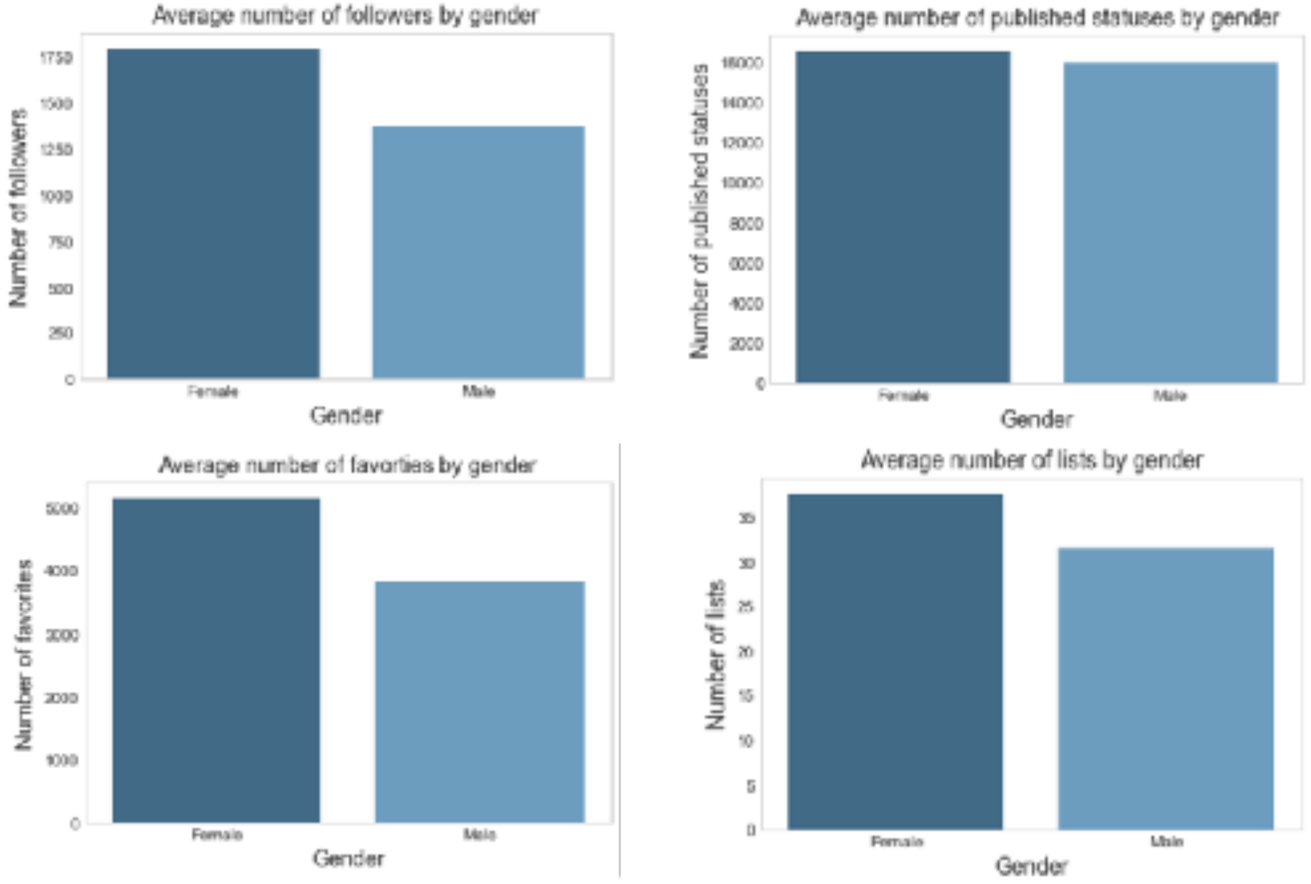

Figure 2: Meta-data differences by gender.

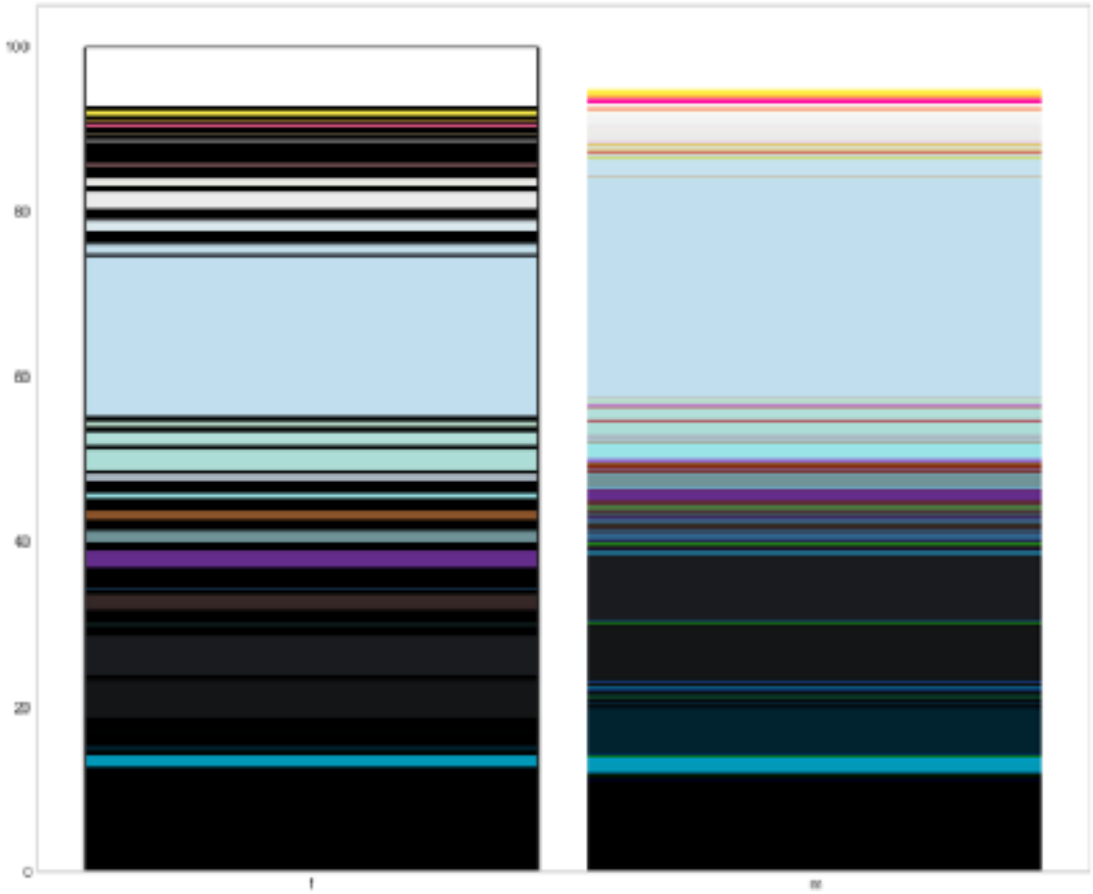

Figure 3: Normalized distribution of background color preferences by gender (f - female, m - male).

The distribution of meta-data characteristics by four personality traits demonstrates that different user characteristics indicate certain types of personality. It seems like more Feeling, Sensing, and Judging indicates more followers. It was also shown that more Perceiving, Feeling, and Sensing users publish more statuses and like more statuses of other people which indicates higher rate of participation on Twitter. Surprisingly, the number of lists which the user could be a member of is highly correlated with Extroverted, Sensing, Feeling, and Judging users. Additionally, a large following on Twitter might be indicative of Introverted personality.

The obtained results from the data visualization were tested using two-sample t-test for different groups of personality traits to compare whether the average difference between two groups is statistically significant or if it is due to a random chance. The results are shown in Table 2. We can conclude that in majority of the cases, the null hypothesis that indicated the absence of difference between two personality groups could not be rejected. However, it was determined that with 95% confidence, Introverted personality is correlated with high average number of statuses published. It was also established that more Sensing personalities indicate more statuses published by the user. Additionally, more Perceiving personality showed a higher rate of participation (more favorites) than Judging personality. These results can be used for further analysis and determination of the features that predict MBTI using machine learning.

Table 2: Results of the two-sample t-tests for meta-data.

Personality trait	Hypothesis	T-score	p-value < 0.05
Introversion-Extroversion	**Ho:** There is no difference in the mean number of followers for extroverted and introverted twitter users. **Ha:** There is a difference in the mean number of followers for extroverted and introverted twitter users.	0.25	FALSE
Intuitive-Sensing	**Ho:** There is no difference in the mean number of followers for intuitive and sensing twitter users. **Ha:** There is a difference in the mean number of followers for intuitive and sensing twitter users.	-0.98	FALSE
Thinking-Feeling	**Ho:** There is no difference in the mean number of followers for thinking and feeling twitter users. **Ha:** There is a difference in the mean number of followers for thinking and feeling twitter users.	-0.27	FALSE
Judging-Perceiving	**Ho:** There is no difference in the mean number of followers for judging and perceiving twitter users. **Ha:** There is a difference in the mean number of followers for judging and perceiving twitter users.	1.45	FALSE
Introversion-Extroversion	**Ho:** There is no difference in the mean number of statuses for extroverted and introverted twitter users. **Ha:** There is a difference in the mean number of statuses for extroverted and introverted twitter users.	1.92	TRUE
Intuitive-Sensing	**Ho:** There is no difference in the mean number of statuses for intuitive and sensing twitter users. **Ha:** There is a difference in the mean number of statuses for intuitive and sensing twitter users.	-2.09	TRUE
Thinking-Feeling	**Ho:** There is no difference in the mean number of statuses for thinking and feeling twitter users. **Ha:** There is a difference in the mean number of statuses for thinking and feeling twitter users.	-0.44	FALSE

Judging-Perceiving	**Ho:** There is no difference in the mean number of statuses for judging and perceiving twitter users. **Ha:** There is a difference in the mean number of statuses for judging and perceiving twitter users.	-1.62	FALSE
Introversion-Extroversion	**Ho:** There is no difference in the mean number of favorites for extroverted and introverted twitter users. **Ha:** There is a difference in the mean number of favorites for extroverted and introverted twitter users.	1.51	FALSE
Intuitive-Sensing	**Ho:** There is no difference in the mean number of favorites for intuitive and sensing twitter users. **Ha:** There is a difference in the mean number of favorites for intuitive and sensing twitter users.	-1.28	FALSE
Thinking-Feeling	**Ho:** There is no difference in the mean number of favorites for thinking and feeling twitter users. **Ha:** There is a difference in the mean number of favorites for thinking and feeling twitter users.	-1.68	FALSE
Judging-Perceiving	**Ho:** There is no difference in the mean number of favorites for judging and perceiving twitter users. **Ha:** There is a difference in the mean number of favorites for judging and perceiving twitter users.	-2.4	TRUE
Introversion-Extroversion	**Ho:** There is no difference in the mean number of lists for extroverted and introverted twitter users. **Ha:** There is a difference in the mean number of lists for extroverted and introverted twitter users.	-1	FALSE
Intuitive-Sensing	**Ho:** There is no difference in the mean number of lists for intuitive and sensing twitter users. **Ha:** There is a difference in the mean number of lists for intuitive and sensing twitter users.	-0.81	FALSE
Thinking-Feeling	**Ho:** There is no difference in the mean number of statuses for thinking and feeling twitter users. **Ha:** There is a difference in the mean number of lists for thinking and feeling twitter users.	-1.36	FALSE

Judging-Perceiving	**Ho:** There is no difference in the mean number of statuses for judging and perceiving twitter users. **Ha:** There is a difference in the mean number of lists for judging and perceiving twitter users.	0.95	FALSE

For analysis of the separate MBTI types, it was determined that ISFJ had the biggest number of followers and was the member of the most public lists. The results are shown in Figure 4.

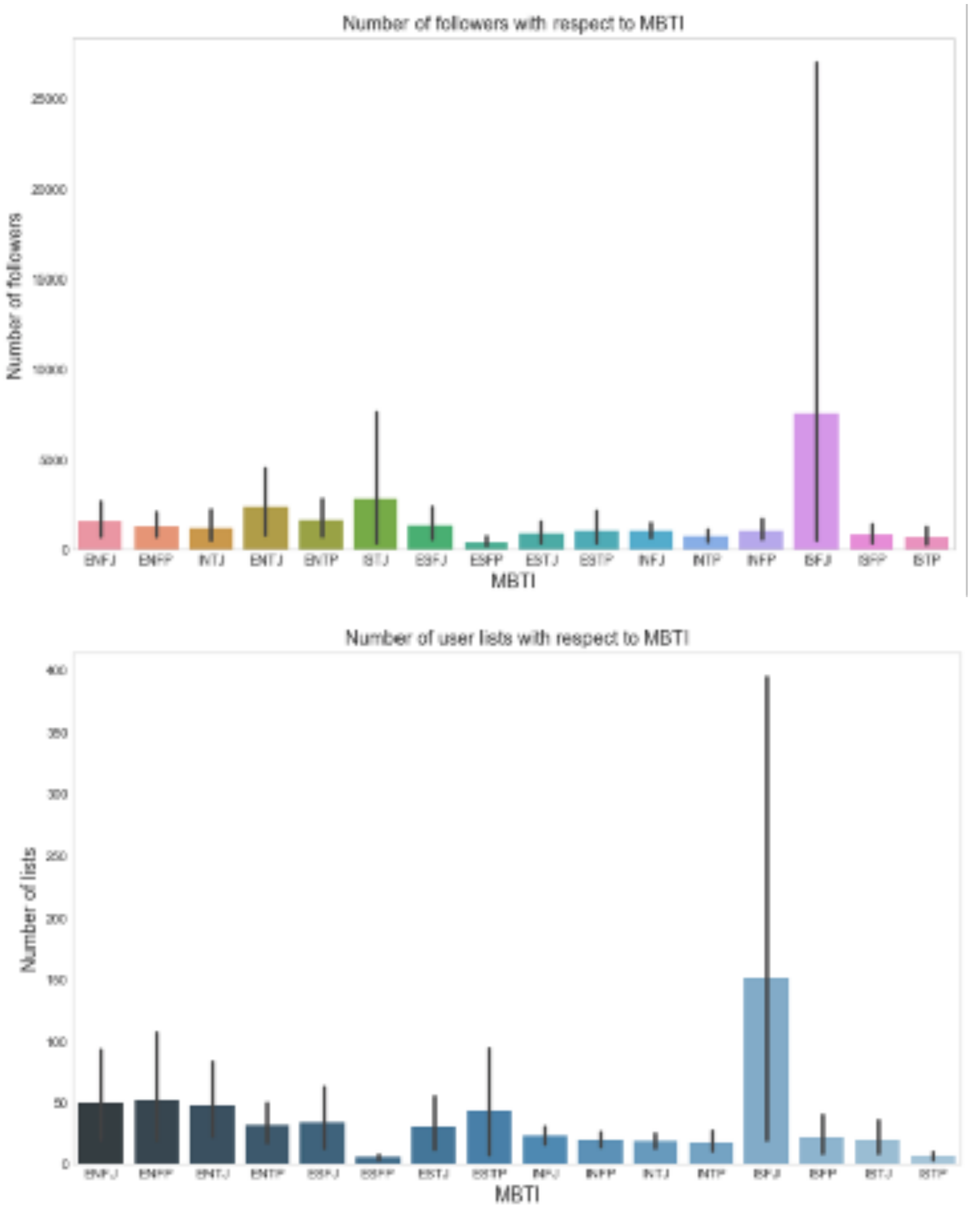

Figure 4: Analysis of follower and statuses counts with respect to MBTI.

Finally, the WordCloud visualization which demonstrates how frequently words appear in a given body of text by making the size of each word proportional to its frequency demonstrated a significant linguistic discrepancy between Introverted and Extroverted personalities as shown in Figure 5. Extroverts tend to use more adjective which provides a subjective description for a situation (“violent”, “sweet”) while introverts talk more about things or actions (“thinking”, “math”, “sleep”).

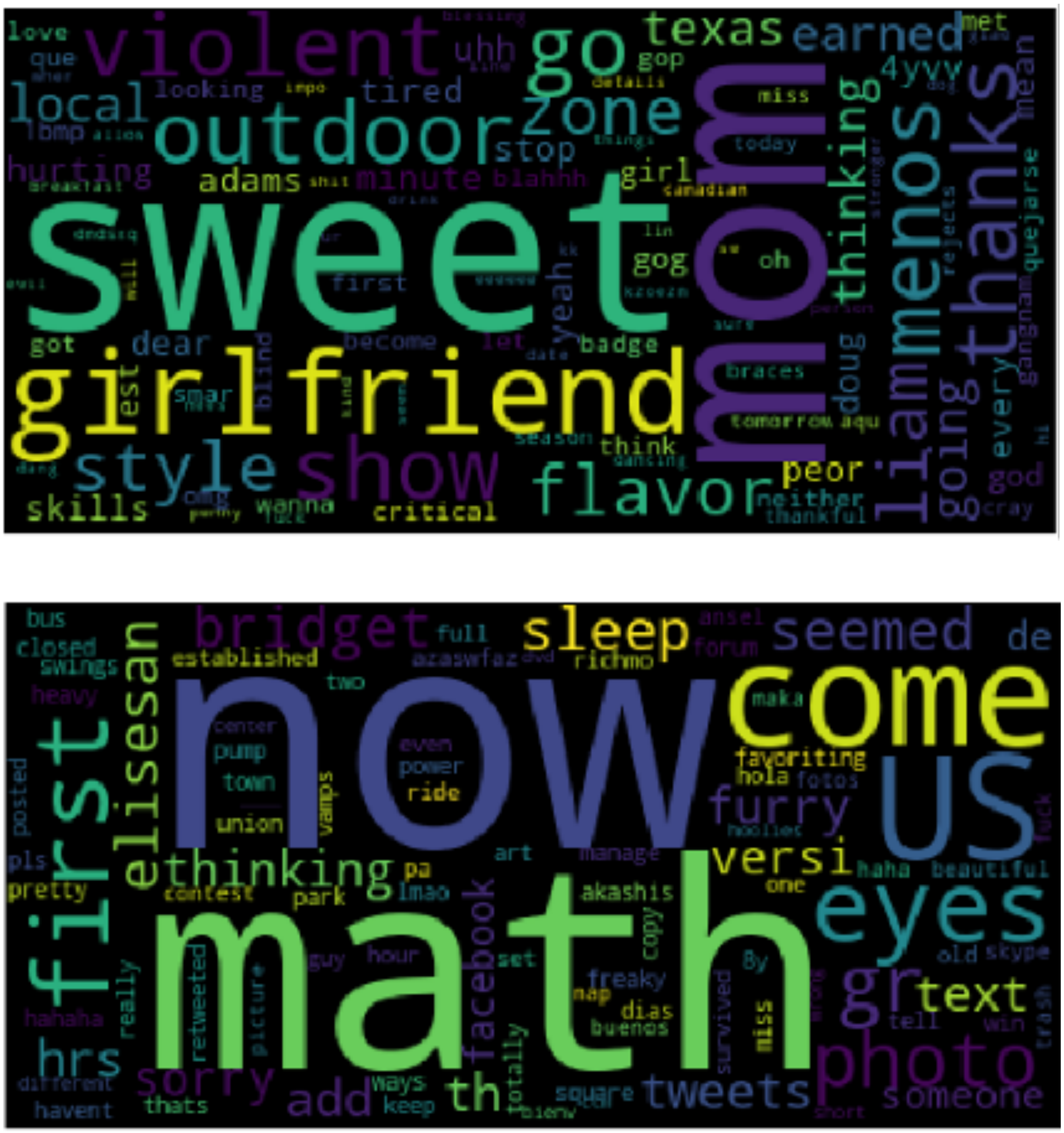

Figure 5: WordCloud visualization of tweets for Extroverts vs. Introverts.

Conclusion

The results obtained from the initial data exploration and statistical analysis of personality traits indicate that different meta-features are affected by certain personality traits. Additionally, gender was found to have an influence on the distribution of personality traits demonstrating that Thinking vs. Feeling can be potentially predicted using users' gender characteristics. However, it was also demonstrated that the majority of the meta-features are too complex to be predicted using the binary scale of the four personality traits through two-sample t-tests. Surprisingly, the profile background color was affected by gender which is indicated by female preference for a darker color scheme. This aspect of the analysis could be used in the future for predicting of MBTI as it was not explored by the creators of the corpus. Finally, it was determined that the linguistic signatures of different personality types have distinctive features including the frequency of varying words used. This indicates that there is more to explore in the future stages of NLP analysis.