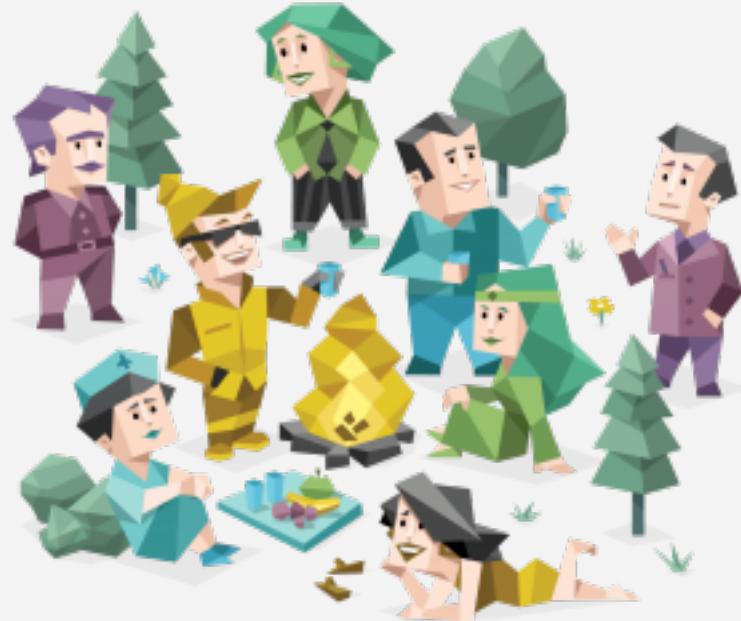


# MBTI TEXT CLASSIFICATION

Can we determine personality only based on written text?



© All images are ownership of 16personalities.com

# CONTENT

01

Goal

02

Data

03

Methodology

04

Data Wrangling

05

EDA and Feature Engineering

06

Algorithms and Machine Learning

07

Predictions & Conclusions

08

Improvements and next steps

# THE GOAL



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

# THE DATA

KAGGLE  
Dataset

8600+  
Observations

USER POSTS  
Predictive  
Column

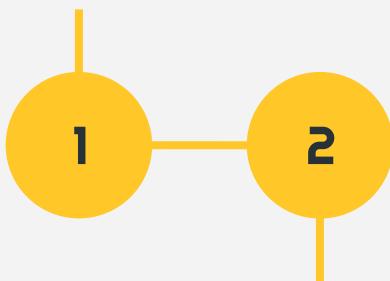
MBTI TYPES  
Target  
Column

## Example first rows

	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...

# METHODOLOGY

Identify the problem



Data Wrangling

Exploratory Data Analysis



Feature Engineering

Model Creation



Predictions

Presenting Results



# DATA WRANGLING

The dataset contained 0 null values and did not require any correction. But we could create new features from the existing data. Here are examples:

## Dichotomies & Keirsey

We extracted the individual dichotomies (I/E, N/S, T/F, J/P) and the Keirsey Temperaments (NT, NF, SP, SJ)

## Text Characteristics

Average length of the words used by the person, number and type of emoticons they use, use of 1<sup>st</sup> person pronouns vs. 2<sup>nd</sup> and 3<sup>rd</sup> person...

## Qualitative Measures

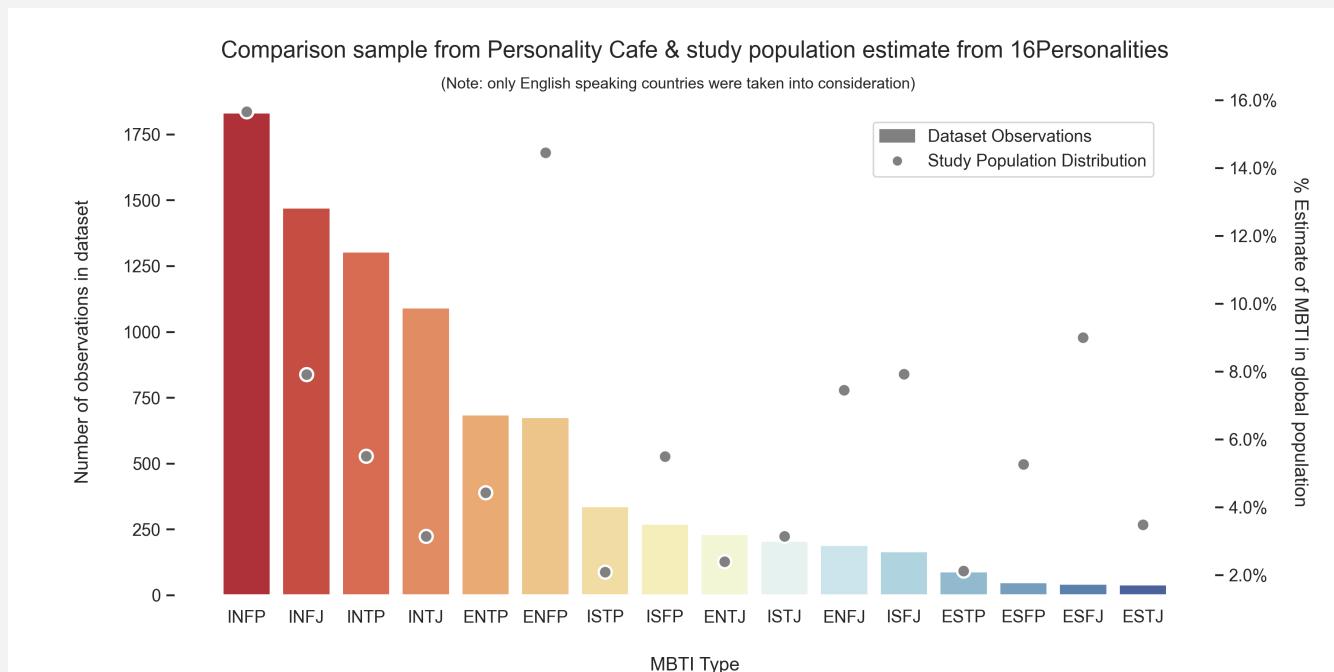
The use of positive and negative words, readability consensus (Flesch Reading Ease score, Coleman-Liau Index), ...

# Exploratory Data Analysis and Feature Engineering



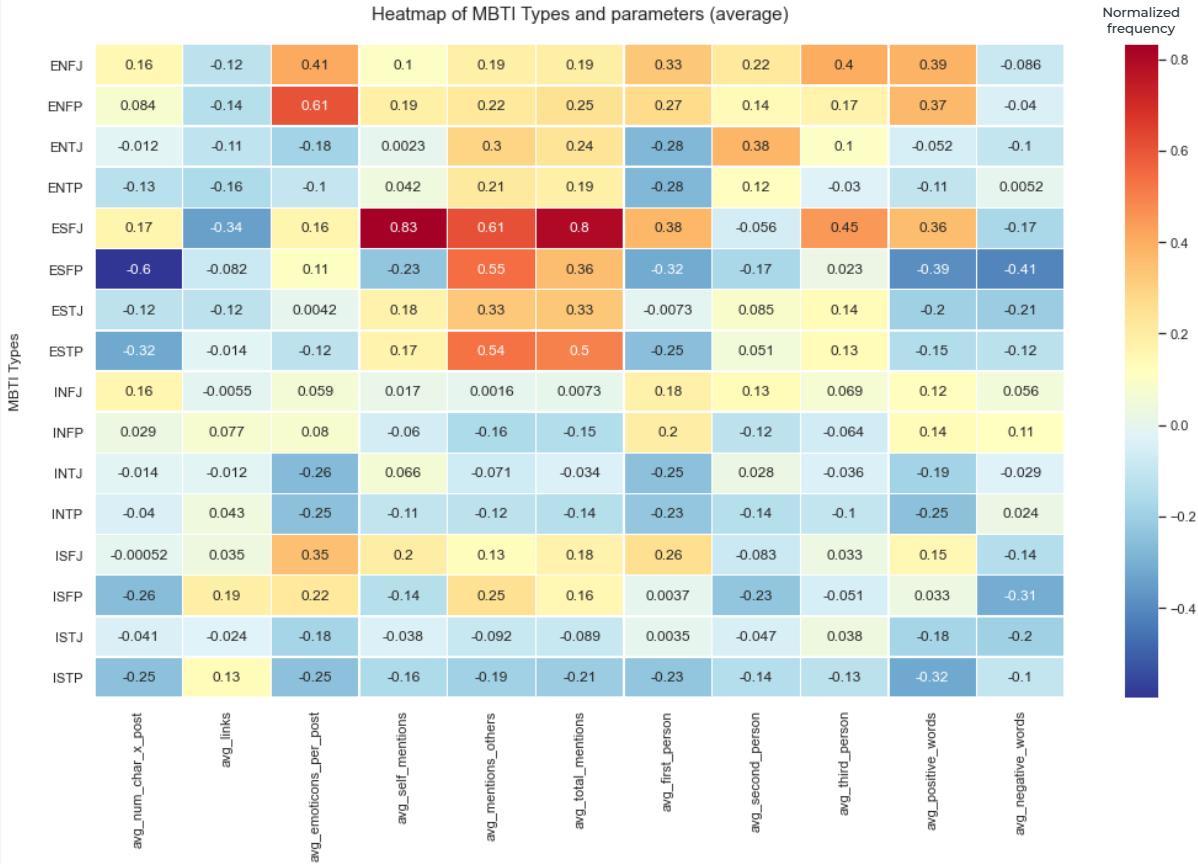
# IMBALANCED DATASET

Strong imbalance towards “intuition” (N) and “introvert” (I) types. While the “sensing” (S) and “extraverts” (E) are a bit more underrepresented. Not completely representative of study population.



# FEATURES and TYPES HEATMAP

We find some interesting differences amongst MBTI types when it comes to certain features.



# FINDINGS

## Characters x post

ESFP use less characters per post than other MBTI types while INFJs and ENFJs use slightly more characters

## Emoticons

Our results confirm what one might expect that “Thinking” types use less emoticons than “Feeling” types.

## Use of Pronouns

“Thinking” types use less 1st person pronouns than “Feeling” types. For 2nd and 3rd person pronouns there were no relevant differences

## Positive-Negative

ENFJs and ESFJs use on average more positive words and INTPs, and ISTPs use less. ESFPs seem to have strong preference for both

# FINDINGS

# CRITICAL WORDS BY DICHOTOMIES



# INTROVERSION

concept eat form dead reality anxiety written  
dislike wish game outside world science internet  
quiet water dream space public  
art afraid religion father soul cat  
listening rarely grade lost state  
beautiful society  
english suppose nature light  
computer born memory family mother  
particular particularly comfortable music human  
entire

# EXTRAVERSION

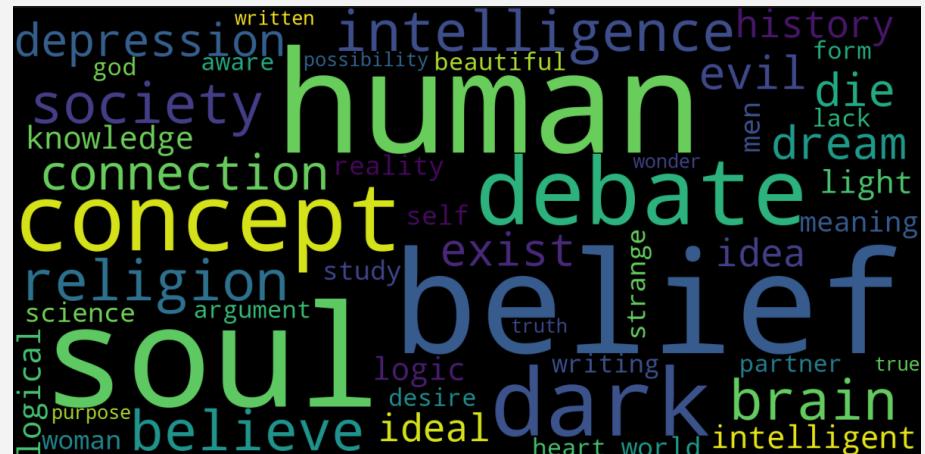
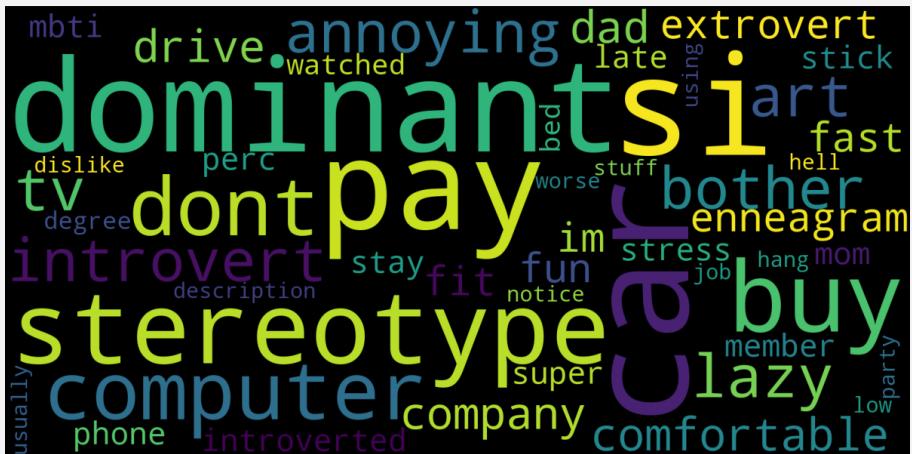
# FINDINGS

## CRITICAL WORDS BY DICHOTOMIES

SENSING



INTUITION



# FINDINGS

## CRITICAL WORDS BY DICHOTOMIES

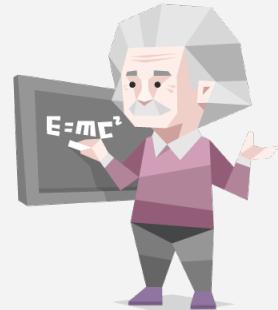
FEELING



A word cloud representing feelings, with words like love, heart, song, dream, beautiful, amazing, and glad being the most prominent. The words are colored in various shades of green, blue, and yellow.

love voice music sensitive relate wish writing sweet  
xd heart relate wish wow thank anxiety  
shy soul feeling hurt sorry sad felt hope  
glad dad eye hurt helped happy hug  
loved haha proud romantic lol lately tired  
beautiful cute totally depressed  
amazing dear boyfriend feel

THINKING



# FINDINGS

## CRITICAL WORDS BY DICHOTOMIES

JUDGING



A word cloud visualization showing critical words associated with the Judging dichotomy. The words are primarily in shades of purple, blue, and green. Key terms include: friendship, behavior, plan, dear, dominant, intuition, goal, rare, process, trust, enneagram, welcome, past, father, research, environment, clearly, advice, assume, tendency, desire, simply, sister, mistake, specific, action, degree, hi, enneagram, perspective, decision, cold, certainty, given, understanding, purpose, information, mother, male, personal, comment.

PERCEIVING



A word cloud visualization showing critical words associated with the Perceiving dichotomy. The words are primarily in shades of yellow, green, and blue. Key terms include: bored, basically, gotwriting, game, started, ideal, music, yeah, weird, pretty, suck, playing, hey, joke, bedkinda, crazy, stick, kid, class, hell, xd, god, cool, art, night, dont, cute, stuff, fun, write, realized, realized, awesome, song, literally, night, dont, cute, stuff, fun, write, depressed, super.

# Algorithms and Machine Learning



# DATA & ALGOS

## INPUT DATA

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

## ML ALGORITHMS

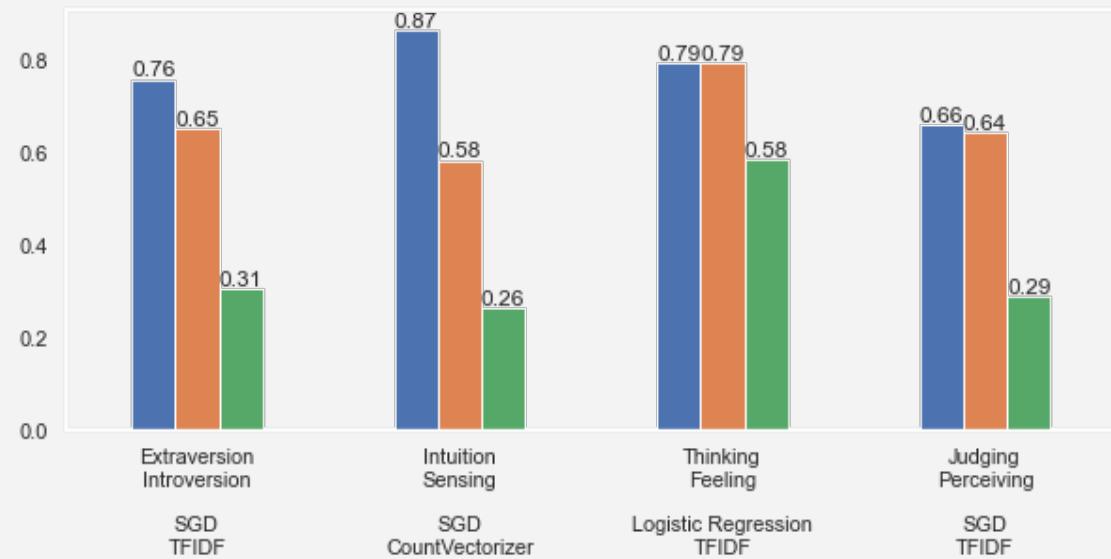
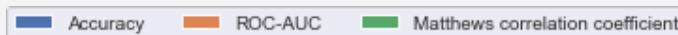
- Naive-Bayes Classifier (NB)
- Logistic Regression (LR)
- Support Vector Machines (SVM)
- k-Nearest Neighbours (kNN)
- Stochastic Gradient Descent (SGD)
- Gradient Boosting (GB)
- XGBoost (XGB)
- Random Forest (RF)

# BEST MODELS

The Stochastic Gradient Descent algorithm was the one that worked the best for three of the dichotomies (E/I, N/S, J/P) while Logistic Regression was the best model for T/F

The results are promising whilst not amazing. Specially good are the T/F followed by the E/I and J/P, while the accuracy of N/S was good, the ROC-AUC score shows there is room for improvement

Best model for every category



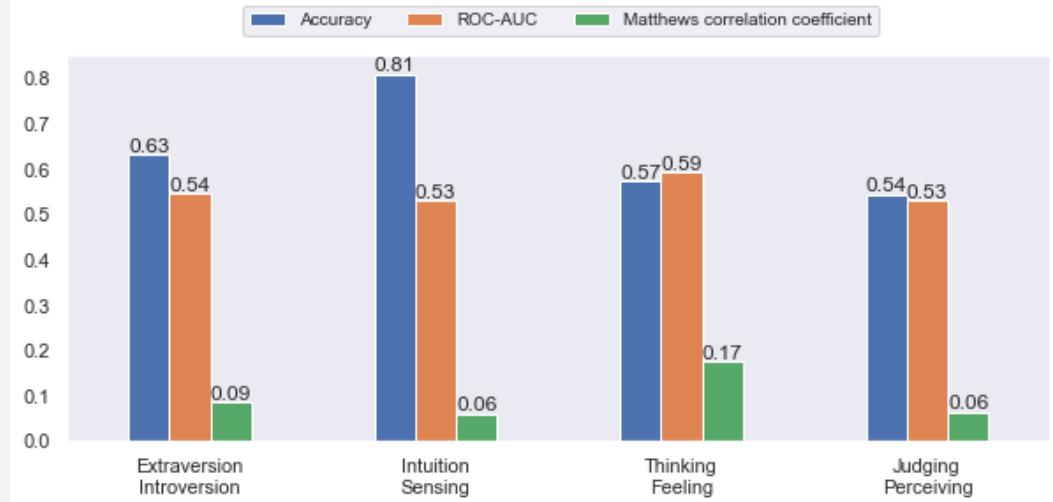
# PREDICTIONS

## TWITTER DATA

We tested our models on completely unseen data. We extracted 100 tweets from 1000 Twitter who had their MBTI type in their bios.

As seen in the graph on the right, the results are not very promising. This is possibly due to people using different ways to communicate in Personality Cafe (where train test data was extracted) and in Twitter (where this data was extracted from).

## Prediction Results Twitter Data



## RECAP

In the EDA we saw there are certain differences in the way different MBTI types express themselves in written language (emoticons, average word length, use of pronouns, etc.).

With a Multinomial NB classifier we were able to extract the words that classified each dichotomy better (word clouds), the results of this analysis and the key words were most surprising.

While our best models show some promising results in the training and test data, our predictions on Twitter data were not optimal. This could be due to the origin of both datasets and how people express themselves differently in each platform where they were extracted from



## CONCLUSION

Our results show that to some extent MBTI personality types can be determined through written classification.

Although there is still a long way to go to have good models that perfectly classify people into the 16 MBTI types, certain characteristics surface as statistically significantly different in how personality types express themselves in written communication.

# IMPROVEMENTS & NEXT STEPS

## IMPROVEMENTS

**Using other features:** currently we only used the original text data.

**Additional data:** scrape new data to train our models. Like using the Twitter data used for predictions or from Reddit's r/multi.

**Deep Learning and Transformers:** we did not use these types of models but they have offered good results for NLP problems

## NEXT STEPS

**Web application:** create a web application where people can put their Twitter screen-name or their Reddit user name and the app returns the MBTI type of the person

**Cross-profiling tools comparison:** do further analyses with other tools like the Big 5 and Enneagram and compare results with our MBTI dataset

# THANKS

Do you have any questions?

[journeydatas@gmail.com](mailto:journeydatas@gmail.com)

<https://github.com/DSJourney/>

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.



© All images are ownership of 16personalities.com