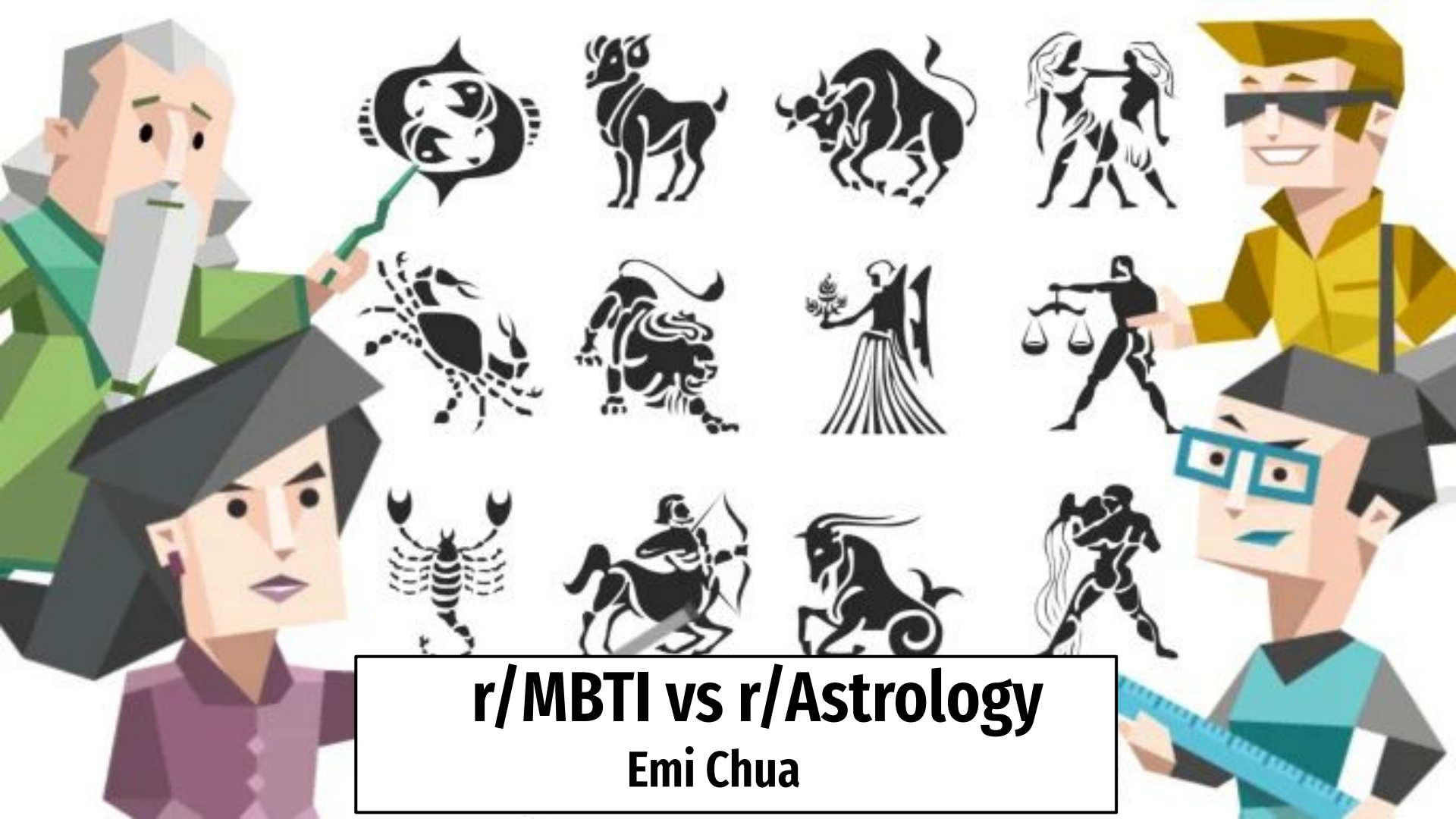


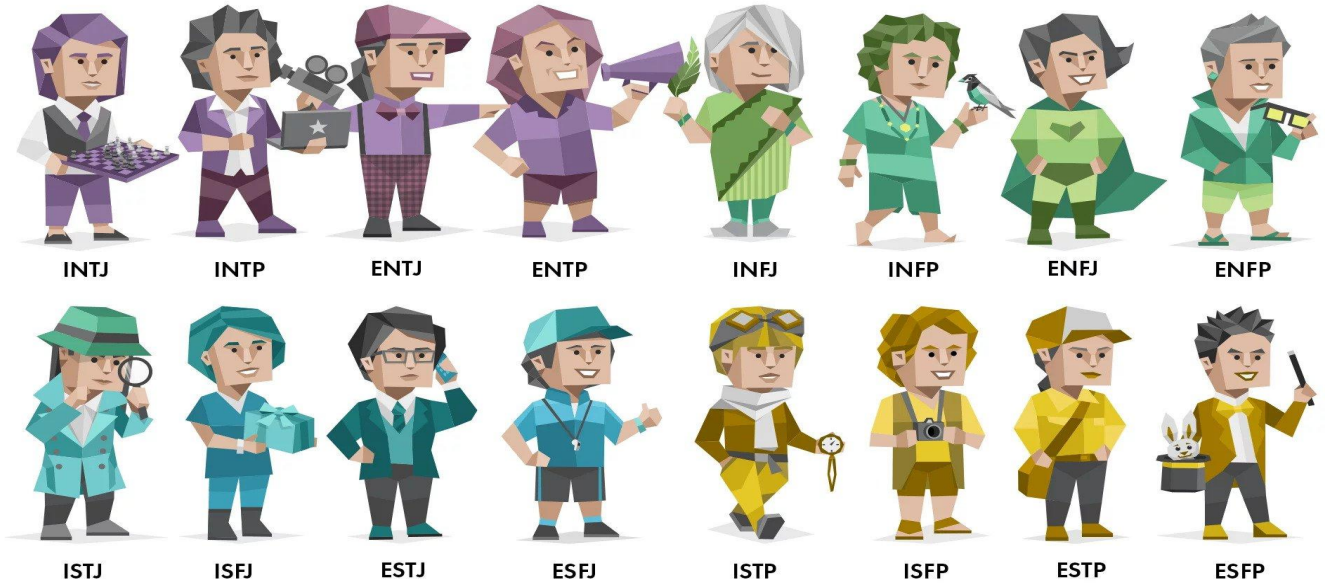
# r/MBTI vs r/Astrology

Emi Chua

# Problem Statement:

- Can we use supervised machine learning to classify content from these two different subreddits?
  - r/mbti
  - r/astrology
- Which model used is the best model?

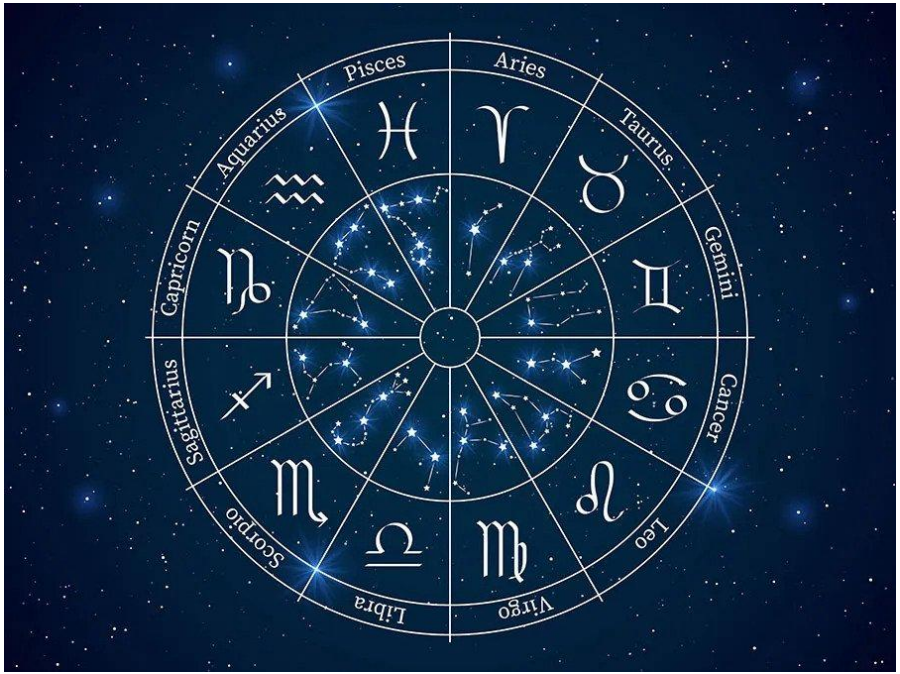
# Background Story




What's MBTI?

- Myers Briggs Type Indicator
- a tool which is frequently used to help individuals understand their own communication preference and how they interact with others.
- 434k subscriber

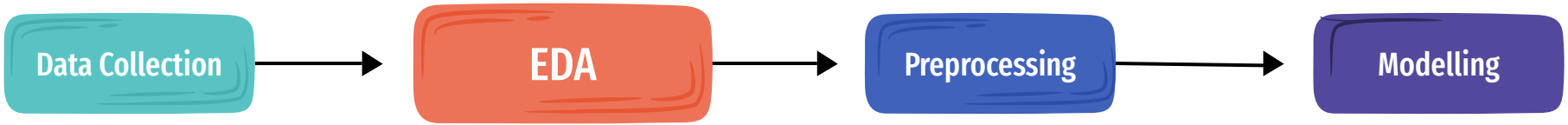
# Background Story




What's astrology?

- the study of the movements and relative positions of celestial bodies interpreted as having an influence on human affairs and the natural world.
- 287k subcriber

# Let's move on to the data science!

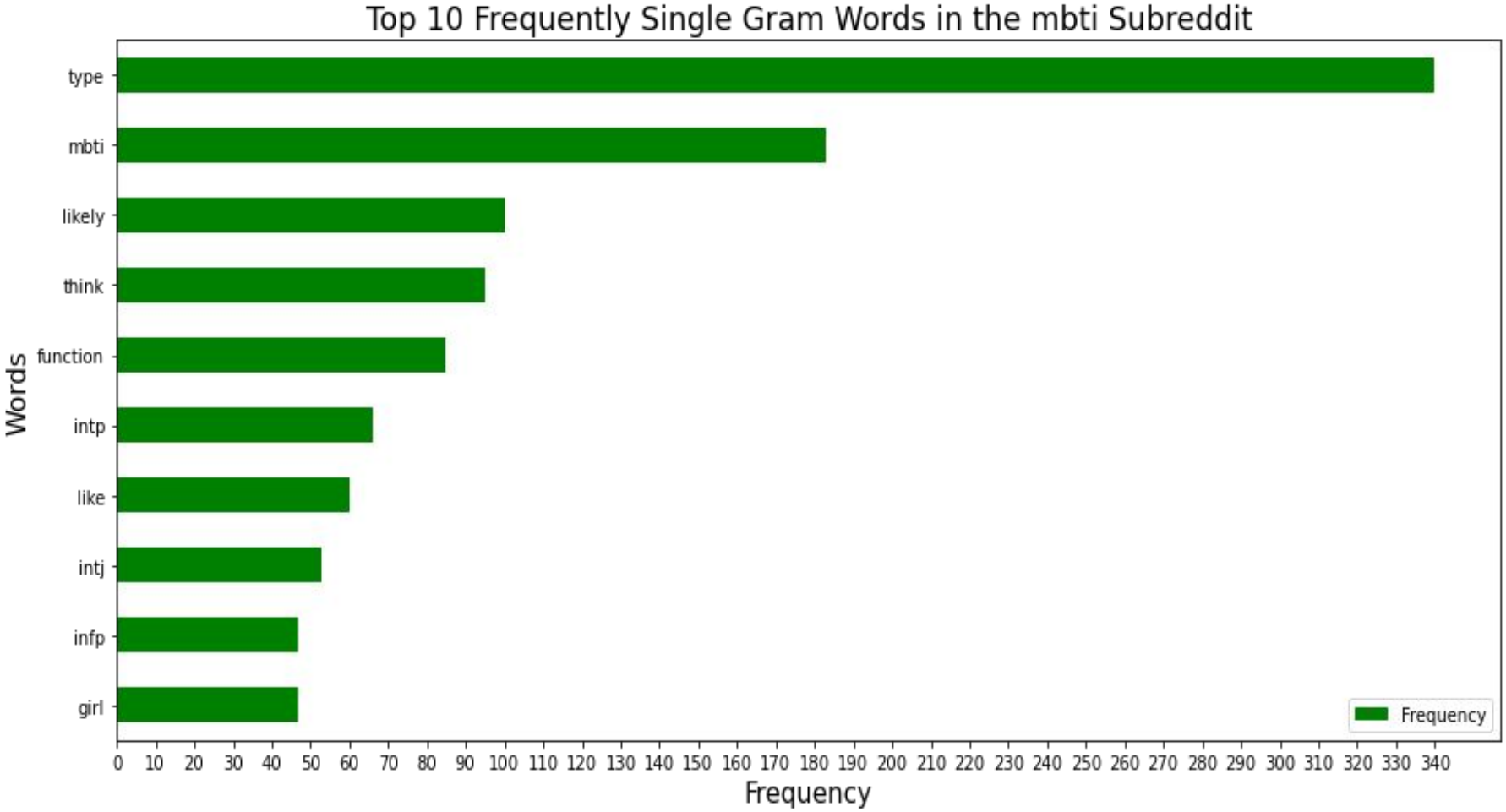


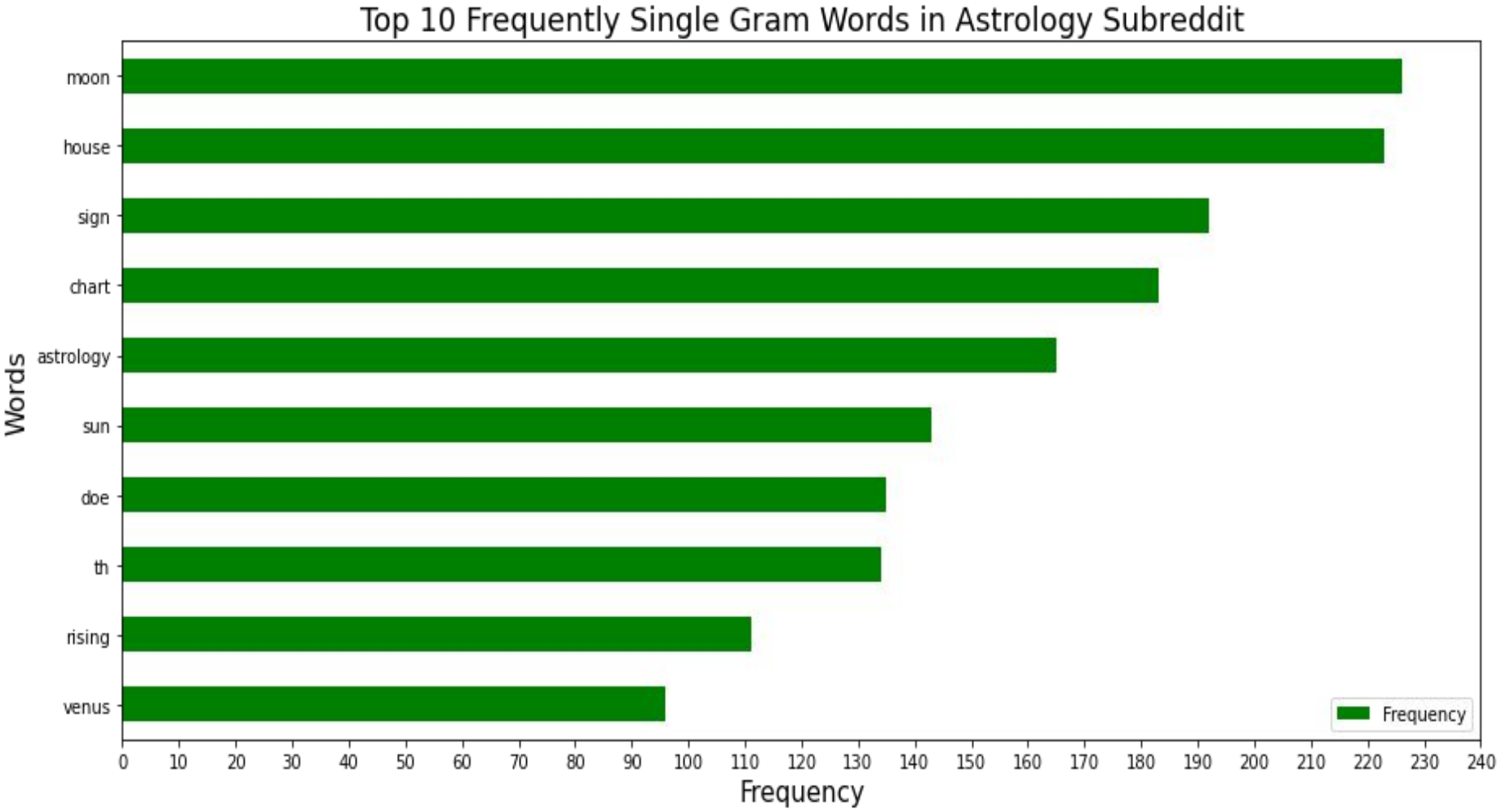

- scraping 2000 posts/subreddit
- Data cleaning
  - Duplicates
  - Null value

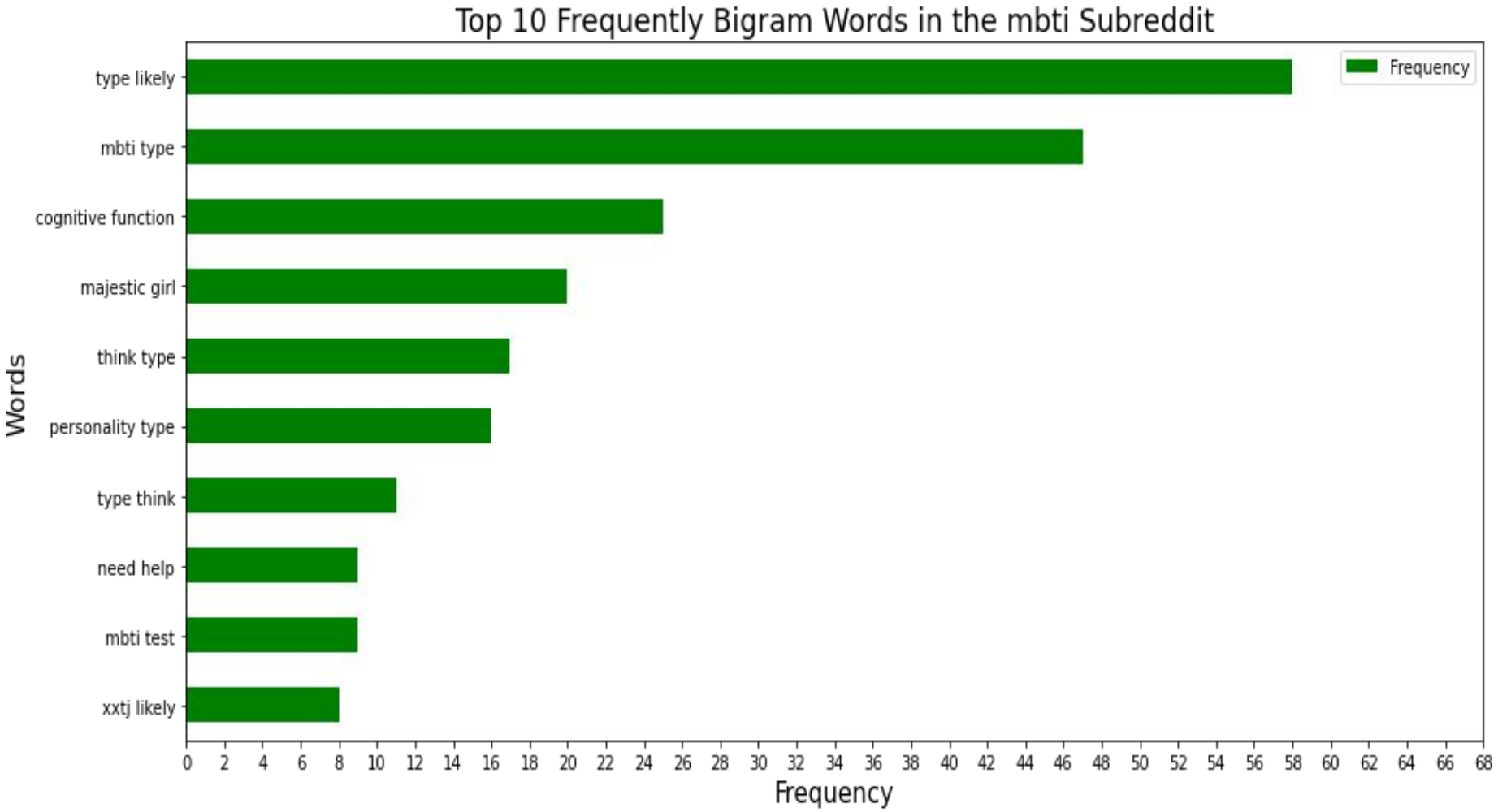
- Date range
- Summary Statistics
- Most common words:
  - single gram
  - bigram

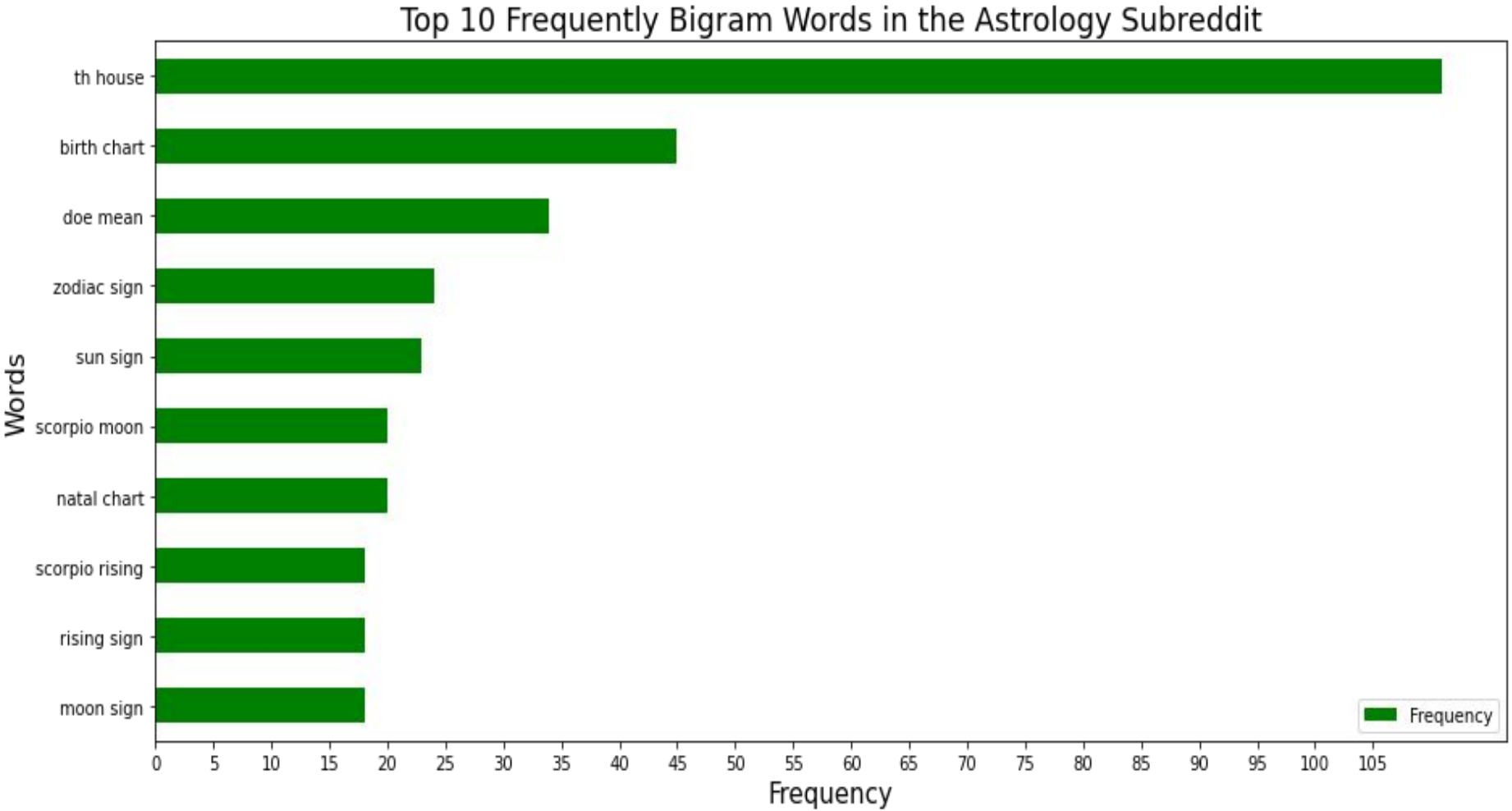
- Mapping the y Variable
- NLP Preprocessing
  - Lemmatization
  - Stemming
  - Stopwords
- Train-test Split
- Determing the Baseline Score

- Logistic Regression
- Bernoulli Naive Bayes
- Bagged Decision Tree
- Random Forest


Top 10 Frequently Single Gram Words in the mbti Subreddit
Words
type
mbti
likely
think
function
intp
like
intj
infp
girl
Frequency
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340
Frequency


Top 10 Frequently Single Gram Words in Astrology Subreddit
moon
house
sign
chart
astrology
sun
doe
th
rising
venus
Words
Frequency
Frequency
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240


Top 10 Frequently Bigram Words in the mbti Subreddit
Frequency
type likely
mbti type
cognitive function
majestic girl
think type
personality type
type think
need help
mbti test
xxtj likely
Words
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68
Frequency


Top 10 Frequently Bigram Words in the Astrology Subreddit
th house
birth chart
doe mean
zodiac sign
sun sign
scorpio moon
natal chart
scorpio rising
rising sign
moon sign
Words
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
Frequency
Frequency

# Preprocessing

- Tokenised
- Join method to make it lowercase
- Split method  for lemmatisation

# Modelling

- Logistic Regression Model (Tfidf)
  - Determine pipeline by using the .get_params attribute via trial & error
- Bernoulli Naive Bayes (Tfidf)
  - Bags of words
- Bagged Decision Tree
  - Baggingclassifier
- Random Forest
  - Determine parameters by using the .get_params attribute via trial & error

# Evaluation of Model

| Model | Training Accuracy Score | Testing Accuracy Score | Correctly Classified | Misclassified | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.963 | 0.93 | 945 | 71 | 0.7902 |
| Bernoulli Naives Bayes | 0.974 | 0.948 | 971 | 54 | 0.8437 |
| Bagged Decision Tree | 0.991 | 0.895 | 918 | 107 | 0.9799 |
| Random Forest | 0.998 | 0.913 | 936 | 89 | 0.9822 |

# Recommendation

- Increase the stopword list with more nouns to have a better predictable model (i.e. ‘like’)
- Each of the subreddits change over time so result might change
- A different model that we had not yet modeled (i.e. SLM, Adaboost)

# Q&A