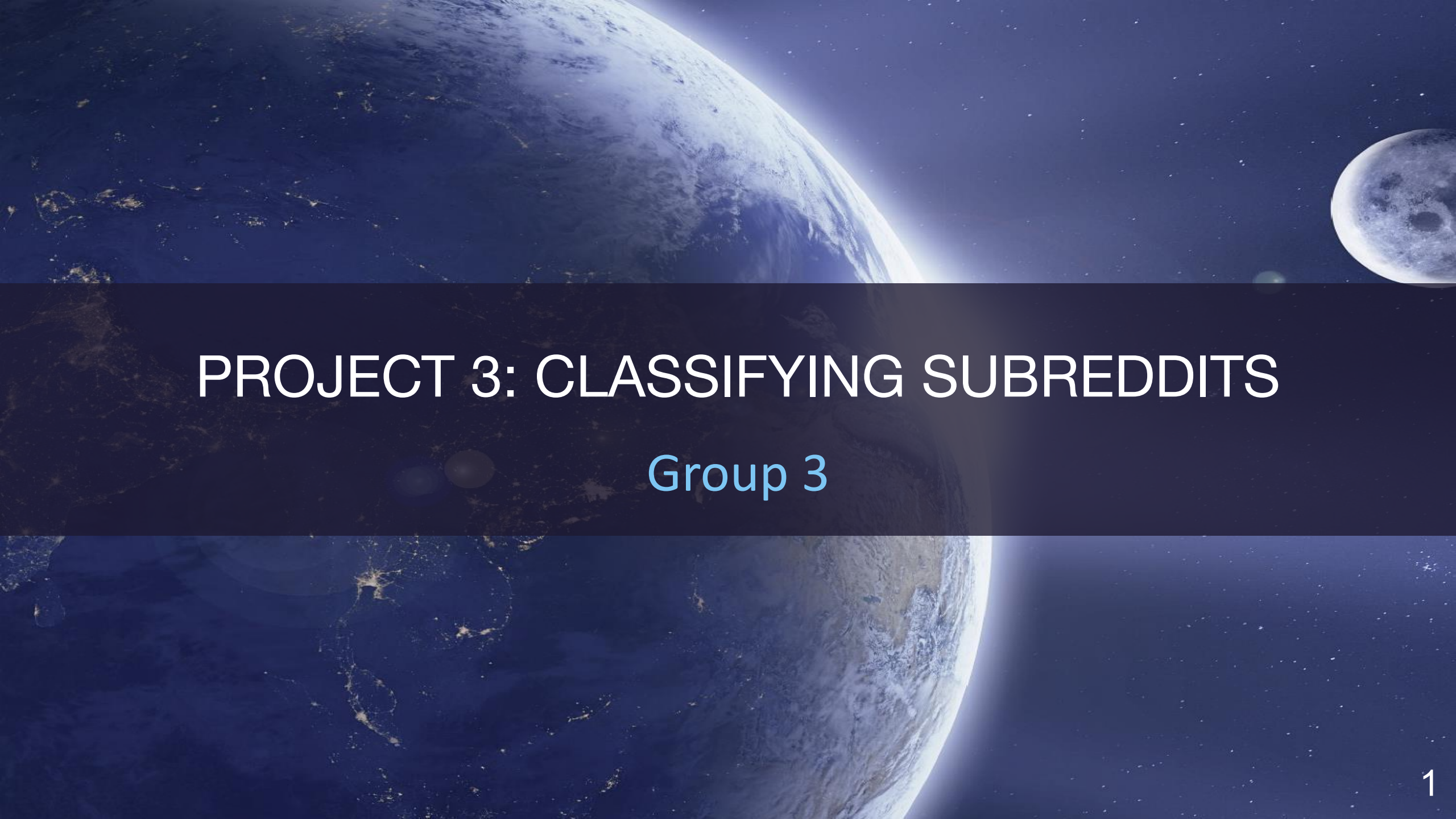
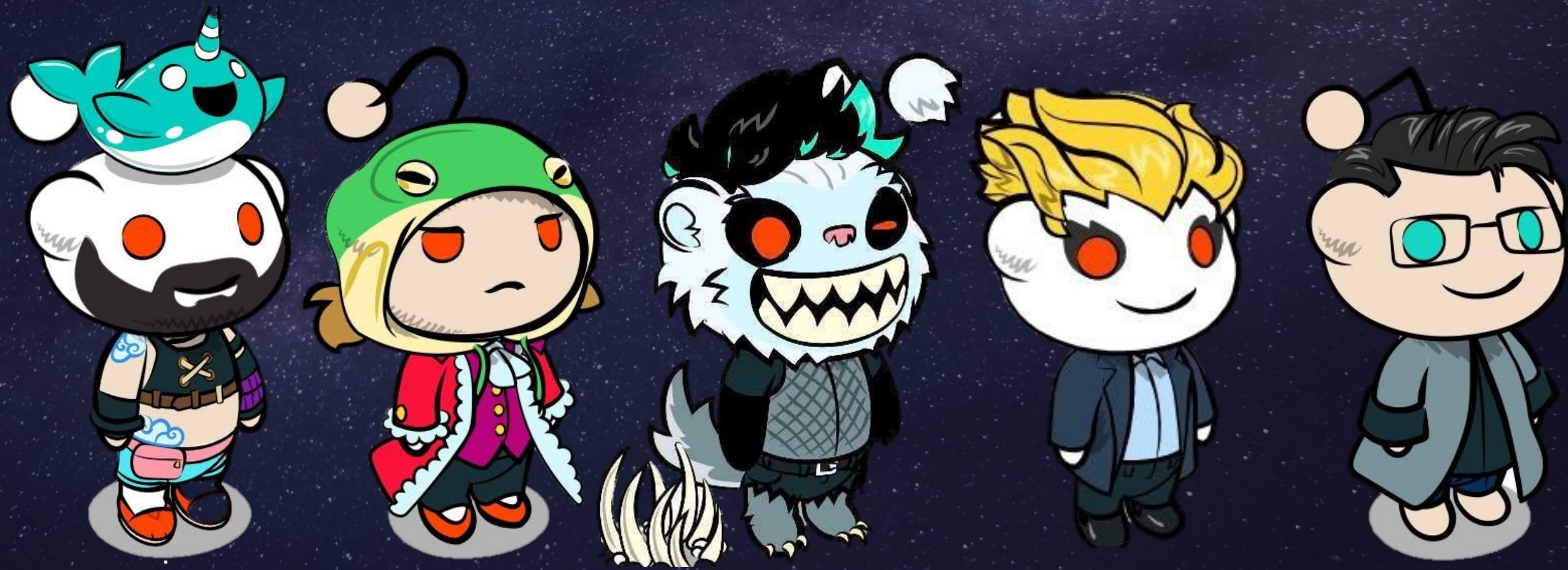


PROJECT 3: CLASSIFYING SUBREDDITS

Group 3

GROUP 3 ALIENS MEMBERS



Darren Tu Cheng Yeow Zavier Soon Thien Sean Chee Tzen

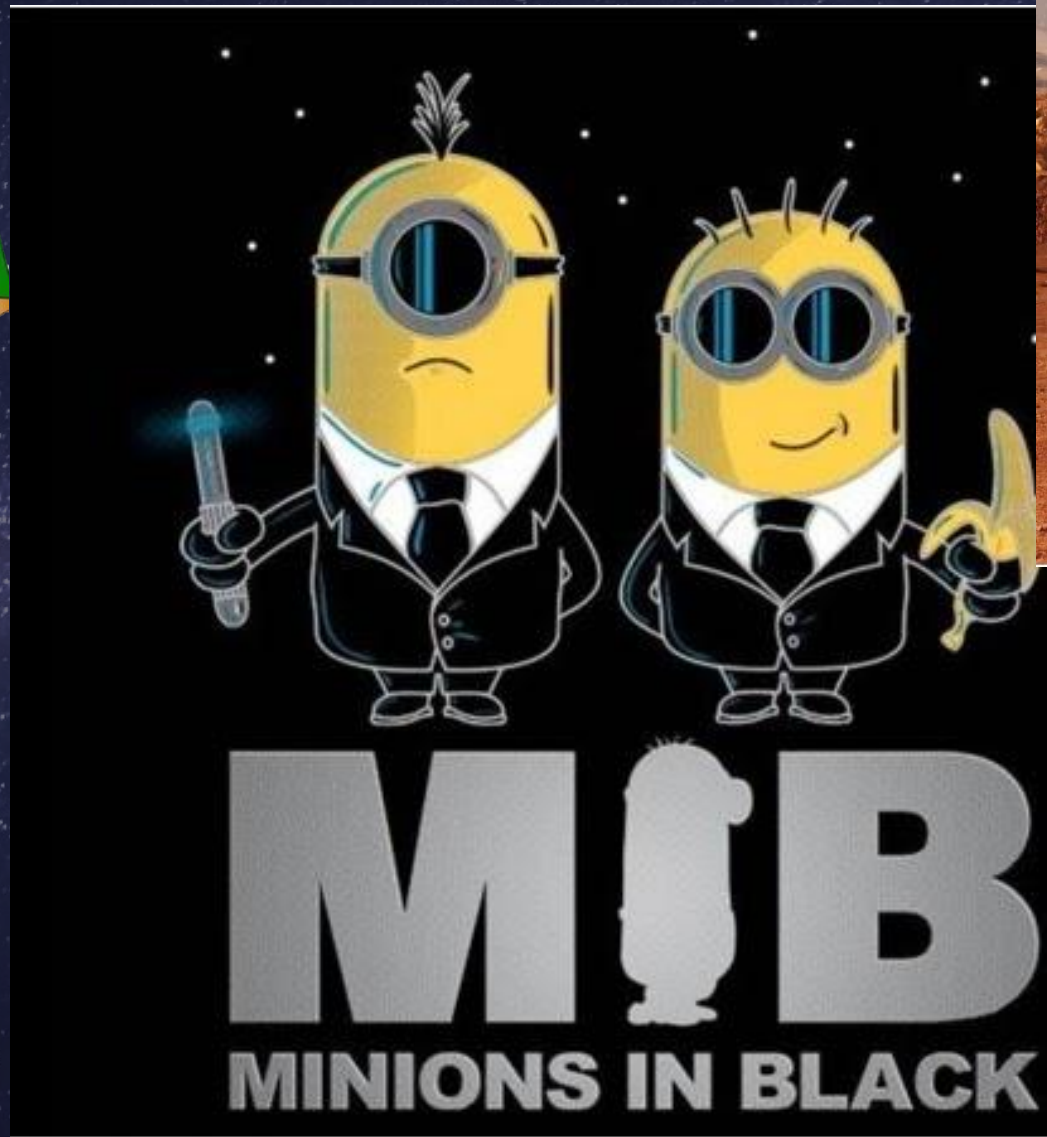
OUR LIST

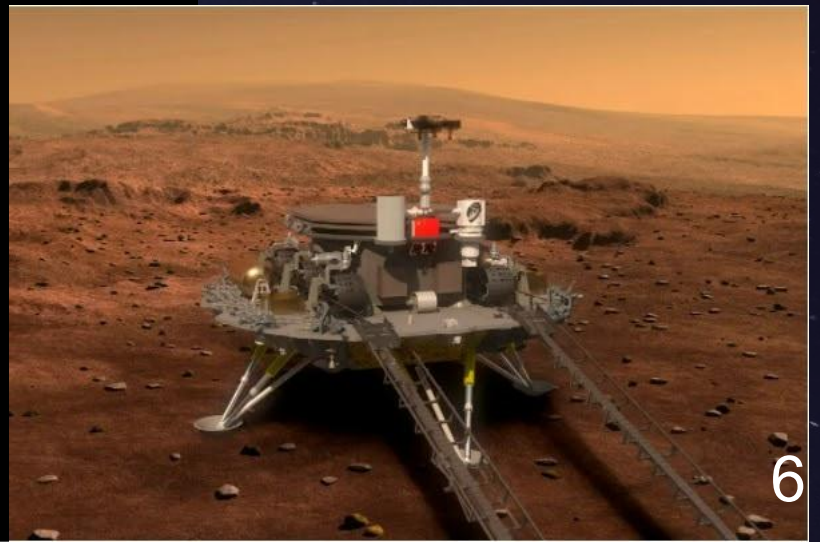
- Problem Statement - Darren
- Data Collection / EDA - ChengYeow
- Preprocessing & Modeling - Zavier
- Evaluation & Conceptual Understanding - Thien Sean
- Conclusion & Recommendations - Chee Tzen

PROBLEM STATEMENT

PROBLEM STATEMENT

PROBLEM STATEMENT





Alien

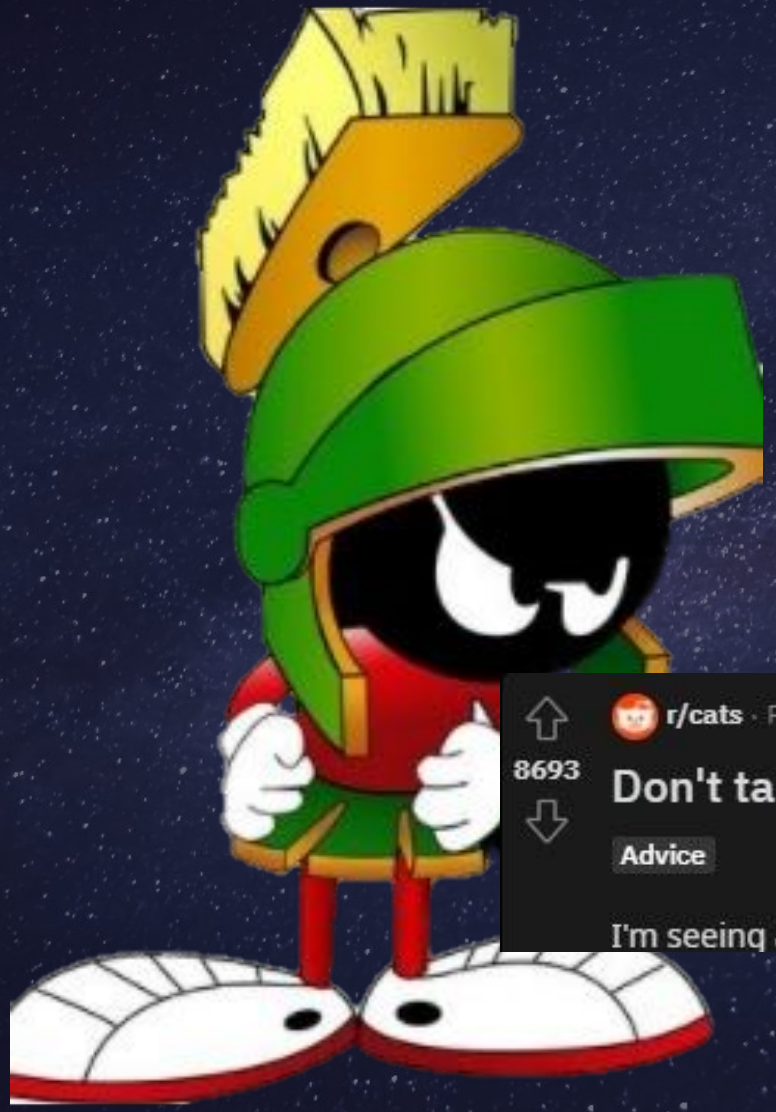
GOAL OF THE PROJECT

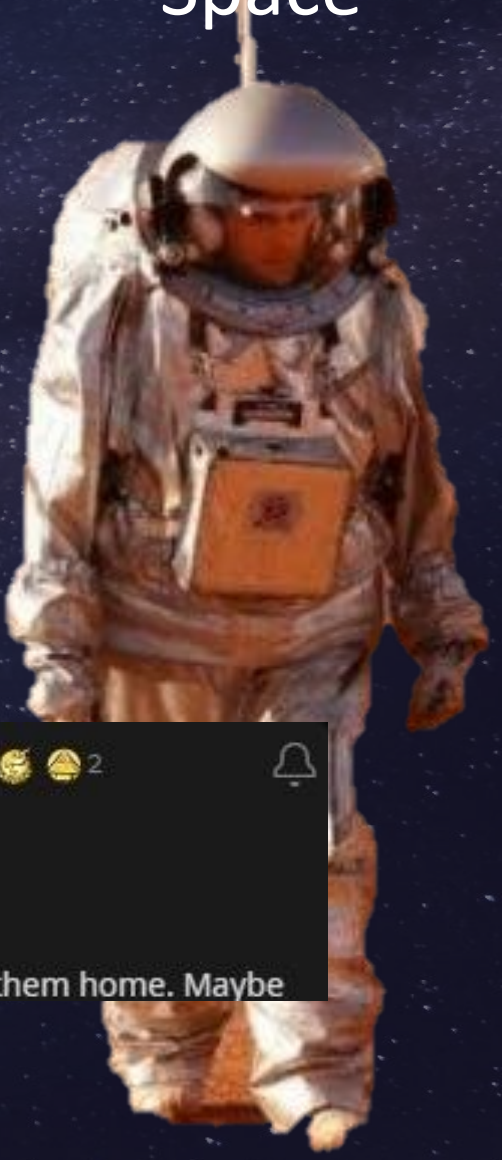
Space

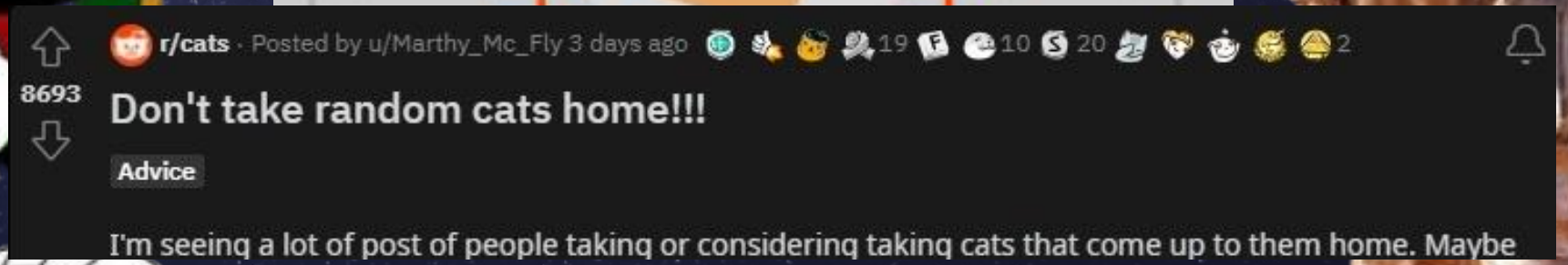
GOAL OF THE PROJECT

Alien

Space



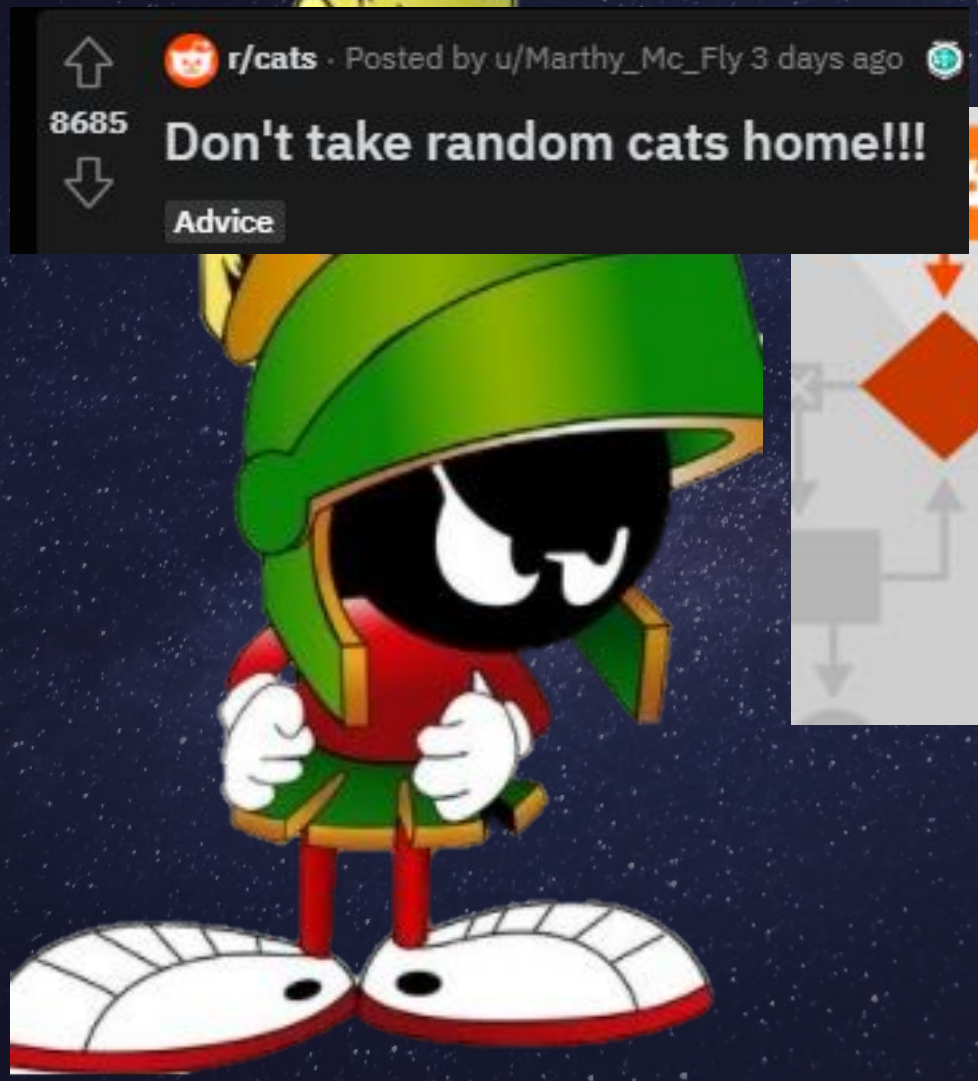




GOAL OF THE PROJECT

Alien

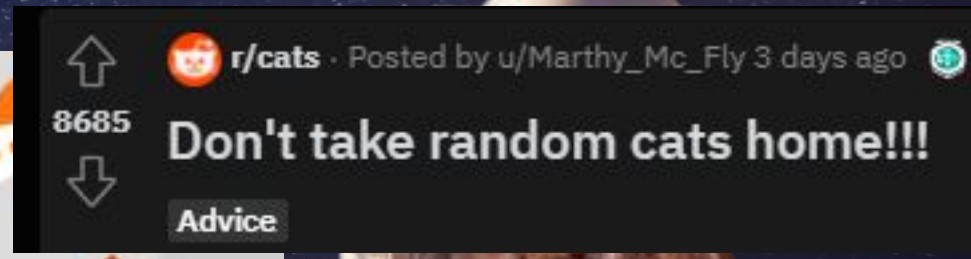
Space



GOAL OF THE PROJECT

Alien

Space



Alien AL OF THE PR Space

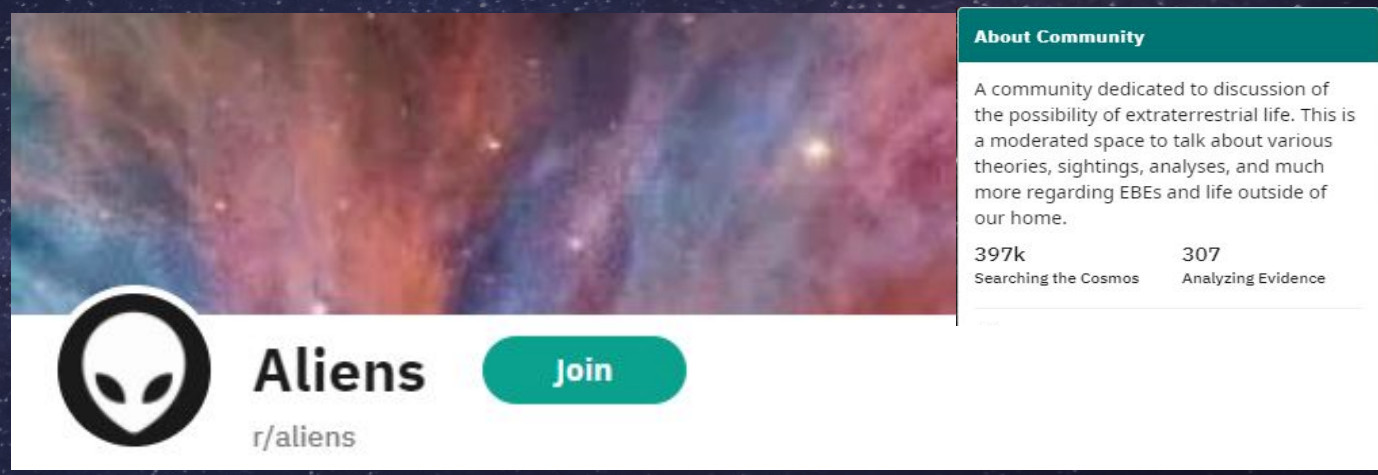
DATA EXTRACTION

- From: 1st Jan 2022, 00:00 Time

- Posts Collected: 2,000

  r/aliens - 1,000 r/space - 1,000

- Features: title, selftext
- Target: subreddit
  (1- aliens and 0- space)



DATA CLEANING

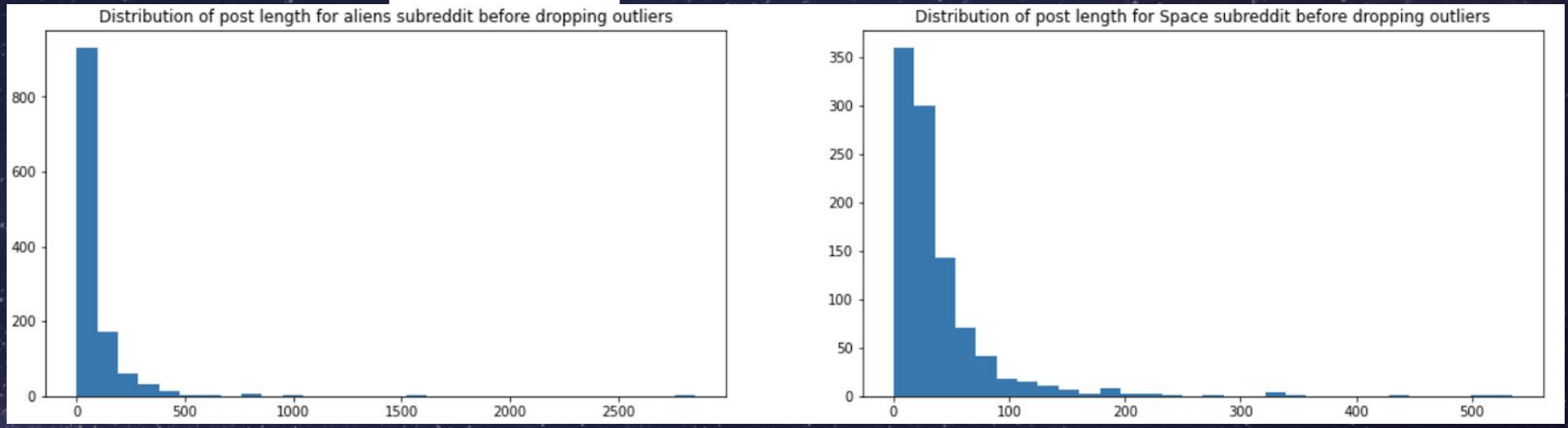
- Remove
  - Null values
  - [remove] & [deleted]
  - Hyperlink (http)
  - Emoji (using Demoji function)
  - Non-English characters
  - Markdown text formatting

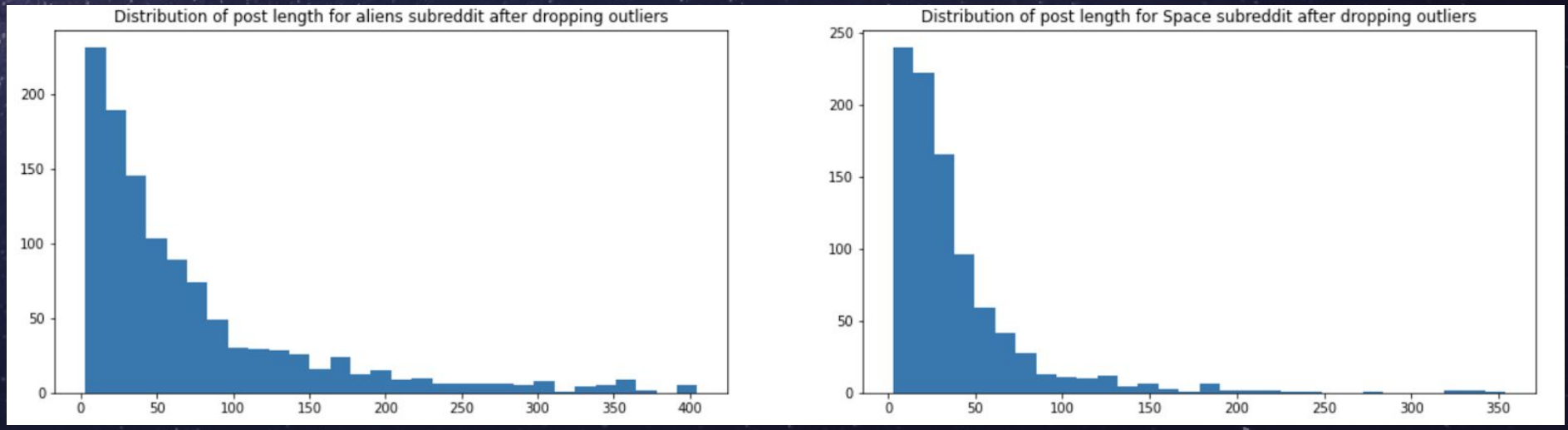
OUTLINERS

Lower Band
Remove words of 2 and lower

Upper Band

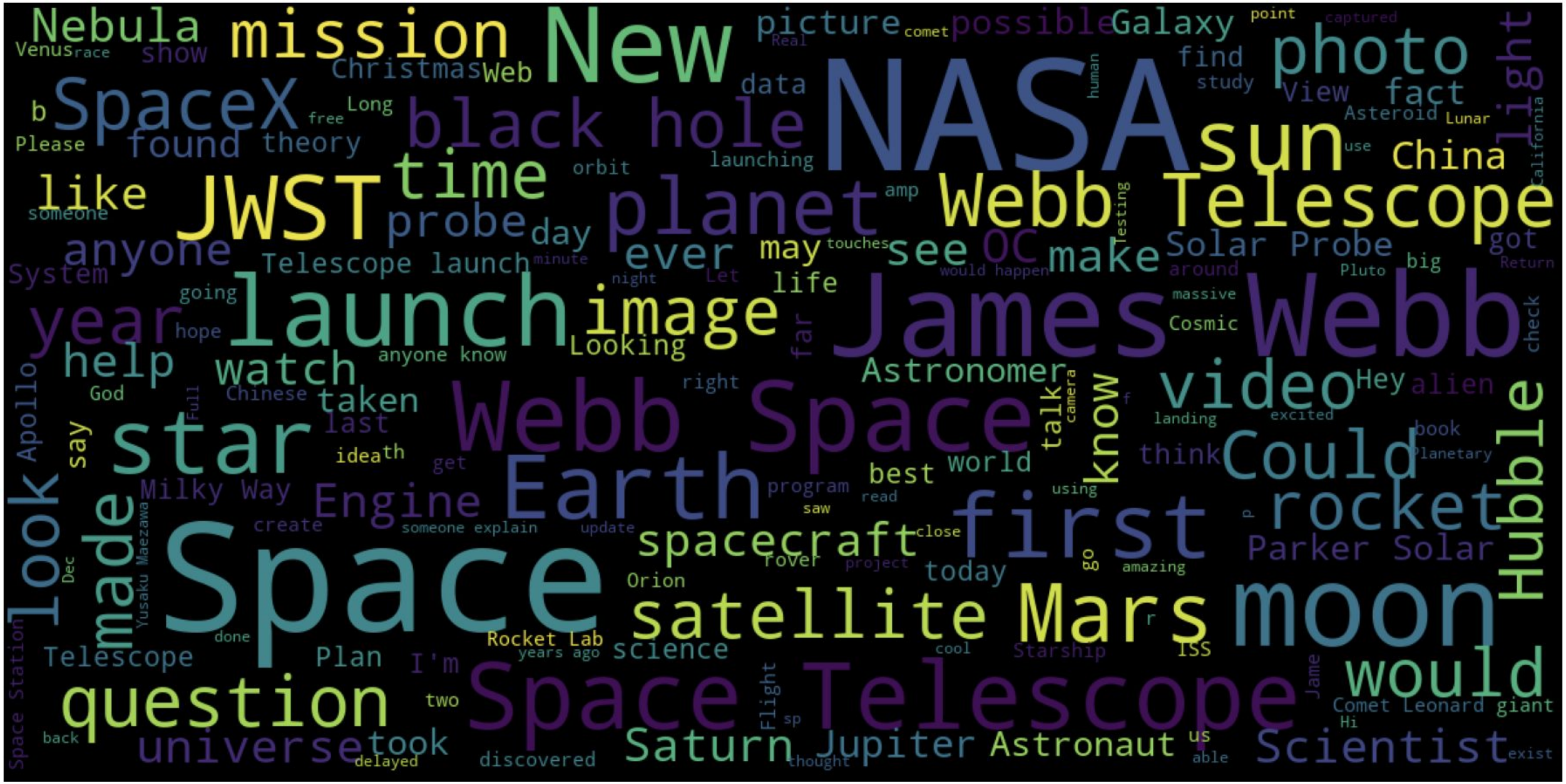
Remove words of 400 and above





Word Cloud for Aliens

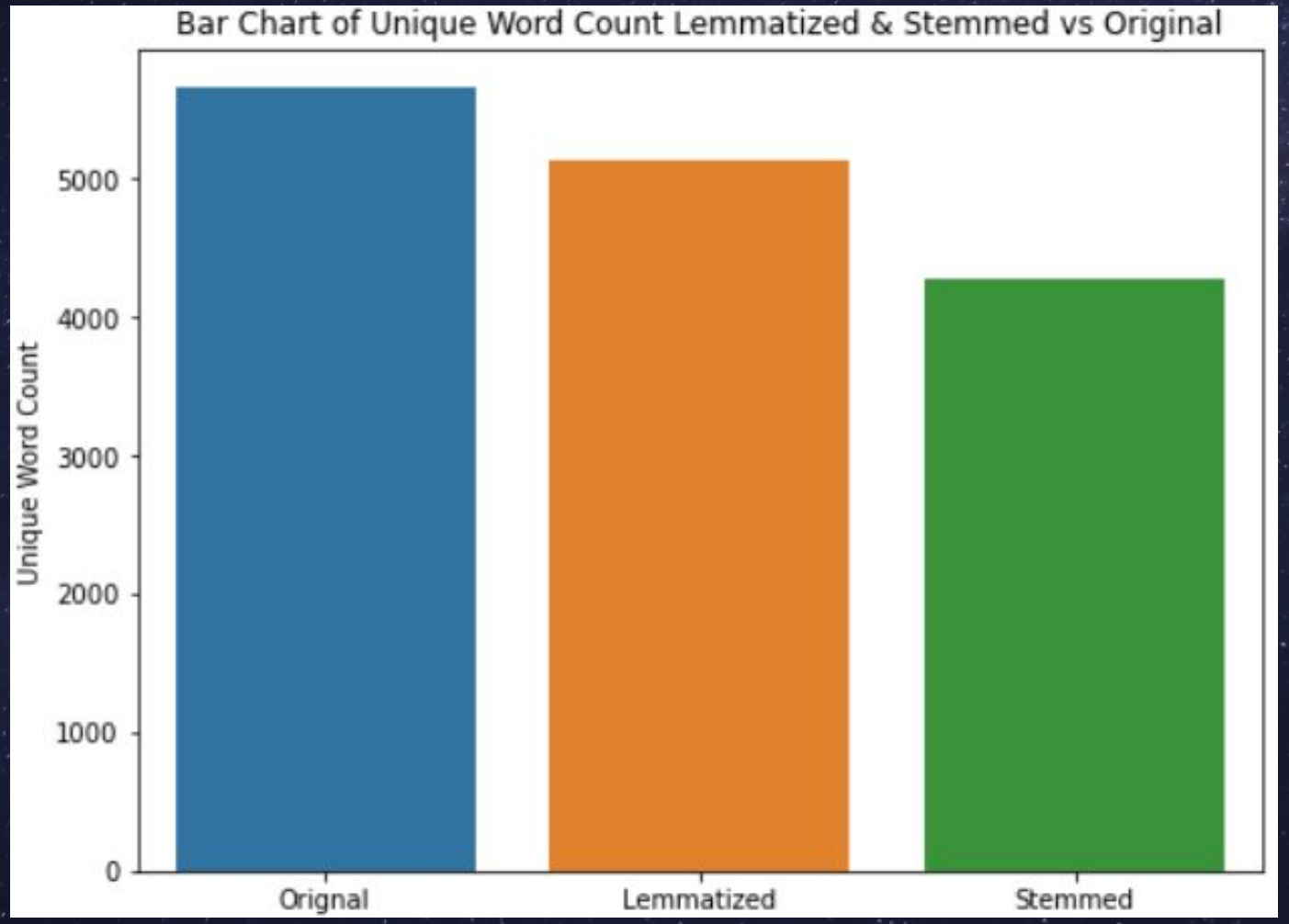
Word Cloud for Space



Preprocessing

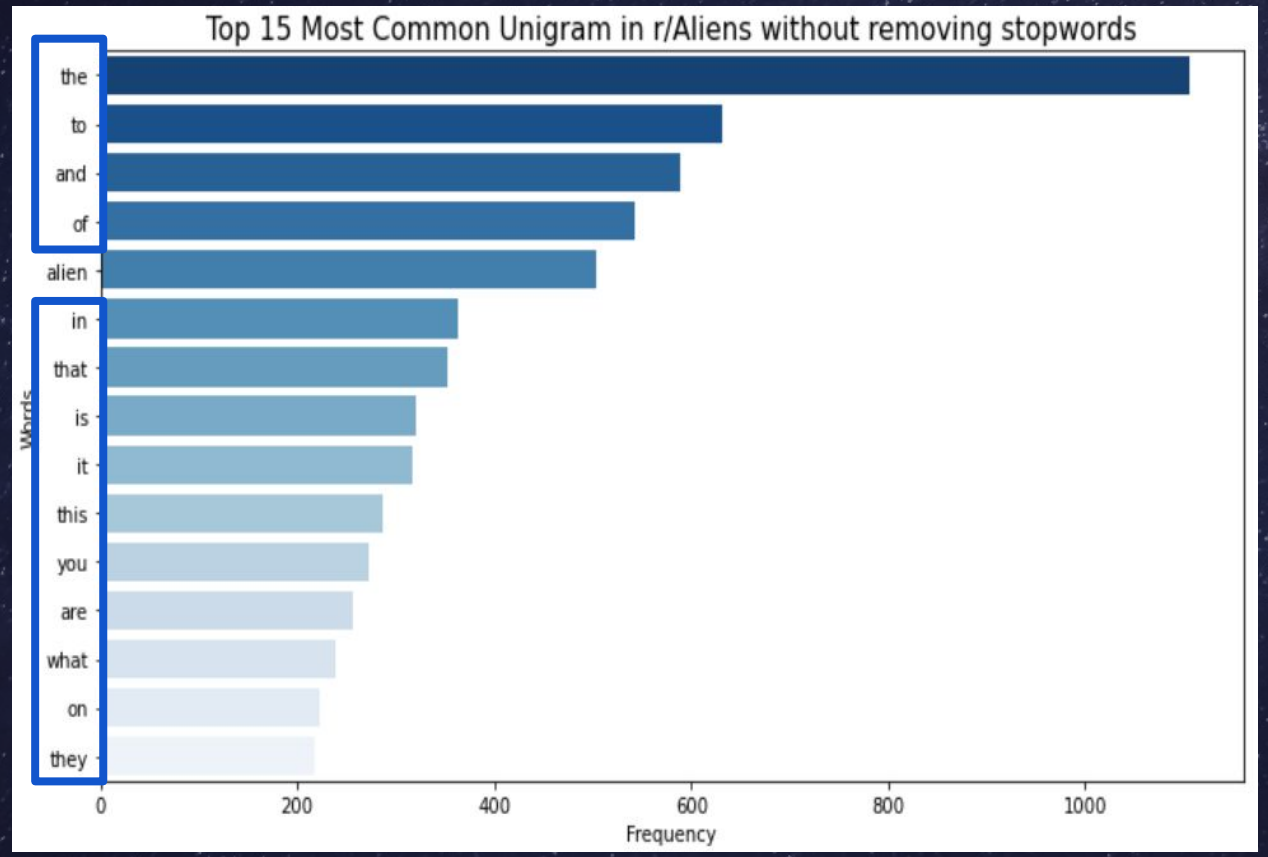
- Methods of preprocessing
  - Lemmatization / Stemming
  - Stop Words

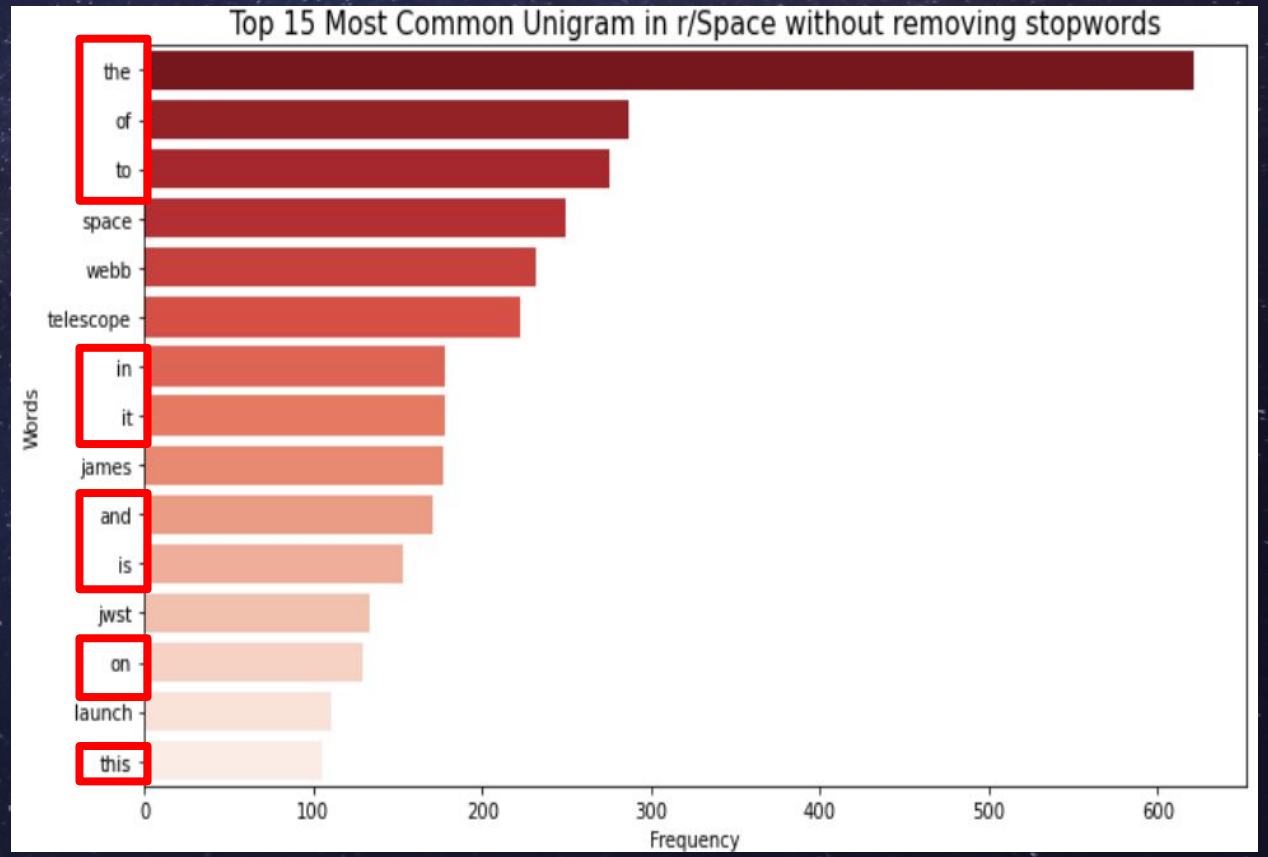
Lemmatization / Stemming



	Lemmatized	Stemmed
0	discovery	discoveri
1	advanced	advanc
2	intelligent	intellig
3	anthropocentric	anthropocentr
4	religious	religi
5	people	peopl
6	alike	alik
7	discussion	discuss
8	this	thi
9	existing	exist
10	automatically	automat
11	disprove	disprov
12	a	as
13	perspective	perspect

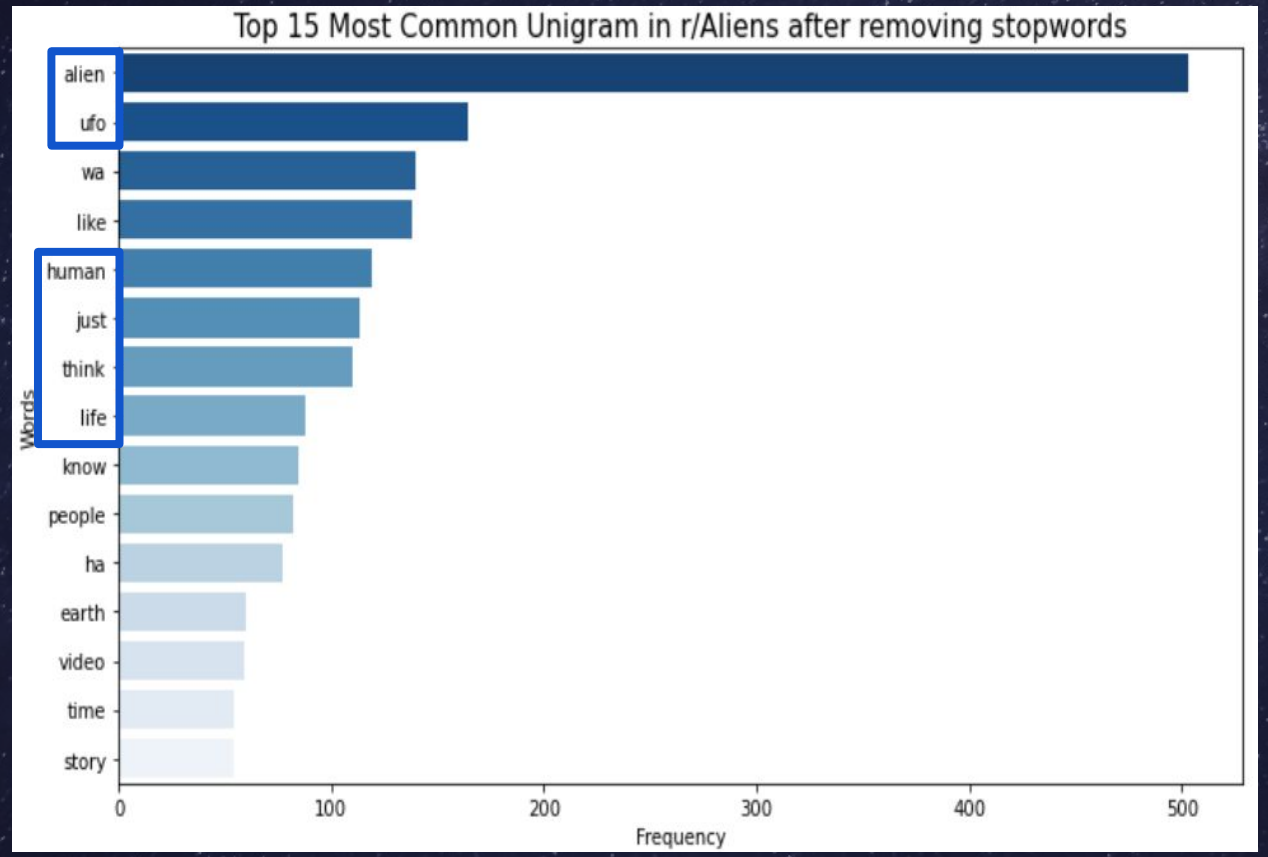
Stopwords

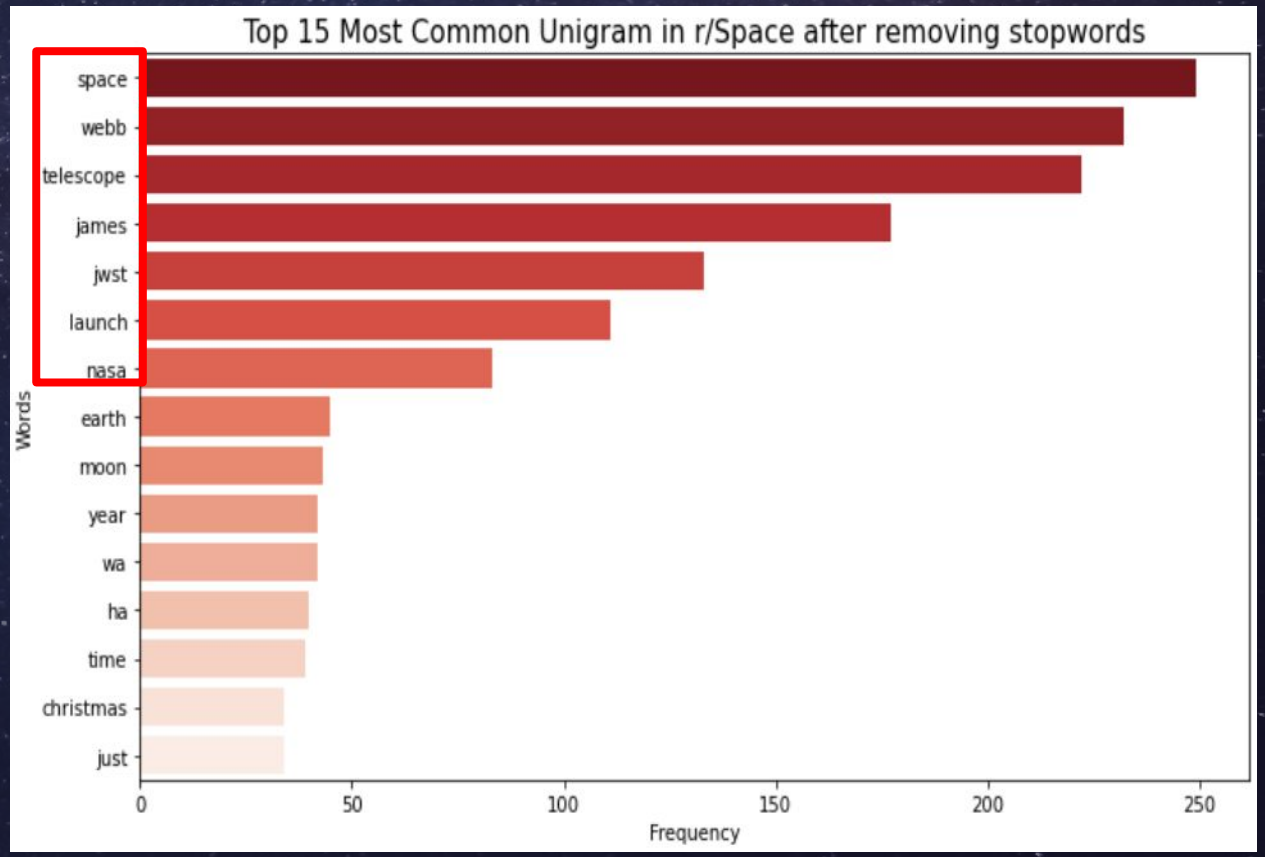




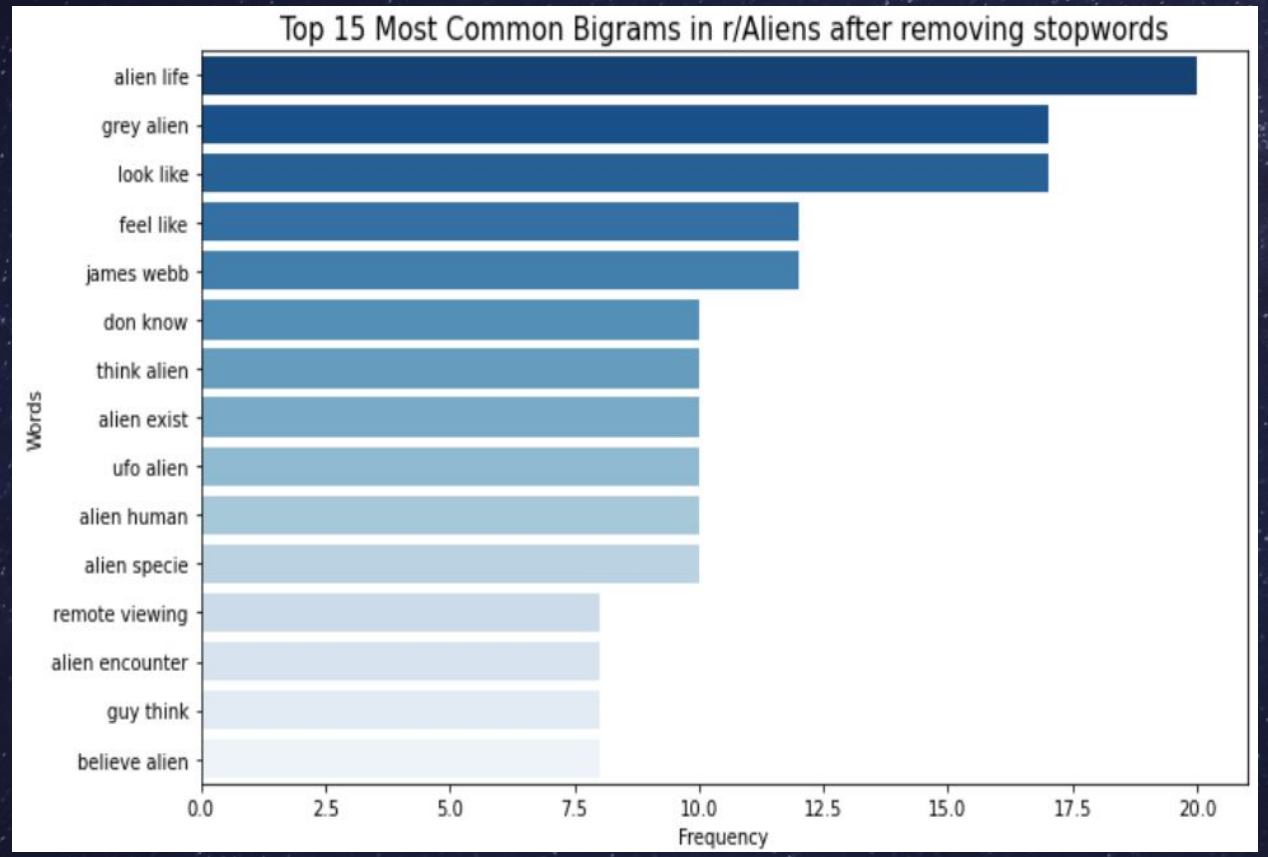
Stop words are words that do not provide any valuable information. They are common words that only add to the grammatical structure and flow of the sentence.

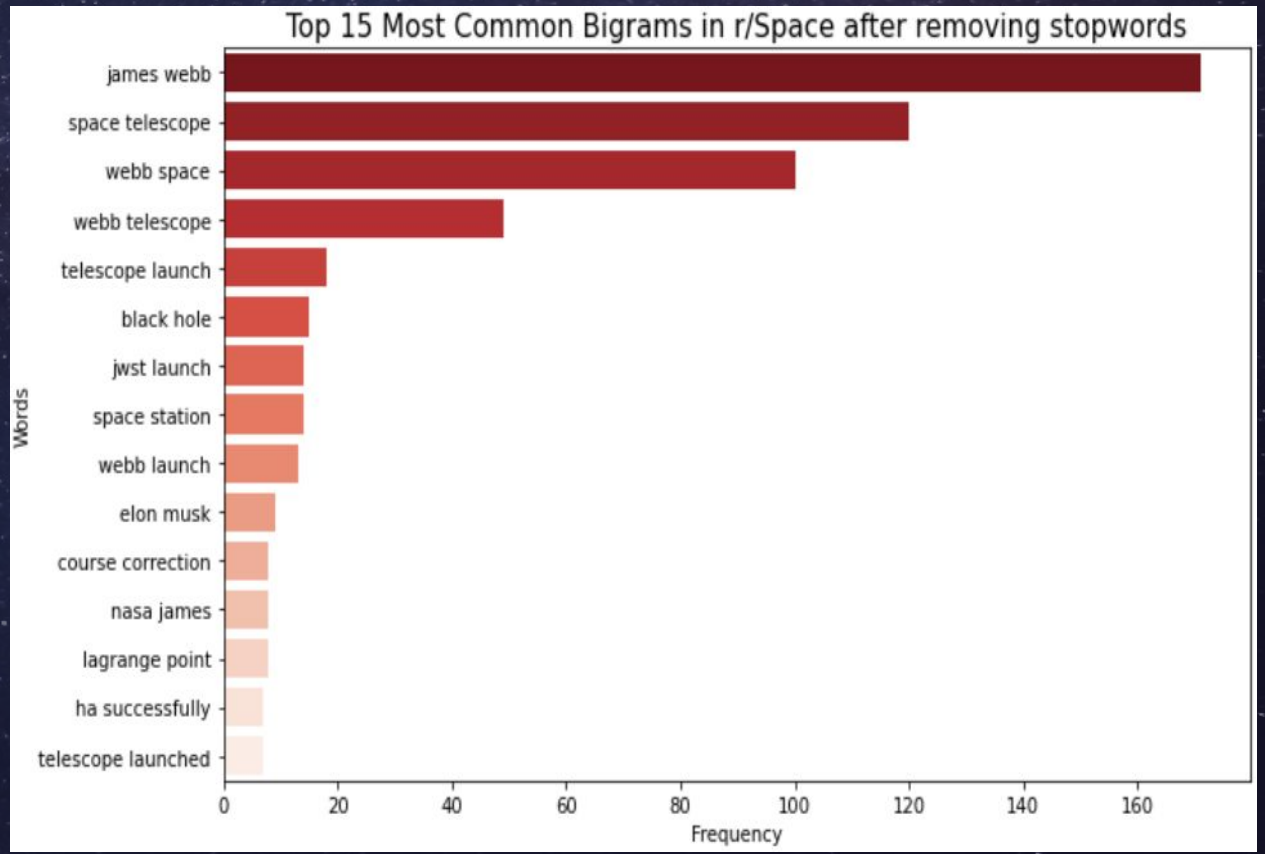
Stopwords


Top 15 Most Common Unigram in r/Aliens after removing stopwords
alien
ufo
wa
like
human
just
think
life
know
people
ha
earth
video
time
story
Words
0
100
200
300
400
500
Frequency


Top 15 Most Common Unigram in r/Space after removing stopwords
space
webb
telescope
james
jwst
launch
nasa
earth
moon
year
wa
ha
time
christmas
just
Words
0
50
100
150
200
250
Frequency

Stopwords

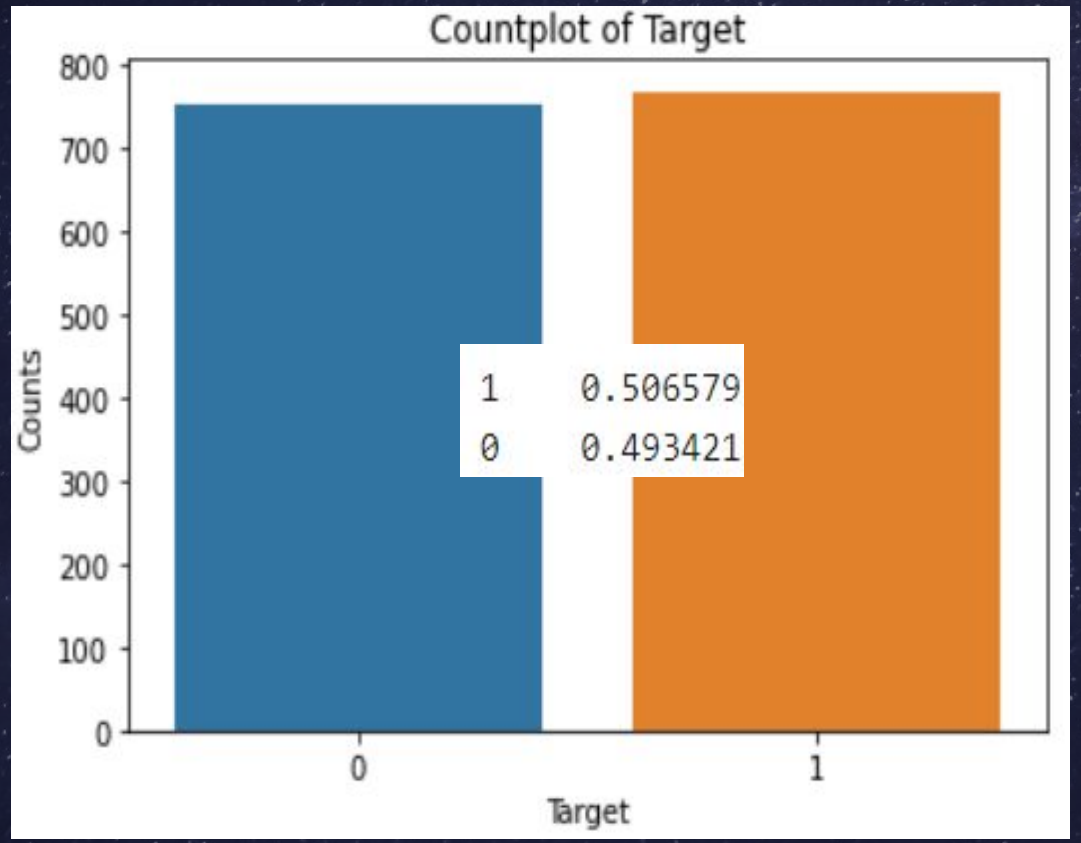

Top 15 Most Common Bigrams in r/Aliens after removing stopwords
alien life
grey alien
look like
feel like
james webb
don know
think alien
alien exist
ufo alien
alien human
alien specie
remote viewing
alien encounter
guy think
believe alien
Words
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Frequency


Top 15 Most Common Bigrams in r/Space after removing stopwords
james webb
space telescope
webb space
webb telescope
telescope launch
black hole
jwst launch
space station
webb launch
elon musk
course correction
nasa james
lagrange point
ha successfully
telescope launched
Words
0
20
40
60
80
100
120
140
160
Frequency

Choice of Model

- Baseline Model
- TFIDF, CVEC
  - LogisticRegression
  - MultinomialNB
  - RandomForest

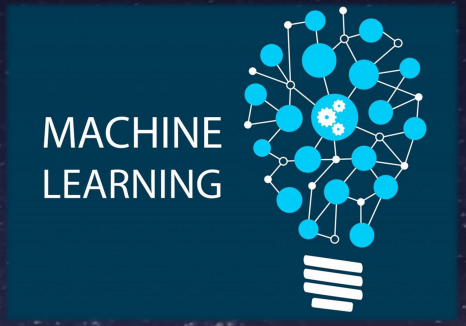
Baseline Model



r/Aliens = 1

r/Space = 0

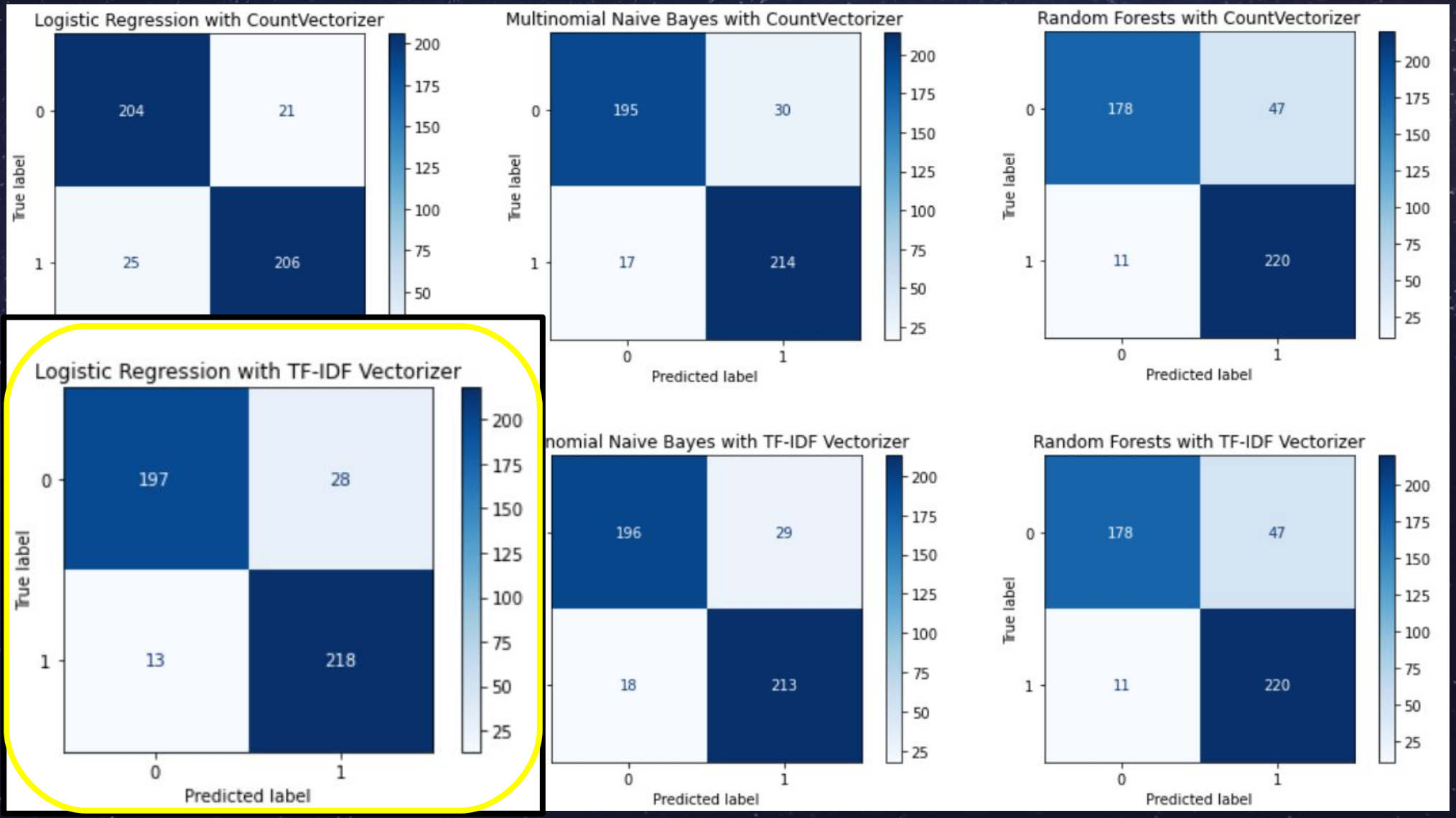
MODEL EVALUATION



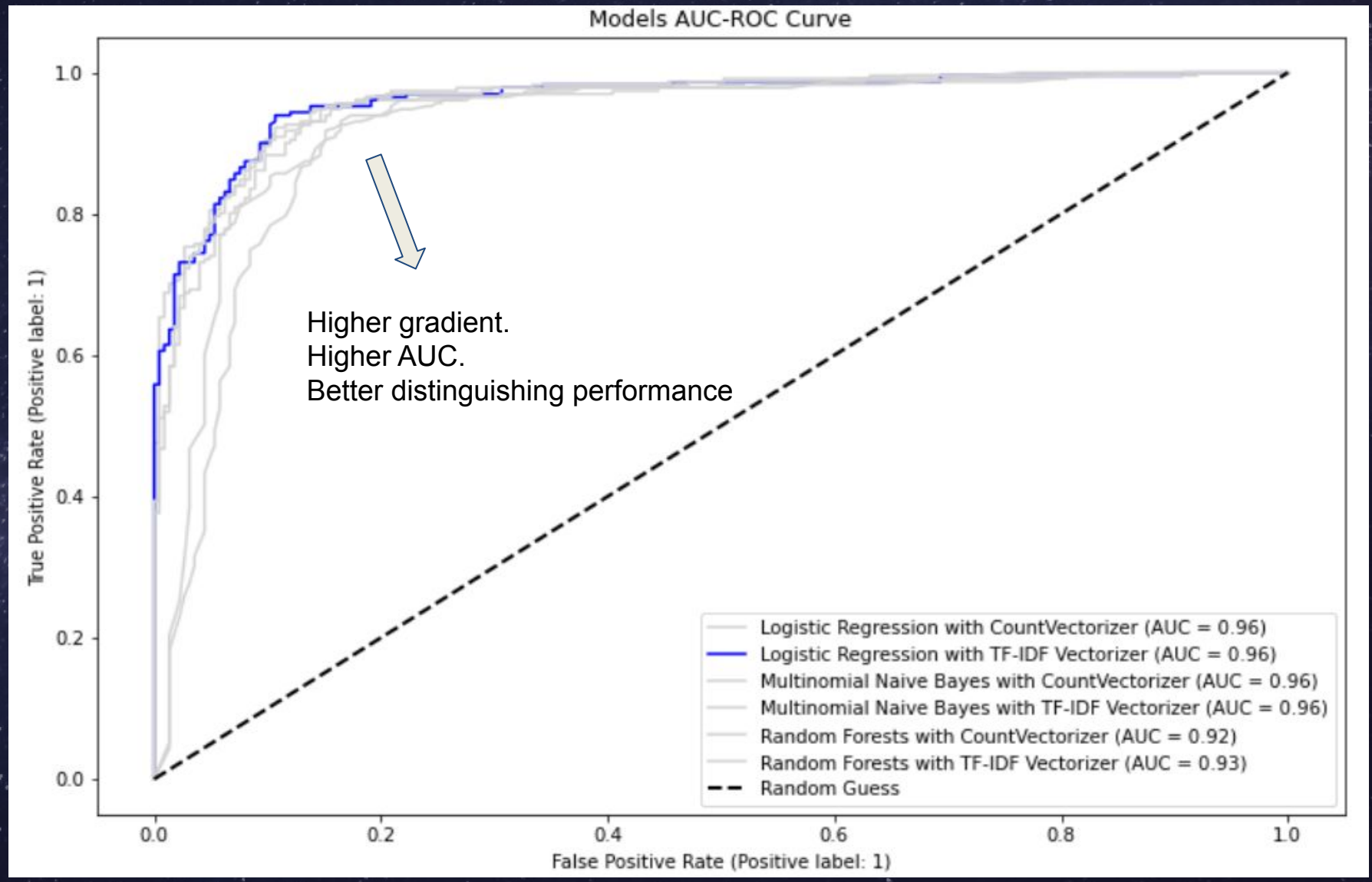
Model	Vectorizer	Train Score	Test Score	F1-score
Logistic Regression	CountVectorizer	0.9774	0.8991	0.8996
Logistic Regression	TfidfVectorizer	0.9718	0.9101	0.9140
MultinomialNB	CountVectorizer	0.9718	0.8969	0.9011
MultinomialNB	TfidfVectorizer	0.9793	0.8969	0.9006
Random Forest Classifier	CountVectorizer	1.0	0.8728	0.8835
Random Forest Classifier	TfidfVectorizer	1.0	0.8728	0.8835

Higher interpretability, lesser overfitting, less processing power.

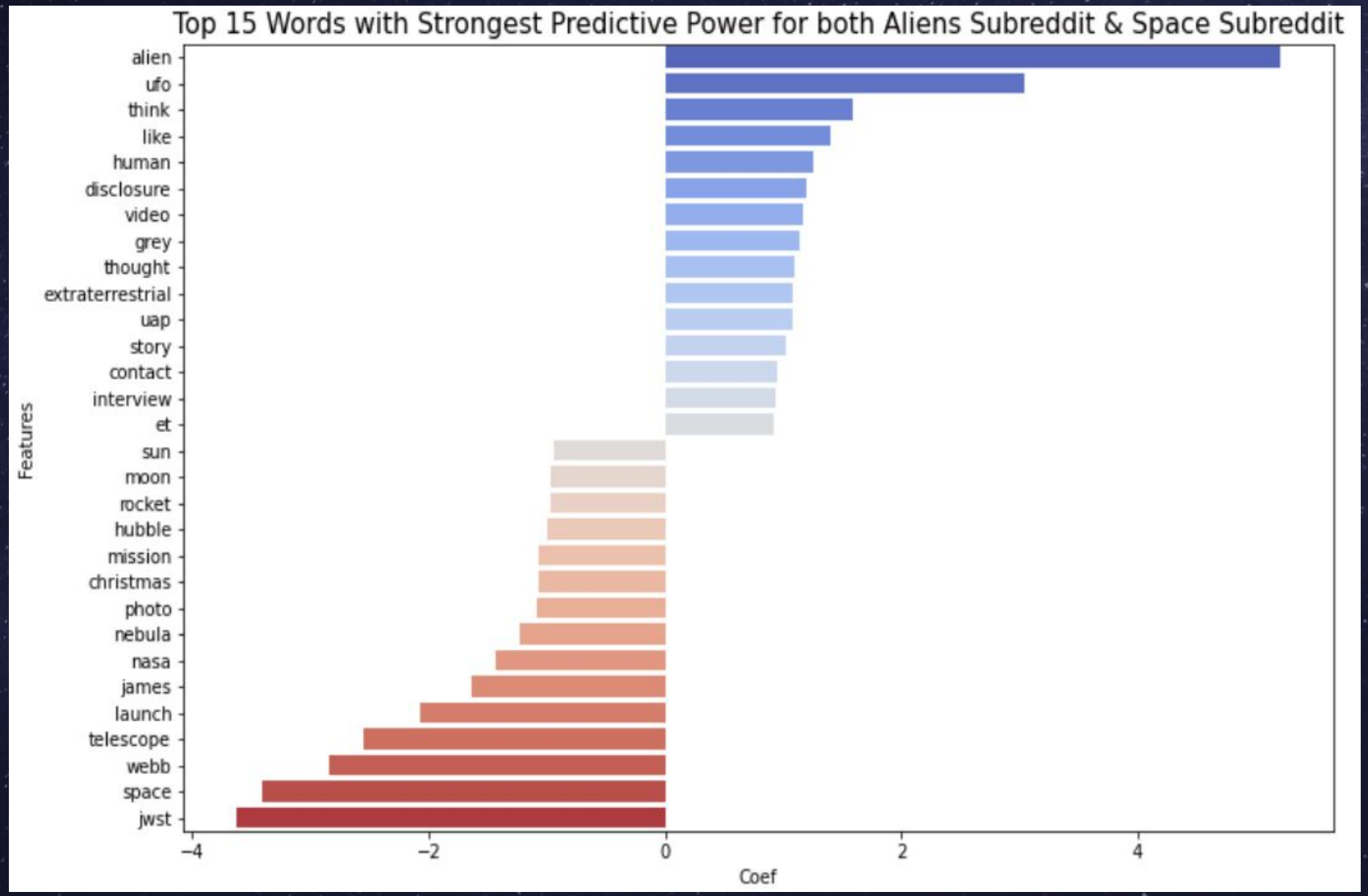
MISCLASSIFICATION



AUC - ROC CURVE



MODEL EVALUATION – LOGISTIC REGRESSION - TFIDF VECTORIZER

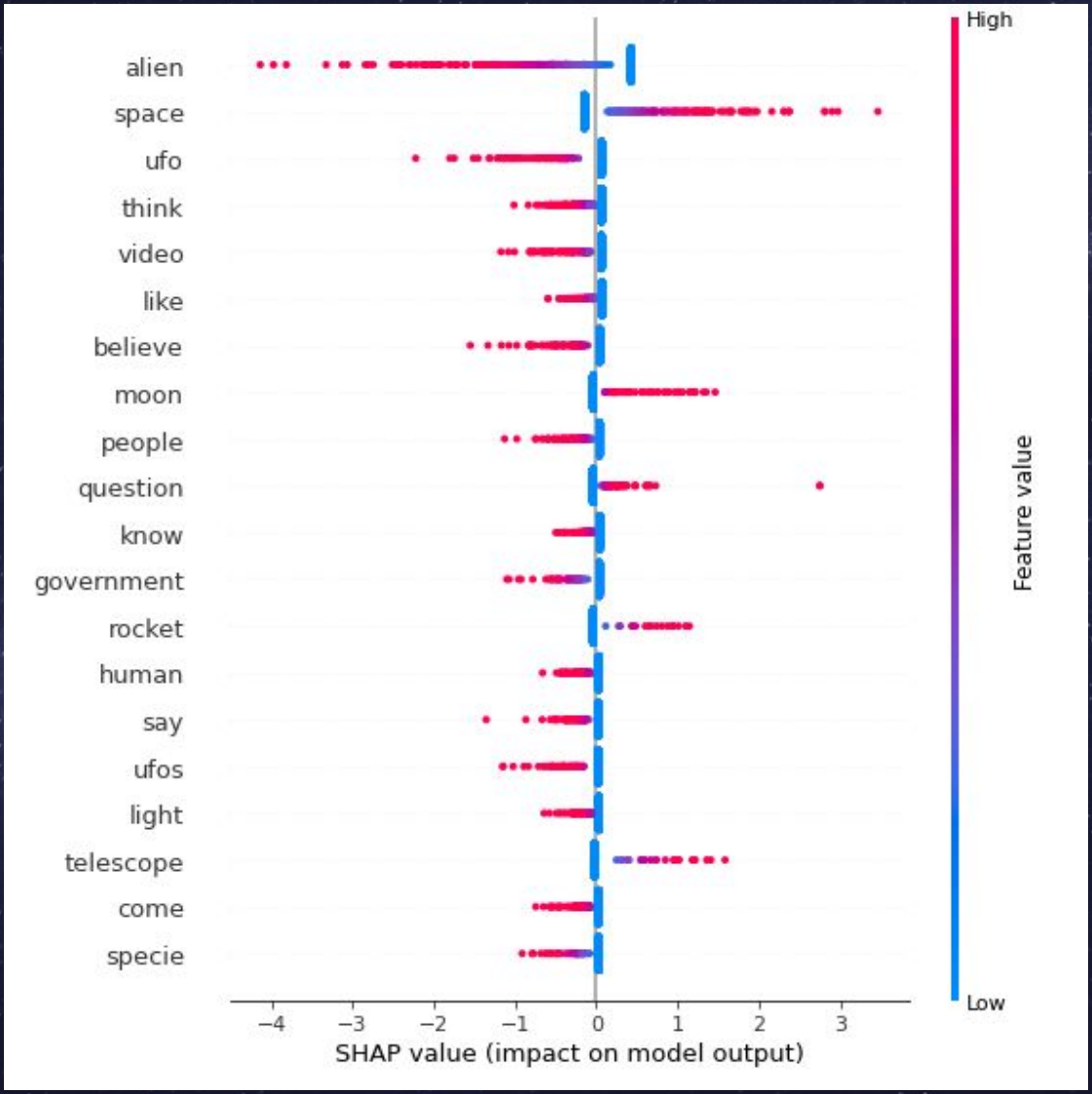


Positive coef= Feature that predicts class 1
Negative coef = Feature that predicts class 0

The coefficients represent the "importance" of each feature.

Logistic regression offers a straightforward method of looking at the coefficients of the features to determine its weight on the outcome.

DEEPER MODEL INTERPRETATION – SHAP

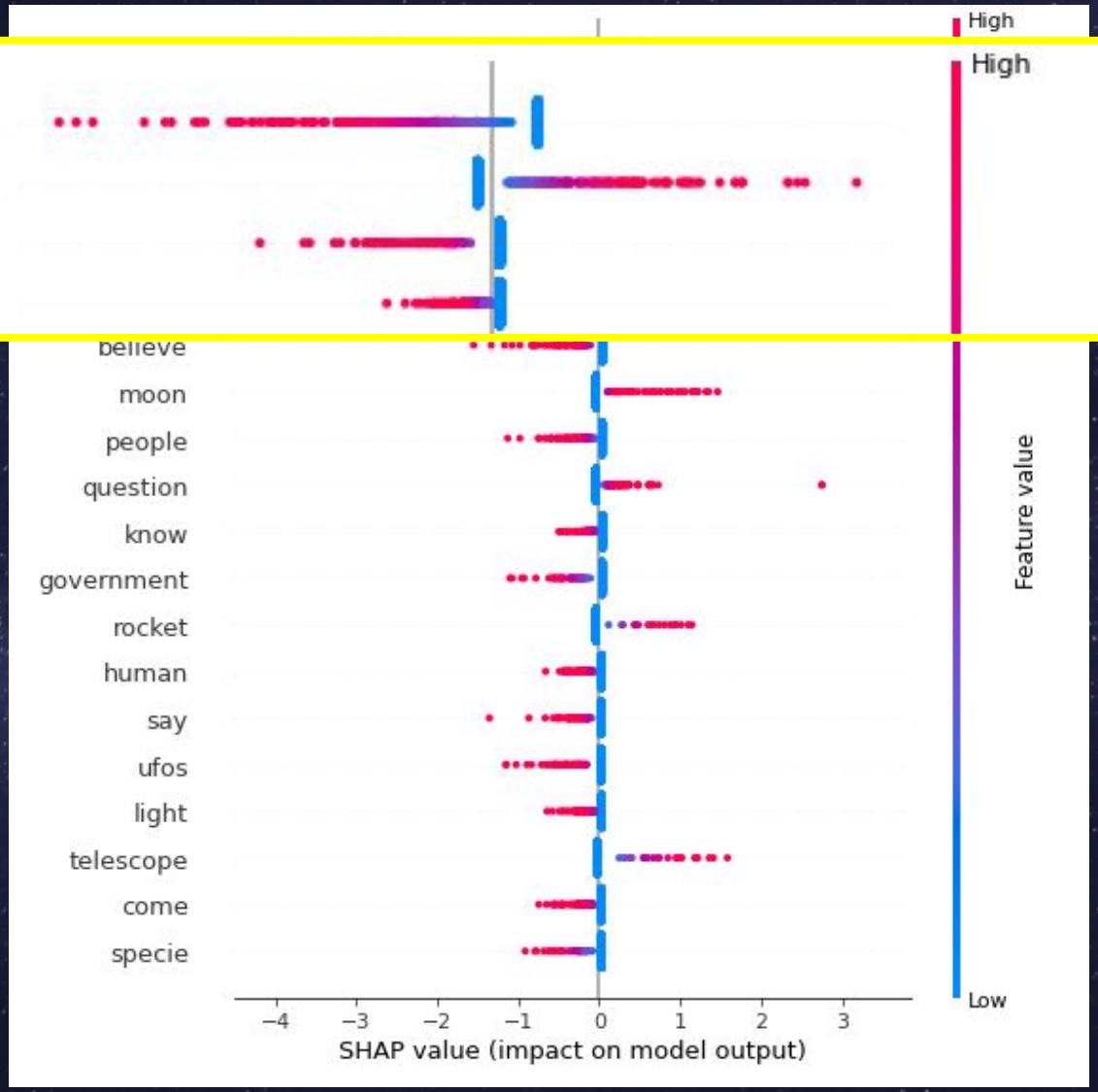


Positive SHAP = Space
Negative SHAP = Aliens

1. Features are ranked in descending order of their importance

2. SHAP's summary plot clearly shows the high impact of words like "alien", "ufo" on classification decision to "Aliens" subreddit.

DEEPER MODEL INTERPRETATION – SHAP



Positive SHAP = Space
Negative SHAP = Aliens

3. This plot also shows the positive and negative correlation “alien” has on both “Aliens” and “Space” subreddit.

DEEPER MODEL INTERPRETATION – SHAP

1. Local interpretability
   a. SHAP allows model to be interpreted locally on each row of the dataframe.
   b. Looking at this data, the model made the decision to classify it as “Aliens” based on the high contribution by “remember”, “alien”, “claim”.

CONCLUSION AND RECOMMENDATION

>90% accuracy & F1 score

CONCLUSION AND RECOMMENDATION

Icon made by noomtah from www.flaticon.com

FUTURE WORK

Image Analysis

Sentiment Analysis

Topic Modelling

THANK YOU

full mark for Group 3 pls, thanks