

PROJECT 3: CLASSIFYING SUBREDDITS

Group 3

GROUP 3 ALIENS MEMBERS



Darren Tu

Cheng Yeow

Zavier Soon

Thien Sean

Chee Tzen 2

OUR LIST

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

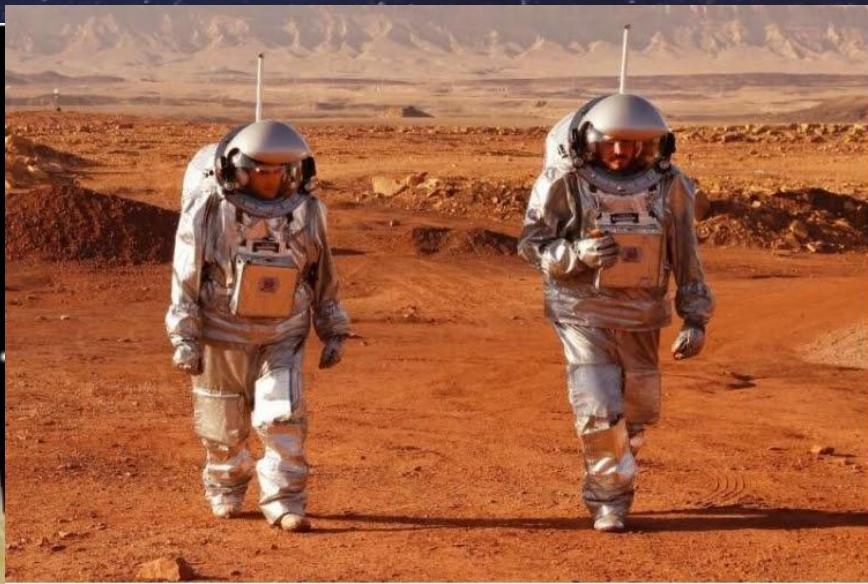
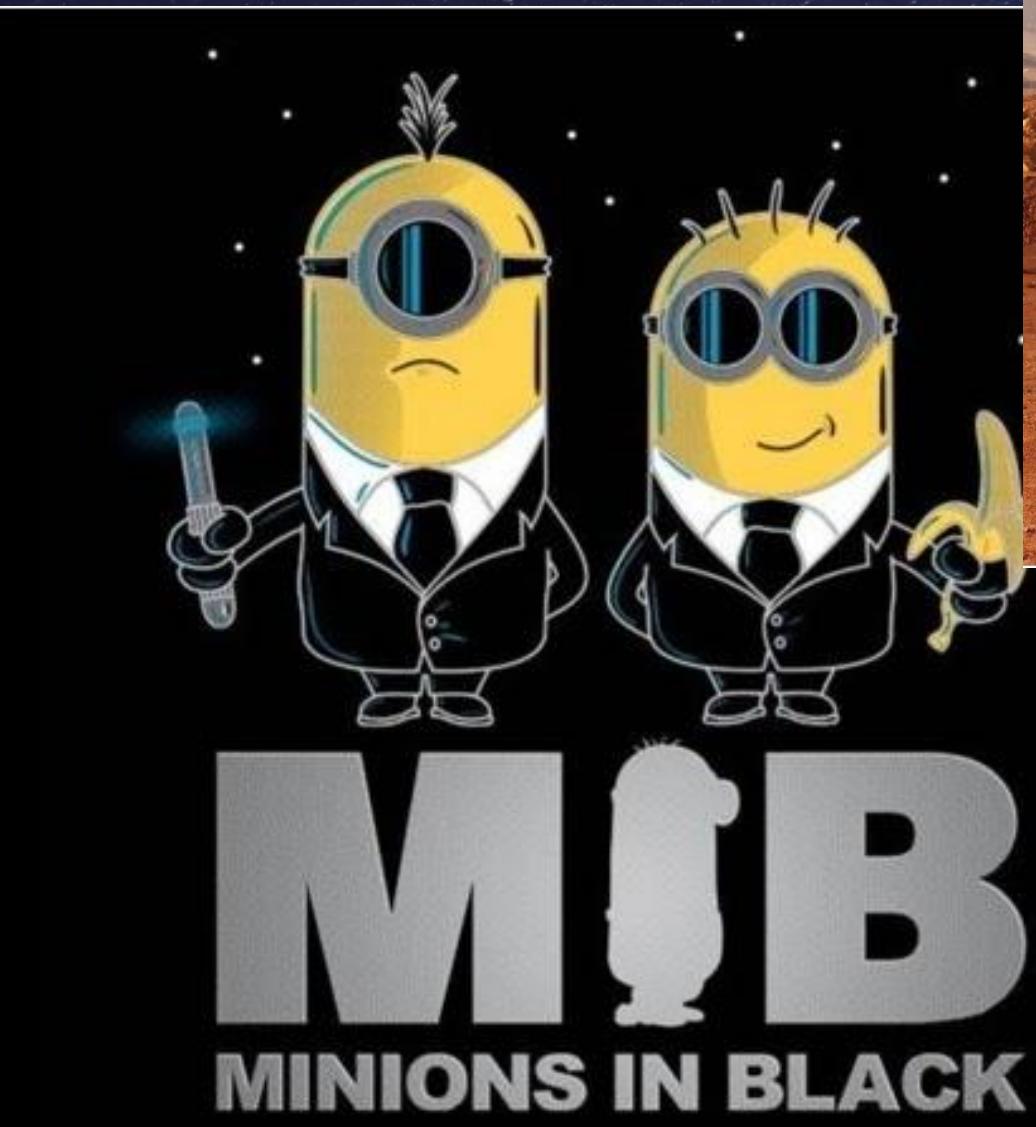
PROBLEM STATEMENT



PROBLEM STATEMENT



PROBLEM STATEMENT



Alien



GOAL OF THE PROJECT

Space



Alien



GOAL OF THE PROJECT

Space



↑ r/cats · Posted by u/Marthy_Mc_Fly 3 days ago 8693 8693

Don't take random cats home!!!

Advice

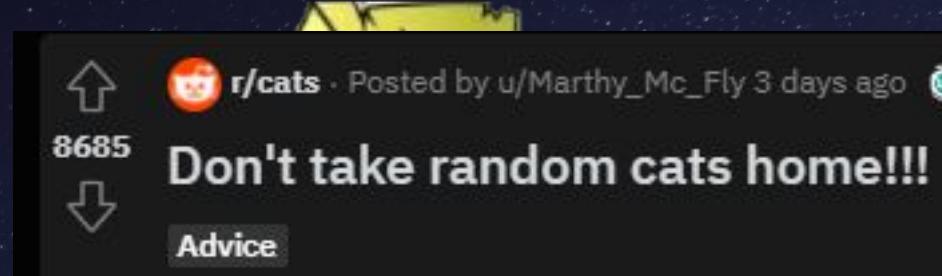
I'm seeing a lot of post of people taking or considering taking cats that come up to them home. Maybe



Alien

GOAL OF THE PROJECT

Space

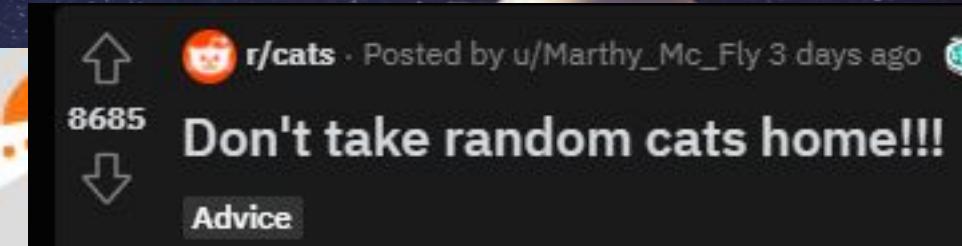


Alien



GOAL OF THE PROJECT

Space



Alien

WORLD 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)

The screenshot shows the homepage of the r/aliens subreddit. At the top right is a green header bar with the text "About Community". Below it is a large, colorful nebula background image. On the left, there's a black alien head icon, the subreddit name "Aliens", a "Join" button, and the URL "r/aliens". To the right of the main content area is a sidebar with the text "About Community" and a description: "A community dedicated to discussion of the possibility of extraterrestrial life. This is a moderated space to talk about various theories, sightings, analyses, and much more regarding EBEs and life outside of our home." Below this are two statistics: "397k" members and "307" posts. A banner at the bottom right features the letters S, P, A, C, E in a stylized font, with a central hexagonal pattern and the text "Celebrating the J.W.S.T".

About Community

A community dedicated to discussion of the possibility of extraterrestrial life. This is a moderated space to talk about various theories, sightings, analyses, and much more regarding EBEs and life outside of our home.

397k 307

Members Posts

Created Jan 26, 2008

Join

Aliens

r/aliens

About Community

Share & discuss informative content on: * Astrophysics * Cosmology * Space Exploration * Planetary Science * Astrobiology

19.6m Members 1.9k Online

Created Jan 26, 2008

/r/space: news, articles and discussion

Join

DATA CLEANING

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



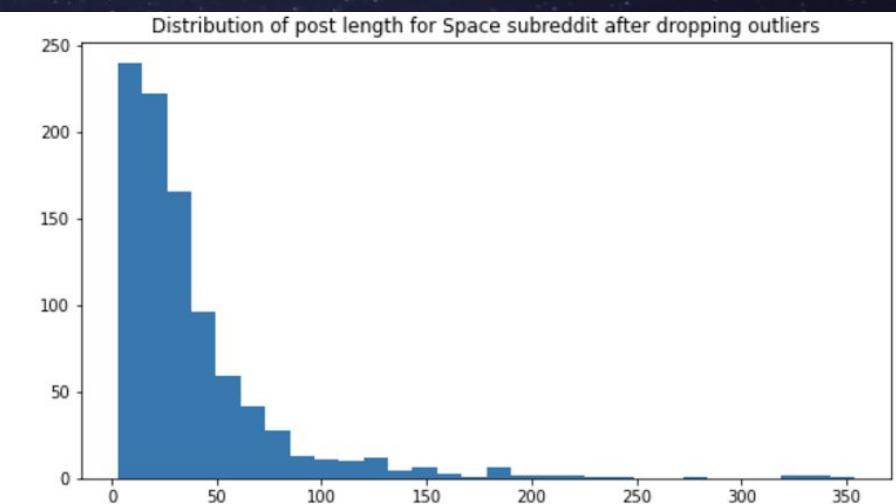
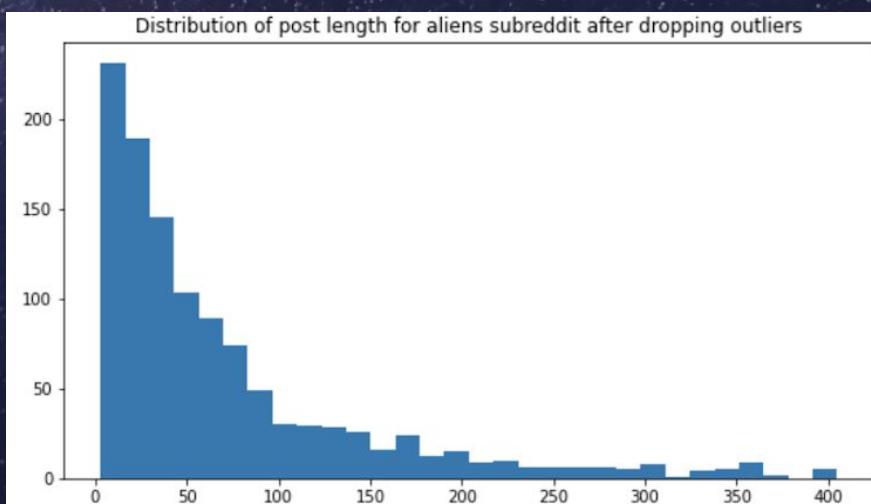
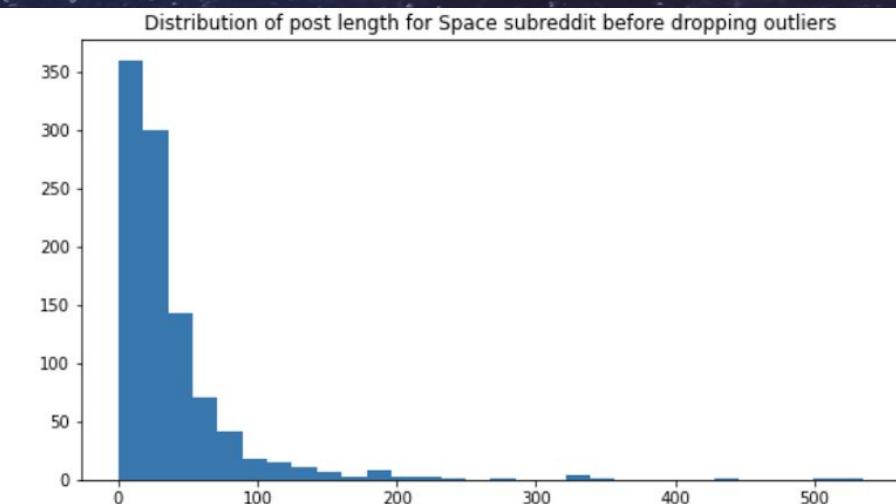
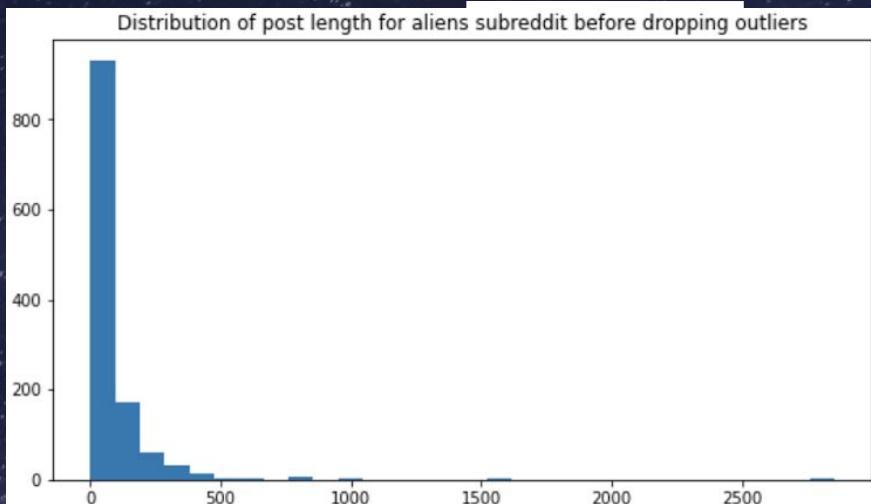
OUTLINERS

Lower Band

Remove words of 2 and lower

Upper Band

Remove words of 400 and above



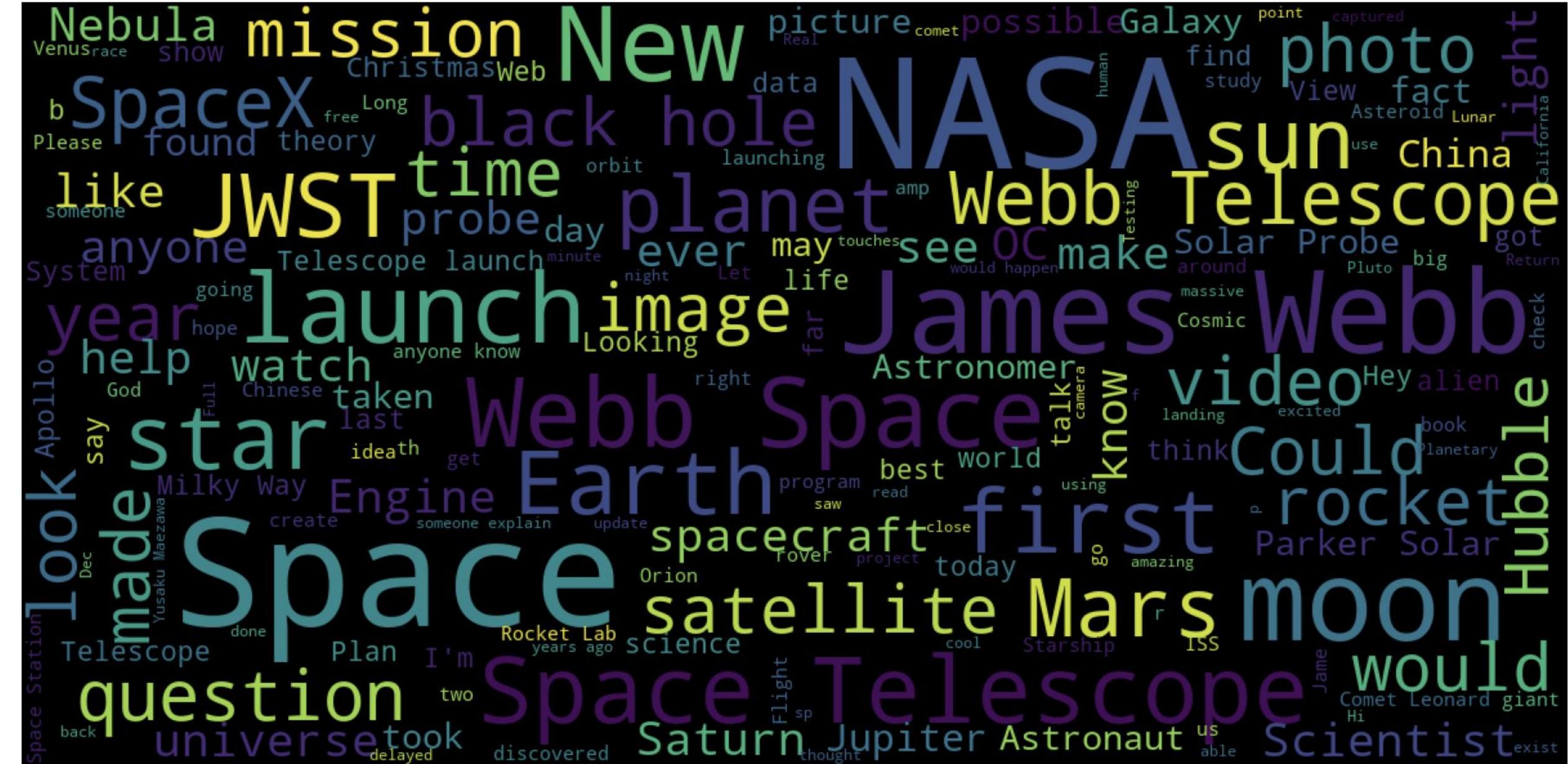
Word Cloud for Aliens



A word cloud generated from text about aliens and extraterrestrial life. The most prominent word is "Alien" in large blue letters. Other significant words include "UFO", "believe", "people", "look", "video", "contact", "interview", "earth", "life", and "think". The words are colored in various shades of green, blue, and purple, and some have small descriptive labels such as "Project" or "Galaxy" next to them. The background is white, and the overall image is a dense cluster of words.

Word cloud for Aliens, showing the most frequent terms related to aliens and extraterrestrial life. The words are color-coded by frequency, with larger and more prominent words being the most common. The most frequent word is "Alien". Other notable words include "UFO", "believe", "people", "look", "video", "contact", "interview", "earth", "life", and "think". Some words have smaller descriptive labels, such as "Project" or "Galaxy" next to them. The background is white, and the overall image is a dense cluster of words.

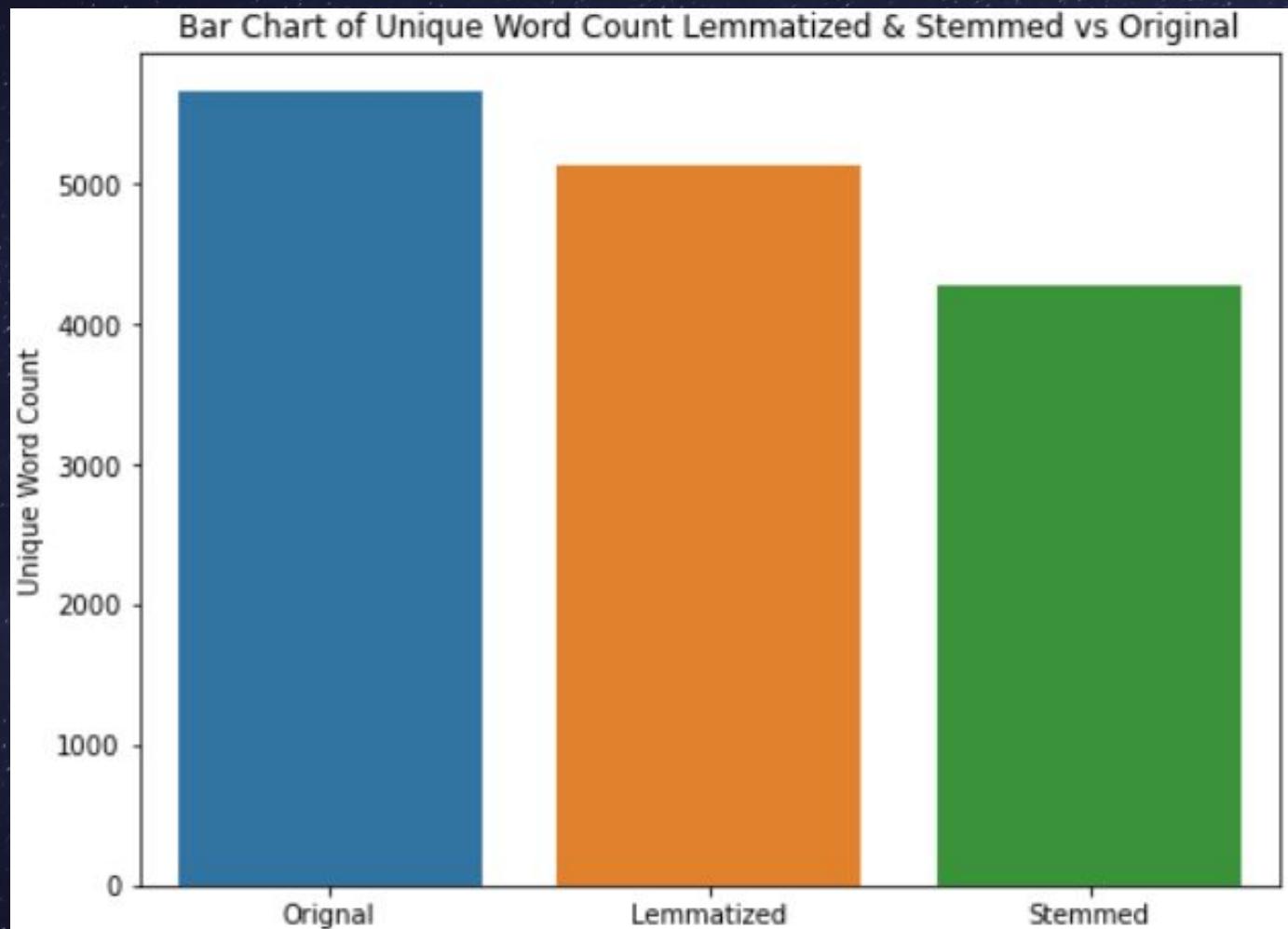
Word Cloud for Space



Preprocessing

- Methods of preprocessing
 - Lemmatization / Stemming
 - Stop Words

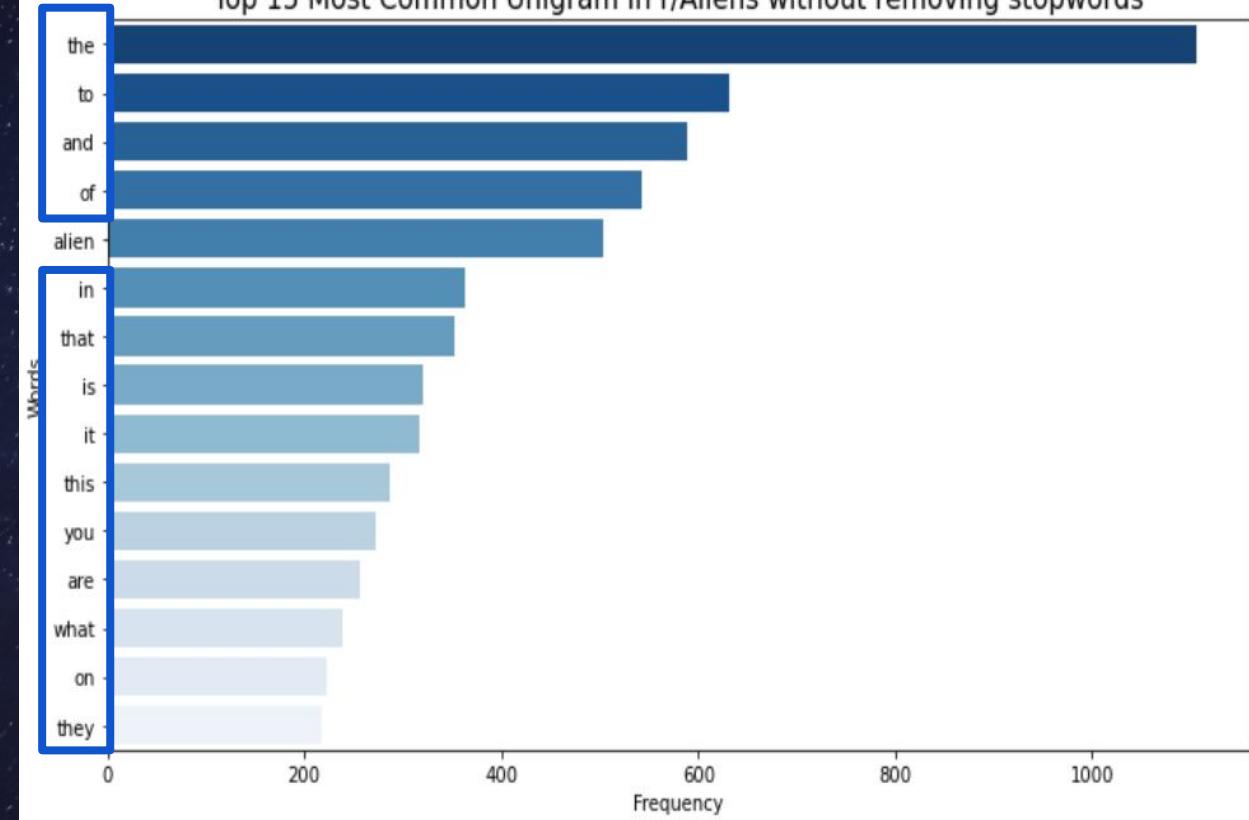
Lemmatization / Stemming



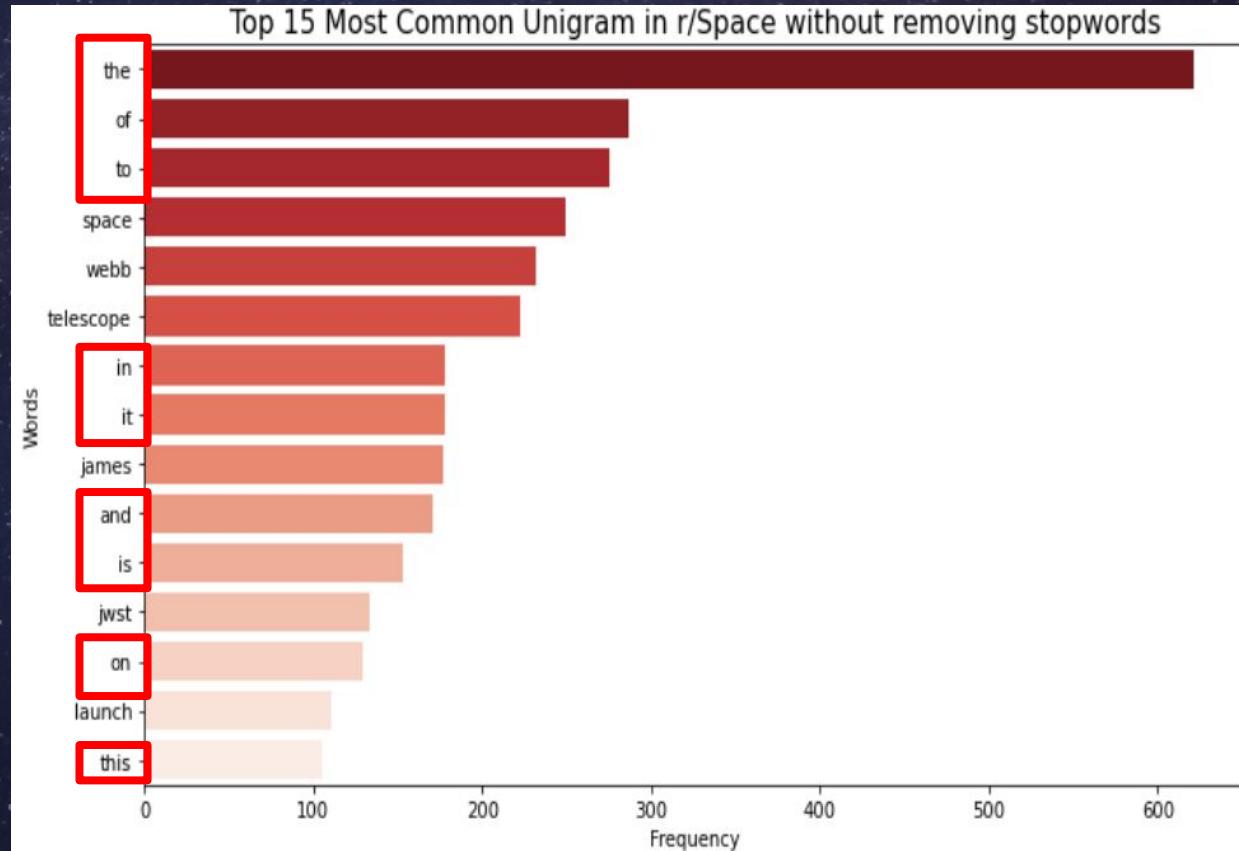
	Lemmatized	Stemmed
0	discovery	discoveri
1	advanced	advanc
2	intelligent	intellig
3	anthropocentric	anthropocentr
4	religious	religi
5	people	peopl
6	alike	alik
7	discussion	discuss
8	this	thi
9	existing	exist
10	automatically	automat
11	disprove	disprov
12	a	as
13	perspective	perspect

Stopwords

Top 15 Most Common Unigram in r/Aliens without removing stopwords



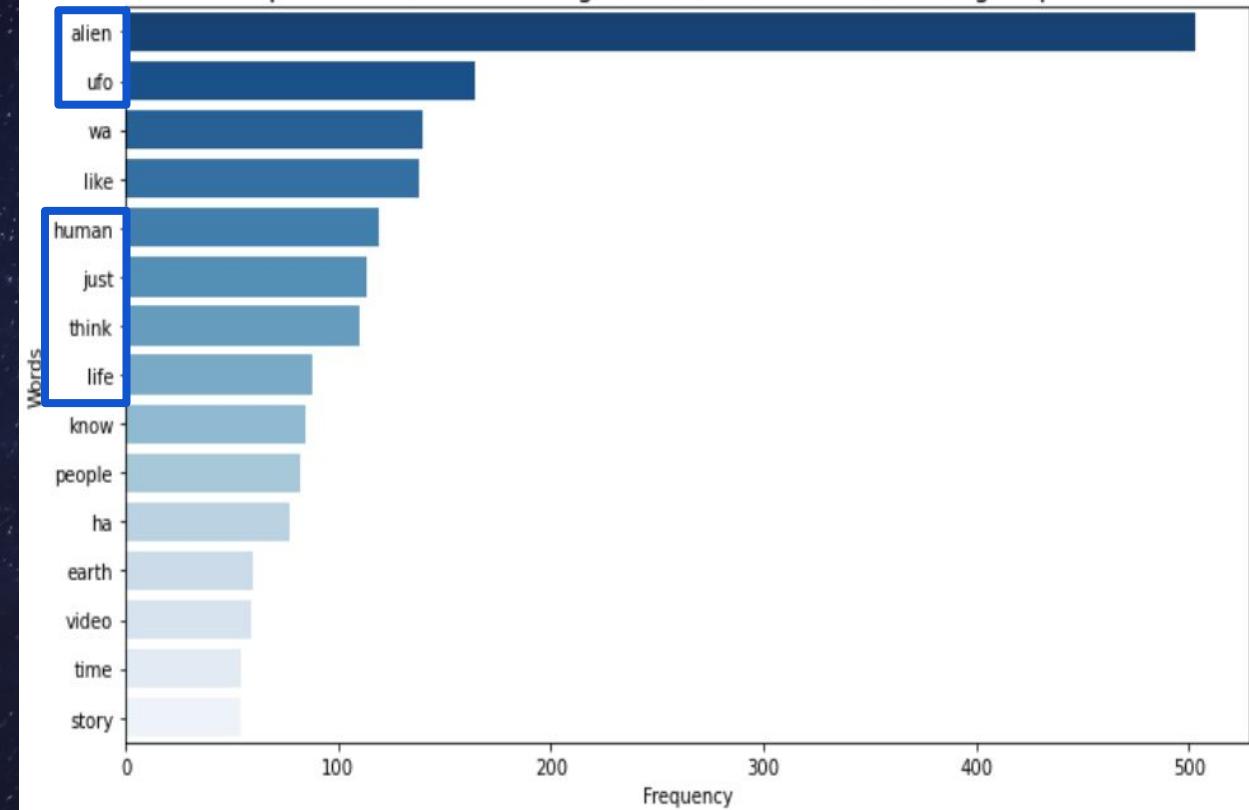
Top 15 Most Common Unigram in r/Space without removing stopwords



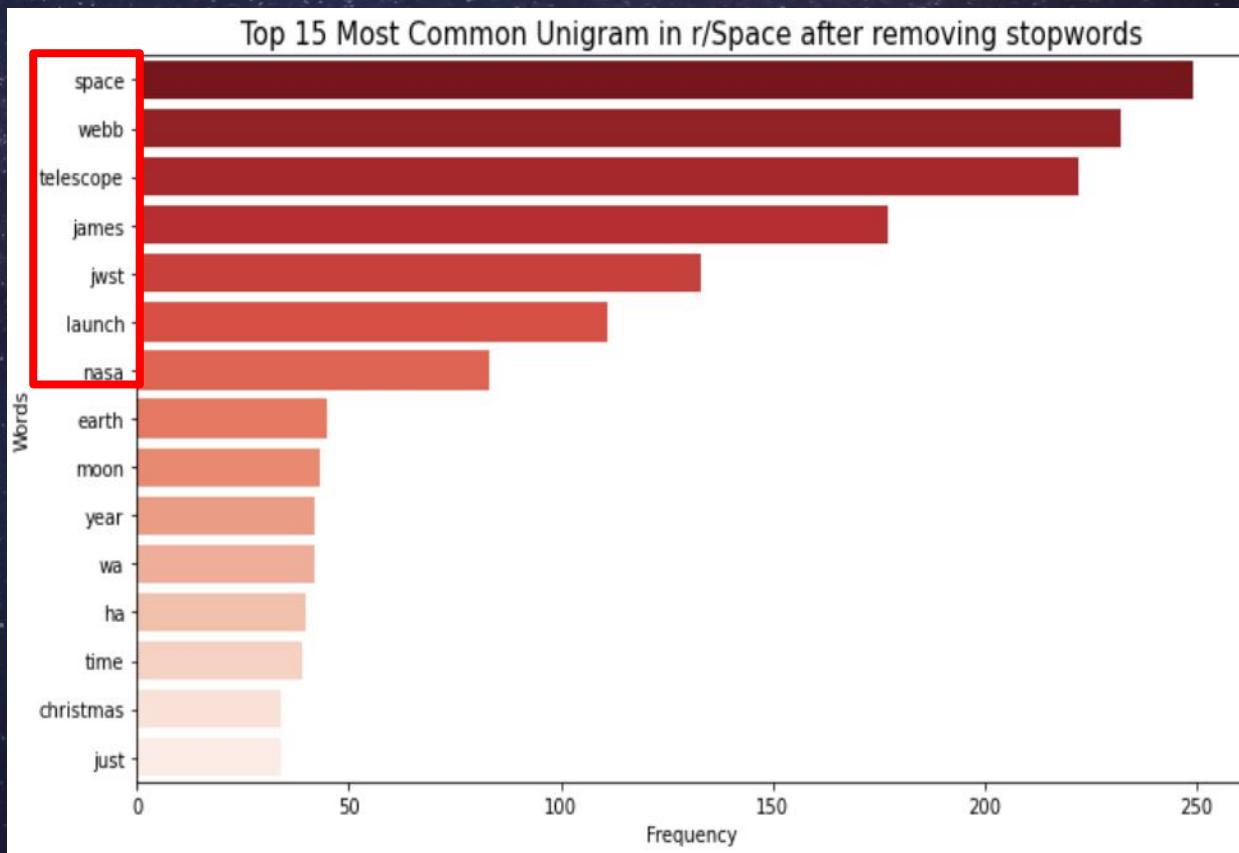
Stop words are words that do not provide any valuable information. They are common words that only add to the grammatical structure and flow of the sentence.

Stopwords

Top 15 Most Common Unigram in r/Aliens after removing stopwords

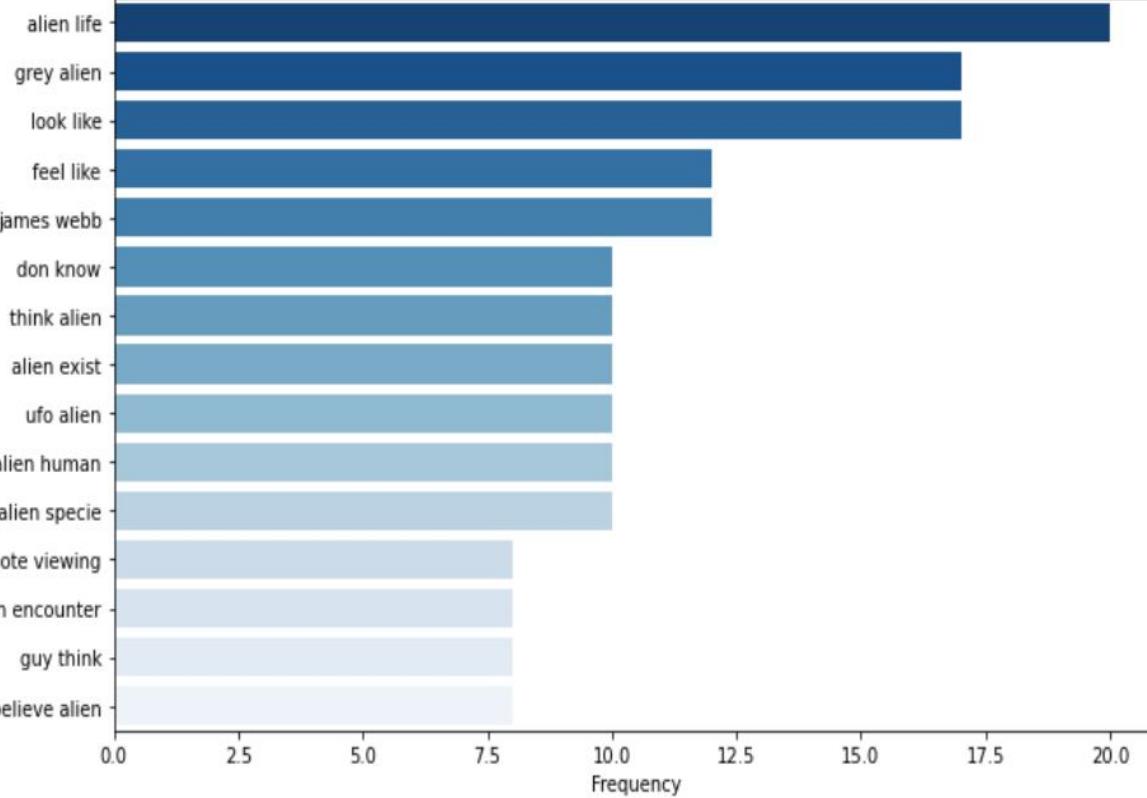


Top 15 Most Common Unigram in r/Space after removing stopwords

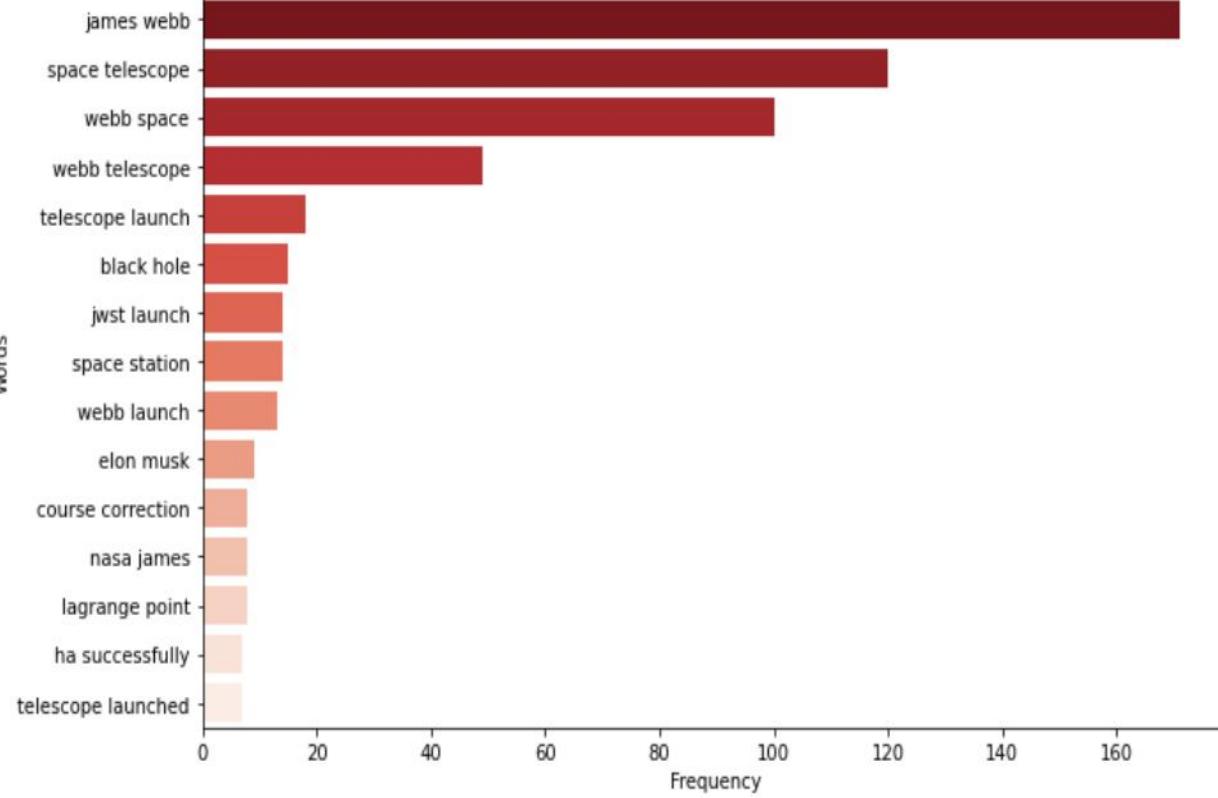


Stopwords

Top 15 Most Common Bigrams in r/Aliens after removing stopwords



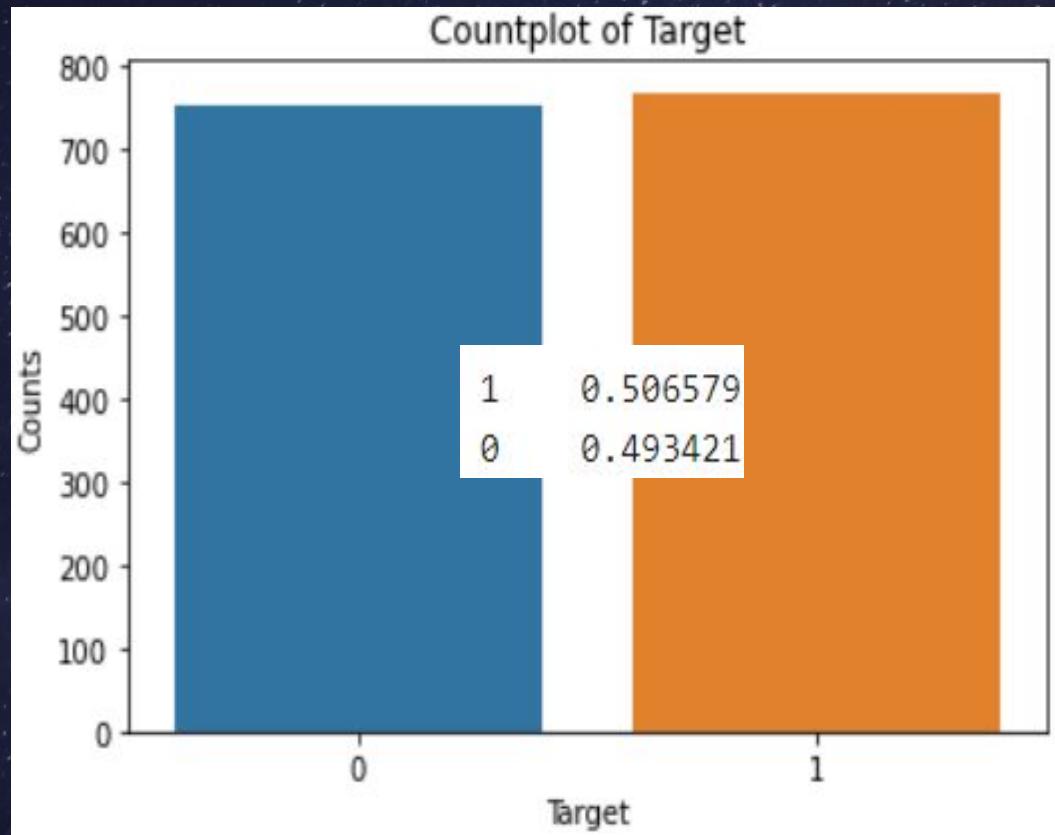
Top 15 Most Common Bigrams in r/Space after removing stopwords



Choice of Model

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

Baseline Model

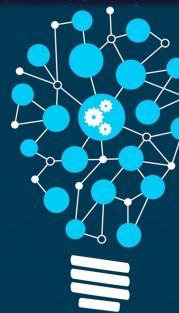


r/Aliens = 1

r/Space = 0

MODEL EVALUATION

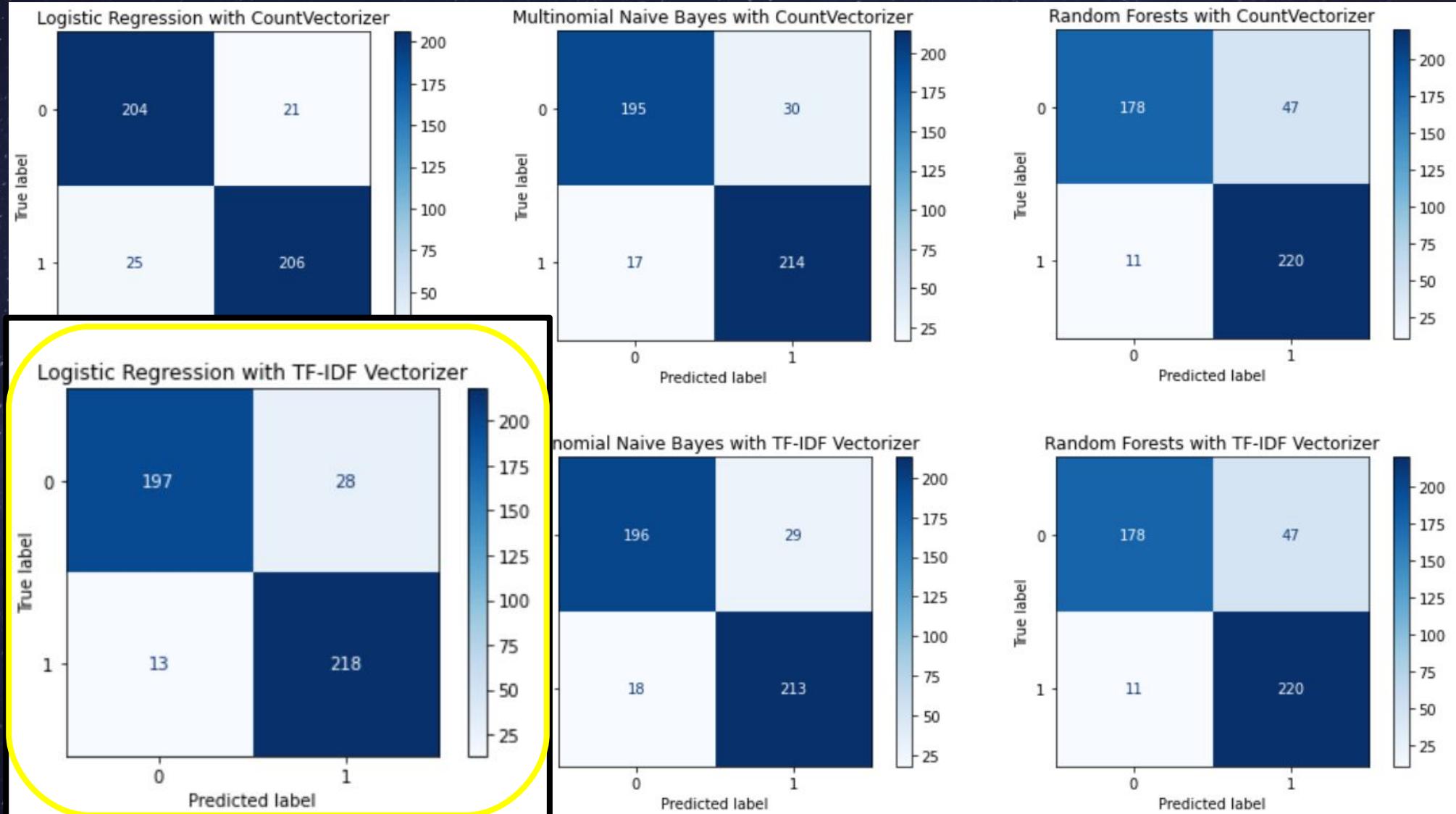
MACHINE
LEARNING



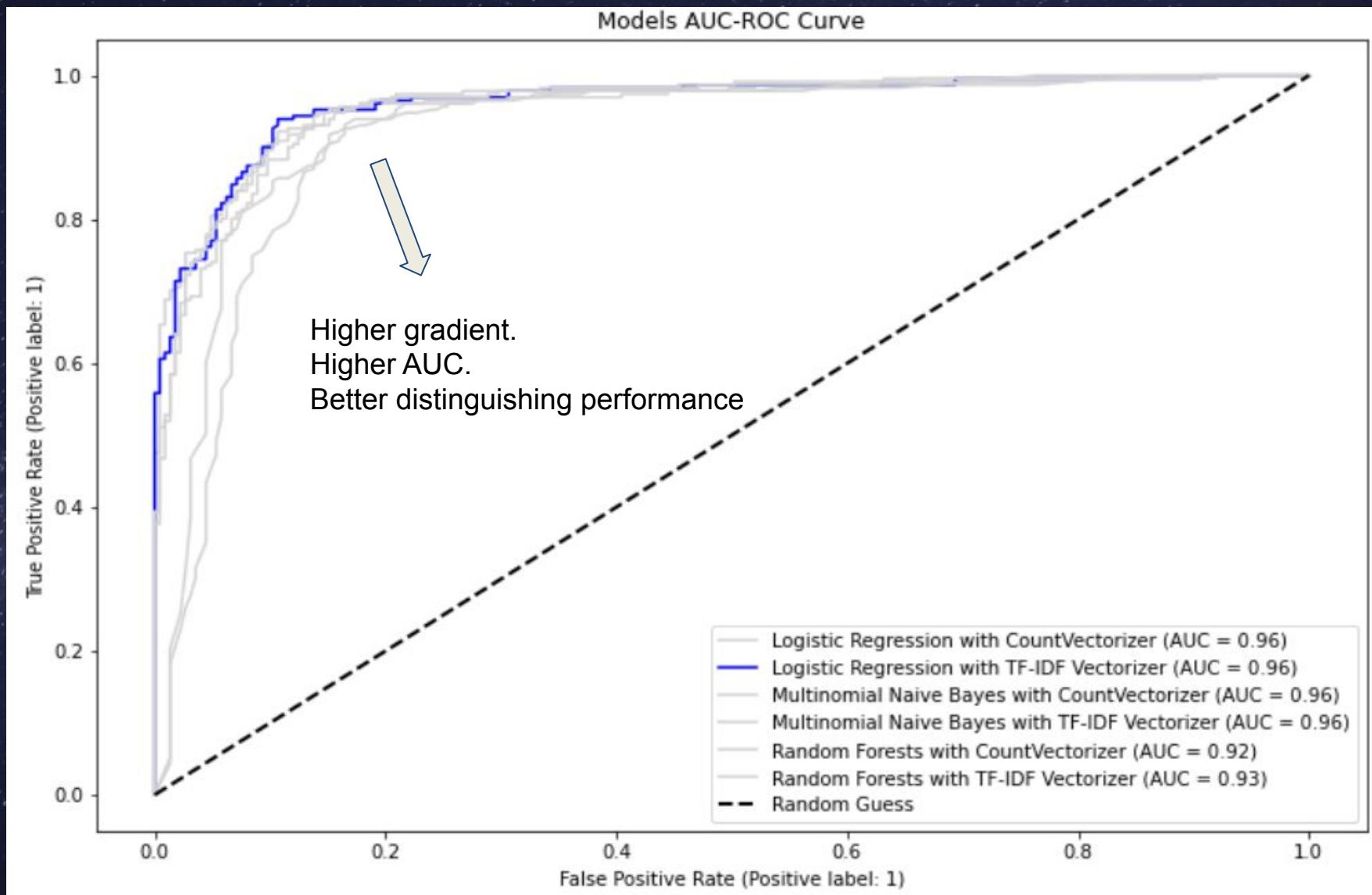
Model	Vectorizer	Train Score	Test Score	F1-score
Logistic Regression	CountVectorizer	0.9774	0.8991	0.8996
Logistic Regression	TfidfVectorizer	0.9718	0.9101	0.9140
MultinomialNB	CountVectorizer	0.9718	0.8969	0.9011
MultinomialNB	TfidfVectorizer	0.9793	0.8969	0.9006
Random Forest Classifier	CountVectorizer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

Higher interpretability, lesser overfitting, less processing power.

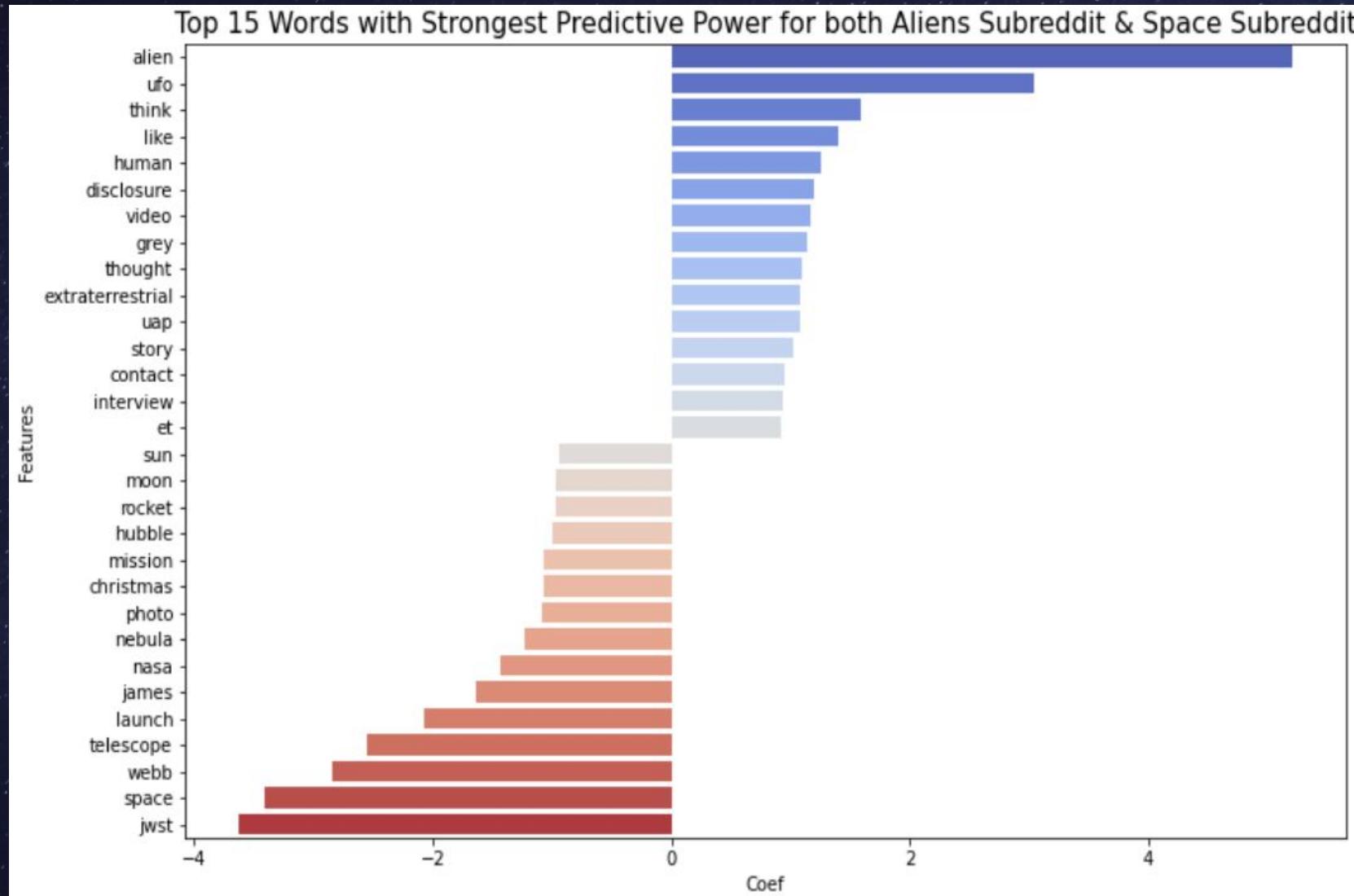
MISCLASSIFICATION



AUC - ROC CURVE



MODEL EVALUATION – LOGISTIC REGRESSION - TFIDF VECTORIZER



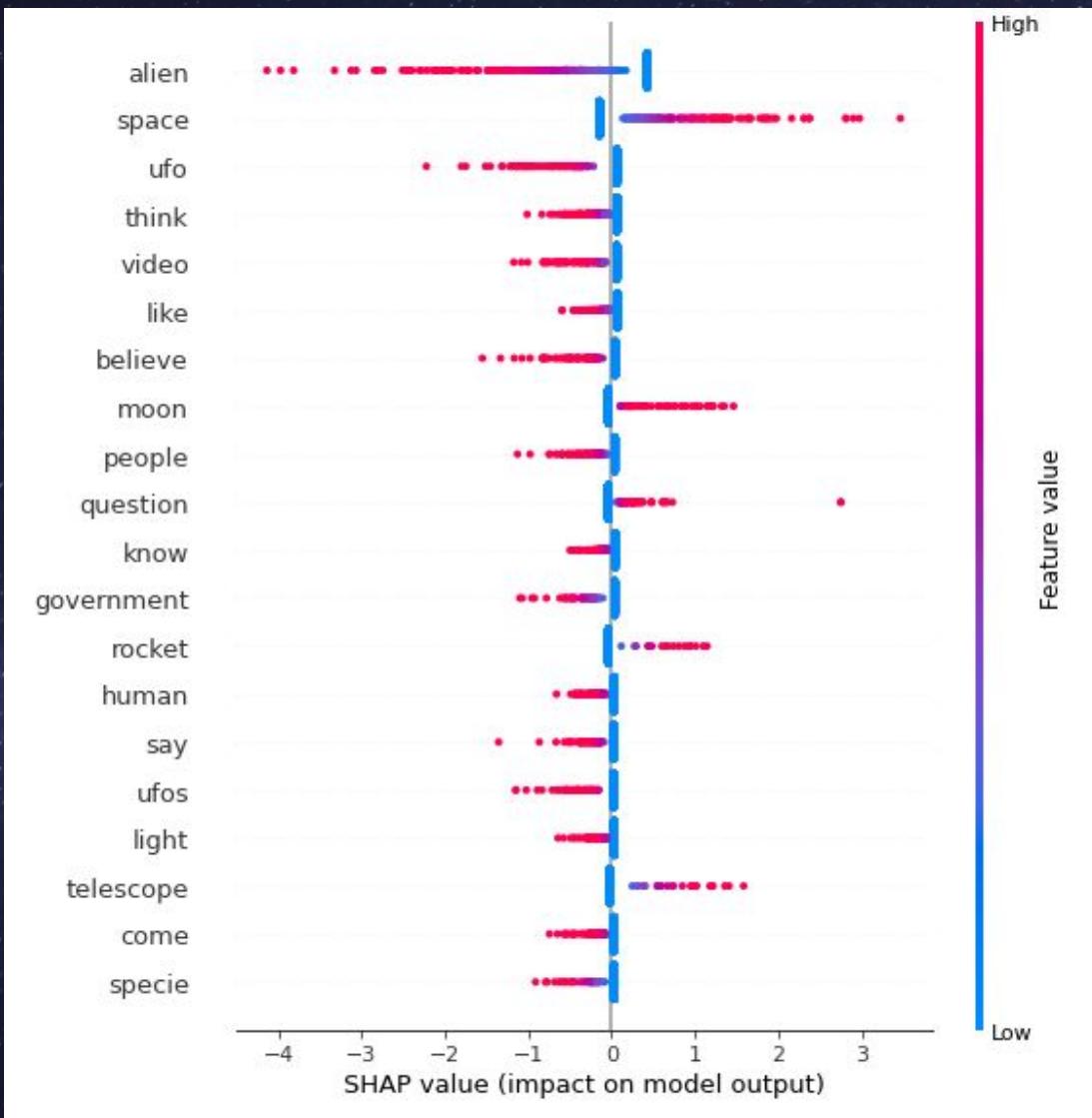
Positive coef= Feature that predicts class 1

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.

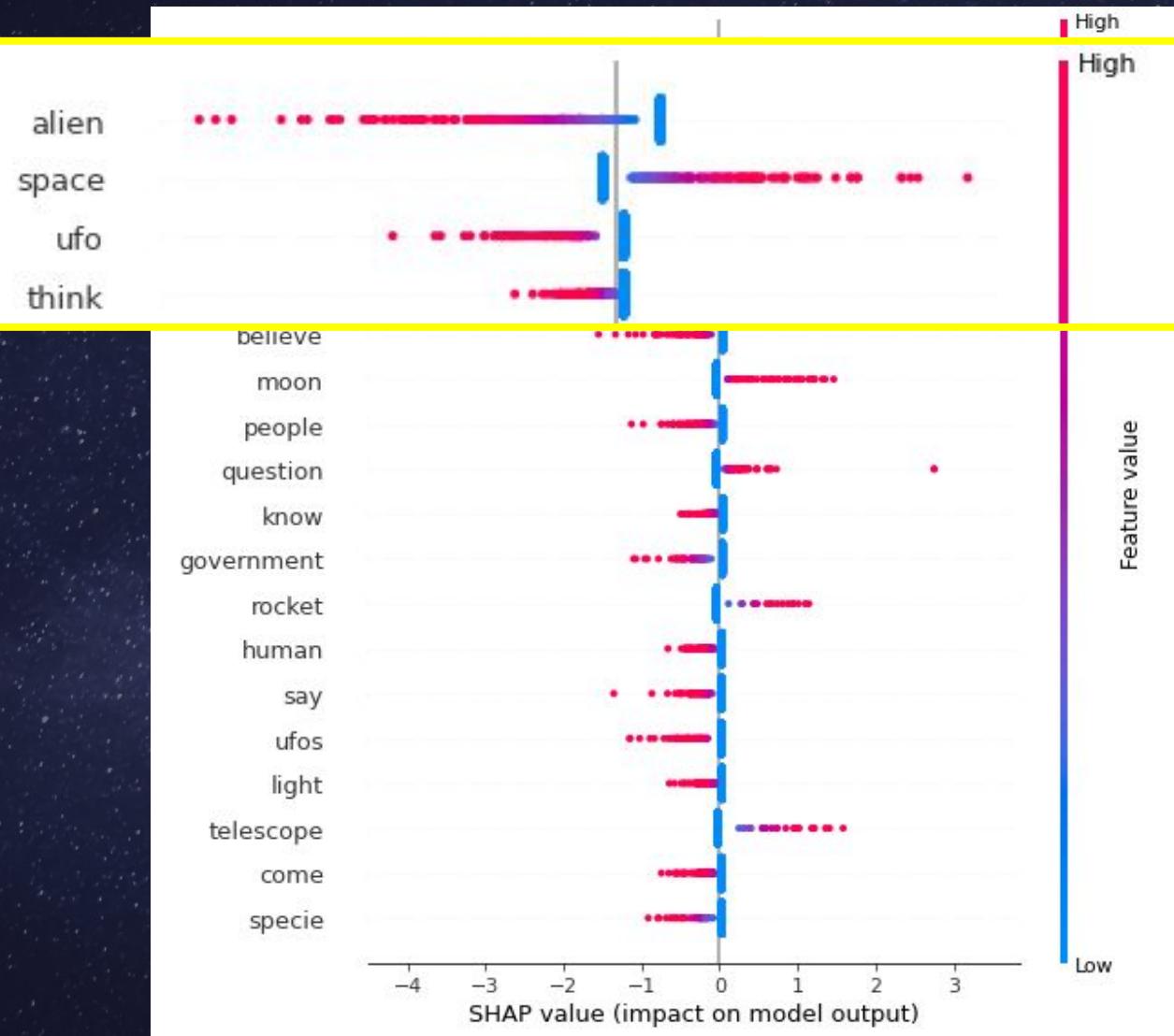
DEEPER MODEL INTERPRETATION – SHAP



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

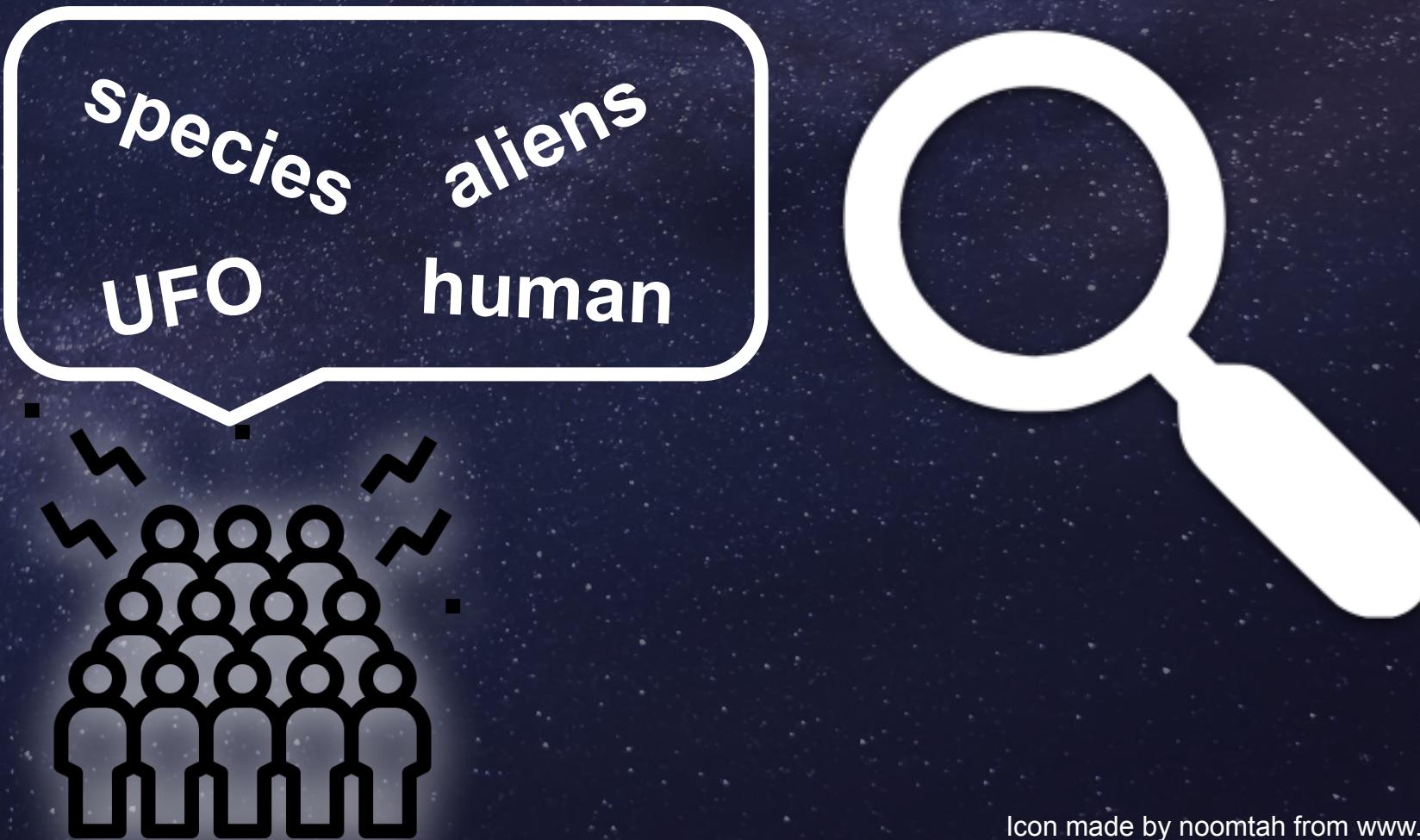


1. Local interpretability
 - a. SHAP allows model to be interpreted locally on each row of the dataframe.
 - b. Looking at this data, the model made the decision to classify it as “Aliens” based on the high contribution by “remember”, “alien”, “claim”.

CONCLUSION AND RECOMMENDATION

>90% accuracy & F1 score

CONCLUSION AND RECOMMENDATION



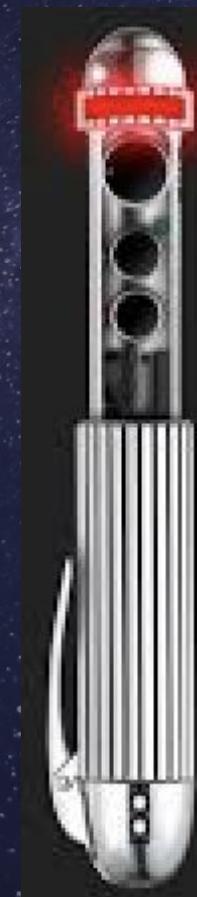
FUTURE WORK

Image Analysis

Sentiment Analysis

Topic Modelling

THANK YOU



full mark for Group 3 pls, thanks

34