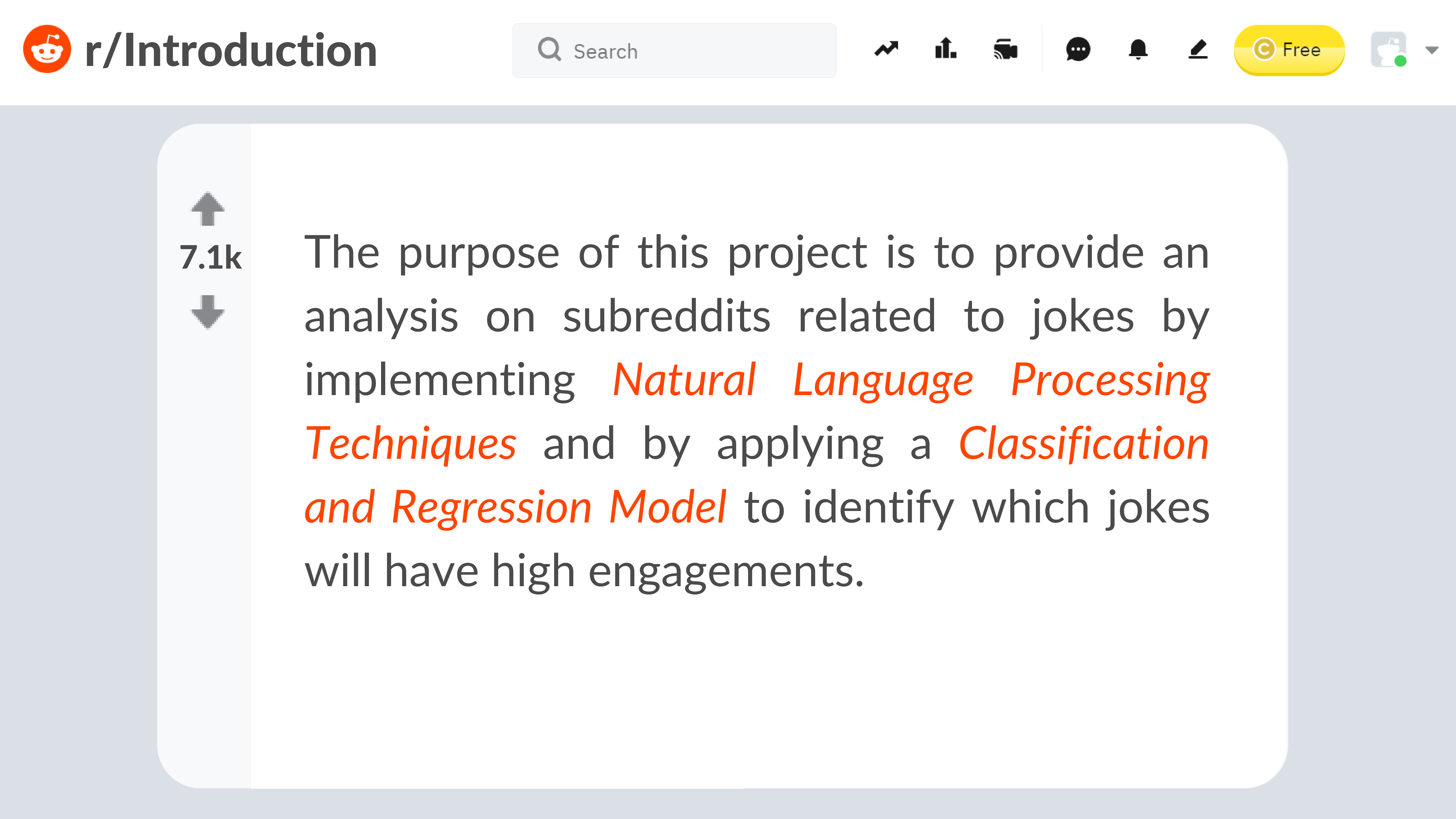
A subreddit analysis on Jokes

TEAM SEPTIC TANK

Edward, Jona, Nico, Rara, Tyron



The purpose of this project is to provide an analysis on subreddits related to jokes by implementing *Natural Language Processing Techniques* and by applying a *Classification and Regression Model* to identify which jokes will have high engagements.

r/Reddit

Search

Free

7.1k

r/What is Reddit?

Reddit is a network of ***communities*** based on people's interests.

Forums on the Reddit platform are called ***subreddits***. Users can use a ***subreddit*** to create new posts on a specific topic. These can be questions and requests for help, but also informative news articles, images, and videos.

r/Jokes

Search

Free

7.1k

r/But why choose Jokes?

Brands can reach their audience through *humor*.

The Influence of Humor Strength and Humor-Message Relatedness onAdMemorability: ADual ProcessModel

"The study shows the impact of humorous advertisements are more on ***recall of ads****, and* ***ad memorability****, the humor get attention and mood when the humor appeal is strong, the positive influence of mood created by humor and product relatedness made."*

Cline and Kellaris, 2007

r/Review of Related Literature

Search

Free

Gets people to listen - *"Let the Good Times Roll Building of a Fun Culture" David Stauffer, Harvard Management*

Increases long term memory retention - *"Relationship between Instructor and Student Learning" M Wanzer, Communication Education*

Helps communicate messages - *"Does Humor Use Enhance SpeakerEthos?" C Ellis Campbell, Association for Applied and Therapeutic Humor.*

Builds trust - *"Humor in the Workplace: A Communication Challenge", Robert A. Vartebedian, Speech Communication Association*

Improves likeability - *"A funny thing happened on the way to the bottom line", B.J. Avollo, Academy of Management Journal*

Home

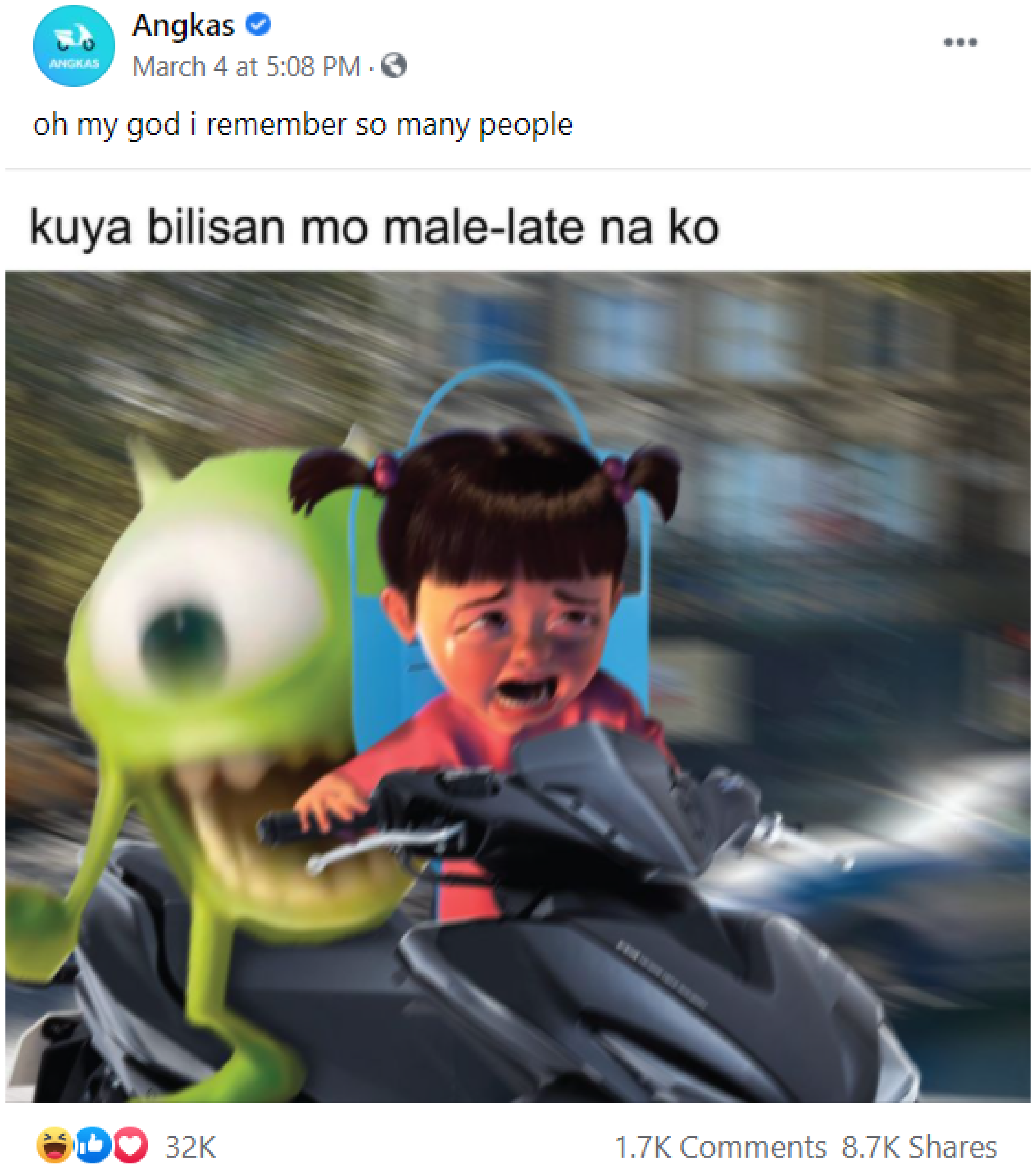
Search

Free

7.1k

r/Angkas

Angkas posted on Facebook last March 4, 2021. This post garnered a total engagements of 32k likes, 1.7k comments, and 8.7k shares



Angkas

March 4 at 5:08 PM ·

oh my god i remember so many people

kuya bilisan mo male-late na ko

32K

1.7K Comments 8.7K Shares

Home

Search

Free

8.1k

r/RC Cola Ph

RC Cola Philippines which tweeted its advertising video on Twitter last November 26, 2020. This tweet garnered a total engagements of 20.6k likes, 1.5 retweets, and 8.6 quote retweets.

r/Objectives

Search

Free

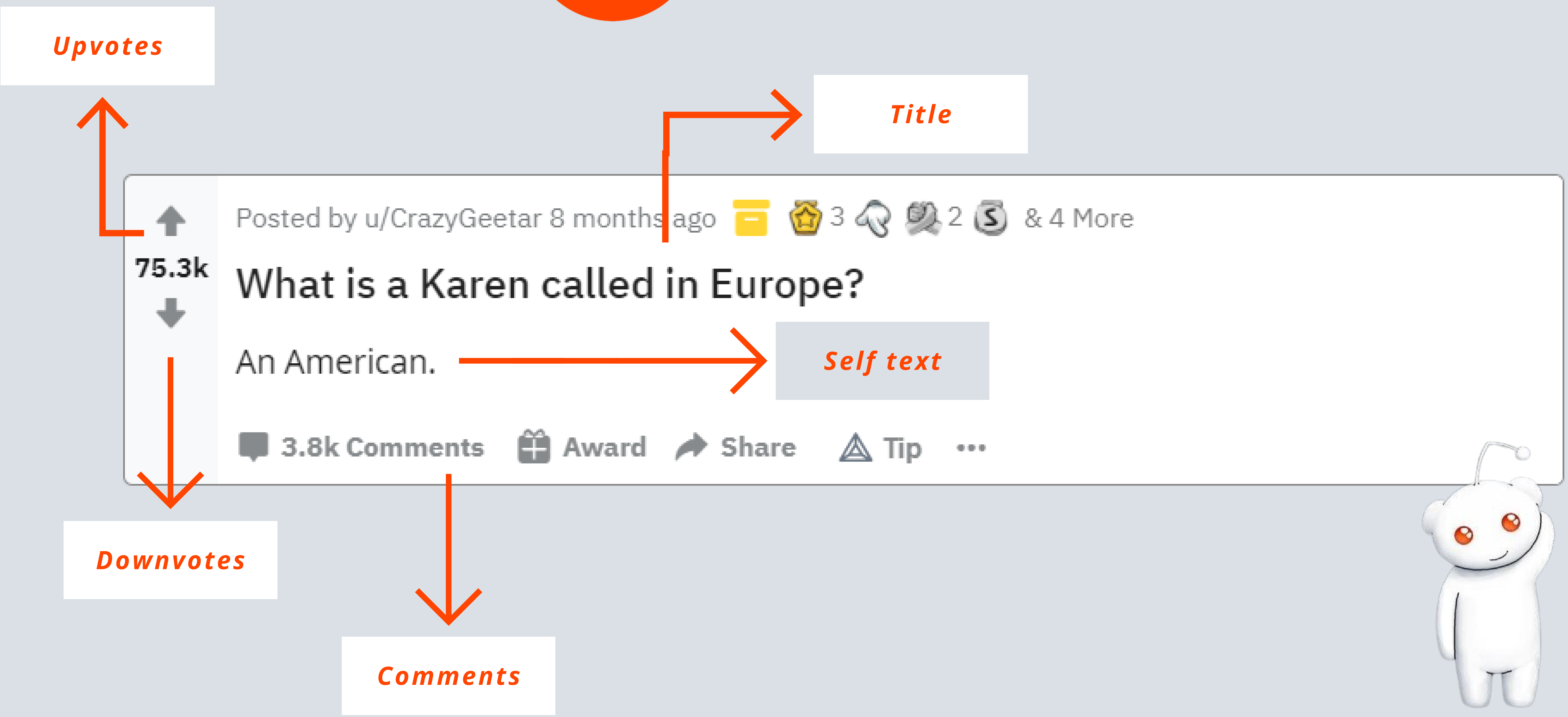
7.1k

- ***Analyze*** the joke subreddits by using Natural Language Processing Techniques
- ***Identify*** which jokes will have high engagements through a Classification and Regression model

r/Data Information

Search

Free



Posted by u/CrazyGeetar 8 months ago 3 2 & 4 More

75.3k

What is a Karen called in Europe?

An American.

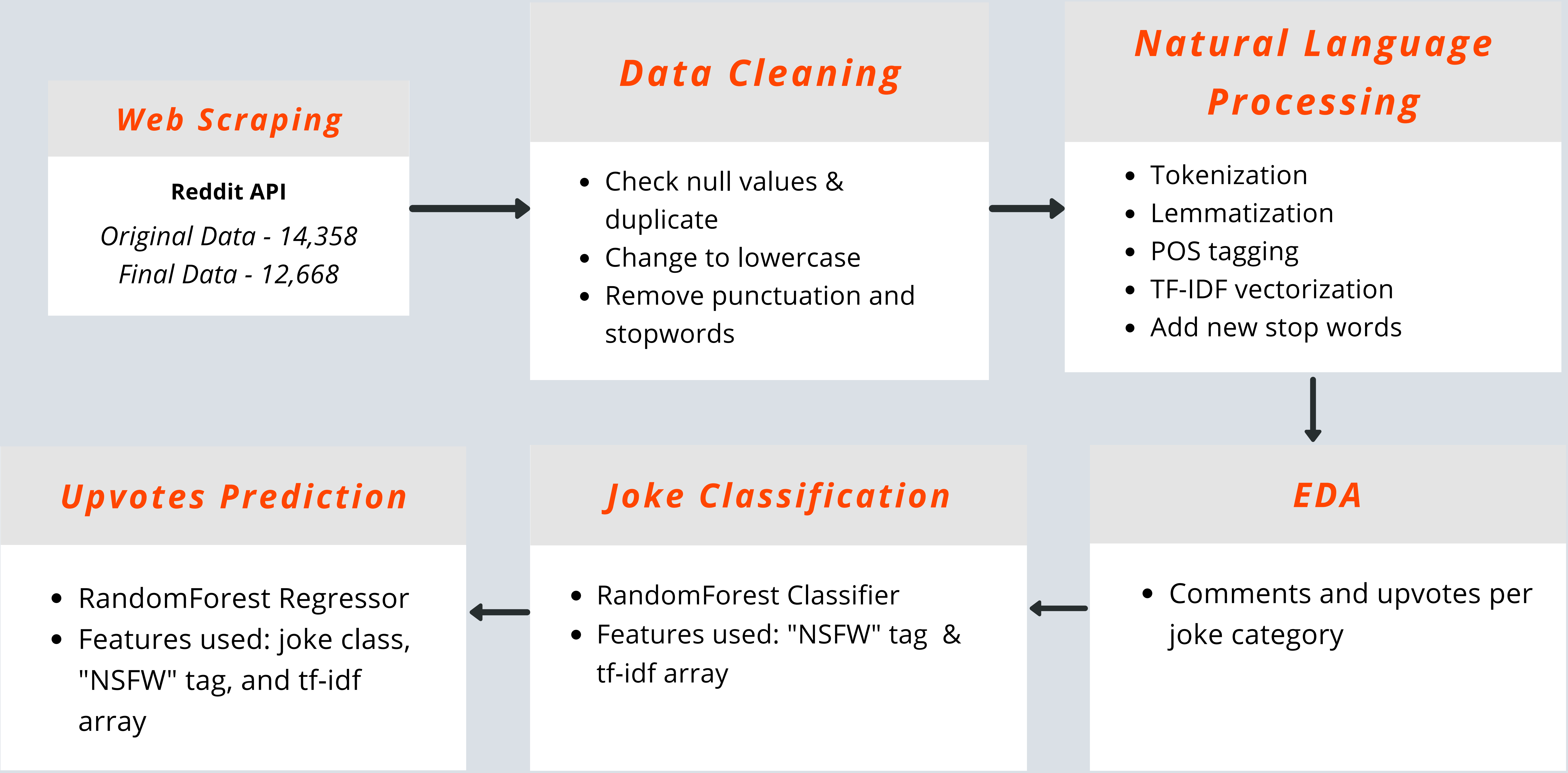
3.8k Comments Award Share Tip •••

r/Methodology

Search

Free

8.7k



Web Scraping

Reddit API

Original Data - 14,358
Final Data - 12,668

Data Cleaning

- Check null values & duplicate
- Change to lowercase
- Remove punctuation and stopwords

Natural Language Processing

- Tokenization
- Lemmatization
- POS tagging
- TF-IDF vectorization
- Add new stop words

EDA

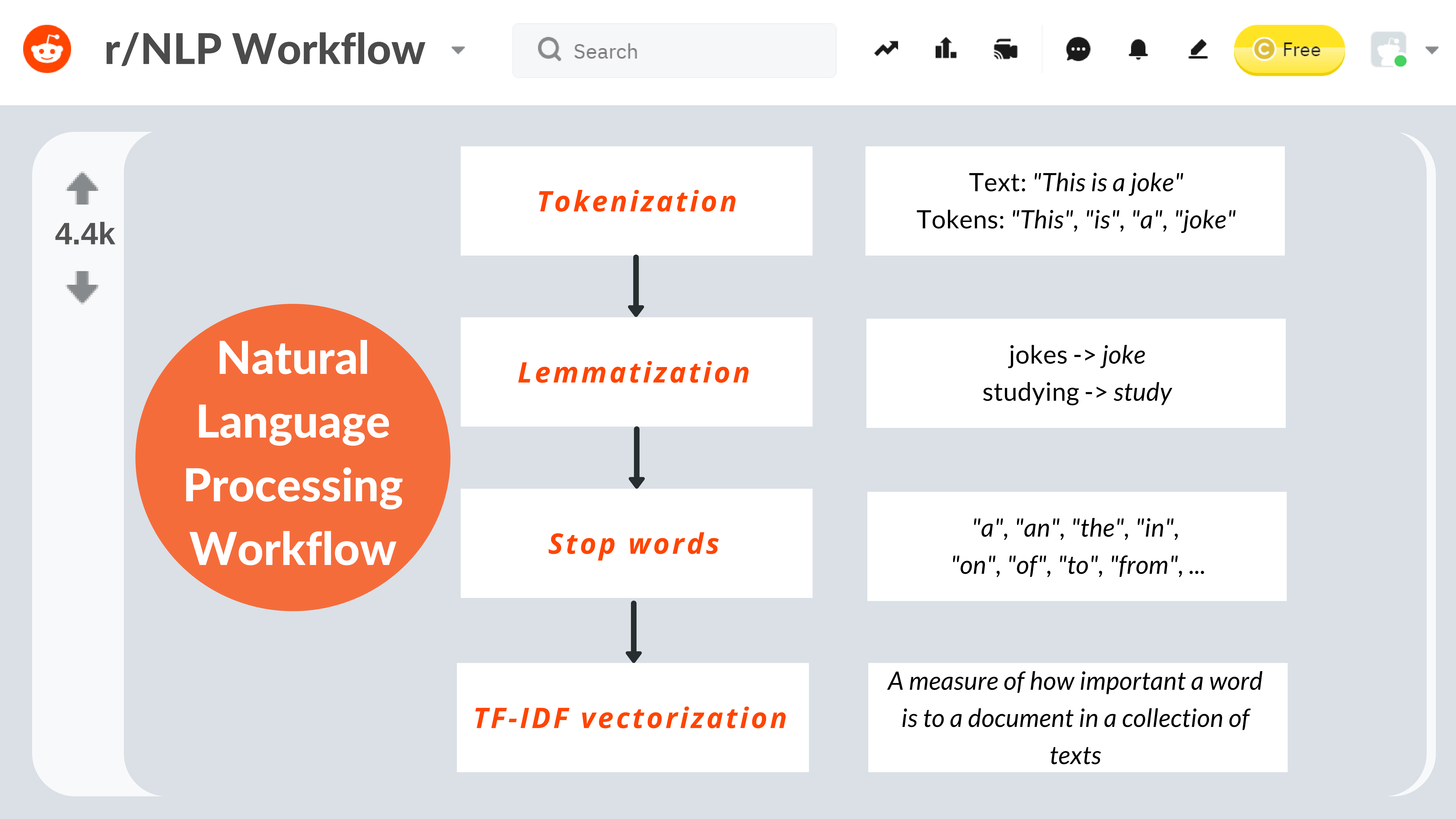
- Comments and upvotes per joke category

Joke Classification

- RandomForest Classifier
- Features used: "NSFW" tag & tf-idf array

Upvotes Prediction

- RandomForest Regressor
- Features used: joke class, "NSFW" tag, and tf-idf array



Natural Language Processing Workflow

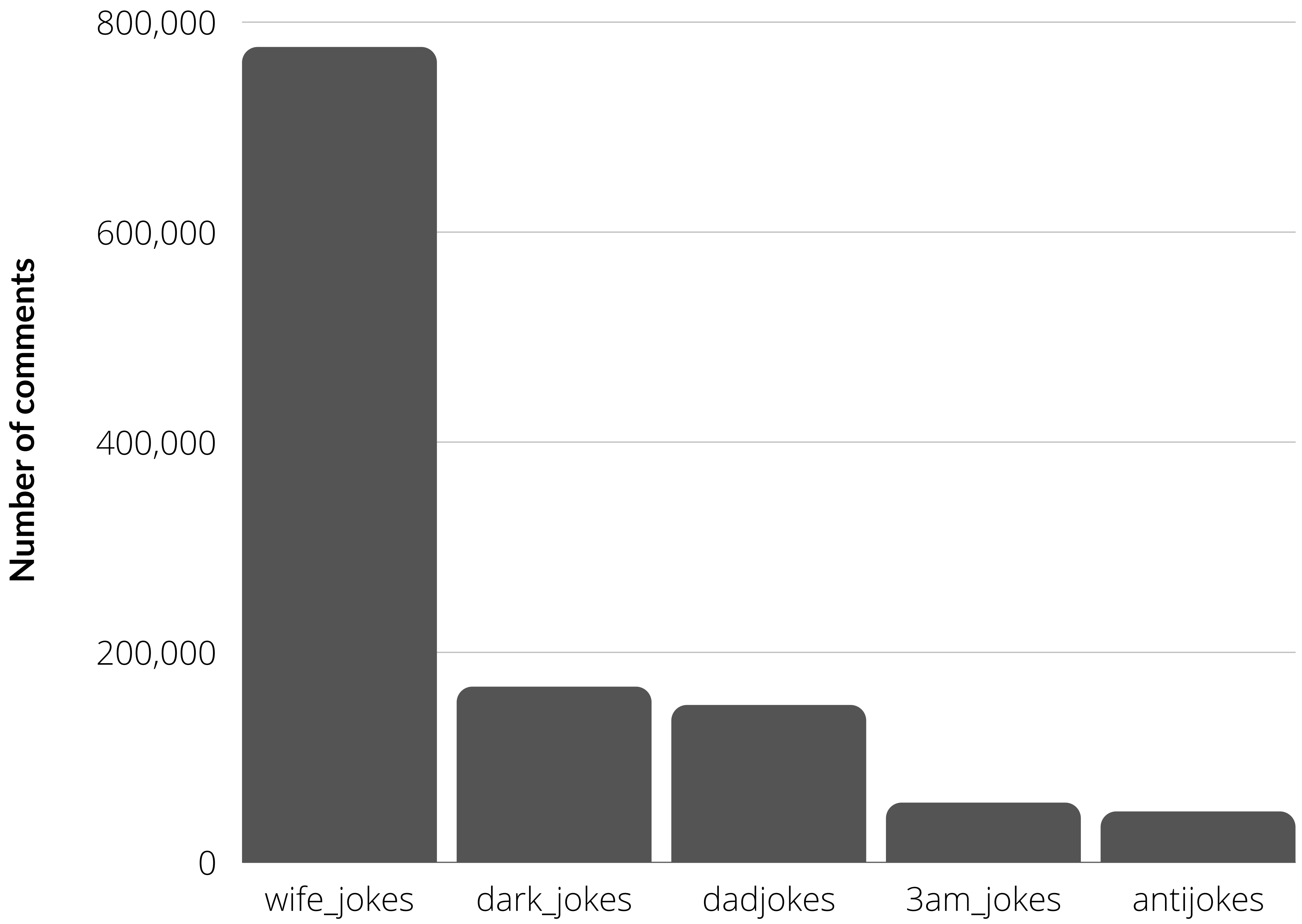
Tokenization	Text: *"This is a joke"* Tokens: *"This", "is", "a", "joke"*
Lemmatization	jokes -> *joke* studying -> *study*
Stop words	*"a", "an", "the", "in",* *"on", "of", "to", "from", ...*
TF-IDF vectorization	*A measure of how important a word is to a document in a collection of texts*

r/EDA

Search

Free

7.1k



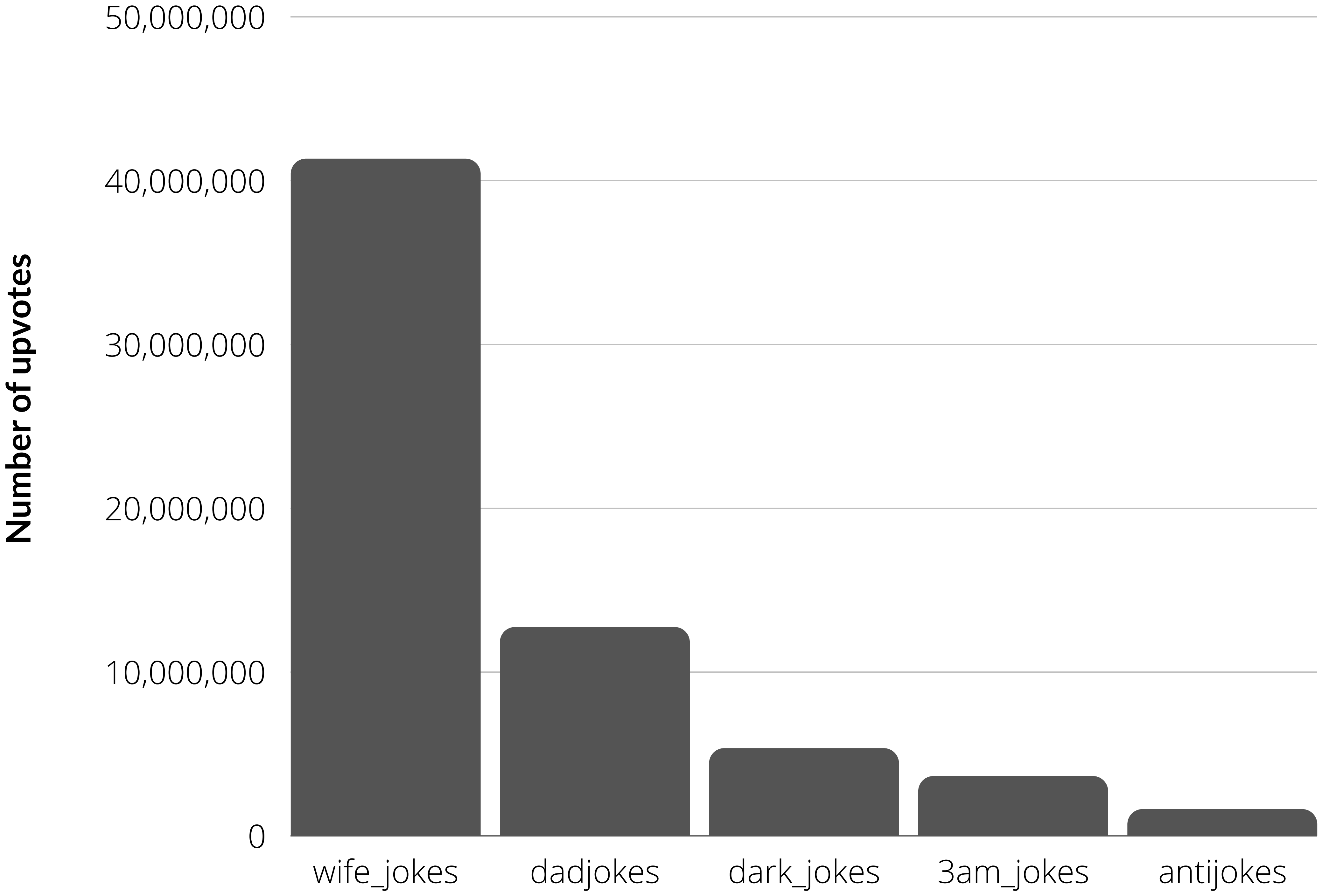
wife_jokes has the ***highest number of comments*** while antijokes has the least number of comments

r/EDA

Search

Free

9.5k



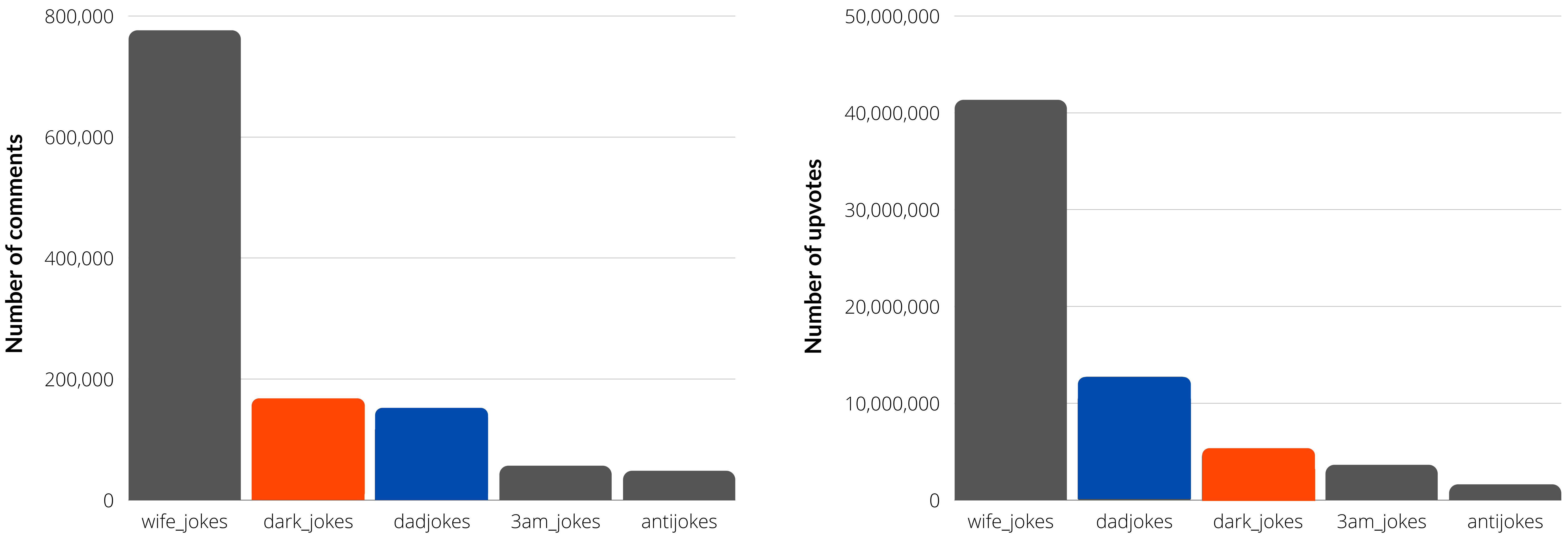
wife_jokes has the ***highest number of upvotes*** while antijokes has the least number of upvotes r/EDA

Search

Free

7.1k

Correlation of Upvotes and Comments

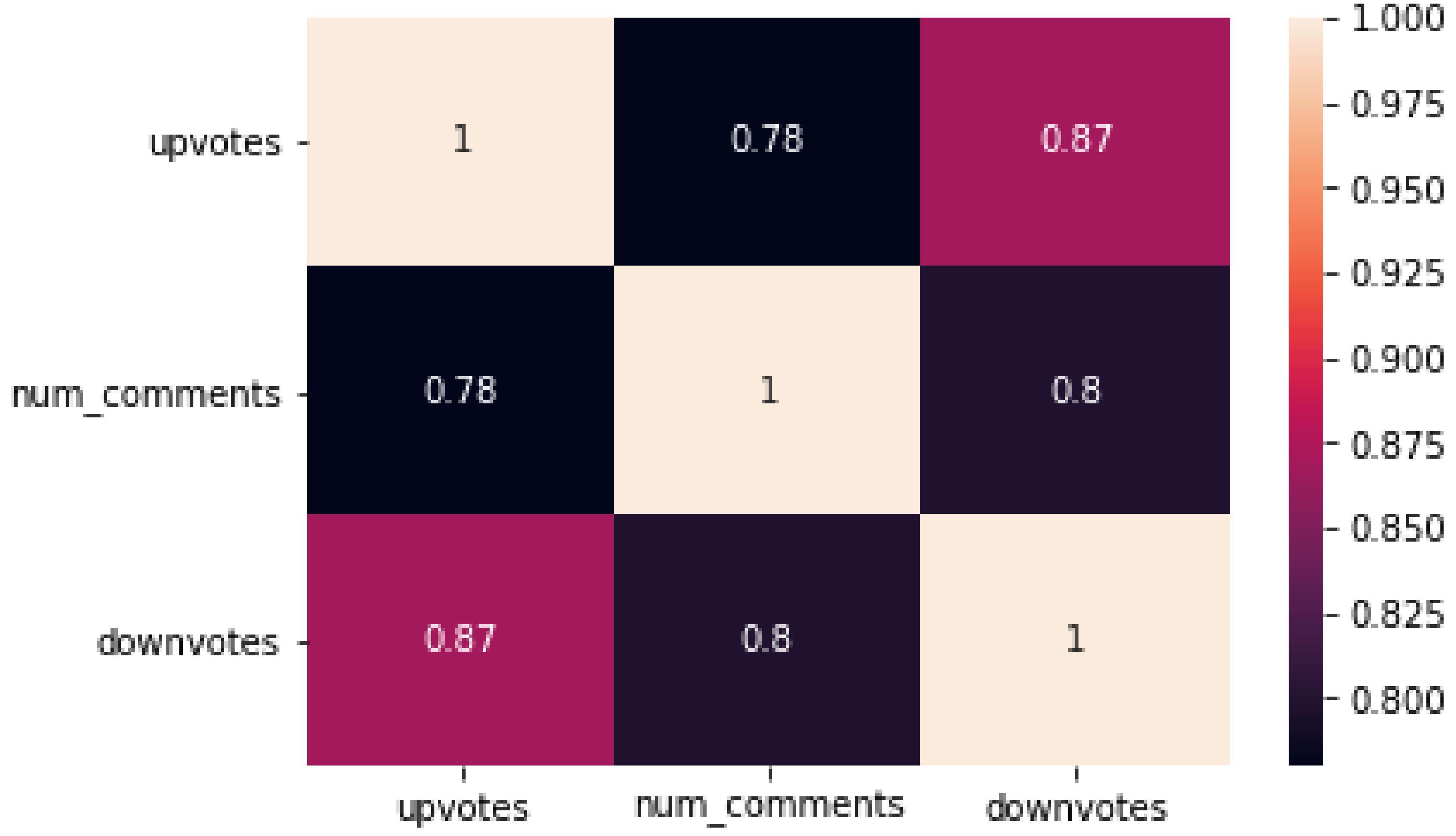


r/EDA

Search

Free

9.5k

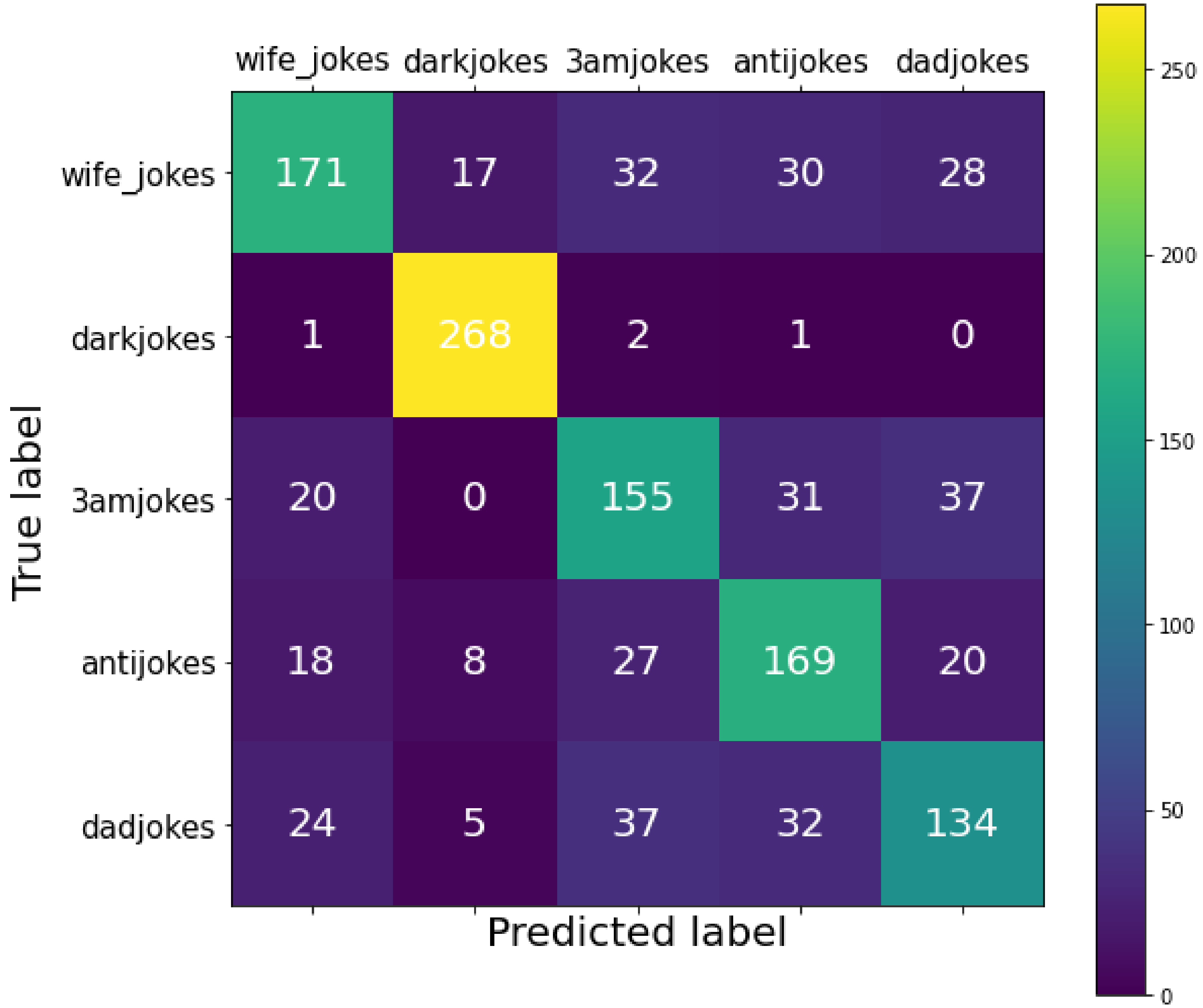


	upvotes	num_comments	downvotes
upvotes	1	0.78	0.87
num_comments	0.78	1	0.8
downvotes	0.87	0.8	1

Although the number of comments and upvotes are ***highly correlated***, we still recommend ***analyzing them separately***.

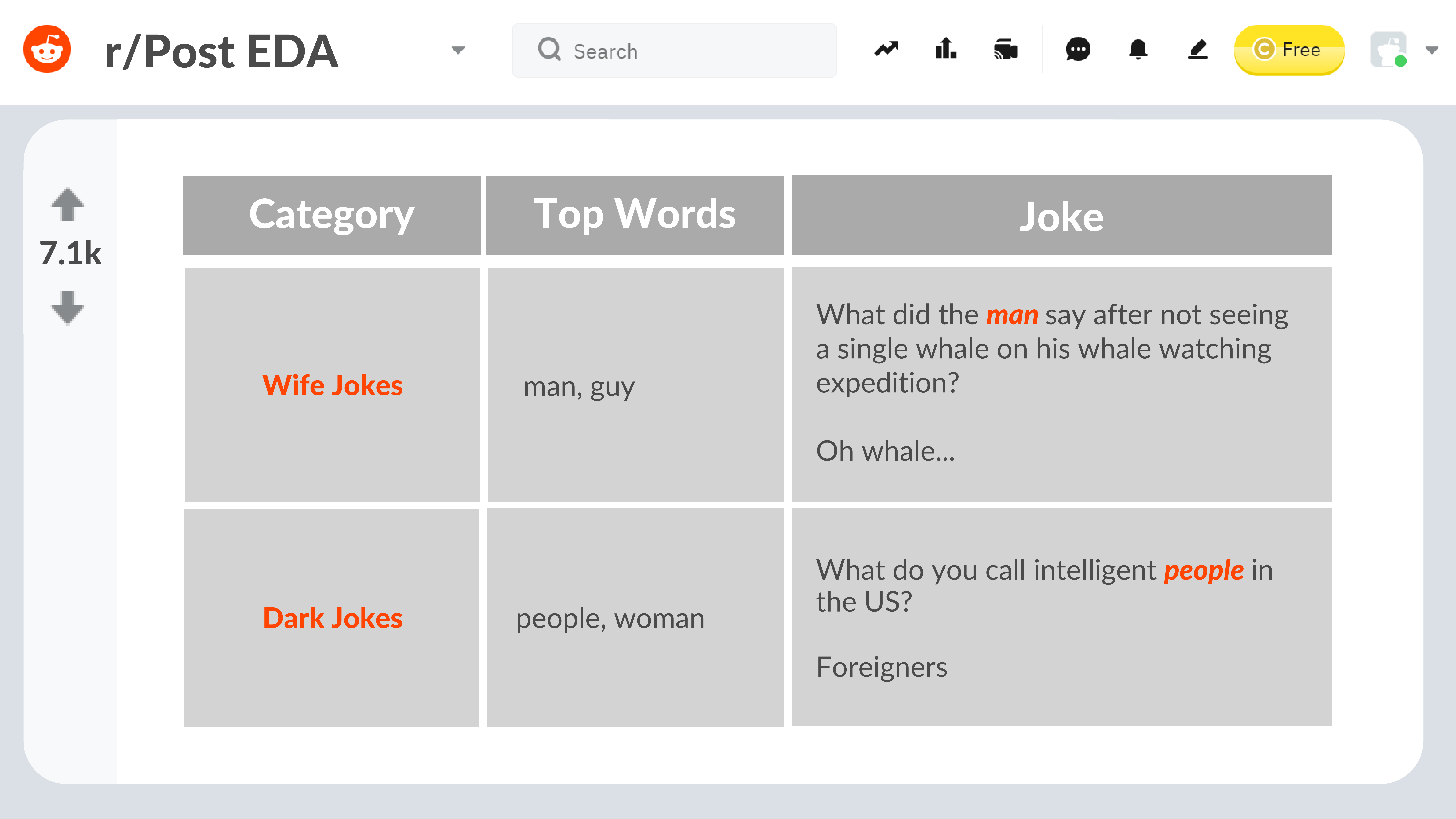
7.1k

r/Classification Model - Random Forest



	wife_jokes	darkjokes	3amjokes	antijokes	dadjokes
wife_jokes	171	17	32	30	28
darkjokes	1	268	2	1	0
3amjokes	20	0	155	31	37
antijokes	18	8	27	169	20
dadjokes	24	5	37	32	134

- Trained on 11,401 and tested on 1,267 jokes
- Accuracy scores
  - 5-Fold CV Training: *69%*
  - Test score: *71%*



Category	Top Words	Joke
Wife Jokes	man, guy	What did the ***man*** say after not seeing a single whale on his whale watching expedition? Oh whale...
Dark Jokes	people, woman	What do you call intelligent ***people*** in the US? Foreigners

r/Post EDA

Search

Free

6.2k

Category	Top Words	Joke
3AM Jokes	call, yesterday	What do you ***call*** a female singer with a humorous gaze? Gwen Starefunny
Anti Jokes	bar, joke	Harry Potter walks into a ***bar*** Just kidding, Harry Potter doesn't exist.
Dad Jokes	dad, wife	When does a joke become a “***dad*** joke”? When it becomes apparent

r/ML Model

Search

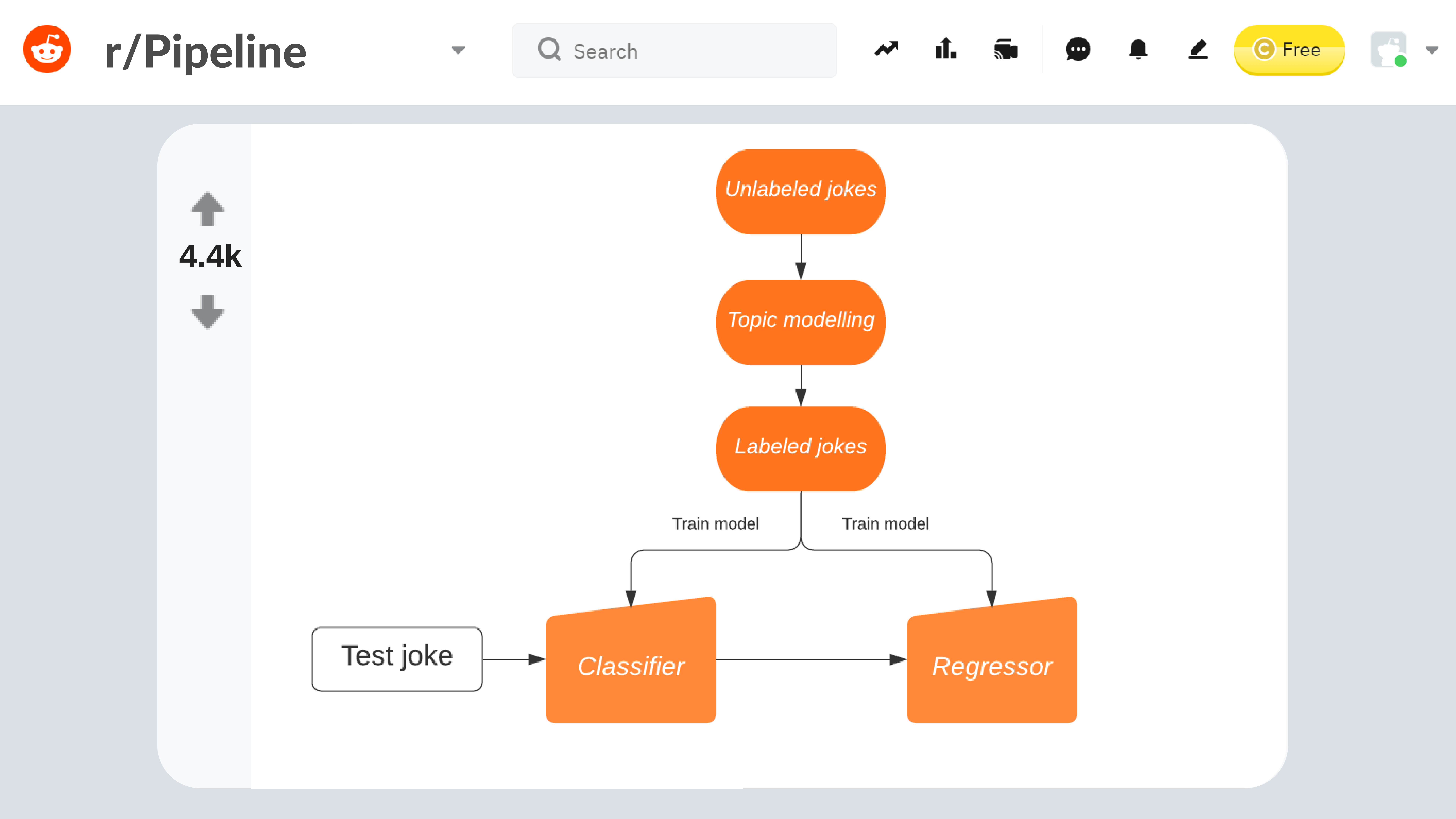
Free

7.1k

r/Regression Model - Random Forest

- Features used
  - joke class
  - "NSFW" tag
  - tf-idf array
- Target variable
  - No. of upvotes
  - May also use the ***no. of comments*** as another target

- Trained on 11,401 and tested on 1,267 jokes
- R^2 scores
  - 5-Fold CV Training: ***59%***
  - Test score: ***61%***


r/Pipeline
Search
Free
4.4k
Unlabeled jokes
Topic modelling
Labeled jokes
Train model
Train model
Test joke
Classifier
Regressor

r/Conclusion

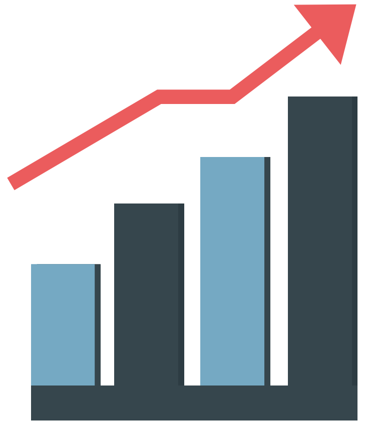
Search

Free

4.4k

wife_jokes have the ***highest*** number of engagements while the antijokes have the least.

Subreddit classification heavily depends on ***key words*** present



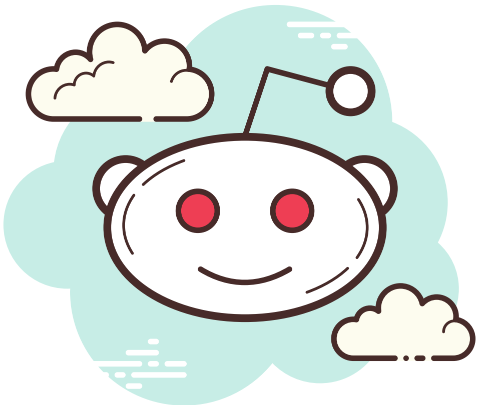
Both ***upvotes*** and ***comment count*** can be used simultaneously as measures of engagement.

r/Recommendation

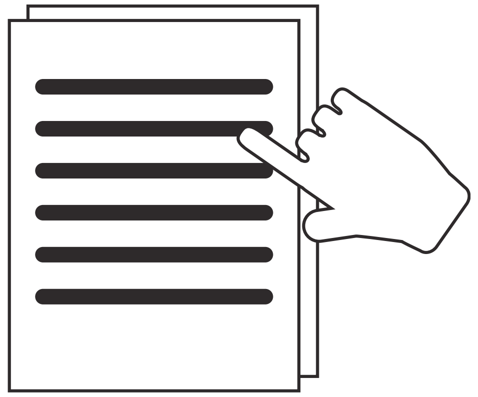
Search

Free

4.1k



Add more data to potentially improve joke classification



Add more features to potentially improve the performance of the models

Clean the data further

Perform image analysis for 'image-type' jokes (e.g. memes)

Home

Search

Free

4.1k

r/Eskwelabs

posted by Team John: Edward, Jona, Nico, Rara, Tyron

reddit