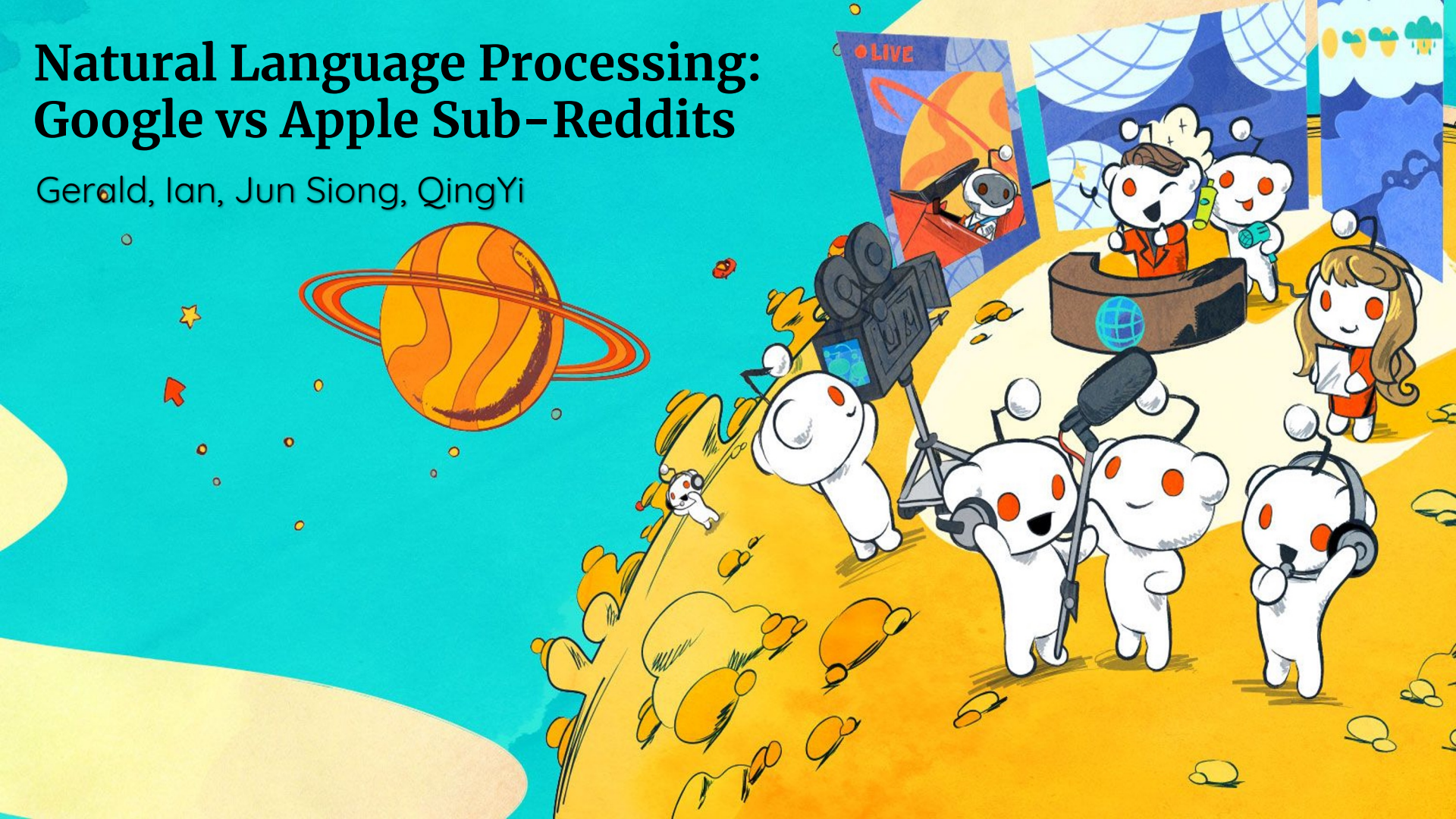
We Juice Apples, All Kinds of Apples

Natural Language Processing: Google vs Apple Sub-Reddits

Gerald, Ian, Jun Siong, QingYi



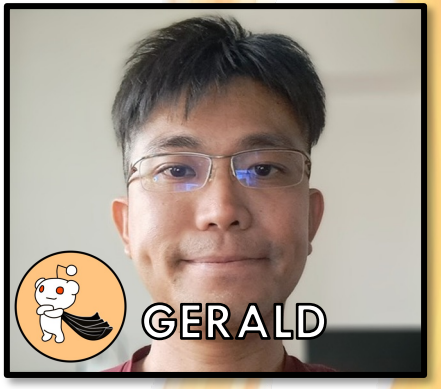
Who Are We?

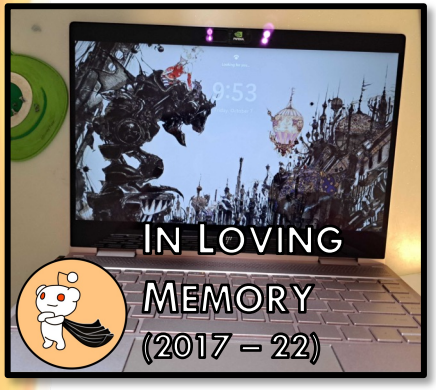
- Data Consulting Firm
- Specialises in NLP

What Can We Give?

- Business Edge through

  ML-Driven Insights





Problem Statement

Can We Accurately Predict

which Sub-Reddit:

GOOGLE or APPLE,

does a Given Post Belongs to?

Background

Data are your **New Oil**

…and your **New Weapons!**

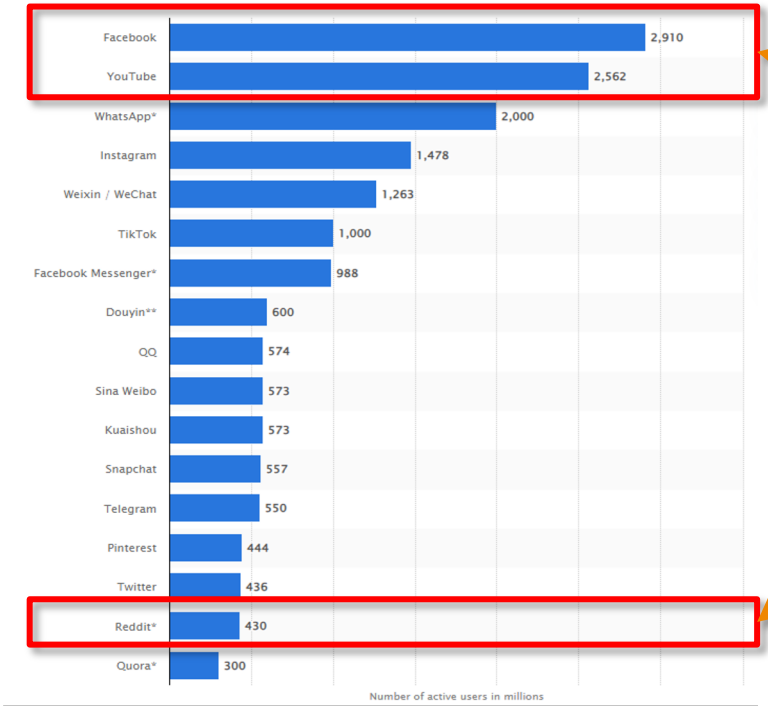
How Trump Consultants Exploited the Facebook Data of Millions

The New York Times

Cambridge Analytica: how did it turn clicks into votes?

theguardian

Social Media Platforms (Ranked by Millions of Monthly Active Users)



The **BIGWIGS**

1st & 2nd

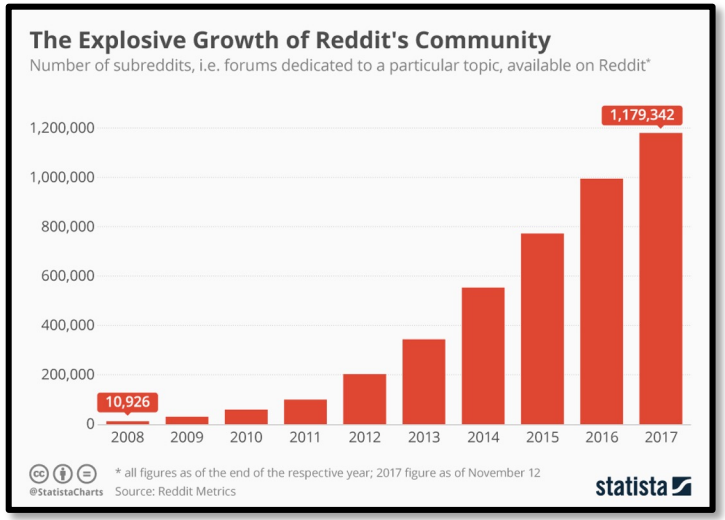
- Top 2 Social Media Platforms
- 54M Combined Active Userbase

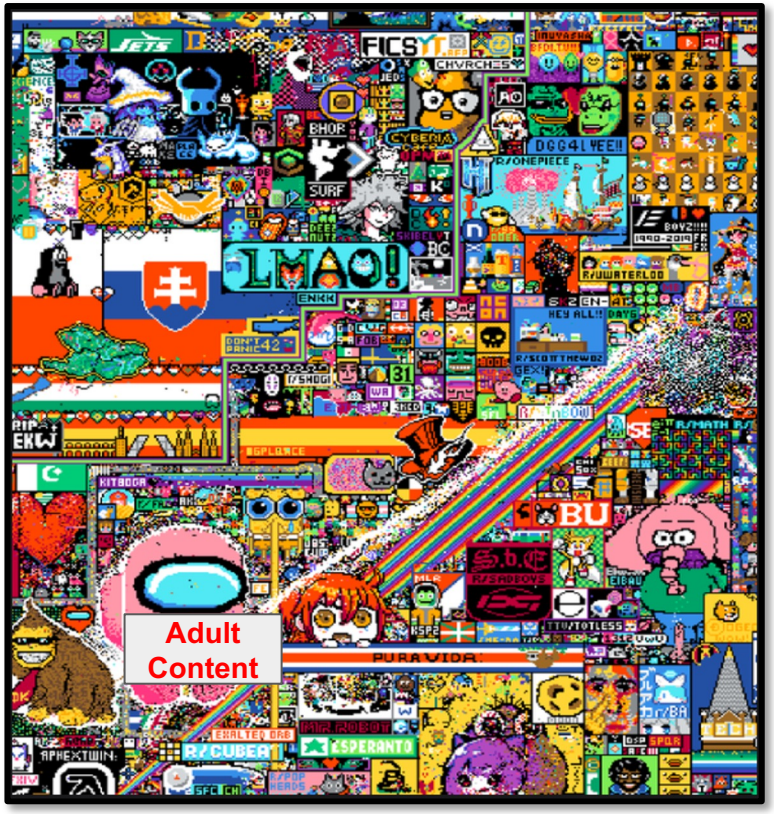
The **UPSTART** reddit

OUR FOCUS

Battlefield REDDIT

Explosive Growth, Unparalleled Collaboration



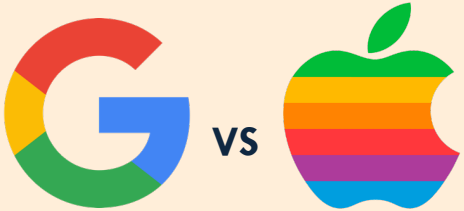


Problem Statement

Can We Accurately Predict which Sub-Reddit:

GOOGLE or APPLE, does a Given Post Belongs to?

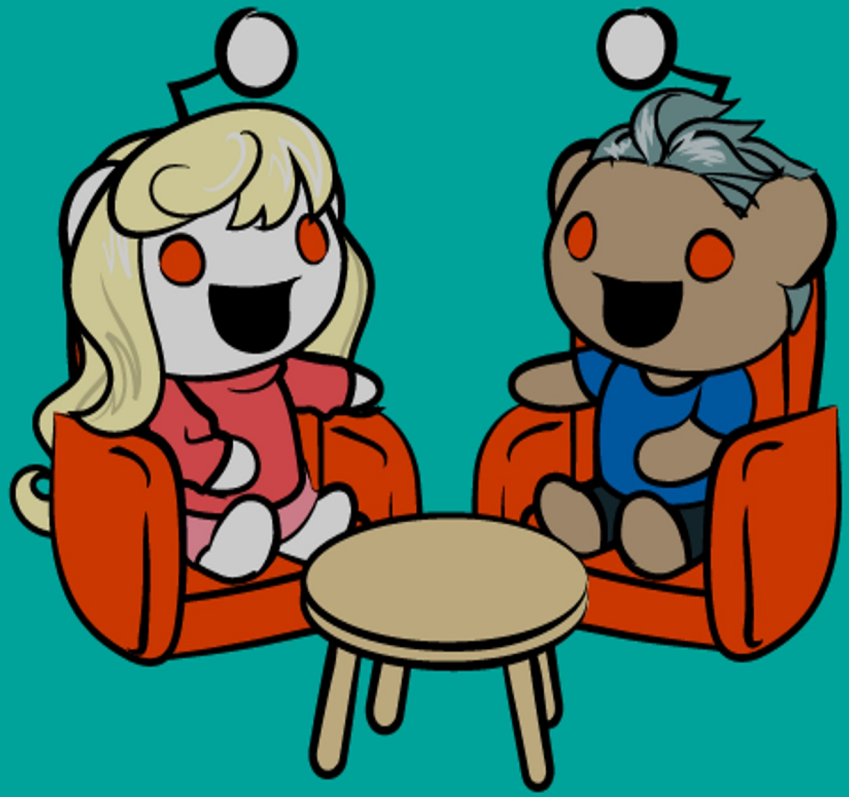
Exploit Reddit Data to Sense Public's Preferred Sub-Reddit to Discuss Certain Topics



"Phones" "Machine Learning" "Streaming"

WHO OWNS THE BUZZ??

"Developer Support" "Coolest Employment"



Exploratory Data Analysis (EDA)

EDA: SIZE OF CORE USERS

Apple: Larger Pool of Reddit Users Providing Sub-Reddit Content

2.0M

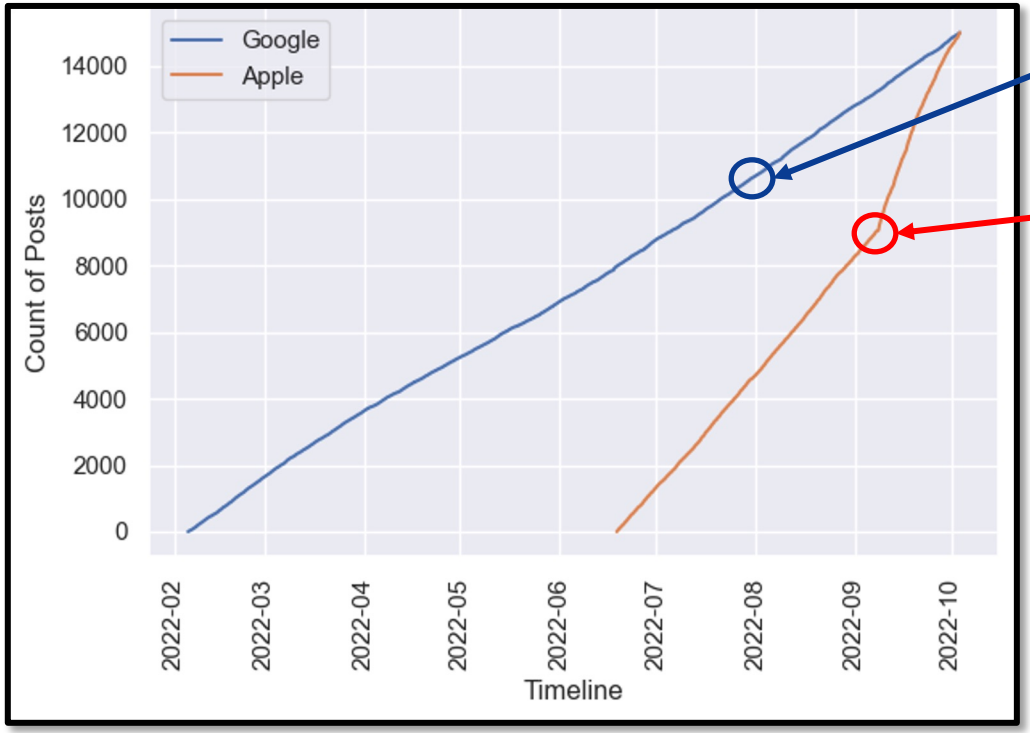
Core Users*: 1.9K

3.8M

Core Users*: 8K

*Core Users: User Count that contributed to 80% of the most recent 15K posts

EDA: POSTING RATE (Time taken to reach 15k in Sub-Reddit)

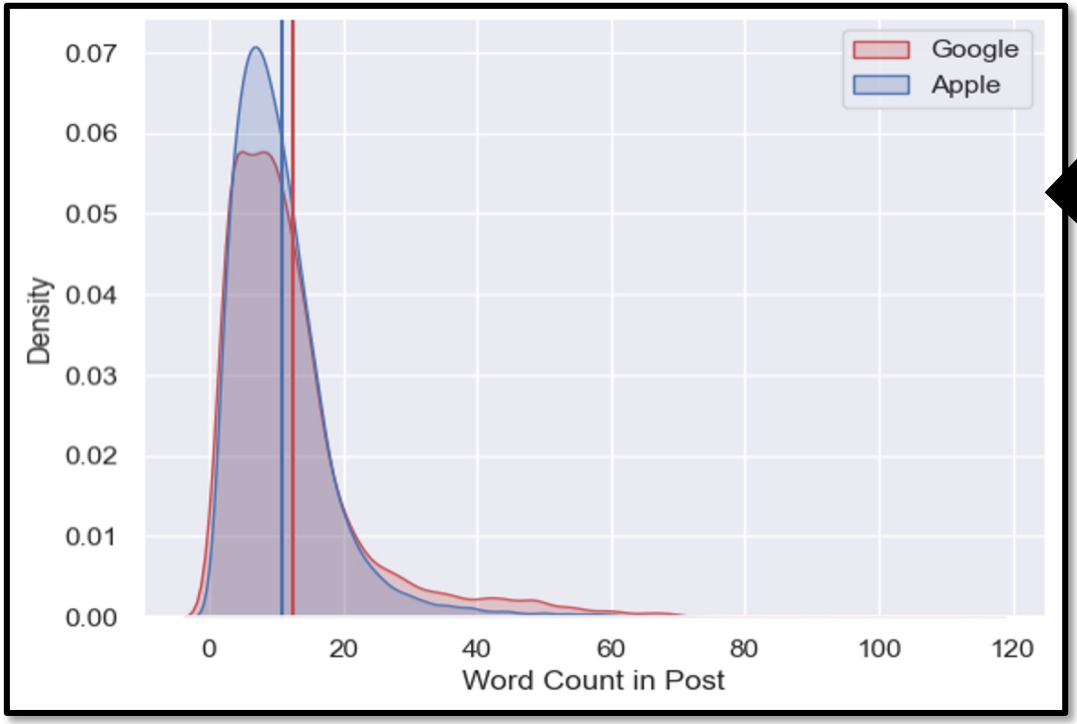


Google Sub-Reddit Fails to be Excited by New OS Release (Aug 22)

Apple Sub-Reddit Registers Sharp Incline with New OS Release (Sep 22)

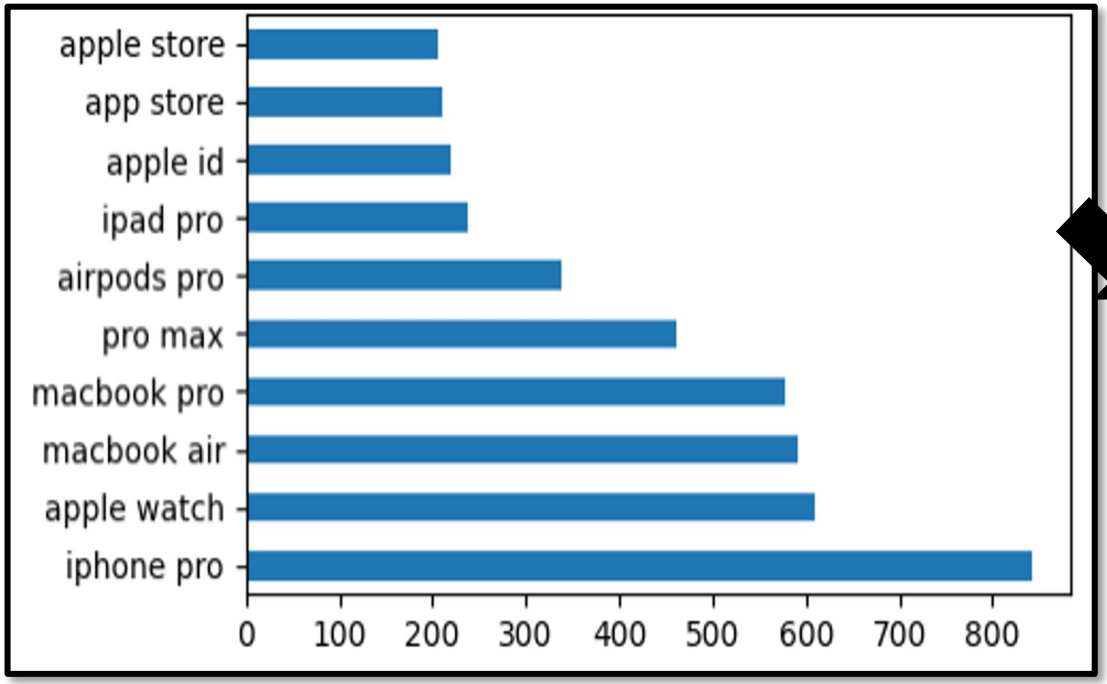
Apple, The Excited:
Uses Half the Time to Reach 15K Posts in Sub-Reddit

EDA: POST LENGTH (Average Words Per Post in Sub-Reddit)



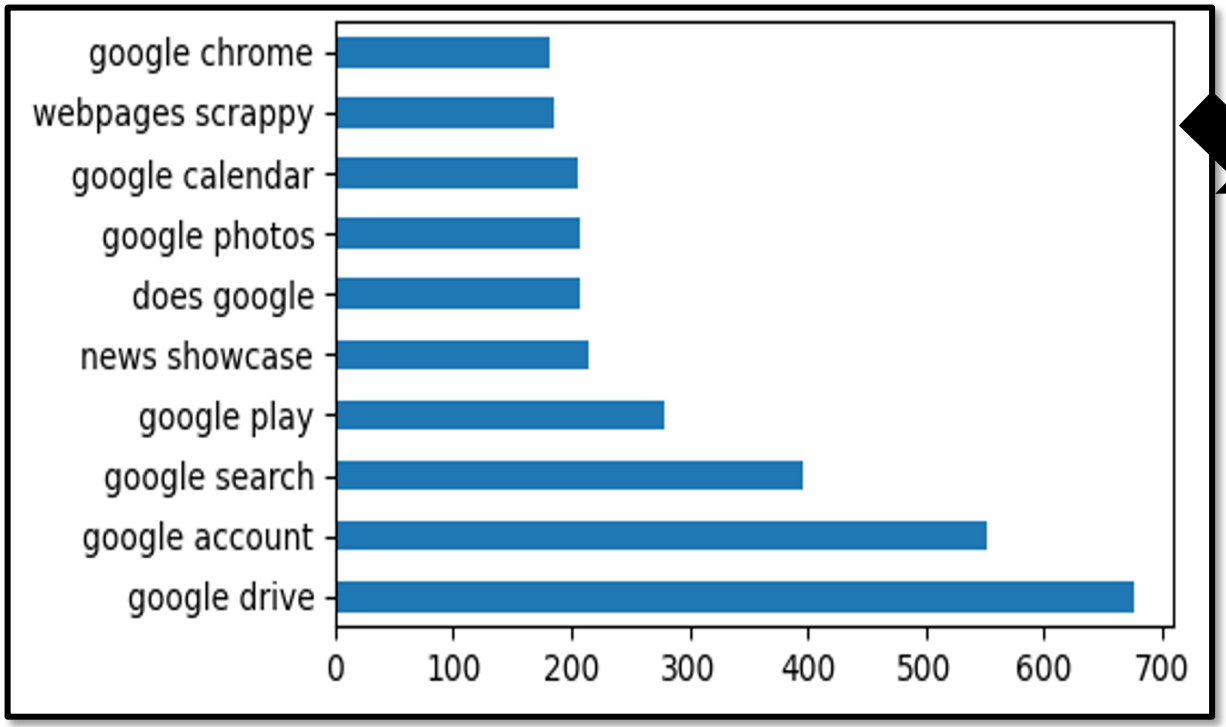
All Square :
No Significant Differences between Sub-Reddits for Average Words/Post

EDA: Most Popular Topics (Apple)



Discussion Centered on Main Apple Products

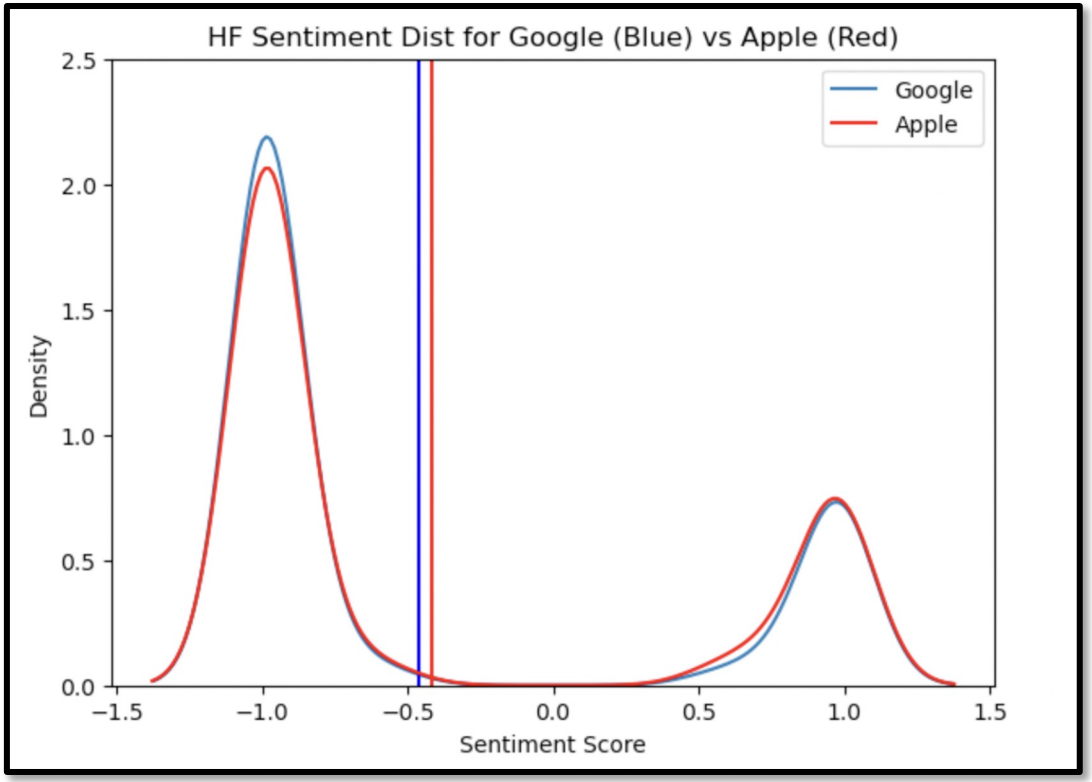
EDA: Most Popular Topics (Google)



Google Play

(Likewise) Discussion Centered on Main Products

EDA: SENTIMENT ANALYSIS PART I

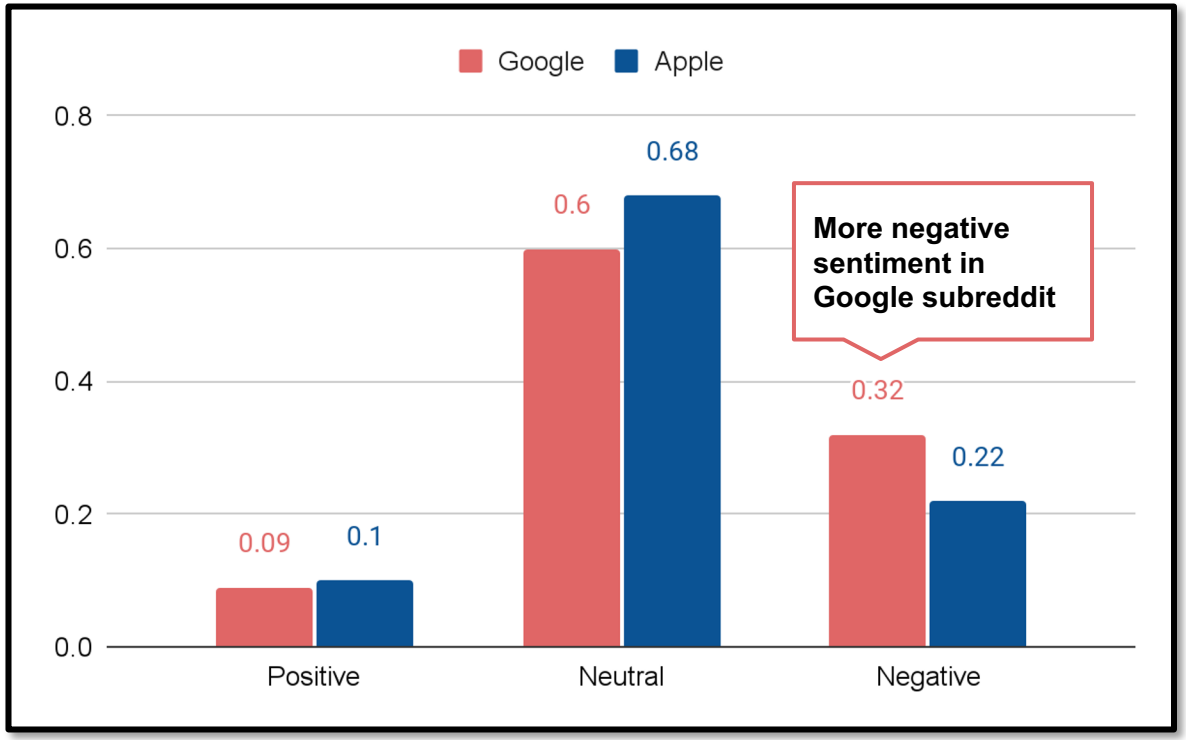


distilbert-base-uncased-finetuned-sst-2-english

Hugging Face

Apple 1 : 0 Google

EDA: SENTIMENT ANALYSIS PART II



Emotion English DistilRoBERTa-base

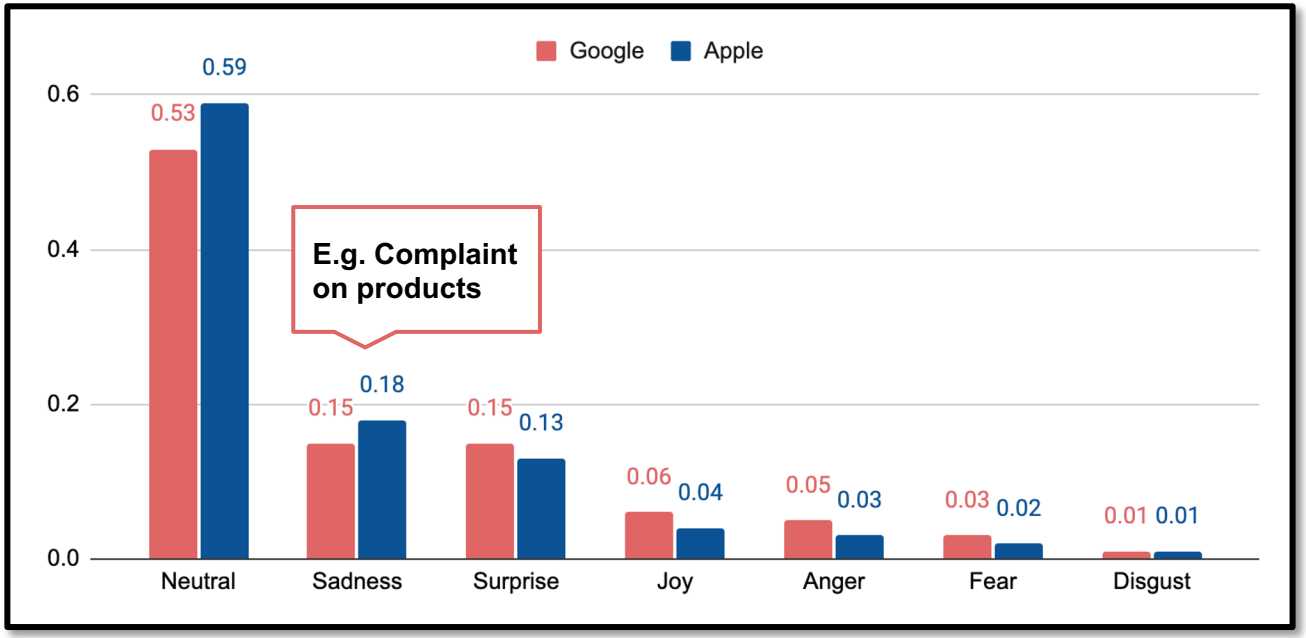
Hugging Face

Apple 2 : 0 Google

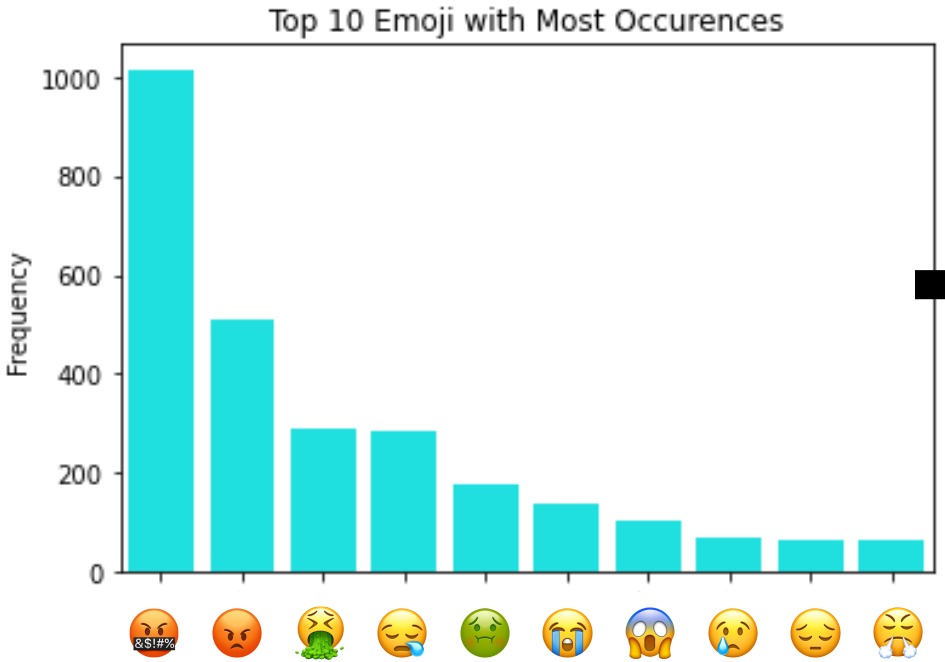
EDA: SENTIMENT ANALYSIS PART III

Emotion English DistilRoBERTa-base

Hugging Face ➡ Apple 2 : 1 Google



EDA: Most Common Emojis



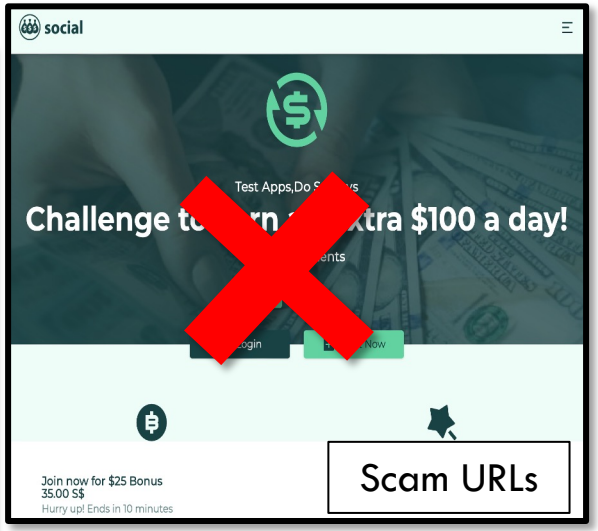
DATA CLEANING, MODELLING, & FINDINGS

DATA CLEANING & PROCESSING

STAGE 1

Remove Scam + Spam Posts and URLs

Auntie Services



Scam URLs

Unrelated News

Tabloids

DATA CLEANING & PROCESSING

STAGE 2

Lower Case, Remove Special Characters + Stop Words,
Remove "Giveaways", Lemmatize

MODEL SELECTION

Search Radius

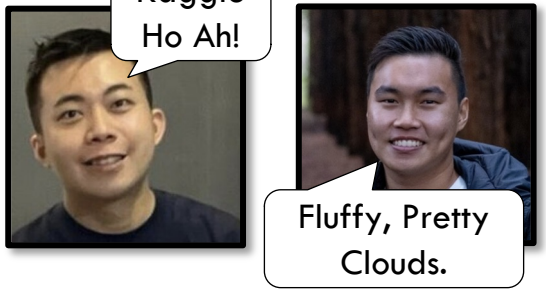
Vectorisation:
Count Vectorisation, TFIDF Vectorisation

Fitting:
16 Classifier Models (with PyCaret)

Criteria

Model Performance *and also*
+ **Explain-ability, + Resistance to Overfit**

kaggle
+
Google Cloud Platform



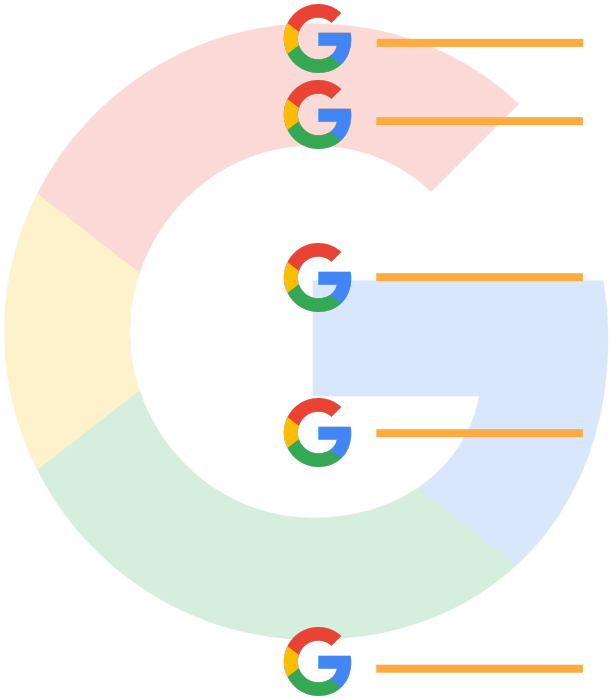
Settled on **Random Forest, Logistics Regression, Naïve Bayes** for Fine-Tuning.

RESULTS & FINDINGS

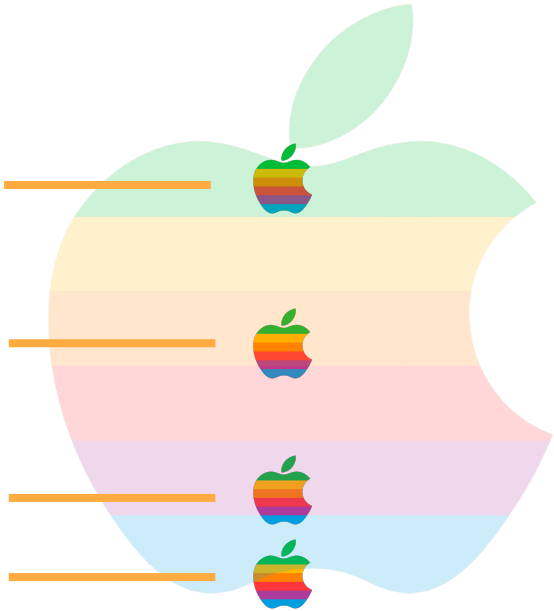
Post Gridsearch, **TFIDF Logistics Regression** was our best Performing Model.

Vectorization Method	Model	Train CV Accuracy	Test Accuracy
Count	Random Forest	98.0%	83.0%
	Logistics Regression	88.7%	83.8%
	Naïve Bayes	84.5%	81.9%
TFIDF	Random Forest	98.0%	82.6%
	Logistics Regression	**87.8%**	**83.9%**
	Naïve Bayes	85.0%	82.0%
Combined	Voting Classifier	92.5%	84.6%

BUZZ GENERATOR



Term Classified
Search Engine
Machine Learning
Video Editing
Job, Employment
Applications, Developer
User Interface
Phone
Gaming
Take Over the World



Conclusion & Recommendations

CONCLUSION

- Model able to classify and label input text as either Apple or Google with an accuracy of 83.9% using Logistic Regression.
- Such a classifier could provide sentiment analysis on overlapping services offered by Apple and Google

  (e.g. Input text: '*Maps is such a godsent for finding places to eat nearby* ' Label: **Google**)

- Black box models optimised for accuracy (Eg. Random Forest); White box models optimised for interpretability (Eg. Logistic Regression).

RECOMMENDATIONS

- Train a classifier on more than 2 subreddits.

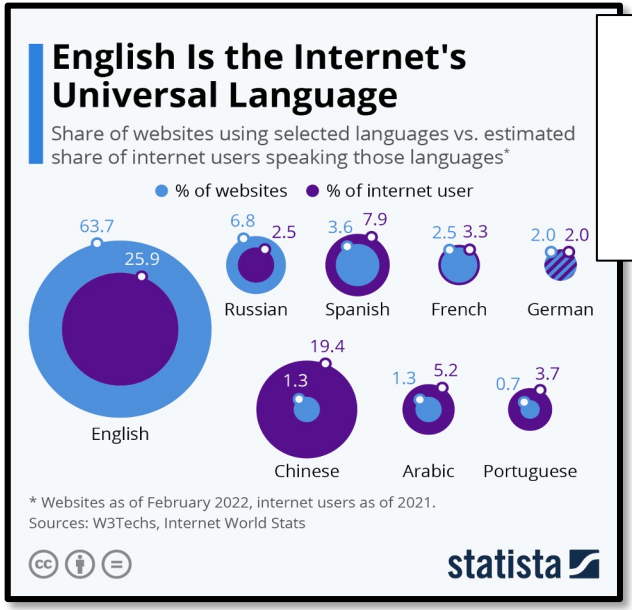
  Use Case: identifies the most relevant subreddit to post a question or comment or seek advice from (1.8M subreddits)

- Include features in the feature matrix that extends beyond text from the posts and title within the subreddits.

  (e.g. Upvotes, posts with a minimum number of comments, timestamps of posts)

LIMITATIONS

Classifier has only been trained on English text and hence does not generalize well for classification of non-English text.



Source: English Is the Internet's Universal Language

Content: 36.3% of all websites use languages that are not English as their content language (Statista, 2022)

Speaker Base: 77.3% of speakers in the top 20 most spoken languages do not speak English (CCJK, 2022)

Source: 20 Most Spoken Languages in the World in 2022

LIMITATIONS

Training data consists of posts from **only one source: Reddit**.

Classifier should be trained on **broader and more diverse text data sources** to create a more robust model that generalises well to different data sources. (e.g. Twitter, Youtube, Baidu, Weibo, Youku), thereby reducing biases that may be implicit in the platform (e.g. media bias, country bias).

Ready to Juice Apple?