Classification of Subreddit Posts

For subreddits: /r/Apple & r/Android

Group 3: Gladys, Jeremy, Yong Khiang, Kayle

Problem Statement

Why?

- Inflow of irrelevant posts
- Complaints from content moderators and users
- Find a permanent fix

Who is it for?

- Admin Team
- Content Moderators

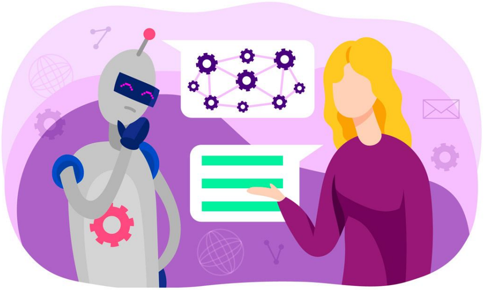
Problem Statement

Our Process

1. Web Scraping

2. Data Cleaning
(e.g. NLP)
with Exploratory
Data Analysis

3. Modelling
(i.e. select and
evaluate model
- ROC AUC)



4. Interpretation and
conclusion

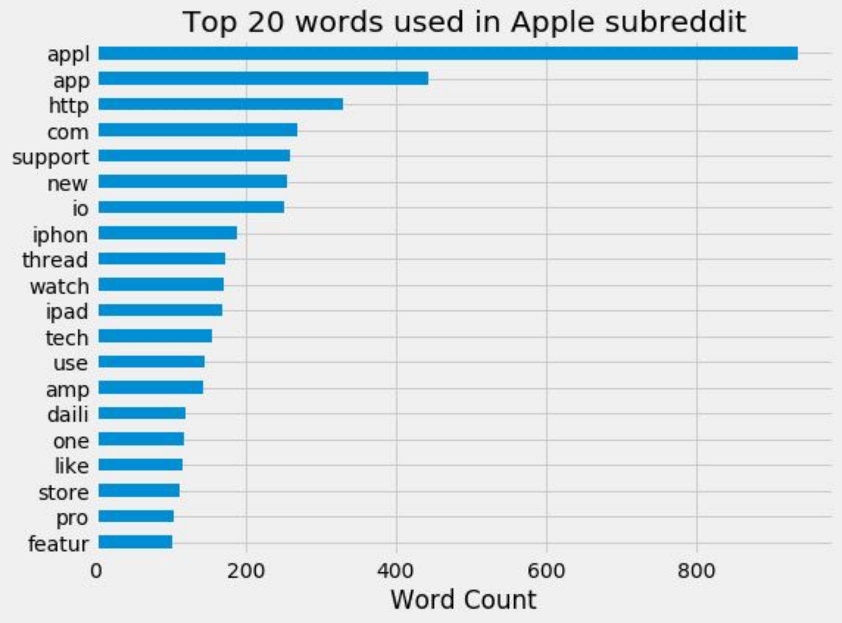
Problem Statement

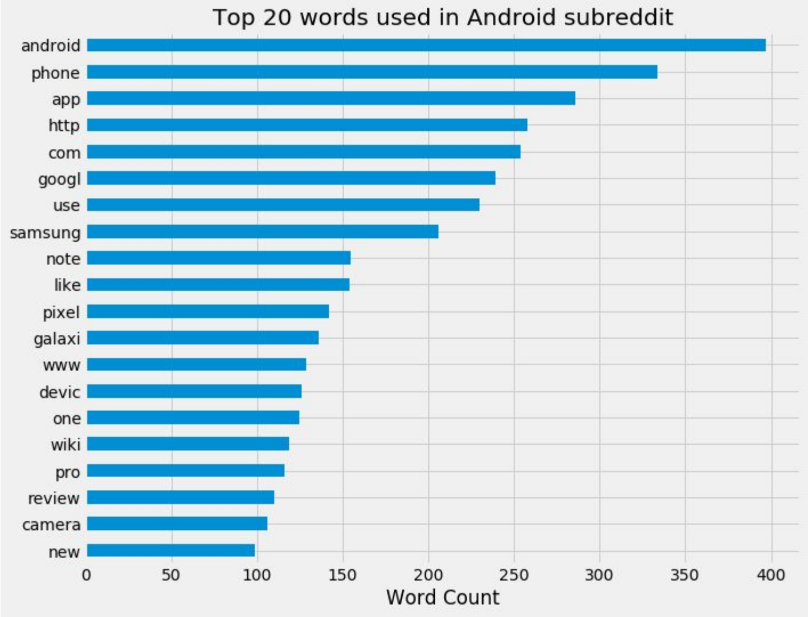
Our Ultimate Aim

- Train a classifier model
- Accurately classify posts
- Gain insights on most important words
- Fix problem of irrelevant posts

Exploratory Data Analysis (EDA)

Common words

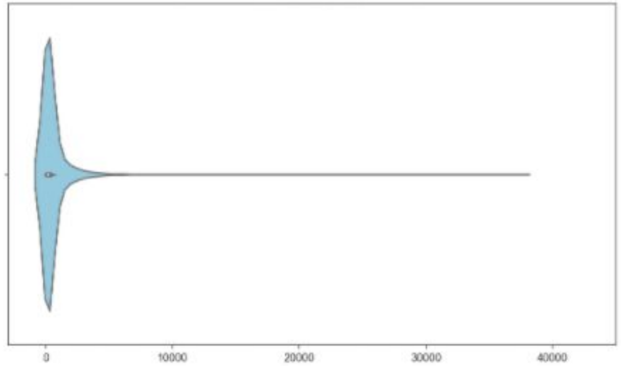




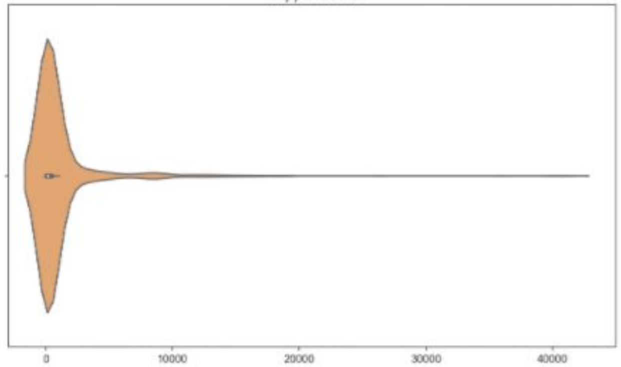
Exploratory Data Analysis

Score

r/Android Score

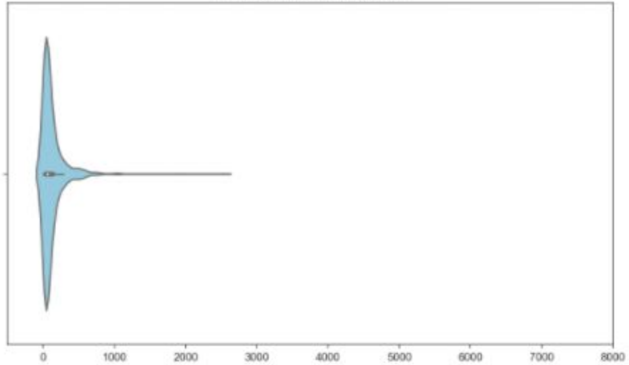


r/Apple Score

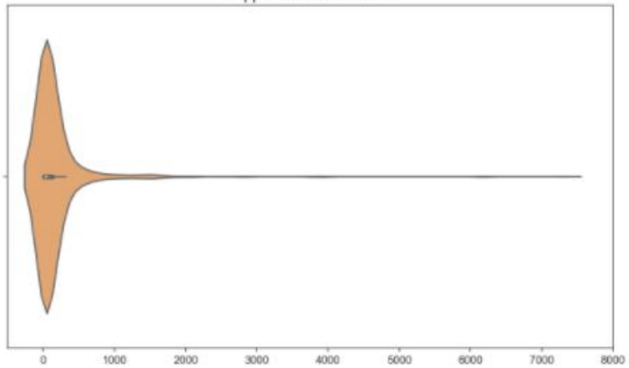


Comment Counts

r/Android Comment Counts



r/Apple Comment Counts

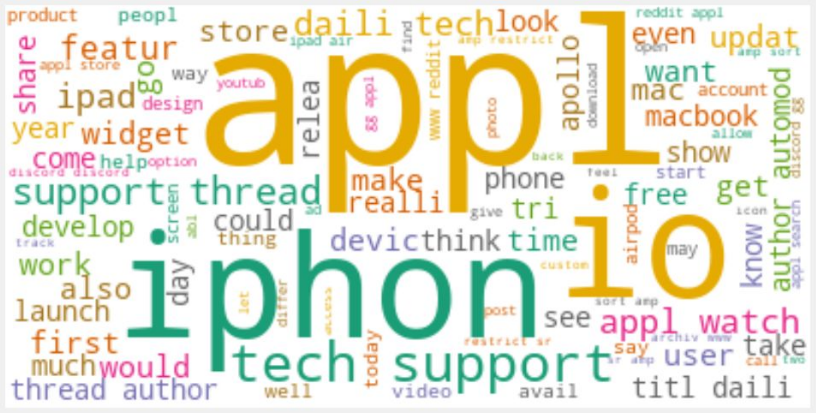


Findings

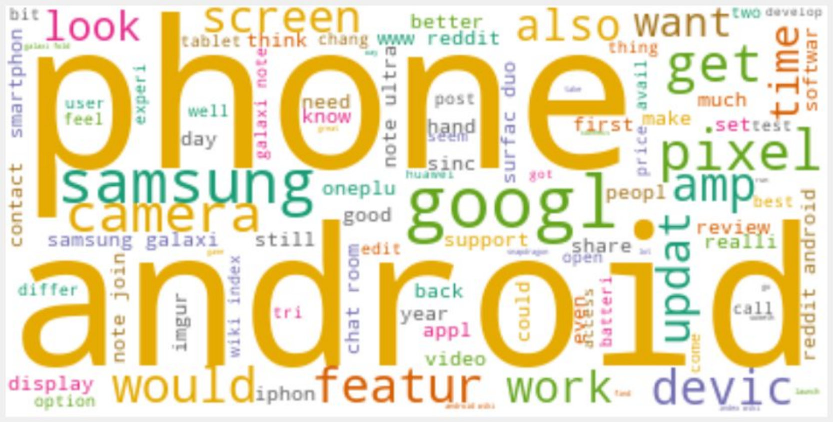
- ‘appl’ appeared 934 times and ‘android’ appeared 397 times
- ‘appl’ has twice as much as ‘android’ even though it has only 3% more posts
- Common words : ‘app’, ‘http’, ‘com’, ‘use’, ‘like’, ‘one’, ‘pro’, and ‘new’
- Remove common words to have better classification accuracy

Top Words

r/APPLE



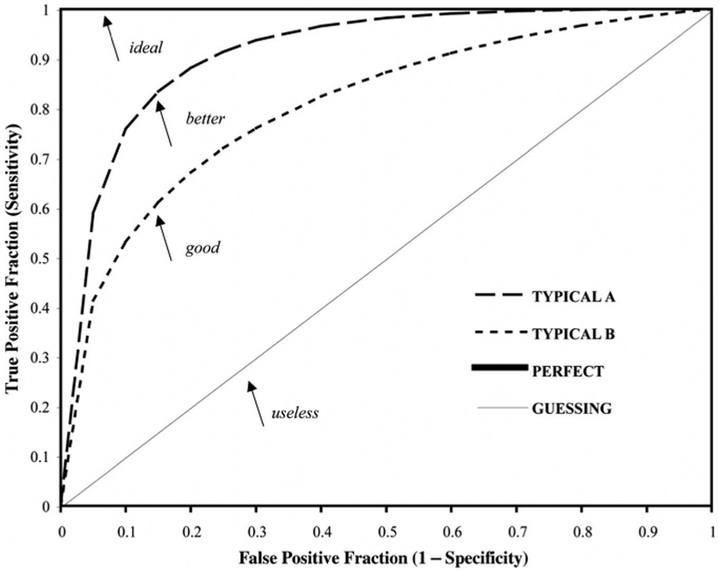
r/ANDROID



Model Evaluation

Predictors	Model	ROC-AUC score (cross validation)
Text-Title	Logistic Regression	0.986
Text-Title	Multinomial Naive Bayes	0.983
Text-Title + Score & Comment Counts	Logistic Regression	0.985
Text-Title + Score & Comment Counts	Multinomial Naive Bayes	0.983

Receiver Operating Characteristic (ROC) Curve

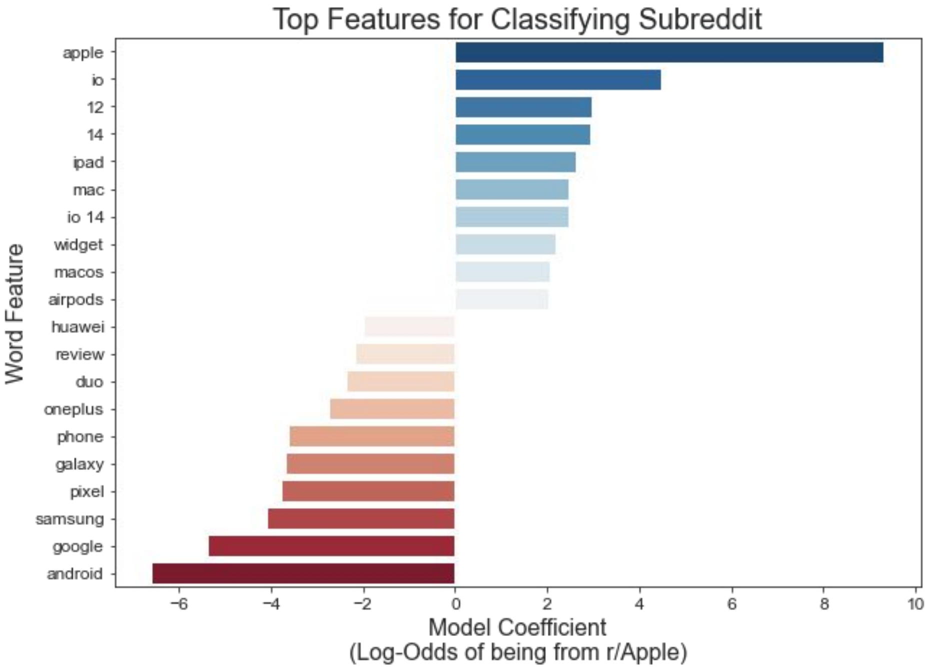


Production Model Performance (Test Data)

	Predict r/Android	Predict r/Apple
Actual r/Android	143	12
Actual r/Apple	6	182

- High ROC-AUC score of 98.1%
- High overall accuracy of 94.8%
    - Misclassification rate : 3.3% (r/Apple)/ 7.7% (r/Android)
- Misclassification reasons
    - common words appearing between the 2 subreddits e.g. comparison of apple & android product features
    - short post without strong word features

Top word features



Interesting finding

- “mac", "12", "widgets" and "airpods" not from the top 10 common words in r/Apple from earlier EDA
    - Due to TF-IDF vectorizer according higher weights to rarer words

Conclusion

Logistic Regression classifier model

Overall Model Performance: > 90% success rate

98%

Distinguish between true positives and true negatives

95%

Accuracy rate in classifying posts

Predicting classes?

97%

Correctly predicts posts that are in r/apple

92%

Correctly predicts posts that are in r/Android

Recommendations

- Use the model as a detector to decrease number of irrelevant and misclassified posts
- Explore additional features
  - e.g. comment text, and sentiment analysis of textual contents
- Train data on different classifier algorithms
  - e.g. Support Vector Classifier, Random Forest Classifier etc.
- Pull data periodically to add to the corpus so as to account for changing trends in technology topics related to Apple and Android