Popcorn War:

Netflix vs Disneyplus

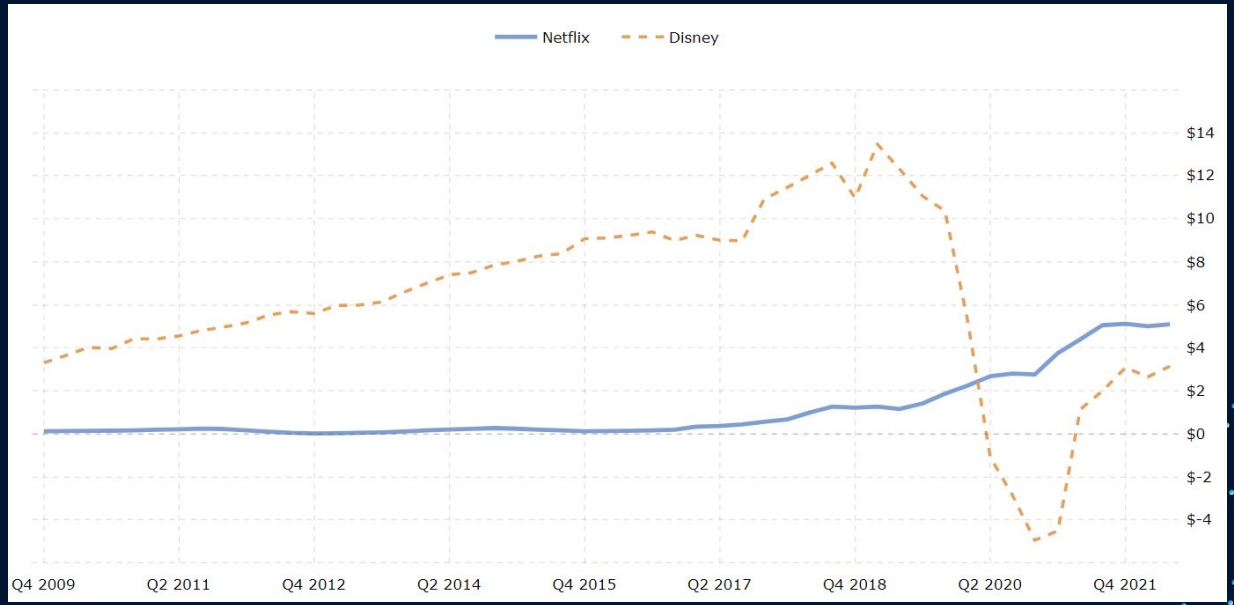
An NLP binary classification project

by:

Ridzuan, Wan Xian, Meriky

Background & Objectives:

- Get insights into market competition between Netflix and Disneyplus
- Optimize machine learning models to classify sub-reddit texts
- Public sentiments and general opinions about them



Net Income (in billion USD)
Source: macrotrends

Stakeholders:

01

CLIENT:

A start-up online movie screening platform

02

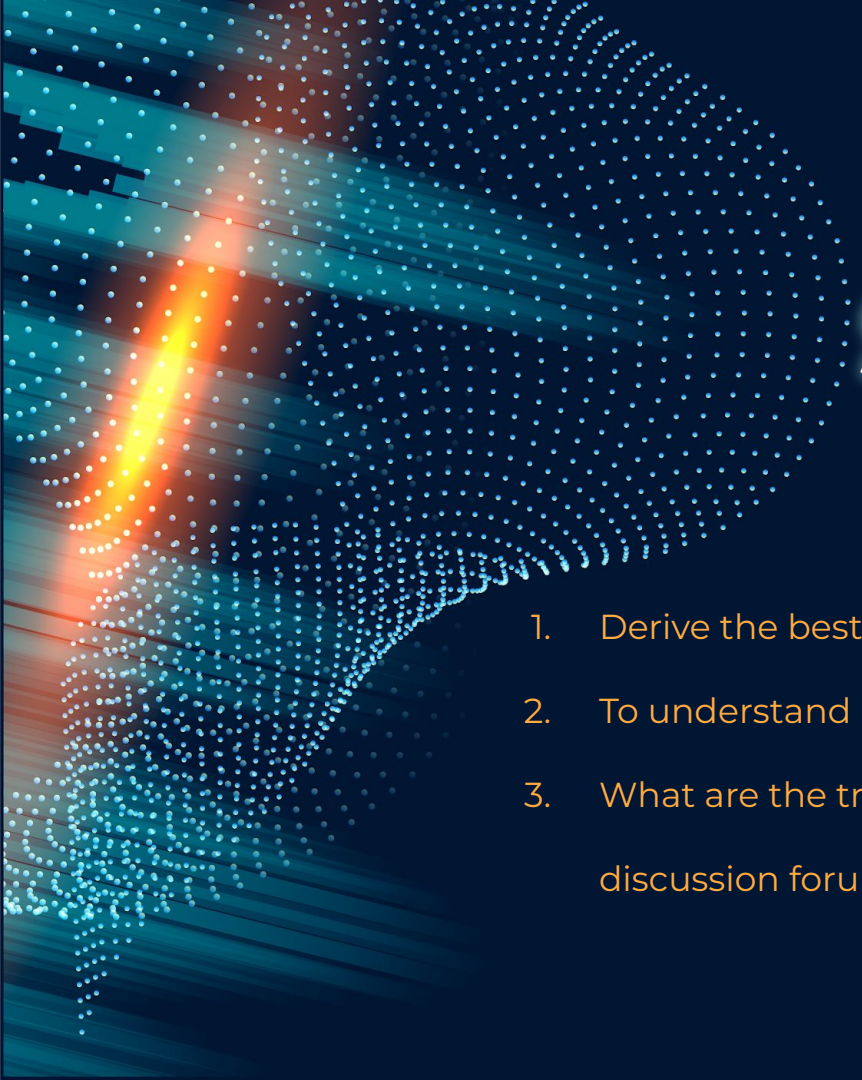
YOU:

A mix of technical & non-tech audience representing potential investors

03

US:

Business consultant hired to give insights on movie platform businesses



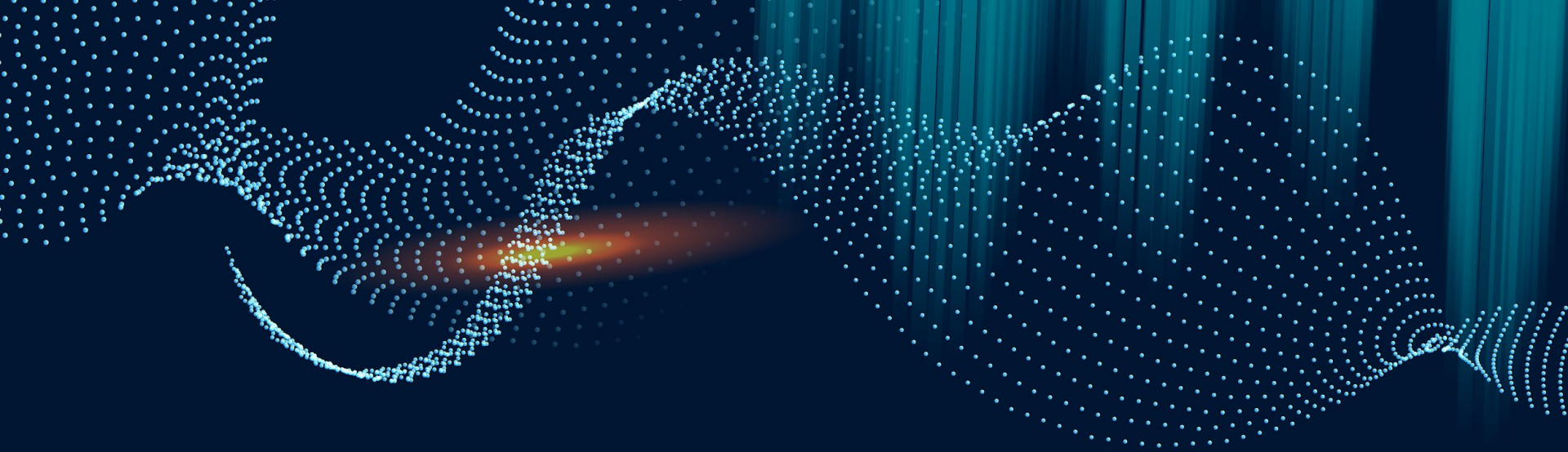
Problem statement

NETFLIX vs DISNEY+

1. Derive the best classification models to predict the source of reddit posts
2. To understand public sentiments of the reddit post
3. What are the trending topics in current movie streaming platform discussion forums?

AGENDA:

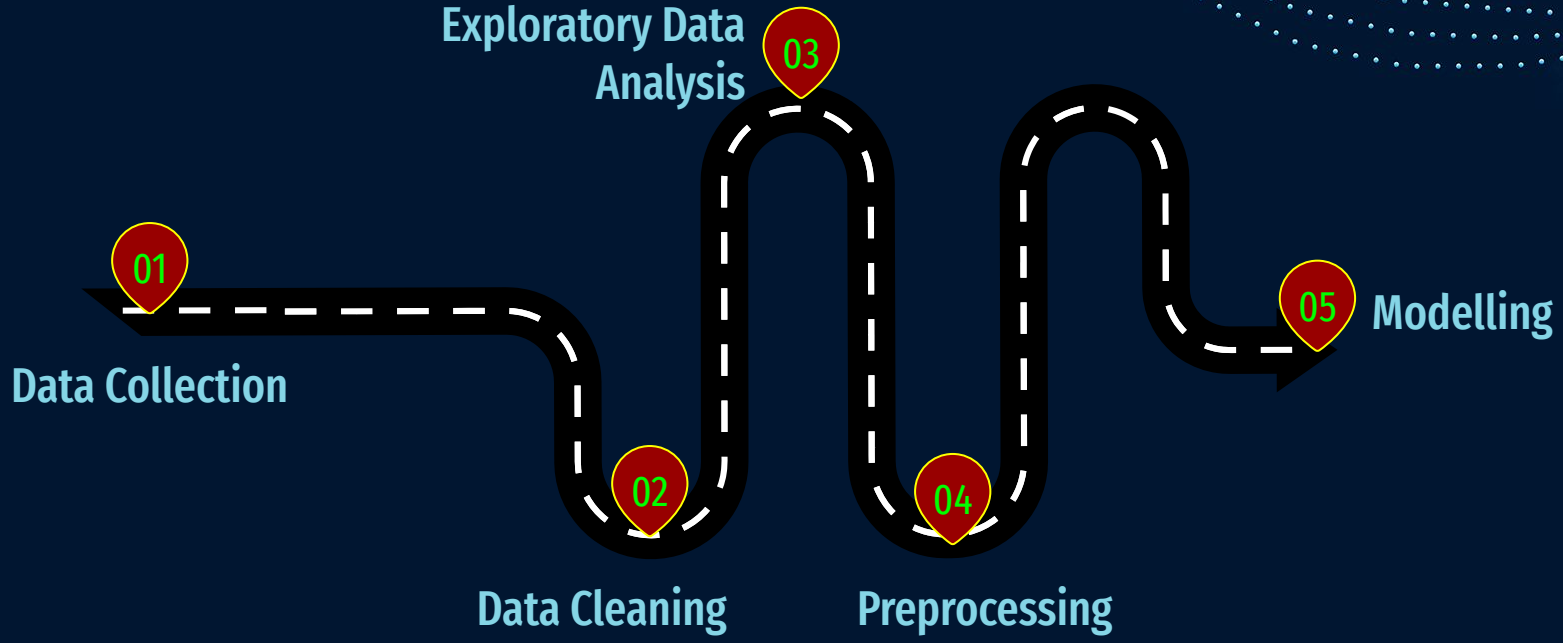

INTRODUCTION
METHODOLOGY
EDA
MODELLING & METRICS
SENTIMENT ANALYSIS
CONCLUSION & RECOMMENDATIONS



02 Methodology & Process flow

Steps taken before modeling process

Process flow



1) Data Collection

Web scraped with Pushshift's reddit API:

- Subreddits: Netflix and Disney Plus
- Features: Title and Post selftext
- Number of post : 15000 each
- Time: After December 27, 2021

  5:00:00 AM

https://medium.com/swlh/choosing-the-right-streaming-service-netflix-vs-disney-5e89fe615df0

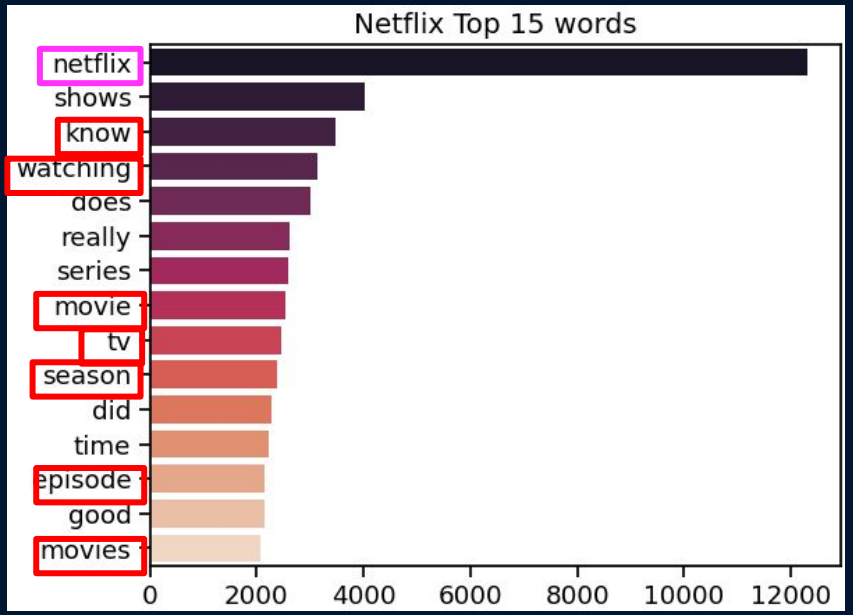
2) Data Cleaning

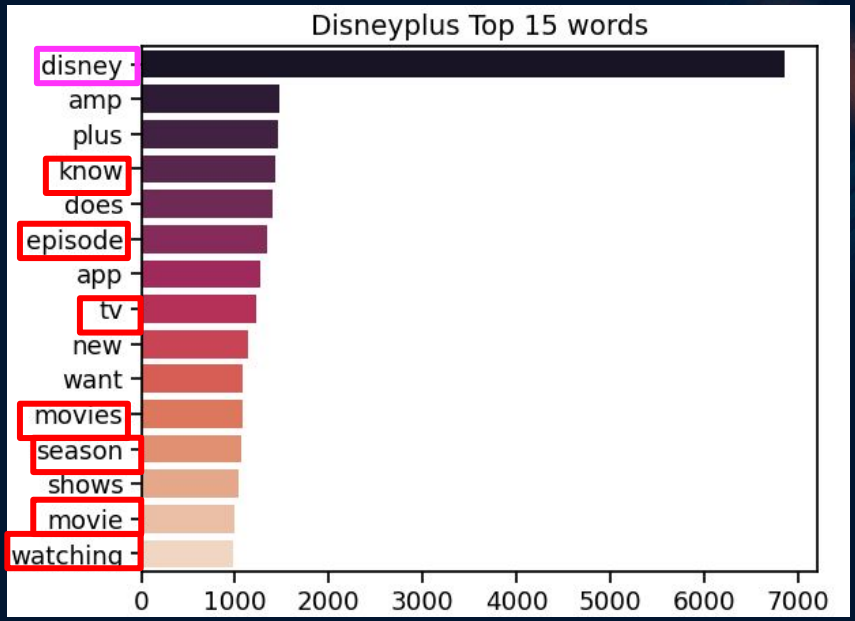
- Remove columns with null
- Exclude rows with selftext containing

  [Removed] & [Deleted]
- Drop duplicated rows

3) Exploratory Data Analysis

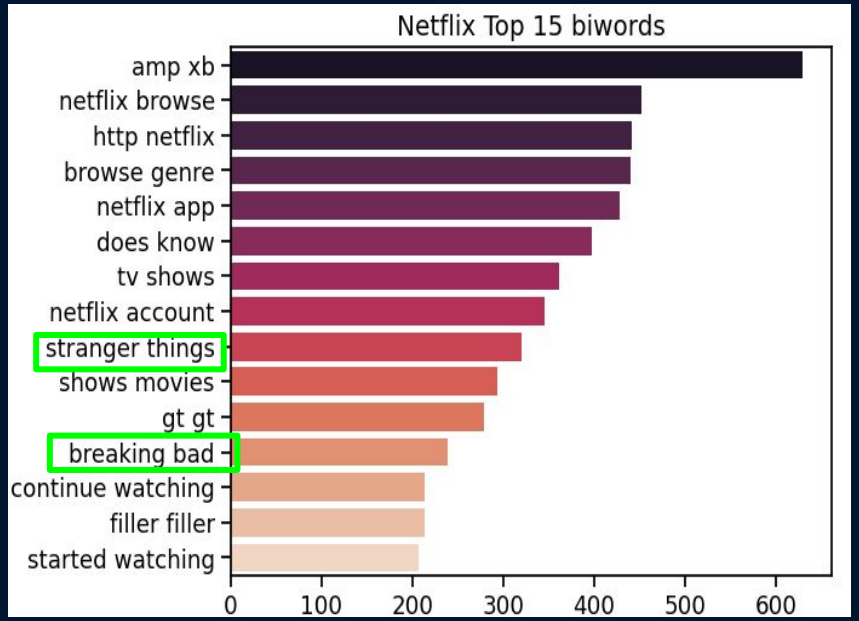
➔ Top 15 Words!

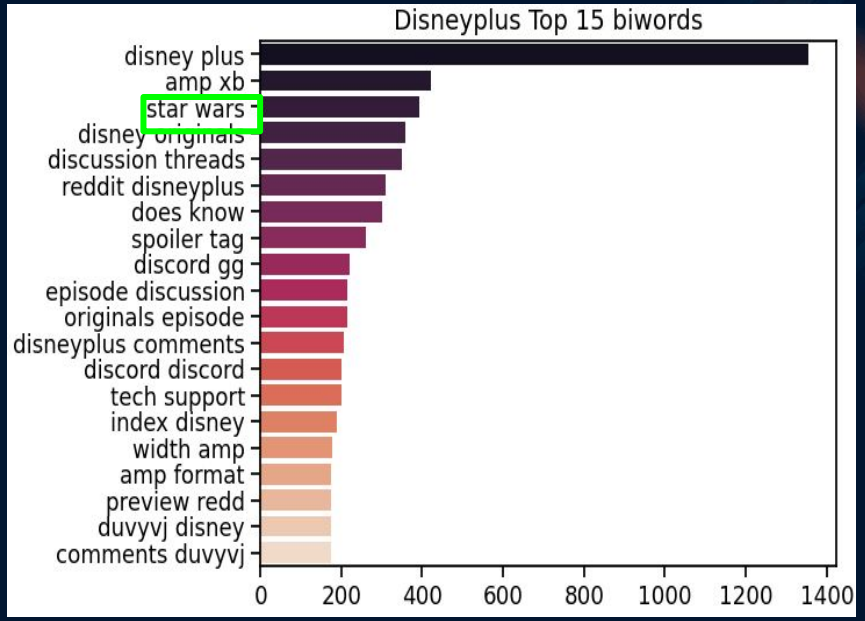




3) Exploratory Data Analysis

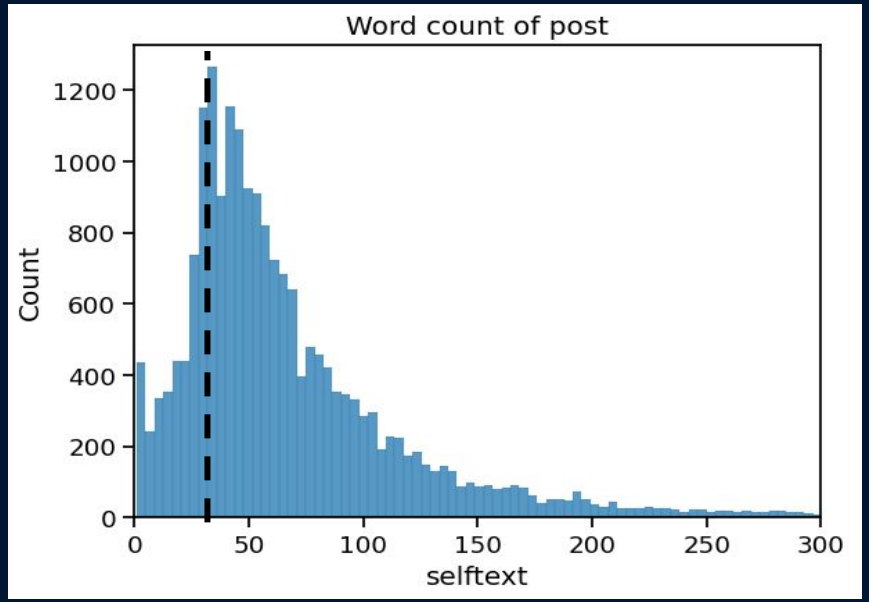
➔ Top 15 Bigrams!

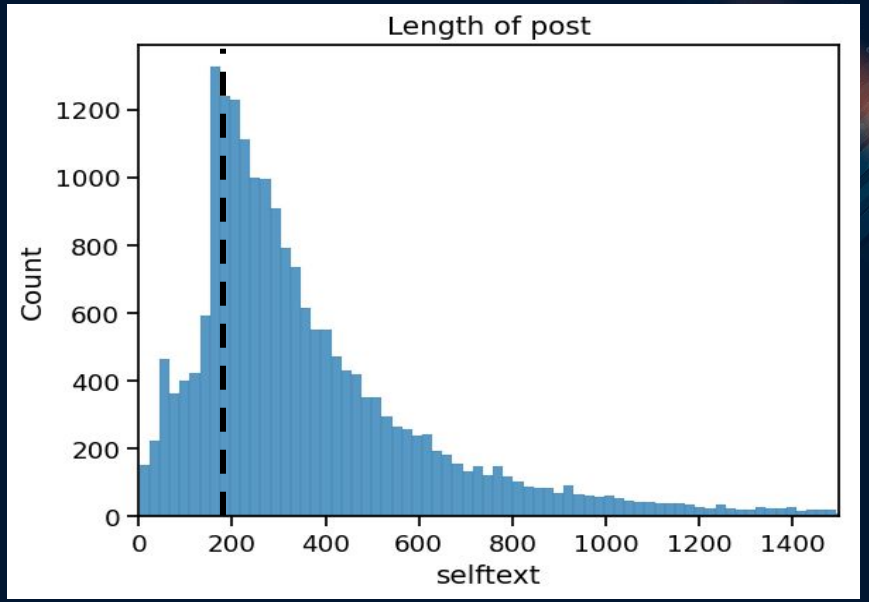




3) Exploratory Data Analysis

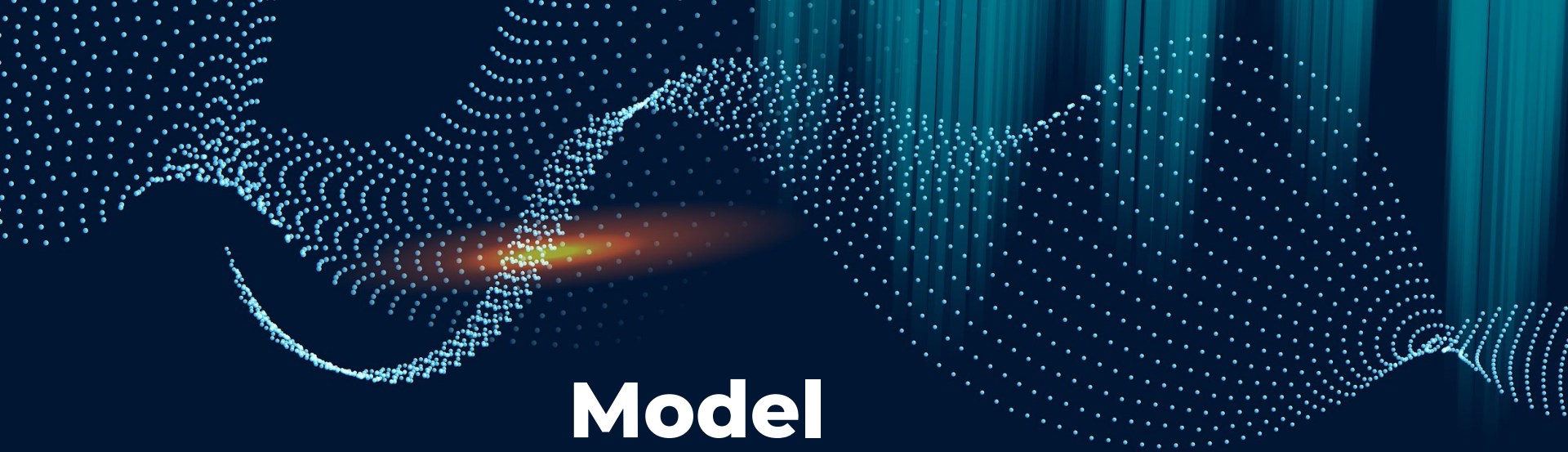
→ Posts text shape





4) Pre-processing

- Demojize emojis (😄 → :smile:)
- Expand contractions (can't → cannot)
- Lower case words (Word → word)
- Keep only words by regular expression
- Lemmatize words (watching → watch)
- Sentence augmentation by Synonym
  Make new sentence with similar words.



03

Model Performance & Evaluation

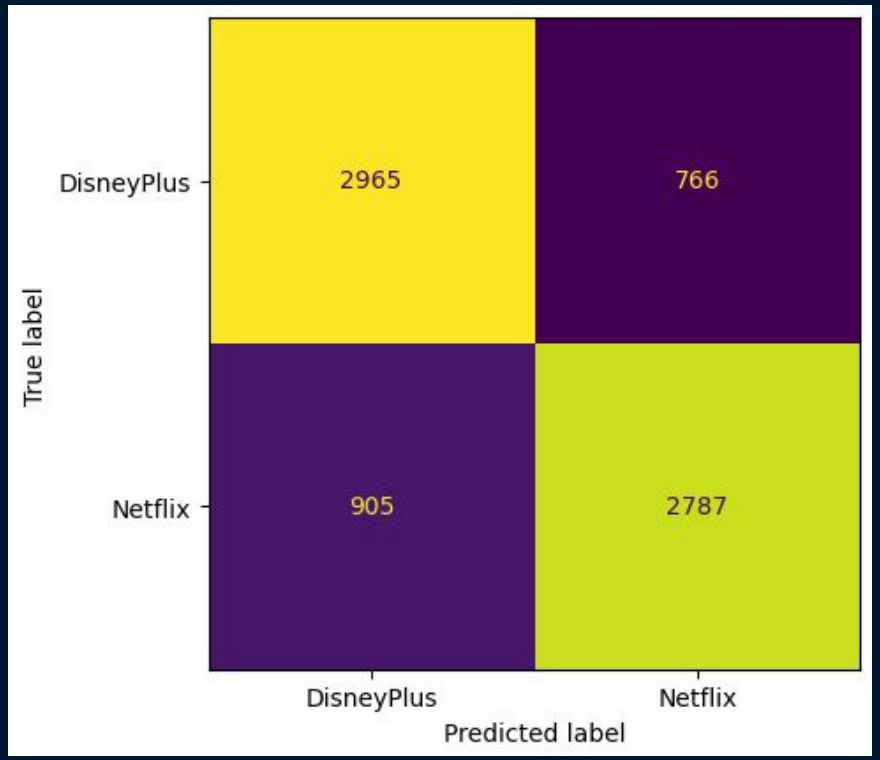
Locate the best performing model
Assess model metrics to the problem

Models used

Model	Vectorizer	Hyperparameter tuned
Naive Bayes	Count	No
Naive Bayes	Count	Yes
Naive Bayes	TF-IDF	Yes
Extra Trees Classifier	TF-IDF	Yes
Logistic Regression	TF-IDF	Yes
SVM - Linear Kernel	TF-IDF	Yes
Random Forest Classifier	TF-IDF	Yes

Baseline Model

Naive Bayes Model with Count Vectorization



Metric	Untuned	Tuned
Accuracy (Train)	0.767	0.766
Accuracy (Test)	0.774	0.774
Recall	0.75	0.75
Precision	0.78	0.78
F1 Score	0.77	0.77
Specificity	0.794	0.795

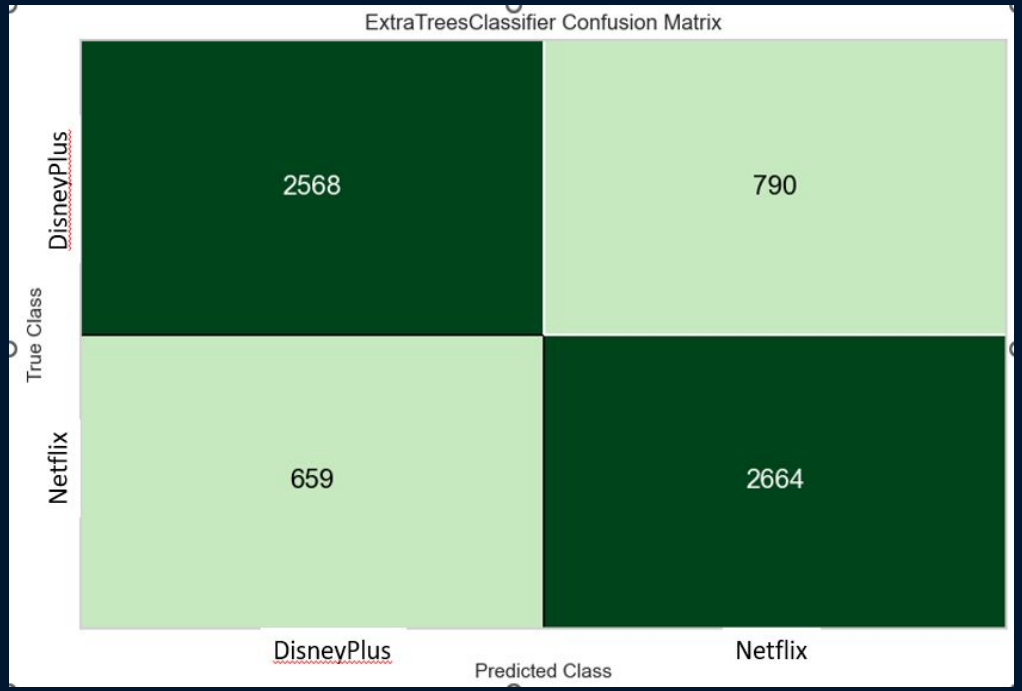
Model is neither overfitted nor underfitted

Multiple Model training

Model	Accuracy	AUC	Recall	Prec.	F1
Extra Trees Classifier	0.7854	0.8587	0.8026	0.7742	0.7881
Logistic Regression	0.7825	0.8698	0.8026	0.7699	0.7859
Ridge Classifier	0.7780	0.0000	0.7962	0.7666	0.7811
SVM - Linear Kernel	0.7773	0.0000	0.8043	0.7622	0.7821
Random Forest Classifier	0.7708	0.8542	0.8115	0.7487	0.7788
Light Gradient Boosting Machine	0.7676	0.8570	0.8300	0.7364	0.7803
Linear Discriminant Analysis	0.7648	0.8423	0.7780	0.7562	0.7669
Naive Bayes	0.7422	0.7981	0.6790	0.7749	0.7237
Decision Tree Classifier	0.7229	0.7307	0.7361	0.7152	0.7254
Gradient Boosting Classifier	0.7048	0.7922	0.9119	0.6434	0.7545
Ada Boost Classifier	0.6557	0.7373	0.9364	0.5984	0.7301
K Neighbors Classifier	0.6518	0.7123	0.8028	0.6150	0.6962
Quadratic Discriminant Analysis	0.5869	0.5881	0.8284	0.6247	0.6234
Dummy Classifier	0.5027	0.5000	0.0000	0.0000	0.0000

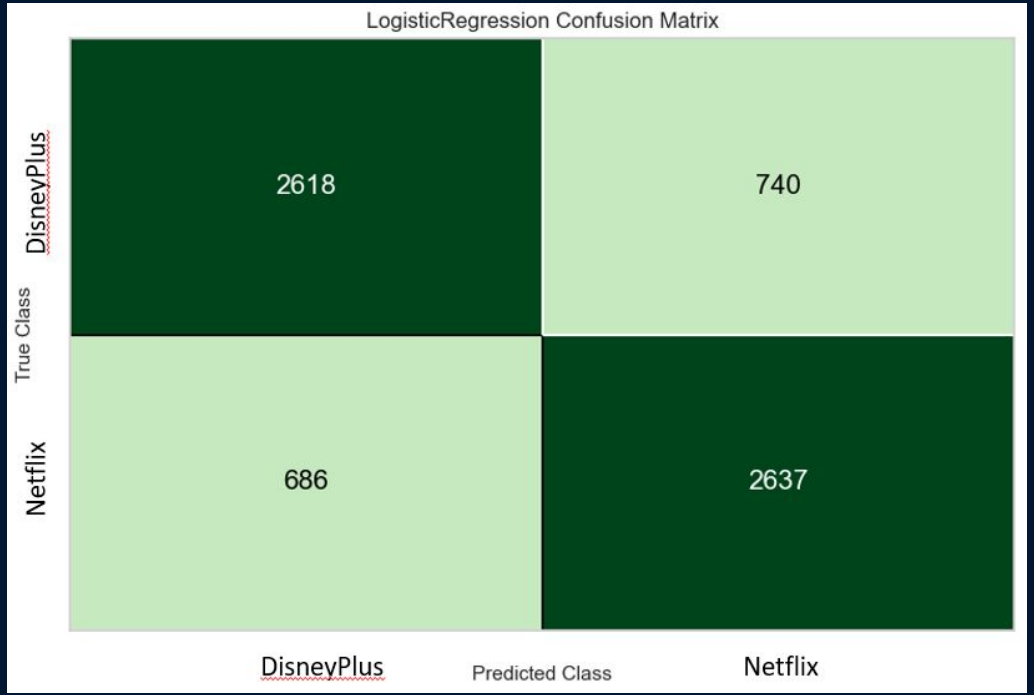
- Leverage on PyCaret to train several classification model & find the best performing one
- TF-IDF Vectorizer is used
  - Grants more score for words with more predictive power
- Below models will be accessed further
  - Extra Trees Classifier
  - Logistic Regression
  - SVM - Linear Kernel

Extra Trees Classifier (Tuned)



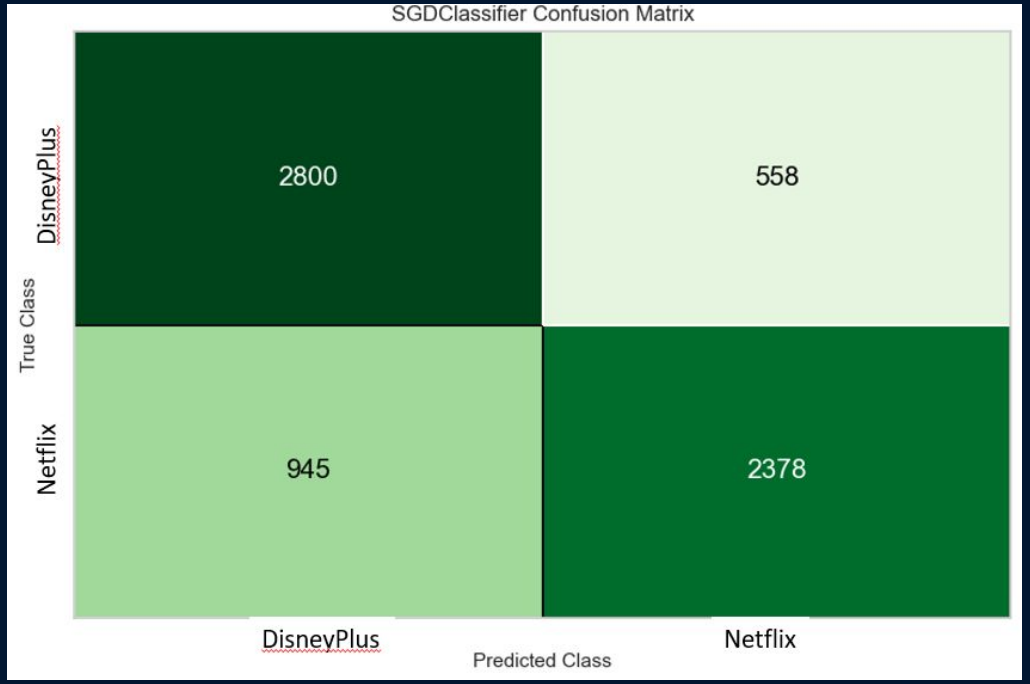
Metric	Tuned
Accuracy (Train)	0.719
Accuracy (Test)	0.792
Recall	0.802
Precision	0.771
F1 Score	0.786
Specificity	0.765

Logistic Regression (Tuned)



Metric	Tuned
Accuracy (Train)	0.783
Accuracy (Test)	0.795
Recall	0.794
Precision	0.781
F1 Score	0.787
Specificity	0.780

SVM - Linear Kernel (Tuned)



Metric	Tuned
Accuracy (Train)	0.767
Accuracy (Test)	0.791
Recall	0.716
Precision	0.810
F1 Score	0.760
Specificity	0.834

Model Evaluation

Model	Vectorizer	Accuracy (Train)	Accuracy (Test)
Naive Bayes (Baseline)	Count	0.767	0.774
Naive Bayes (Tuned)	Count	0.766	0.774
Extra Trees Classifier	TF-IDF	0.719	0.792
Logistic Regression	TF-IDF	0.783	0.795
SVM-Linear Kernel	TF-IDF	0.767	0.791
Random Forest Classifier	TF-IDF	0.688	0.779
Naive Bayes (Tuned)	TF-IDF	0.767	0.777

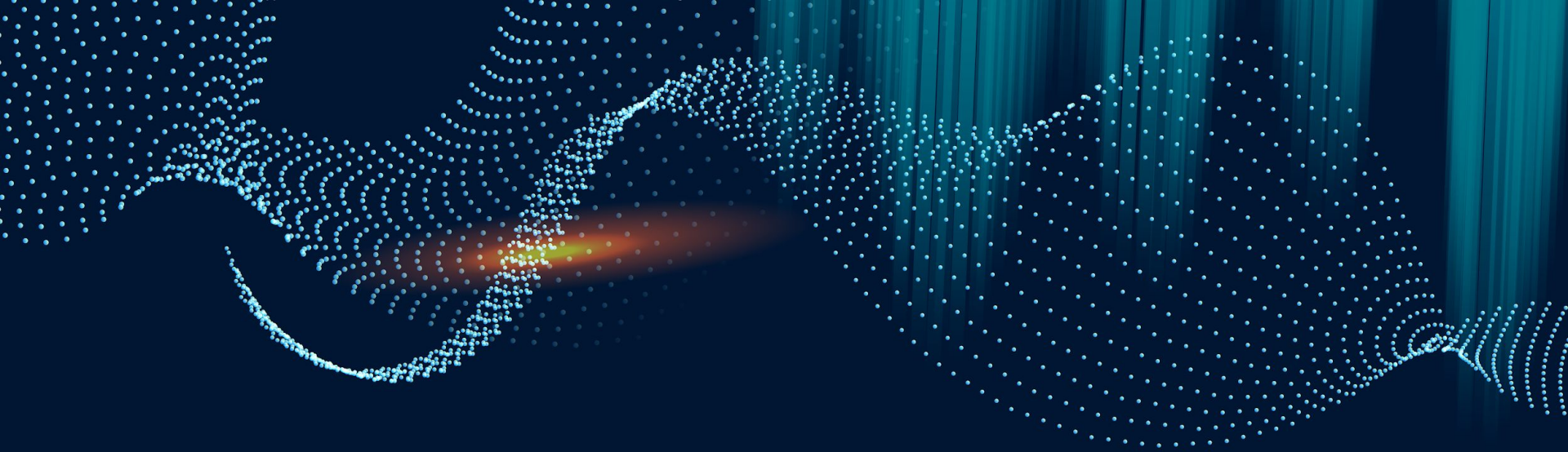
Accuracy score is the most relevant metric for this problem statement

Model Evaluation

Model	Accuracy	F1 score	Recall	Precision	Specificity	AUC
Logistic Regression	**0.783**	**0.787**	**0.794**	**0.781**	**0.780**	0.871
SVM-Linear Kernel	0.767	0.760	0.716	0.810	0.834	0.000
Extra Trees Classifier	0.719	0.786	0.802	0.771	0.765	0.813

- Logistic Regression
  - Overall metrics is the most balanced (Accuracy, F1 score, Recall, Precision, Specificity differ by at most 0.01)
  - Highest AUC score

- SVM-Linear Kernel & Extra Trees Classifier
  - Metrics have a wide variance across each other (usually more than 0.01)

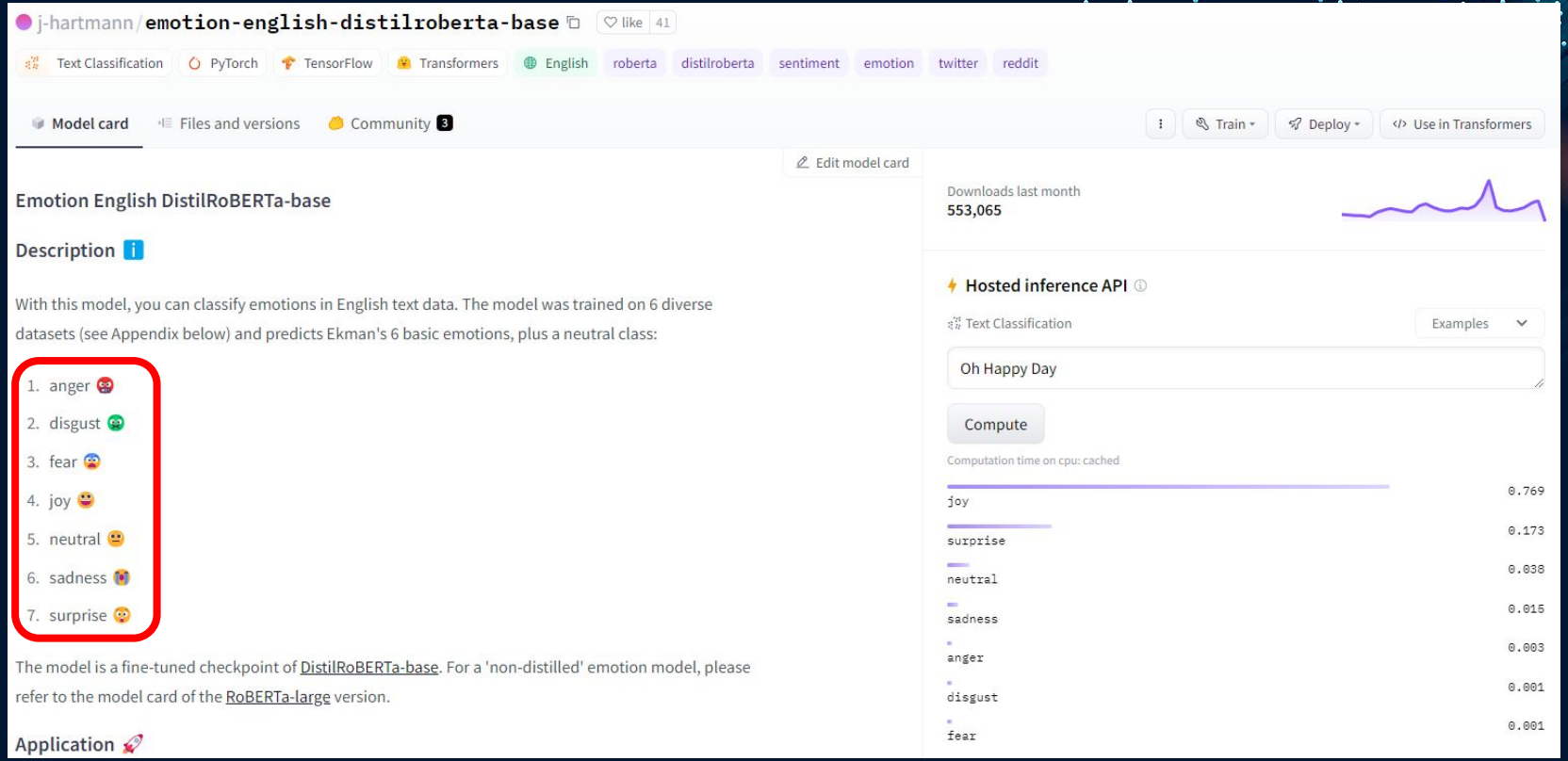


04 Sentiment Analysis

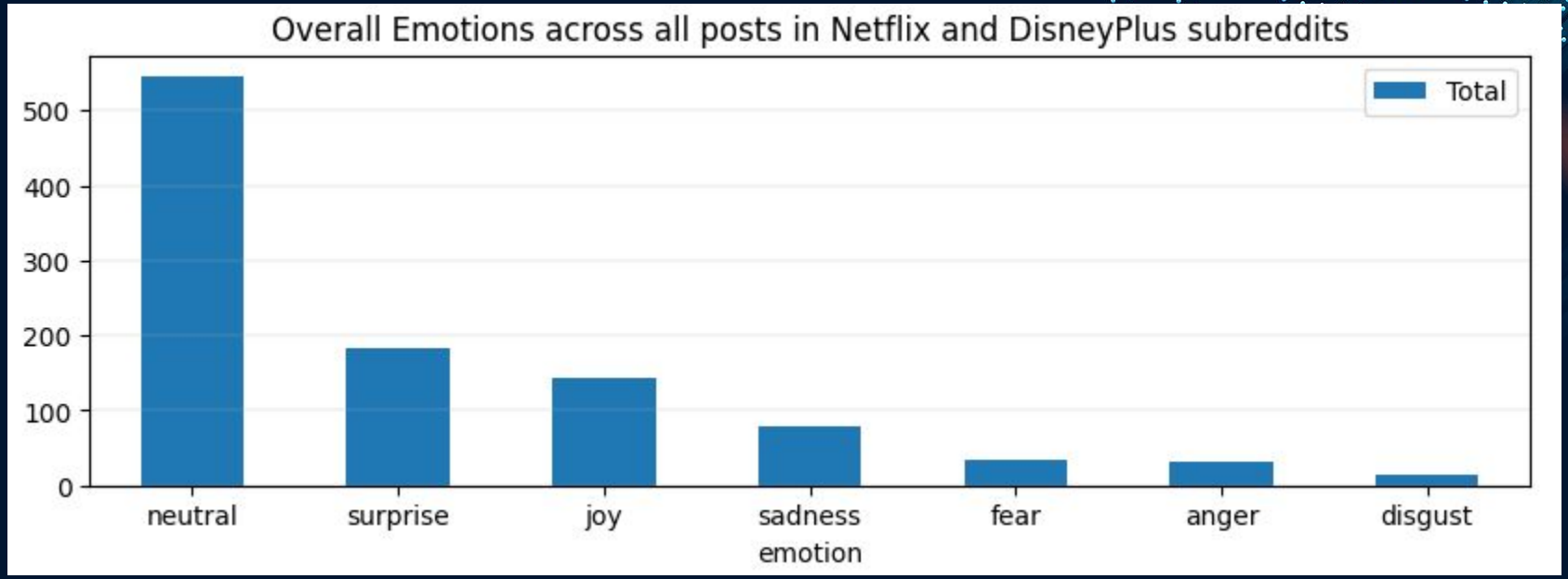
Users' opinion on shows

BERT model used

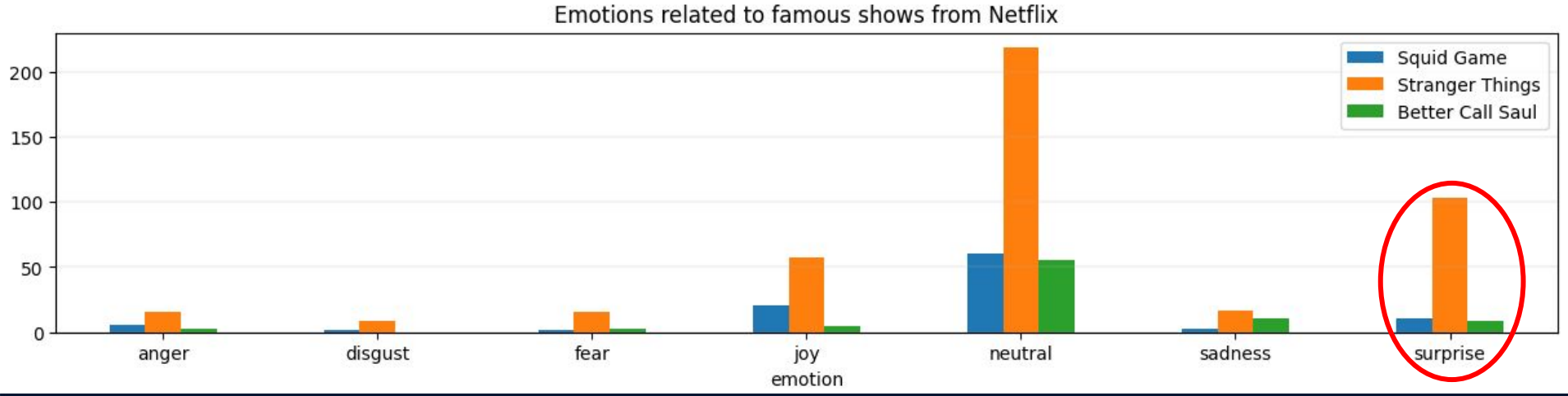
Link to BERT model



Sentiment Analysis

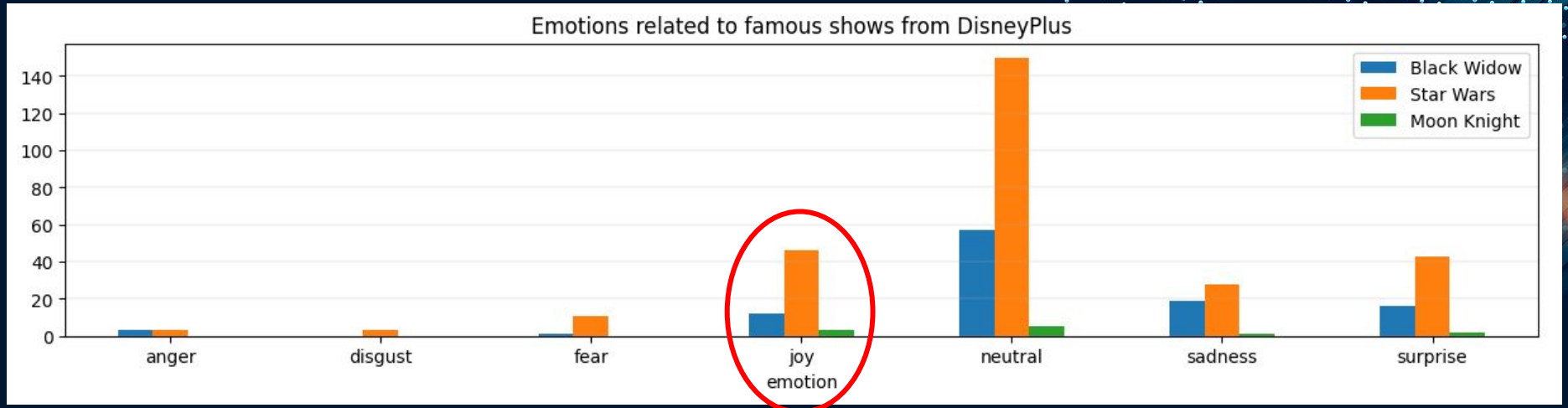


Sentiment analysis (Netflix)



- 2nd most common emotion: Surprise
  - Most posts were related to Stranger Things

Sentiment analysis (Disney)

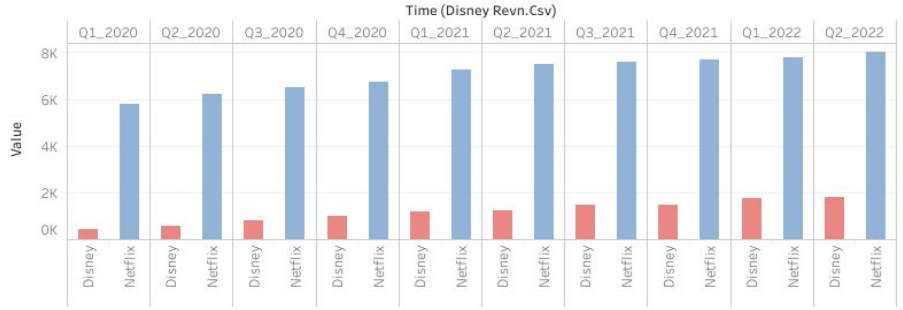


- 2nd most common emotion: Joy
  - Most posts were related to Star Wars

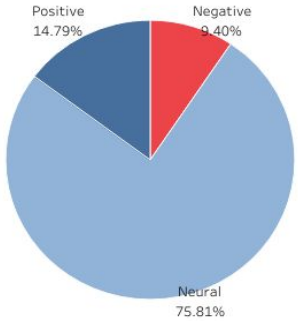
05 Conclusion Dashboard

POPCORN WAR: NETFLIX vs DISNEY+

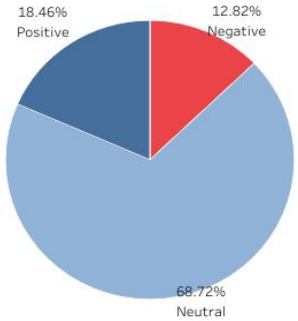
Quarterly revenues (in million USD)
Sources: Statista, Businessofapps



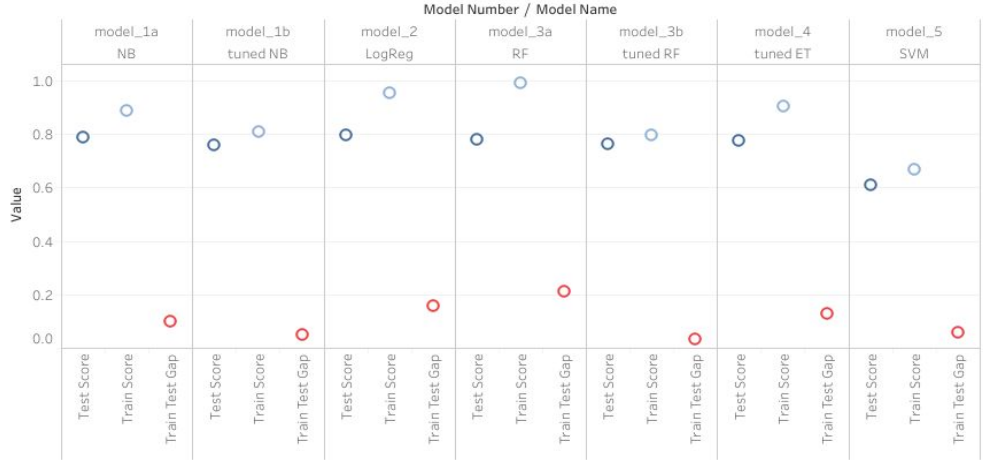
Disneyplus



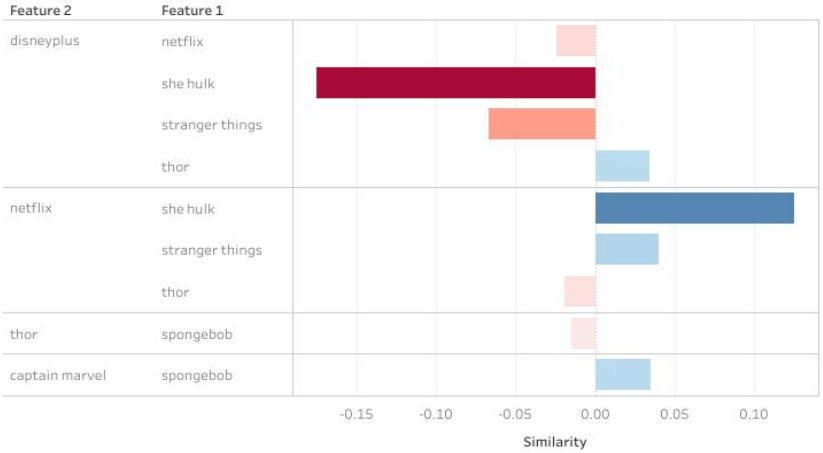
Netflix

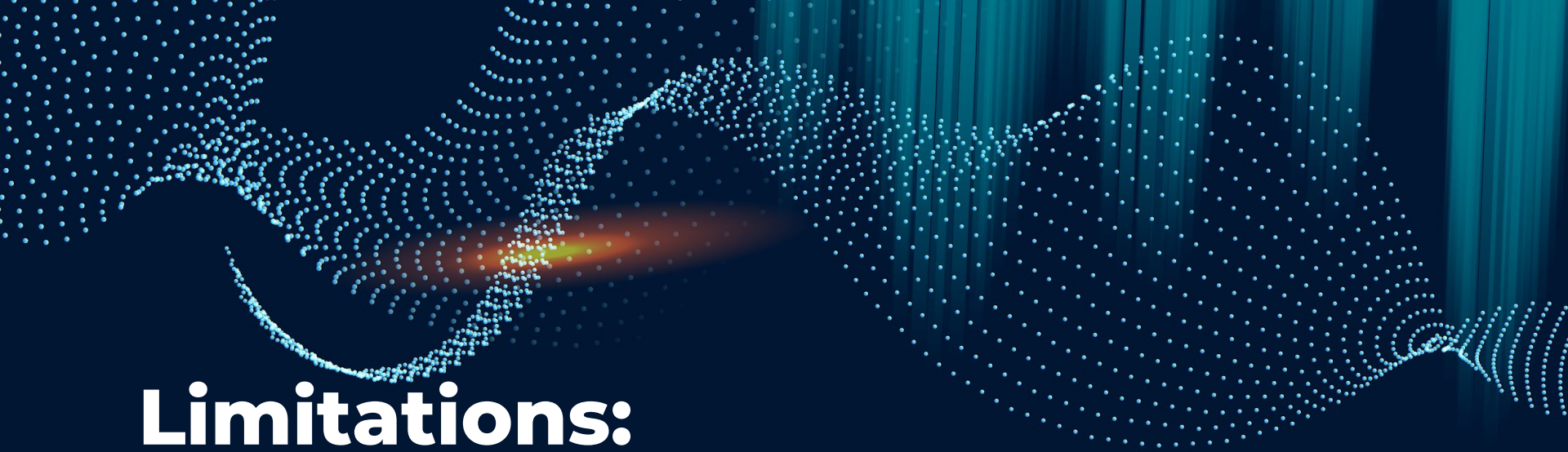


Model score comparison



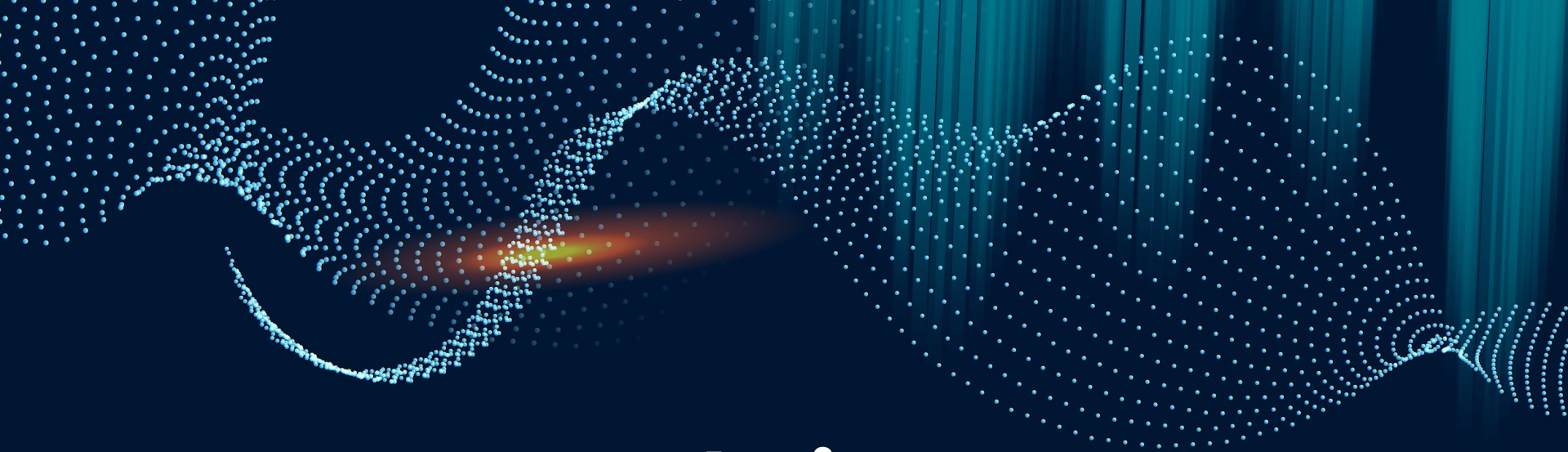
Word2Vec Similarity Scores





Limitations:

- Data limitation
- Model limitation
- Sentiment analysis limitation vs human emotions



Recommendations:

- Logistic Regression
- Data collection: specific timing
- Capture more specific and wider range of emotions: anticipations, conspiracy theories

Thank you