STAR WARS Vs MARVEL STUDIOS

Subreddit Classification

BILL FU
AUGUST 28 2020

Overview

- Objective:

Using natural language processing (NLP) to train binary classifiers to determine whether a reddit post came from “StarWars” or “marvelstudios”.

- Outlines
  - Data preparation
  - Text Vectorizers comparison
  - Classifiers comparison
  - The optimized classifier evaluation
  - Conclusions

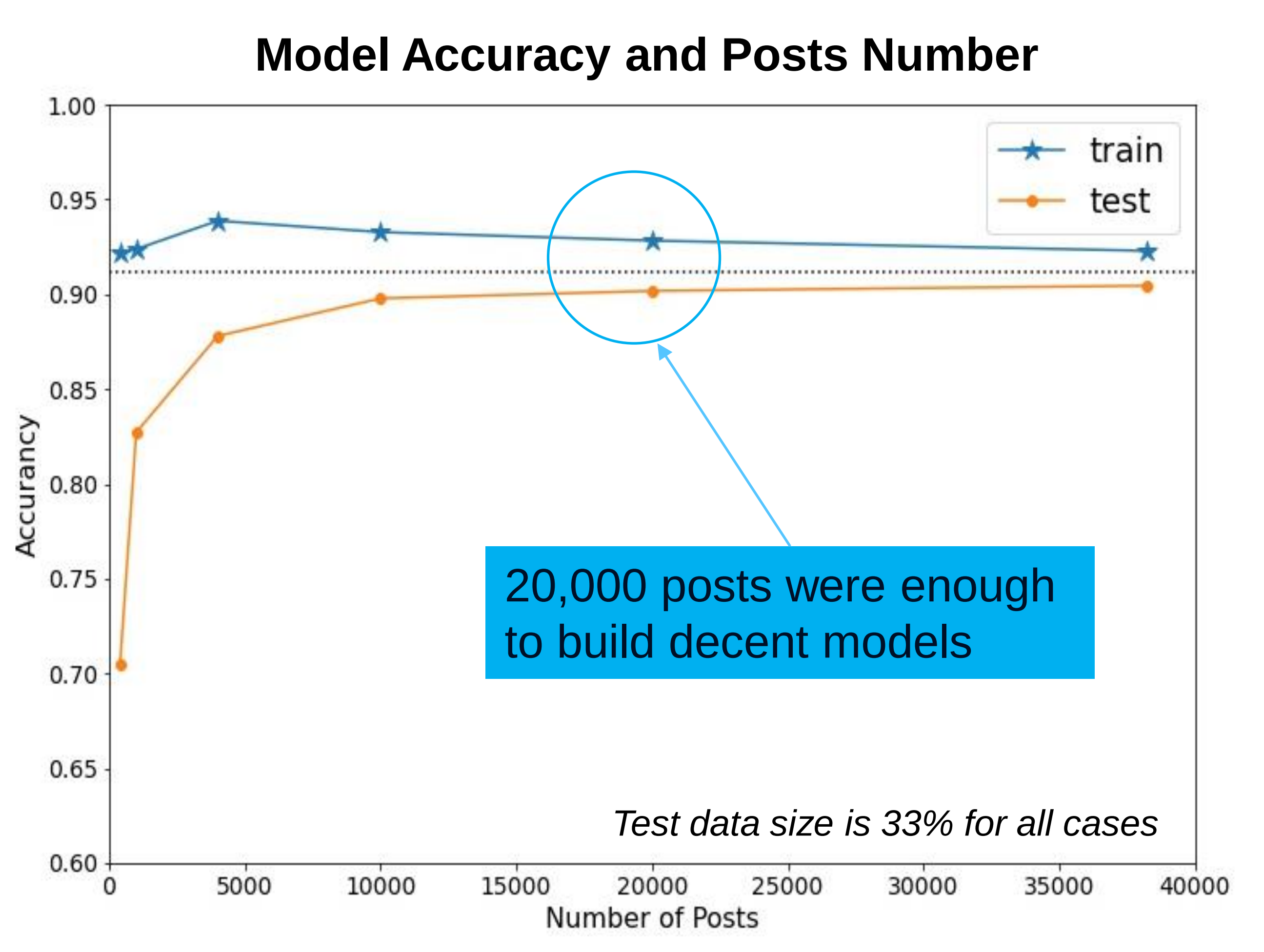
Data preparation

- Data was collected from reddit.com using the “pushshift” application programming interface (API)
- 20,000 posts were collected from each of the two subreddits:
  - “StarWars”
  - “marvelstudios”
- Only the titles of the posts were used, and duplicated titles were discarded
- Final datasets:
  - 19,044 posts from “StarWars”
  - 19,184 posts from “marvelstudios”

How many posts are needed?

Using the Multinomial Naïve Bayes (MNB) classifier, evaluations were carried out for the following cases*:

- 400 posts
- 1,000 posts
- 4,000 posts
- 10,000 posts
- 20,000 posts
- 38,228 posts (all the collected posts)

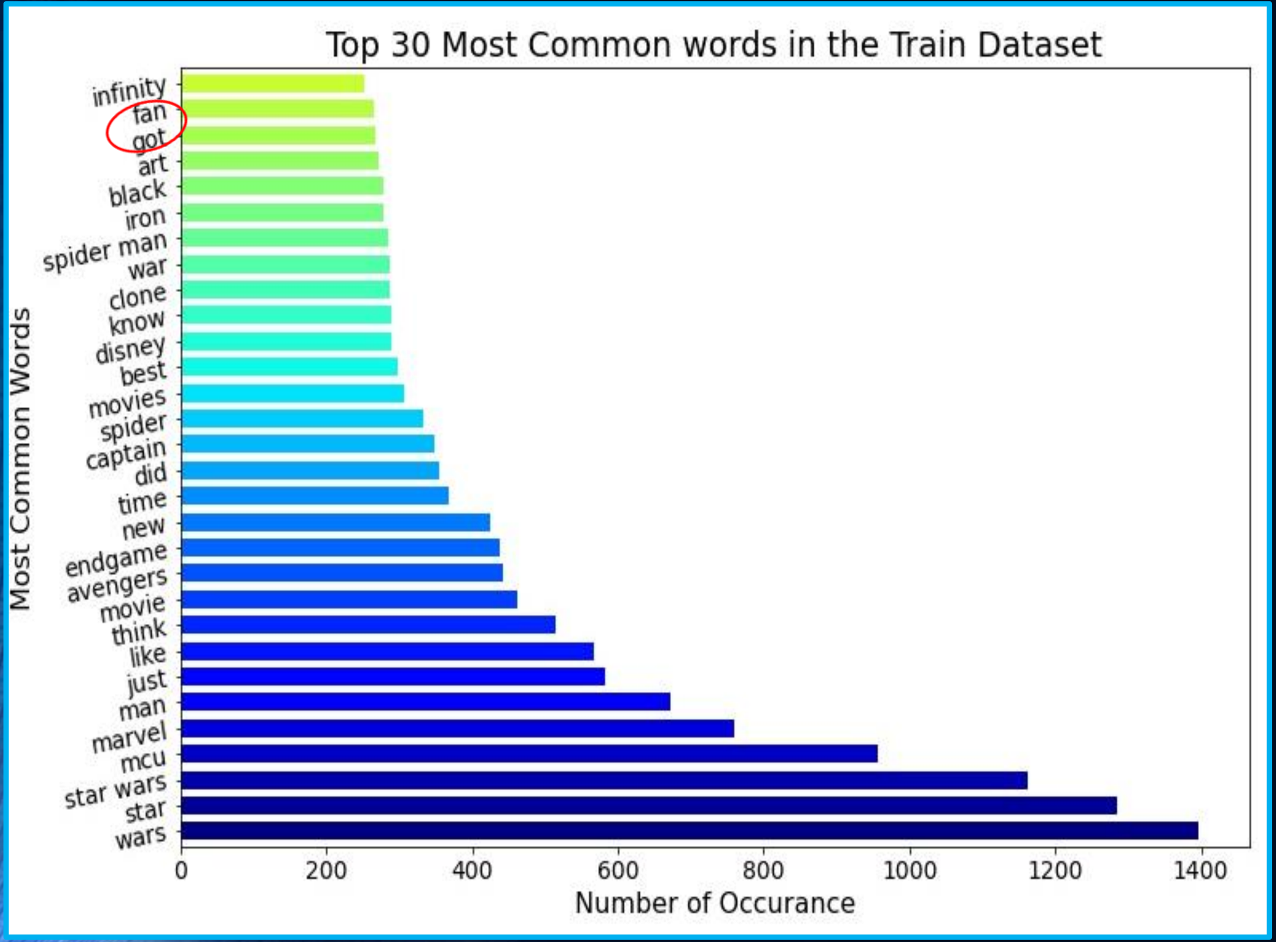


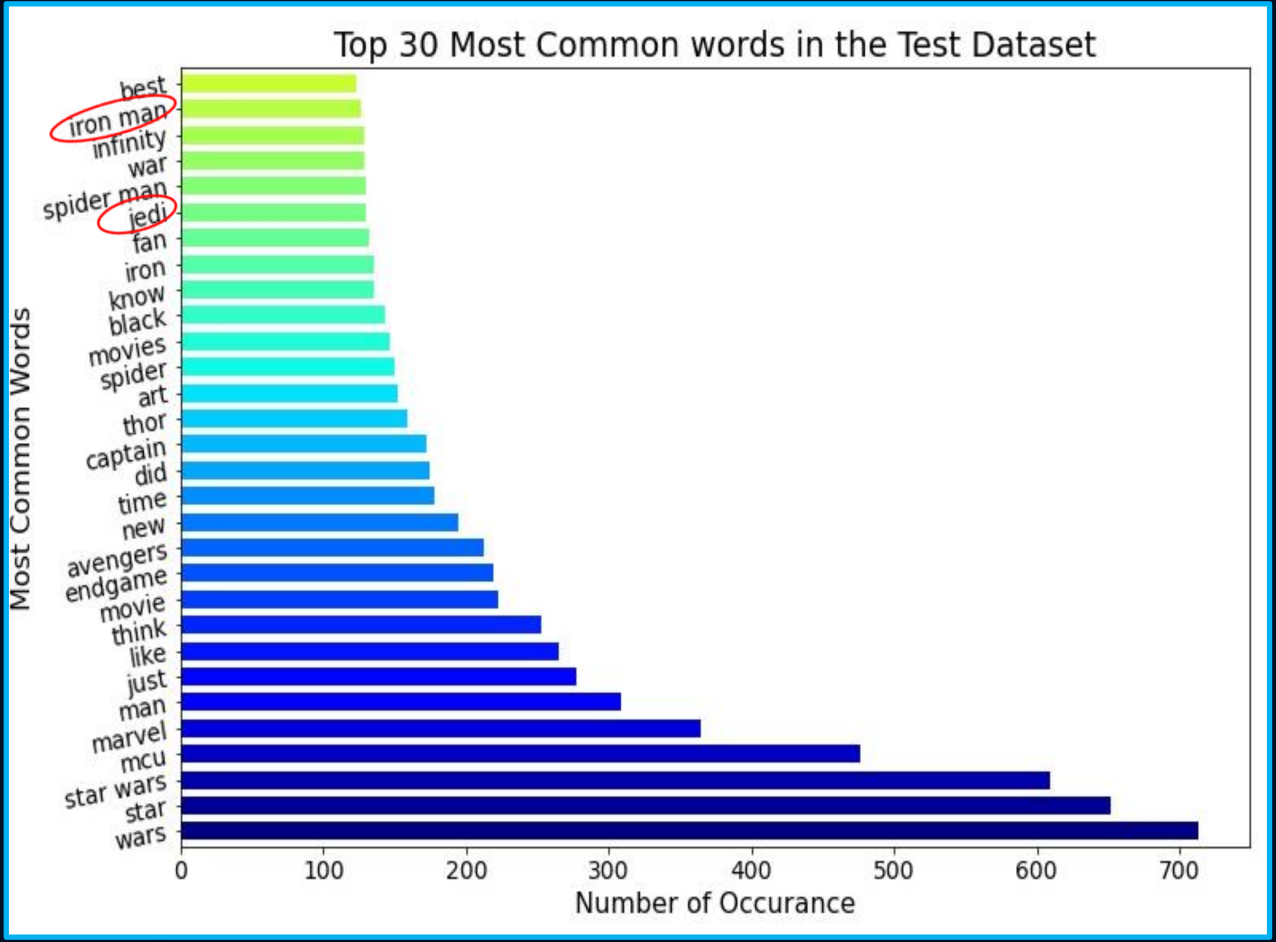
20,000 posts were enough to build decent models

Test data size is 33% for all cases

* The number of posts from each subreddit are the same – Baseline model accuracy is 0.5

Most common words in the posts





- Texts were converted by count-vectorizer with stop_words = ‘english’
- The top 30 most common words are consistent in the training and testing datasets

Count vs Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizers

- Count and TF-IDF Vectorizers had different optimized parameters:
  - Count: English stop words

    single terms and bigrams
  - TF-IDF: No stop words

    single terms

- Count and TF-IDF Vectorizers gave similar prediction results

PREDICTION SCORES*

	Count	TF-IDF
Training	0.928	0.937
Testing	0.902	0.900

* Predicted from MNB classifiers

Classifiers Comparison

PREDICTION SCORES

	MNB	KNN	Logistic Regression	Random Forest	SVC
Train	0.928	0.873	0.957	0.937	0.970
Test	0.902	0.792	0.903	0.888	0.875

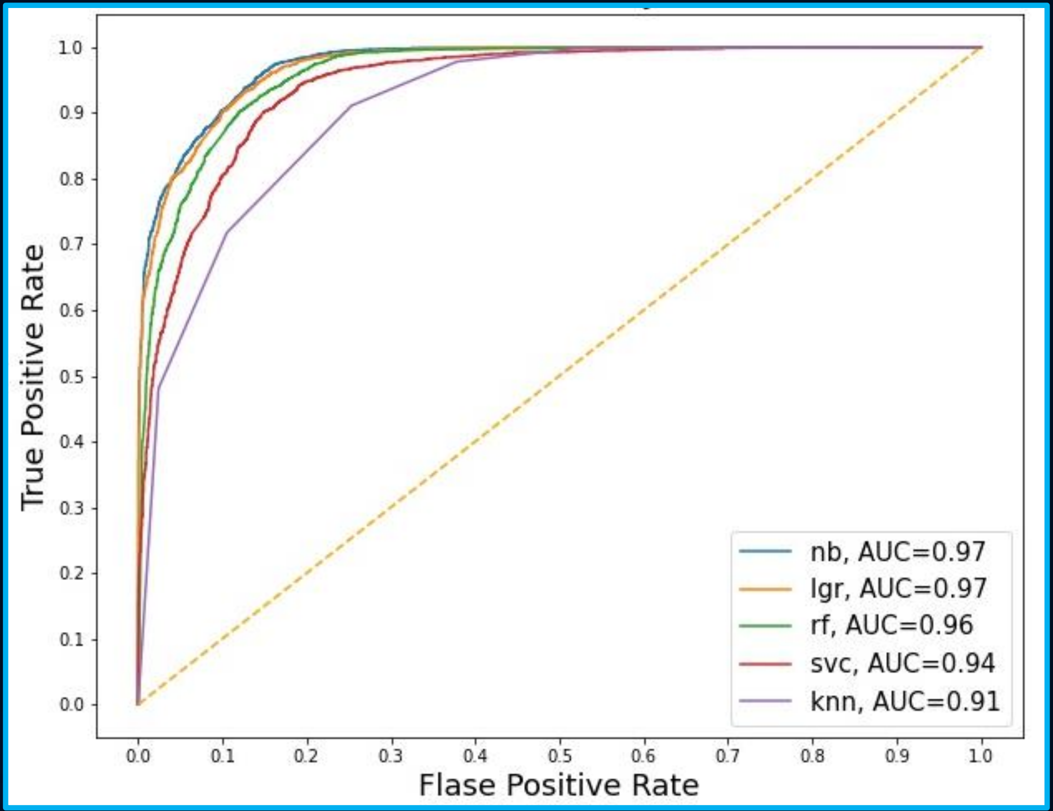
- Logistic Regression (LGR) and MNB classifiers gave good predictions on both training and testing datasets

- Random forest (RF) and support vector machine (SVC) classifiers gave good predictions on training datasets, but not so good predictions on testing datasets.

- KNN classifiers performed not as good as other classifiers

- Receiver operating characteristic (ROC) analysis on different classifiers was consistent with the prediction scores

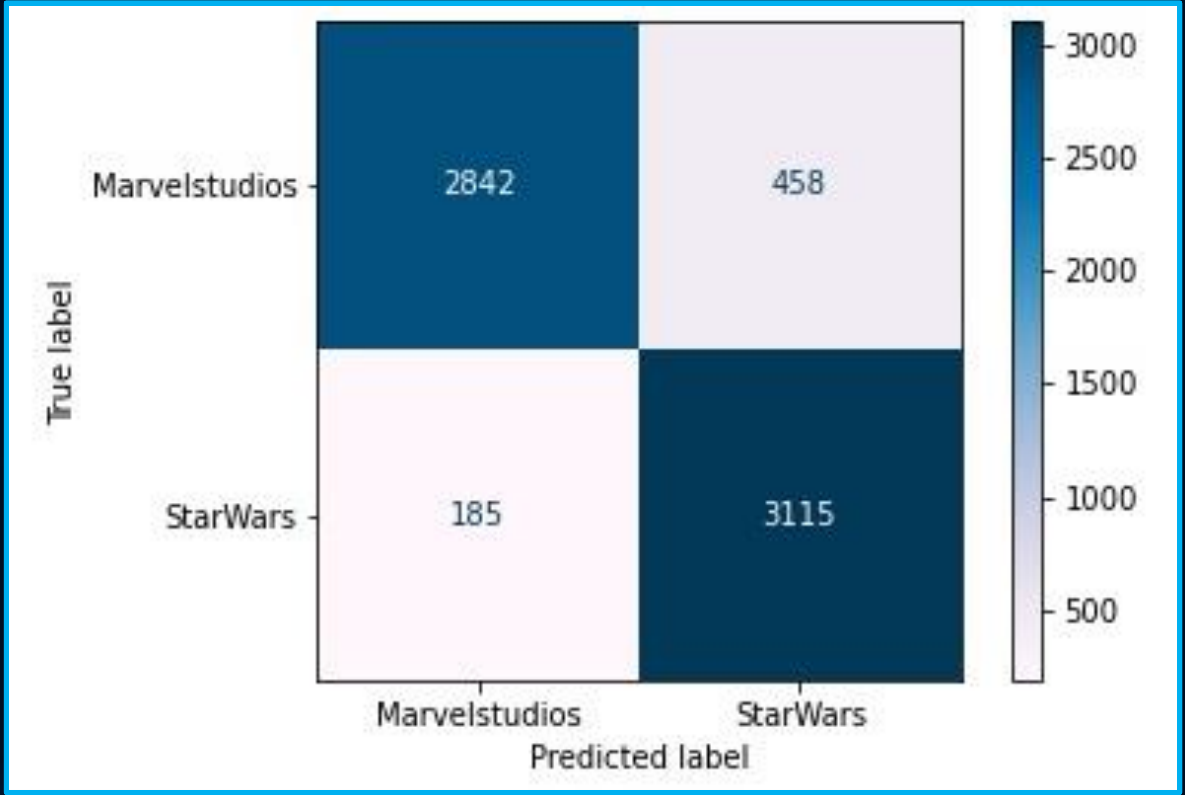
ROC CURVES



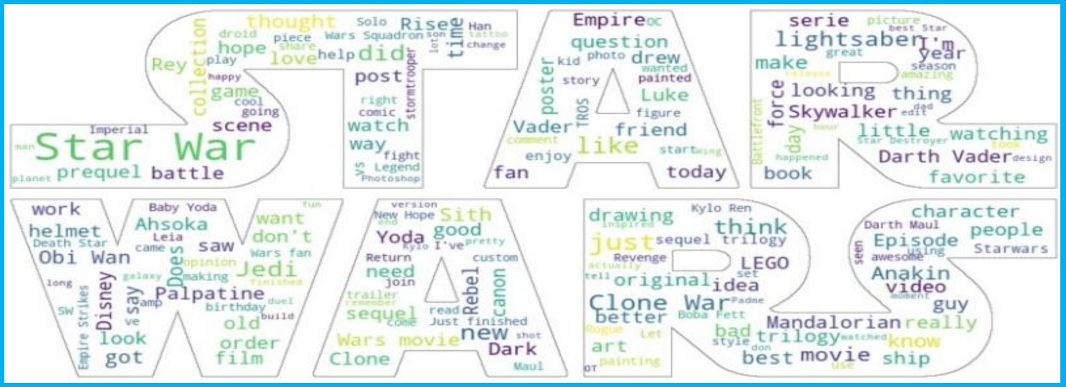
Optimization of Logistic Regression Classifiers

- Best parameters
  - CountVectorizer: max_feature=8000,
    ngram = (1, 2),
    stop_words = ‘english’
  - LogisticRegression: C=1.2
- Prediction scores
  - Training: 0.961
  - Testing: 0.903
- Sensitivity: 0.944
- Specificity: 0.861

Confusion matrix on testing datasets

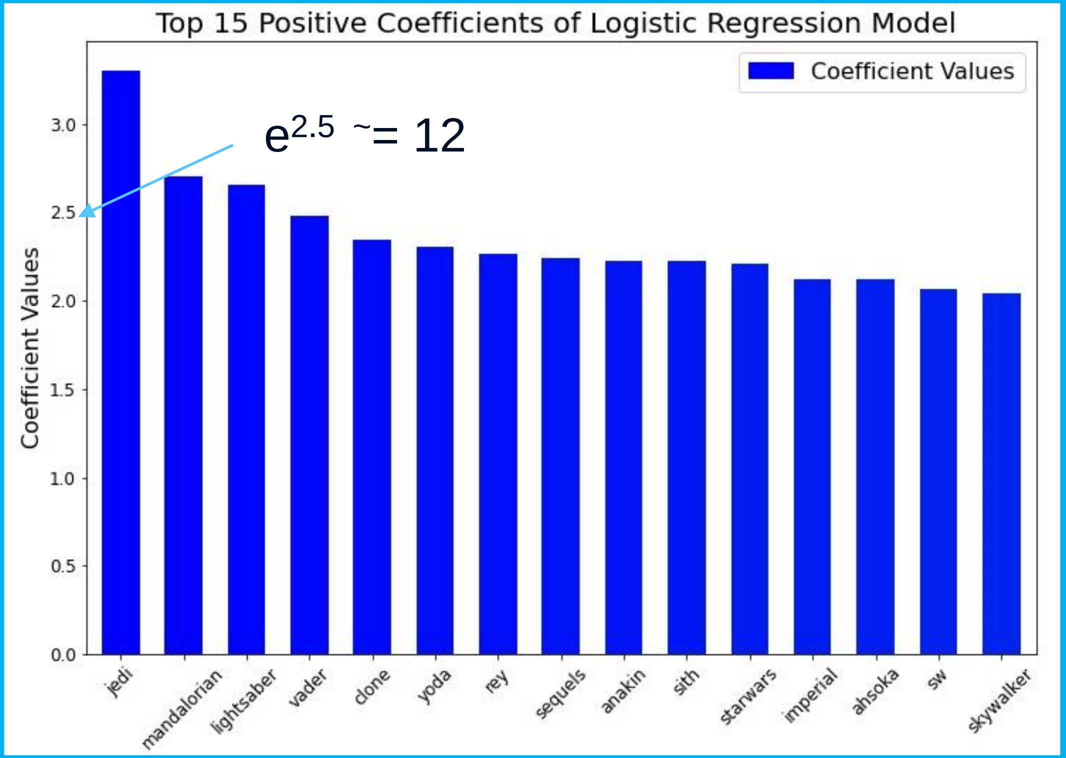


Top driving words for Logistical Regression Classifiers



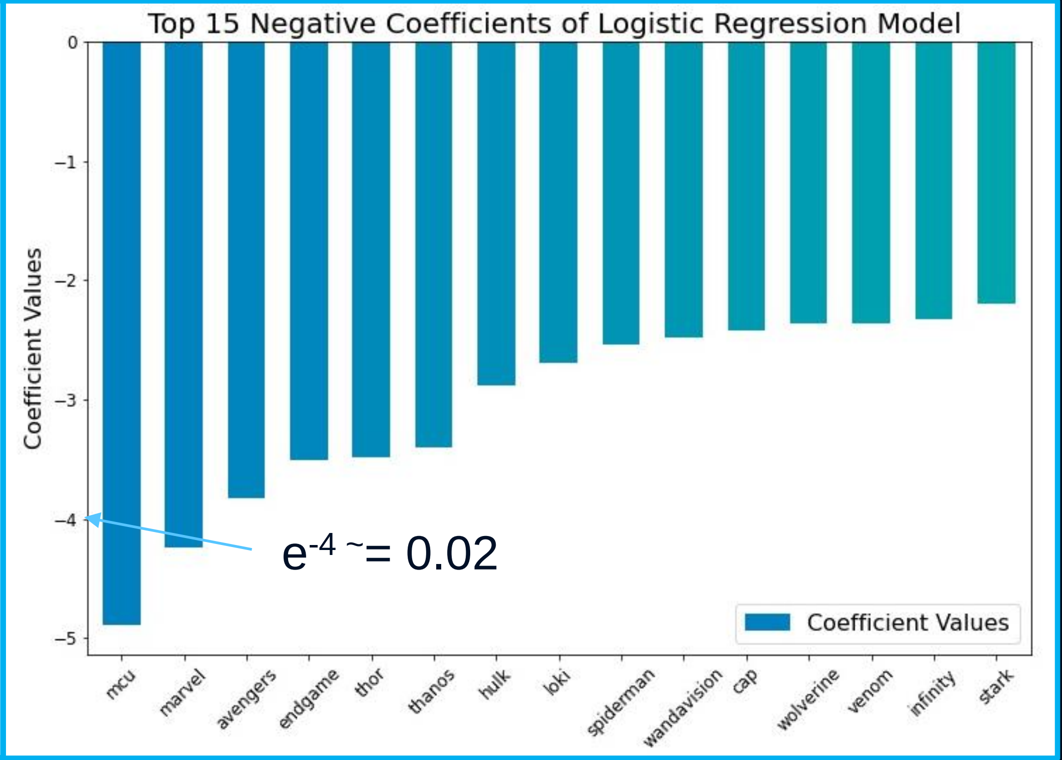
MCU
movie
Marvel
Endgame
Avenger
Spider Man
Black Panther
think
time
Loki
Civil War
villain
Disney
thought
Hulk
Iron Man
Doctor Strange
Black Widow
Thor
Thanos
Captain America
Infinity War
Ultron
fan
drawing
Phase
poster
question
Theory
character
SHIELD
just
new
year
idea

Top 15 Positive Coefficients of Logistic Regression Model



$e^{2.5}$ ~= 12

Top 15 Negative Coefficients of Logistic Regression Model



e^{-4} ~= 0.02

Removing the top 15 driving words:

- Accuracy

  0.903
  ↓
  0.853

- Sensitivity

  0.944
  ↓
  0.878

- Specificity

  0.861
  ↓
  0.829

Conclusions and future work

- Binary classifiers were successfully trained to determine whether a reddit post came from “StarWars” or “marvelstudios”.
- Different text vectorizers and classifiers were compared and the optimized classifier could gave decent predicting results.
- There are still rooms to reduce the overfitting of the classifiers.
  - Bagging the classifiers
  - Reducing the number of features
  - Reducing the model complexity