

**PROJECT 3:**

**NLP REDDIT PROJECT**

**MARVEL**



**UNITED**

Presented by: Daniel | Peter | Supriya | Wei Hua

# ***PROBLEM STATEMENT***

## Who's the Client?

Marvel DC United specialising in video games

## Who are we?

Data scientists from a consultancy firm

## Task:

1. Classify Marvel and DC fans on reddit for targeted advertising of the client's new superhero game
2. Identify and recommend most discussed heroes to be included in the game

# STAKEHOLDERS

## Primary Stakeholder

→ Video Game Creator



## Secondary Stakeholder

→ Video Game Consumers & Reddit Users



# CONTEXT

## Reddit:

An American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members.

## Marvel vs DC:

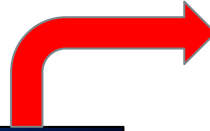
Rival American superhero franchises centered on the stories of two different groups of superheroes.

## Dataset:

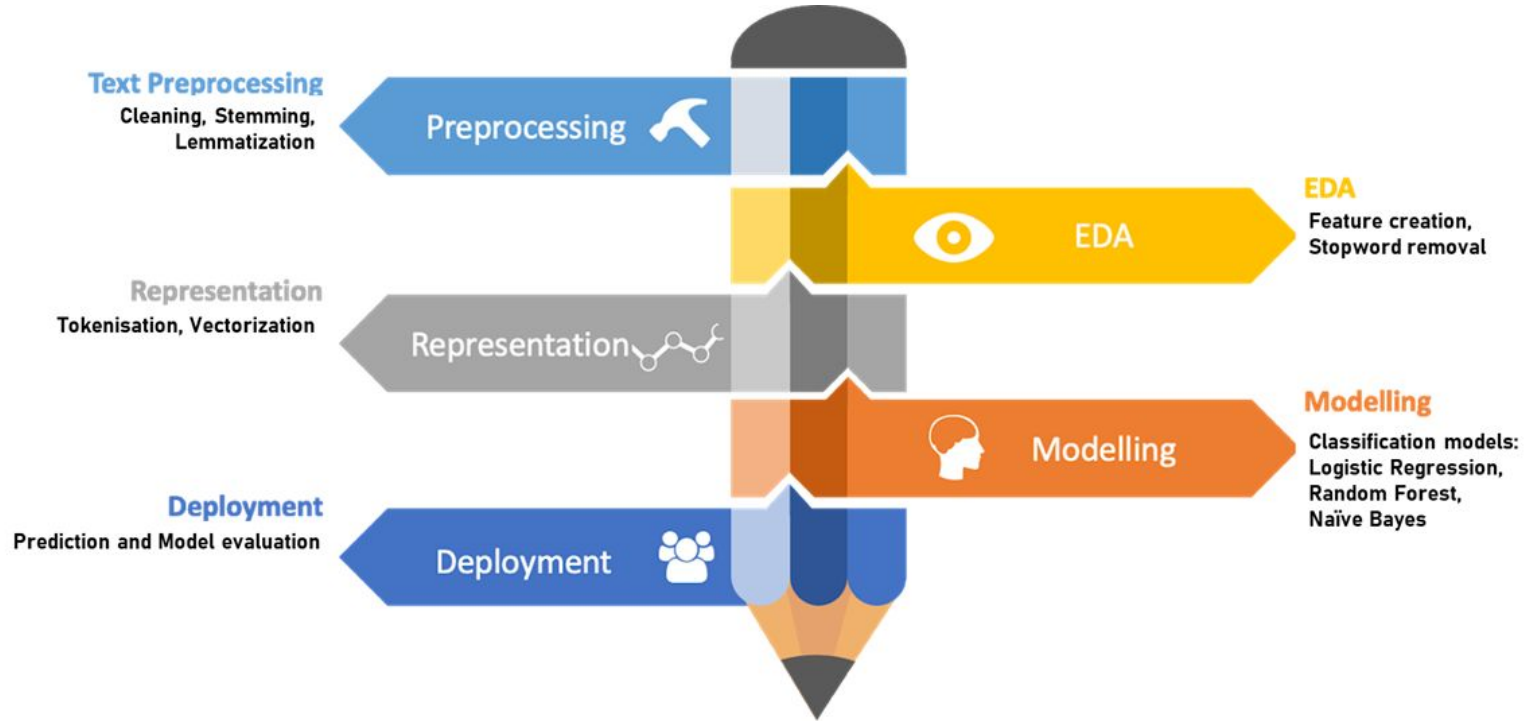
1,000+ posts each from r/marvelstudios and r/DC\_Cinematics



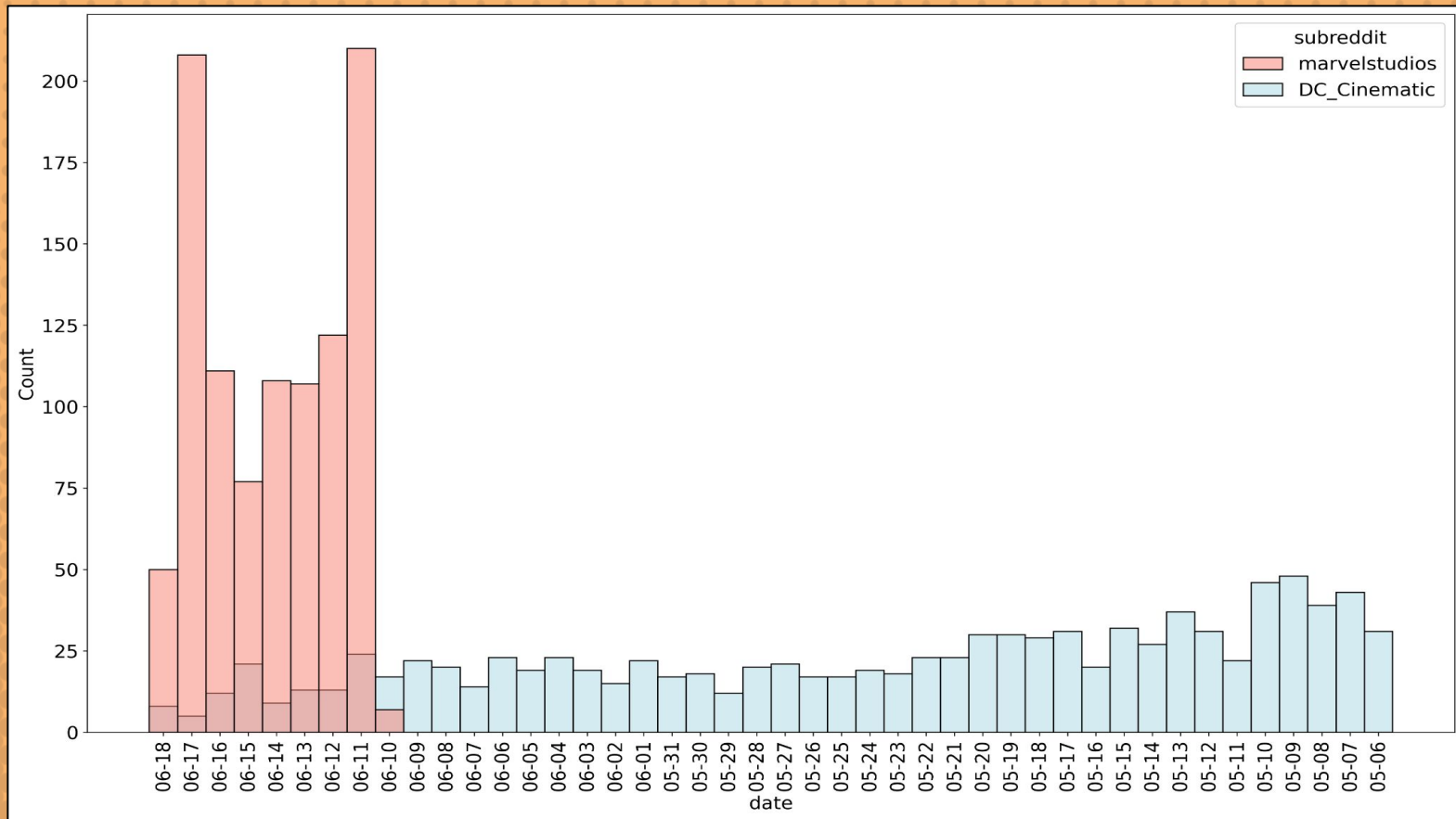
# WHAT WE AIM TO DO



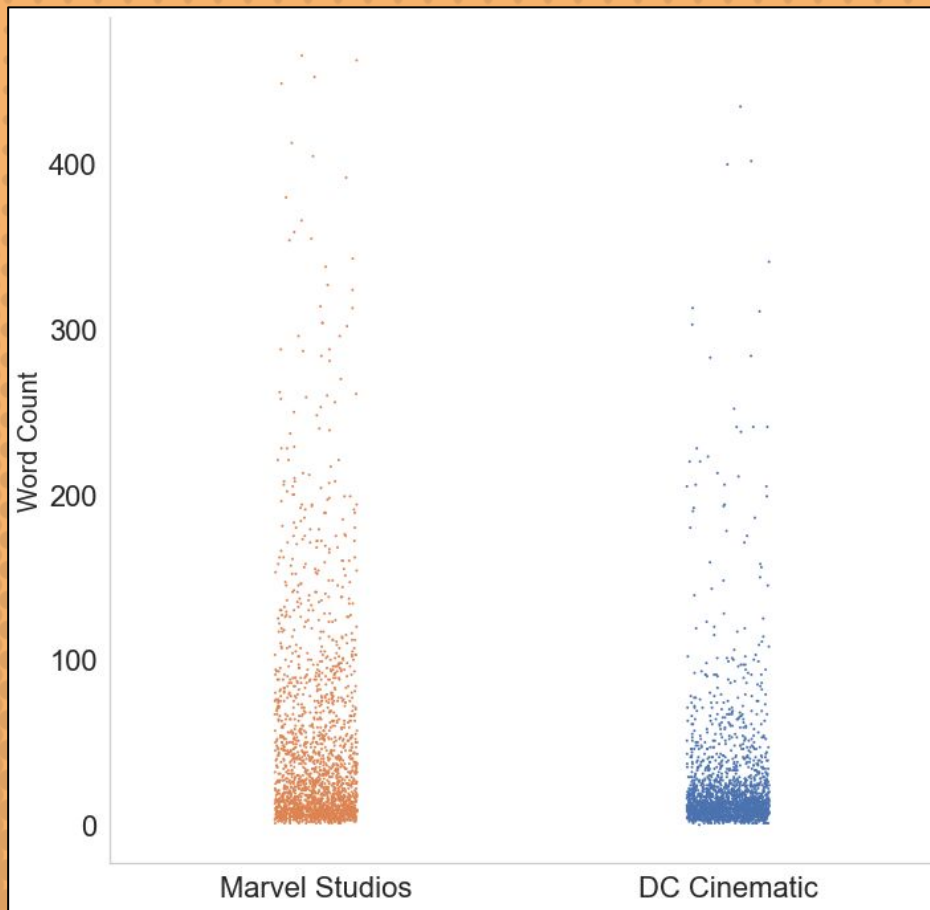
# THE DATA SCIENCE PROCESS



# ***DISTRIBUTION OF 1000+ POSTS PER SUBREDDIT OVER TIME***

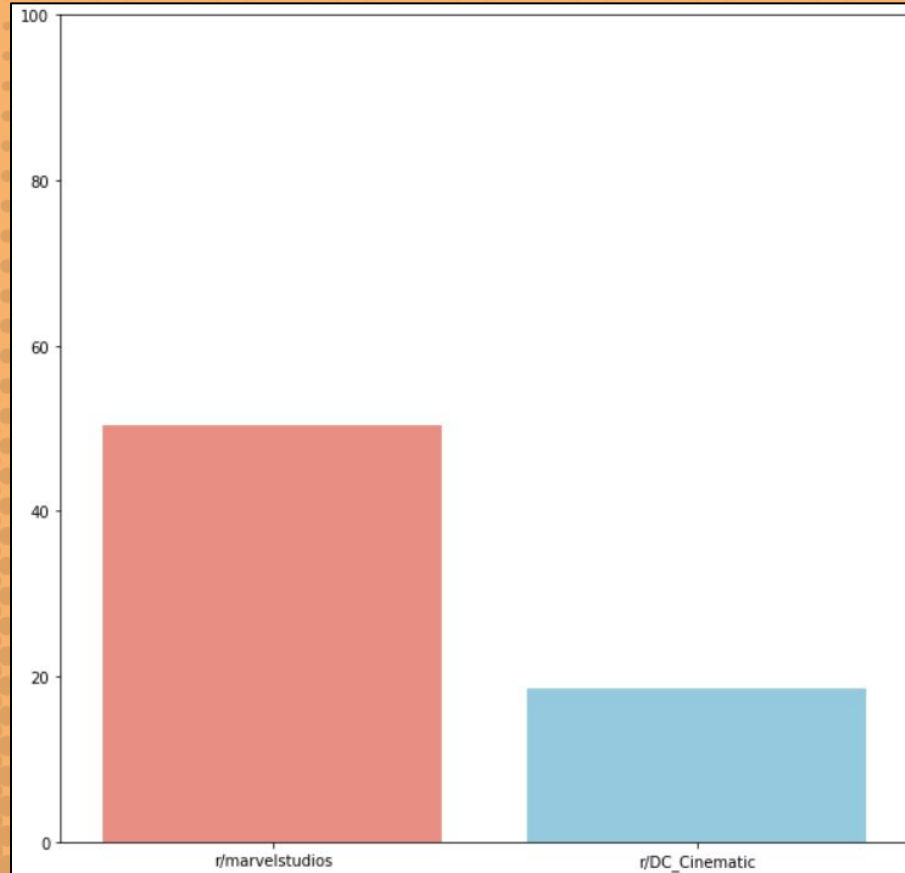


# ***DISTRIBUTION OF POSTS BY WORD COUNT***

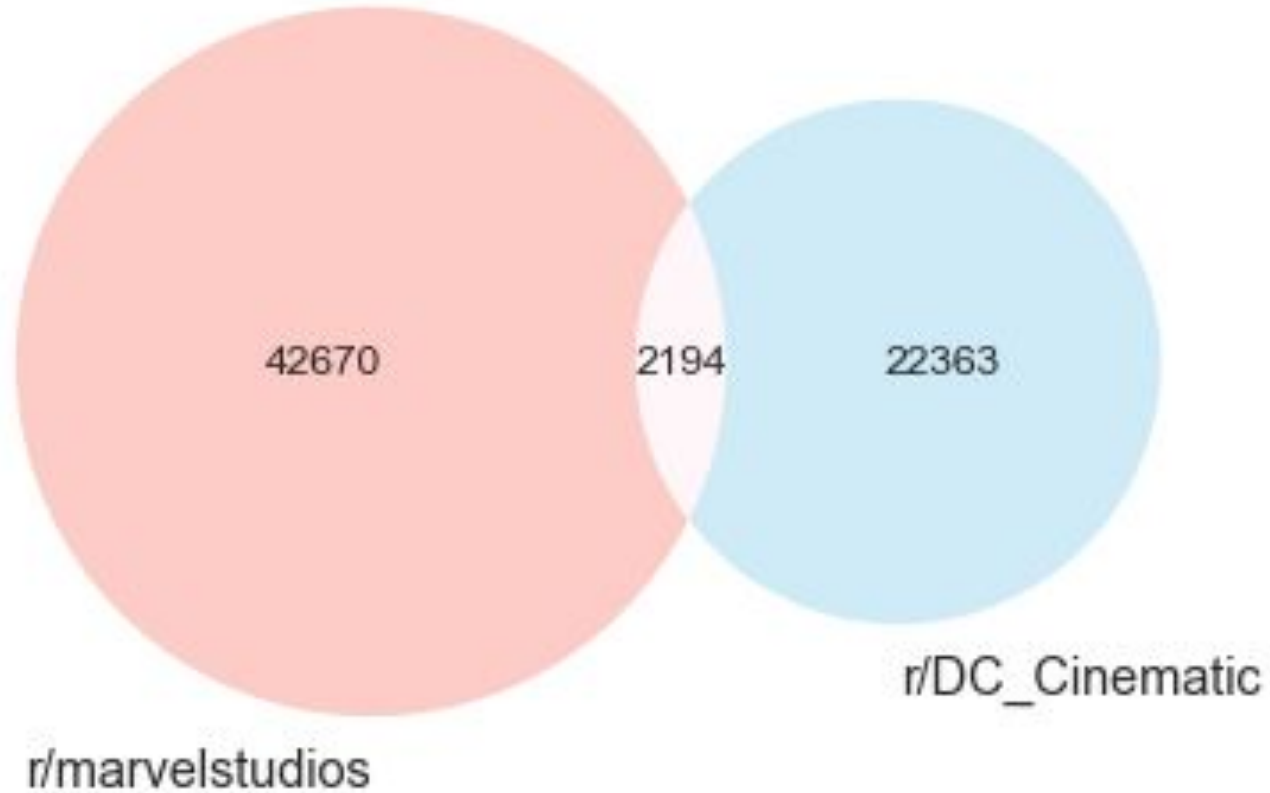




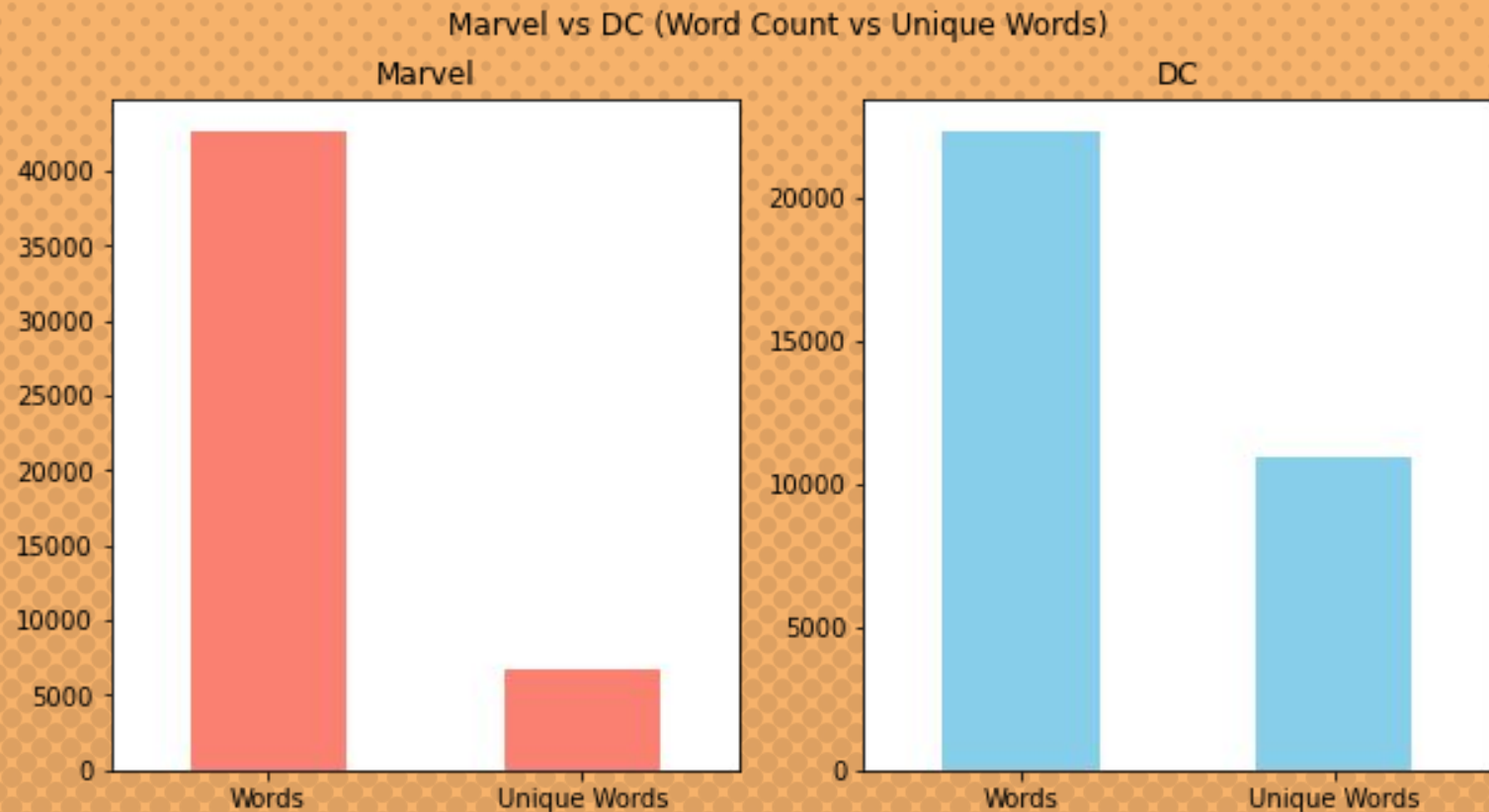
# ***PERCENTAGE(%) OF POSTS WITH SELFTEXT***



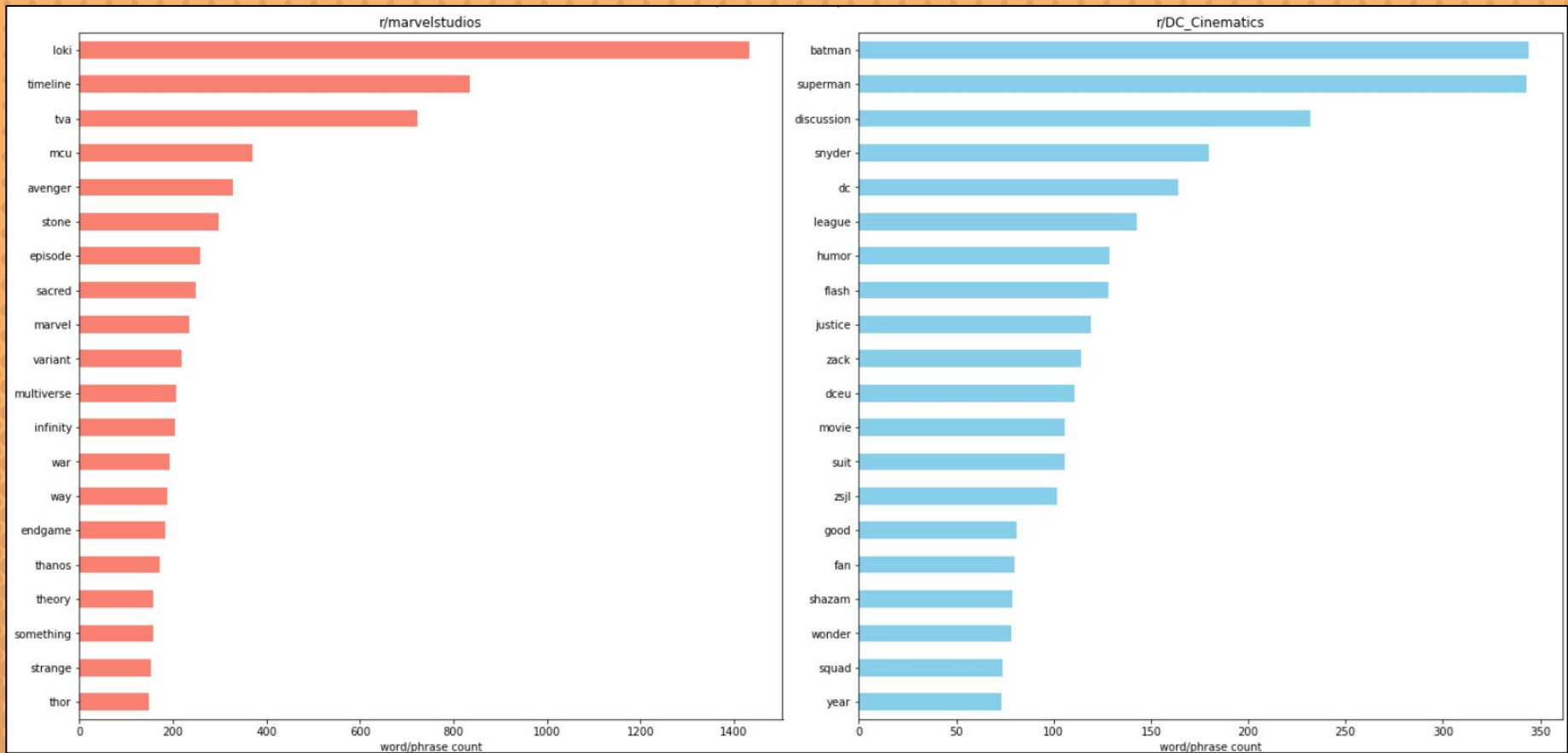
# ***WORDS IN POST***



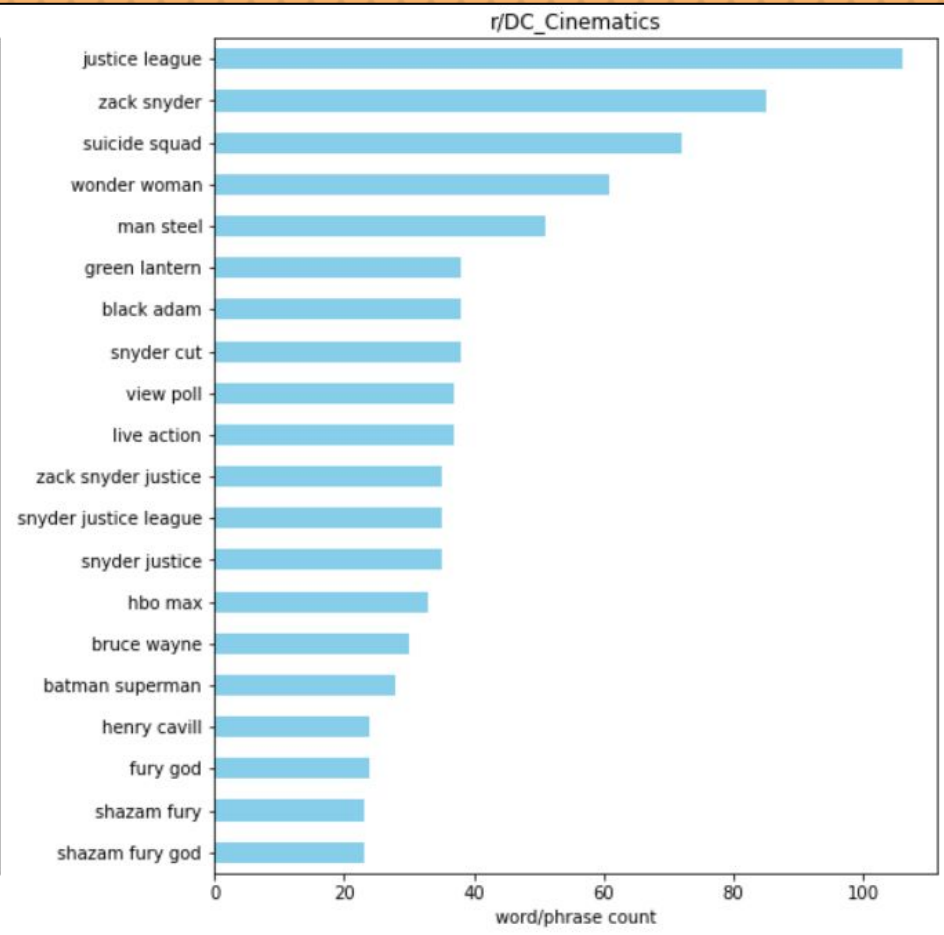
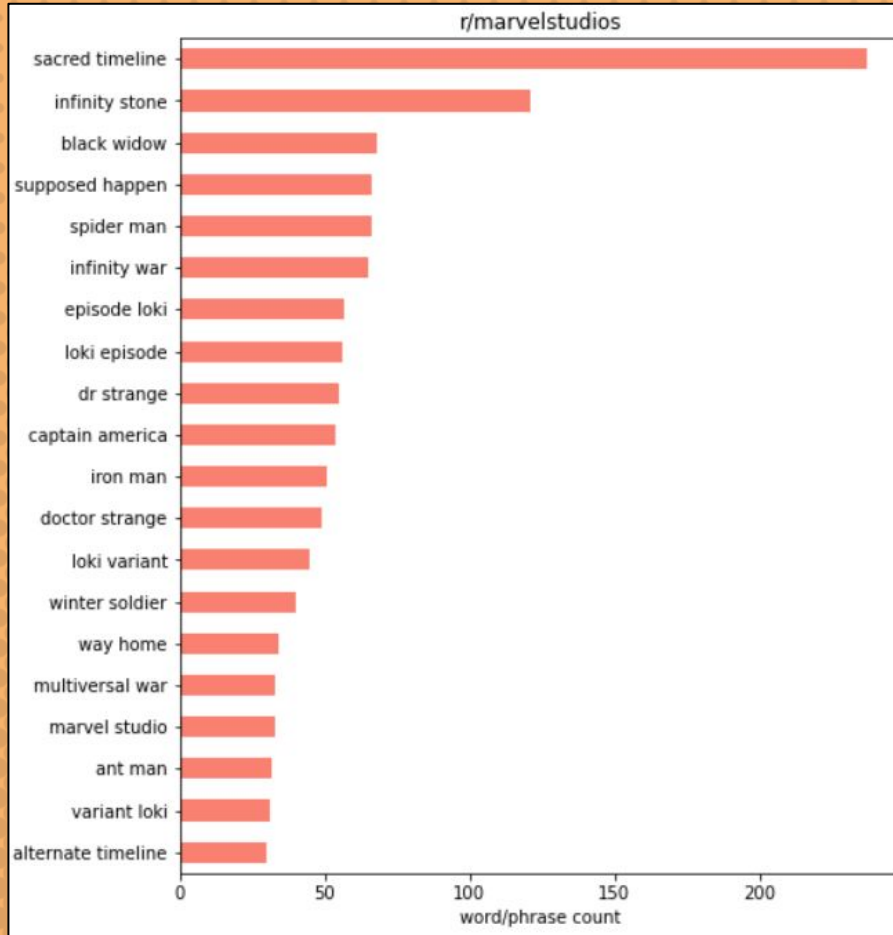
# ***MARVEL VS DC (WORD COUNT VS UNIQUE WORDS)***



# TOP 20 SINGLE WORDS



# TOP 20 COMBINATION OF WORDS





# **MARVEL TOP WORDS**

***CAPTAIN AMERICA***



***SPIDERMAN***



***DR STRANGE***

***LOKI***



***BLACK WIDOW***



# **DC TOP WORDS**

**BATMAN**



**SUPERMAN**



**BLACK ADAM**



**WONDER WOMAN**



**JUSTICE LEAGUE**



# MODELLING

Data Collection,  
Pre-processing & EDA

GridSearch

CountVectorizer  
1. N\_gram : (1,1)  
2. max\_features : 5000

TF-IDF Vectorizer  
1. N\_gram : (1,1)  
2. max\_features : 5000

Random Forest Classifier

Hyperparameters:  
1. Max depth  
2. n\_estimator

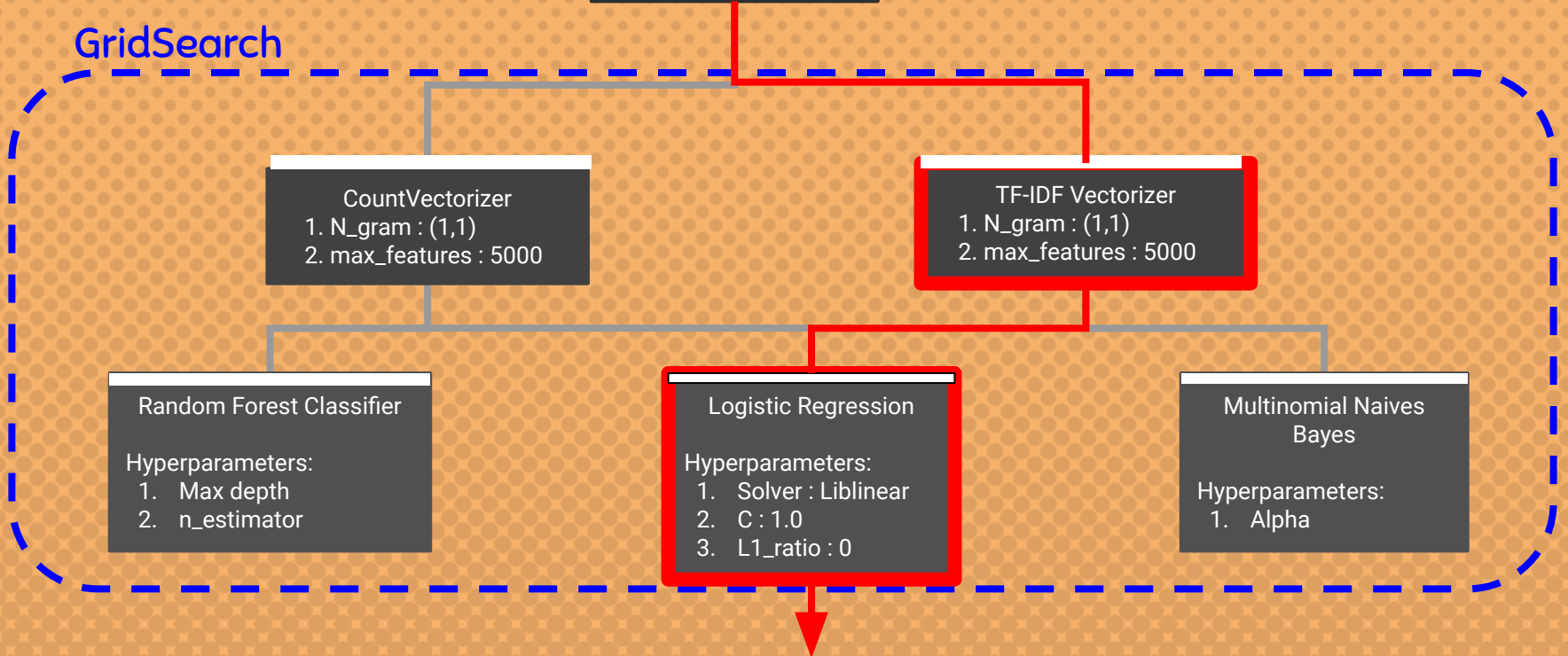
Logistic Regression

Hyperparameters:  
1. Solver : Liblinear  
2. C : 1.0  
3. L1\_ratio : 0

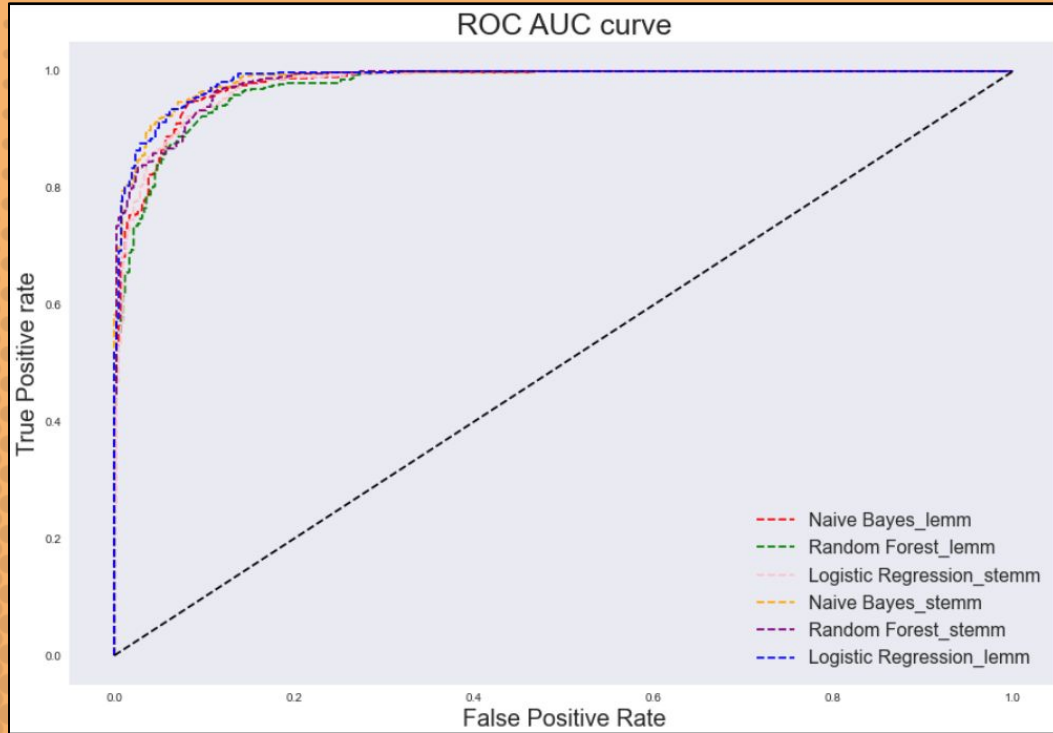
Multinomial Naives  
Bayes

Hyperparameters:  
1. Alpha

Best Choice by  
GridSearch

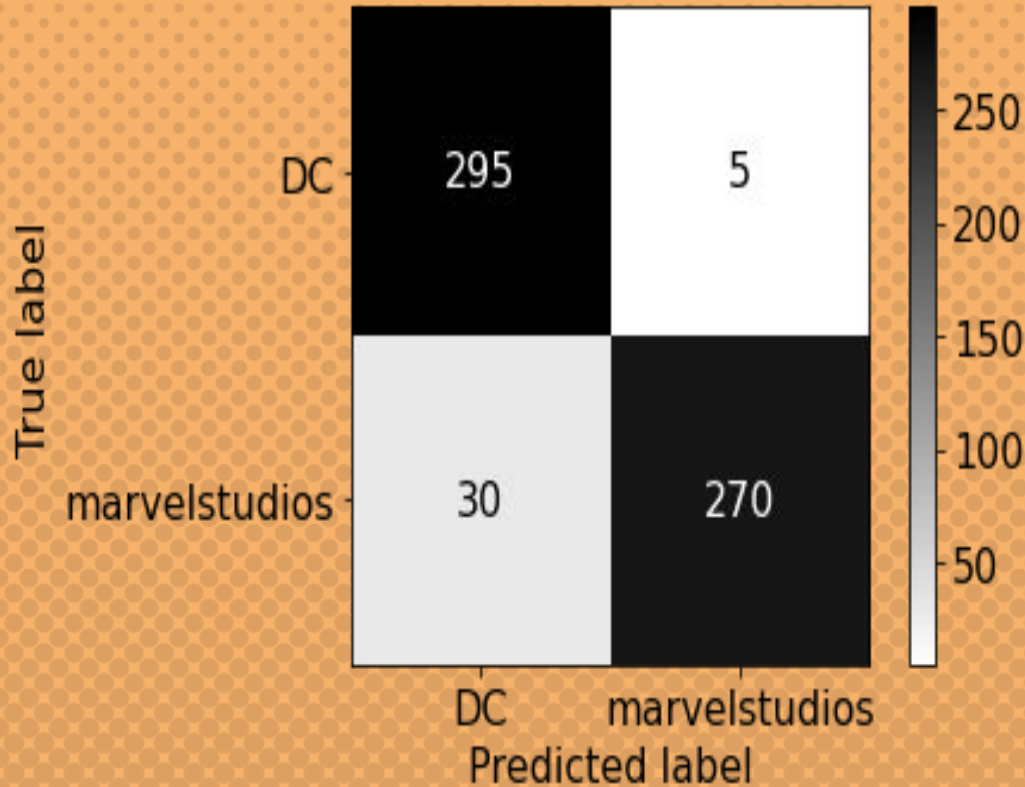


# METRICS



Parameters	Results
Vectorizer	Tfidf Vectorizer
Model	LogisticRegression
Train AUC Score	0.999
Test AUC Score	0.993
Model Train Accuracy	0.990
Model Test Accuracy	0.959

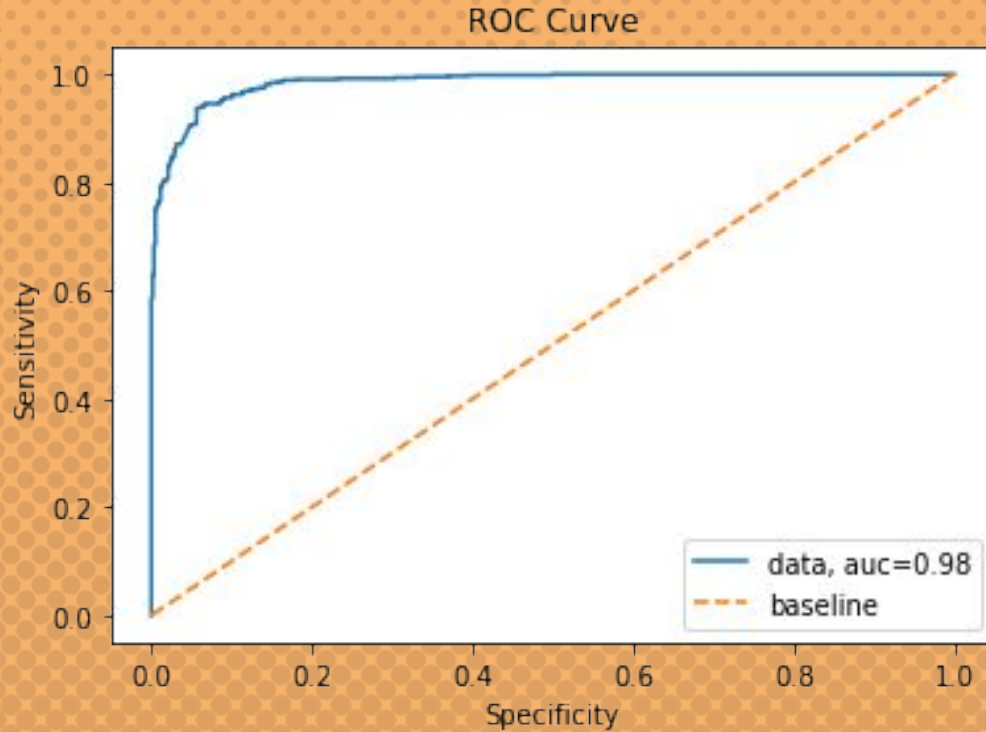
# METRICS



Parameters	Results
Vectorizer	Tfidf Vectorizer
Model	LogisticRegression
Train AUC Score	0.999
Test AUC Score	0.993
Model Train Accuracy	0.990
Model Test Accuracy	0.959

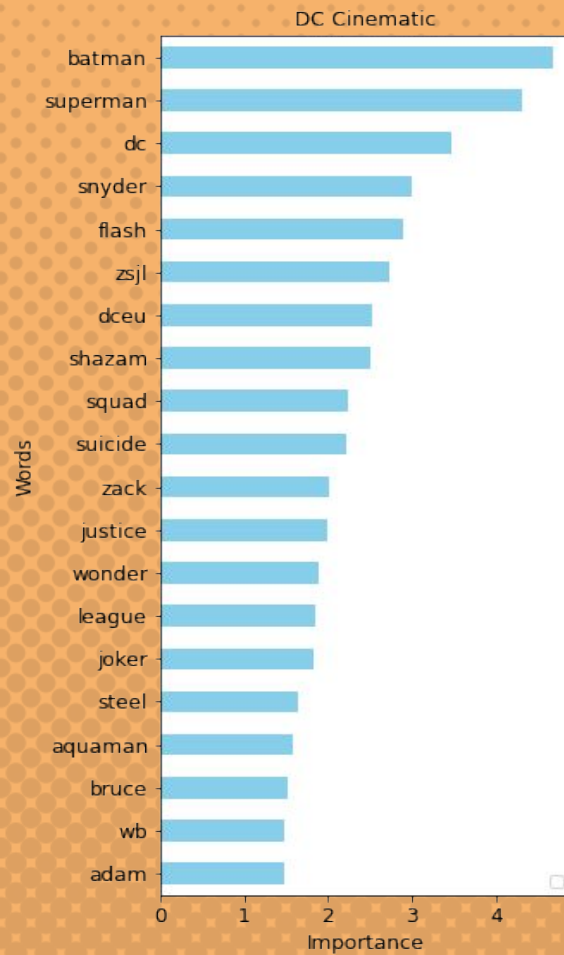
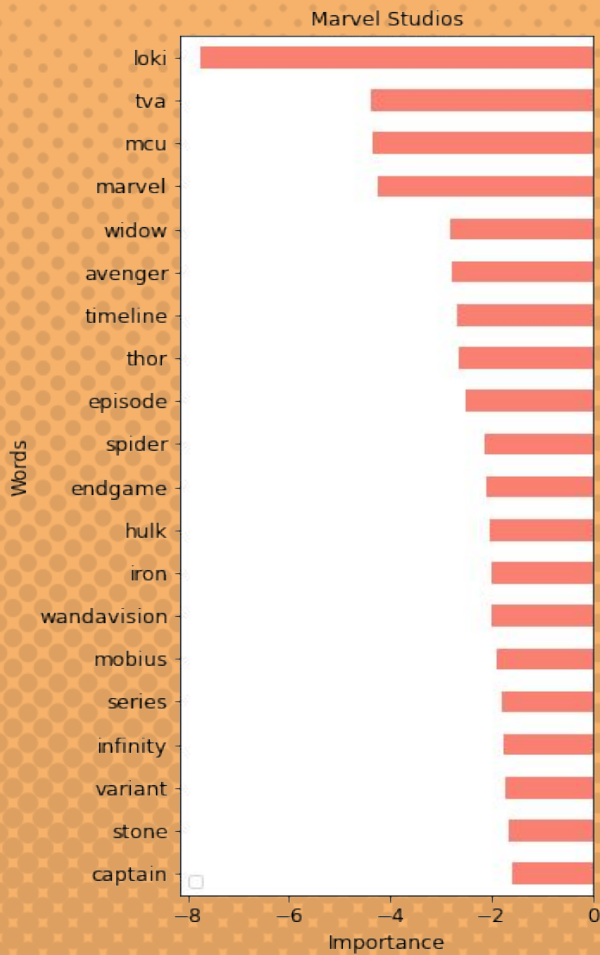


# METRICS



Parameters	Results
Vectorizer	Tfidf Vectorizer
Model	LogisticRegression
Train AUC Score	
Test AUC Score	
Model Train Accuracy	
Model Test Accuracy	

# MODEL EVALUATION - TOP PREDICTORS



# CONCLUSION

Identify Marvel and DC fans on reddit for targeted advertising of the client's new superhero game

Our model is trained on reddit posts and can classify posts as Marvel or DC with 96% accuracy.



# CONCLUSION

The coolest  
superheroes  
recommended  
for the game!!

**MARVEL** STUDIOS



Loki



Black Widow



Spiderman



Batman



Superman



Wonder Woman

# ***LIMITATIONS AND RECOMMENDATIONS***

## Limitations

- Model limited to Reddit data
- Data was limited to 1000+ posts per subreddit
- Sentiment analysis on posts

## Recommendations

- Retraining of models every 2 months (or at every product release)
- Expand into other social media platforms





***THANK YOU!***