Project 3: Marvel/DC Reddit Post Classifier & Sentiment Analysis

Farah Liyana Normal, Karina Kong, Koh Way Keat

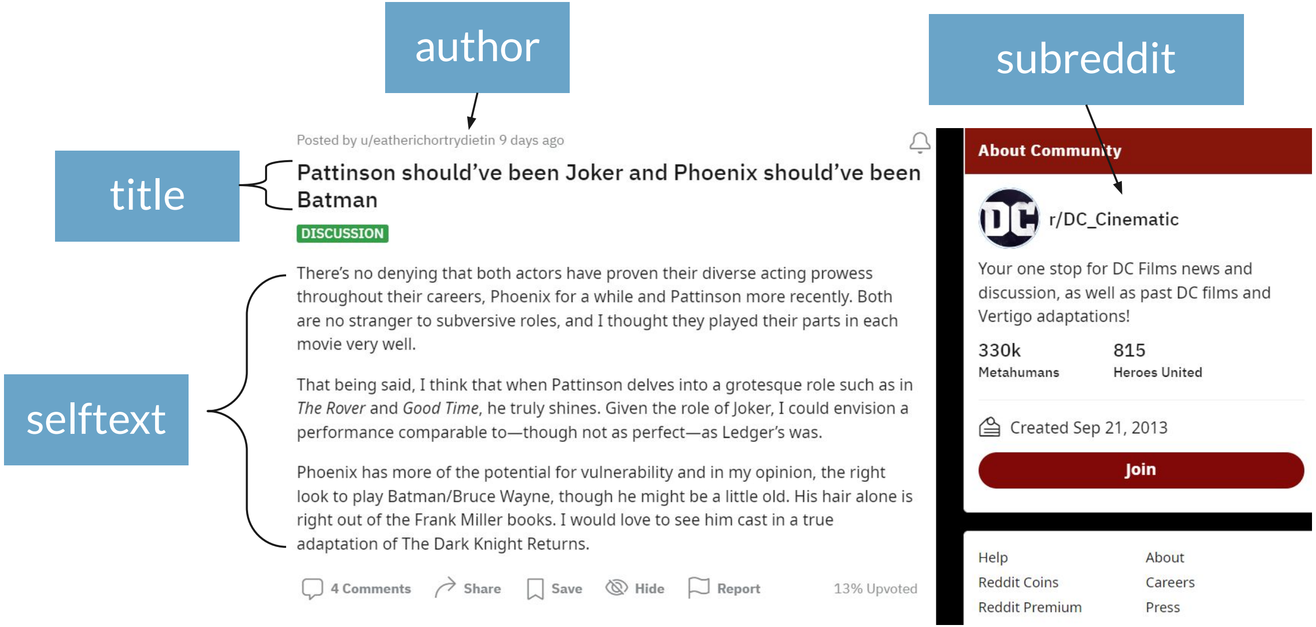
Problem Statement

We are employees of a marketing agency hired by a toy company to perform market research to **classify posts related to either Marvel or DC** movies in order to:

a. Build a classifier model that can be applied to other platforms (e.g. Twitter, Facebook) with text data to determine public interest in either movie franchise

b. Identify which top heroes to create toys that gives most returns

Data Collection & Cleaning - Features



author

subreddit

title

selftext

Posted by u/eatherichortrydietin 9 days ago

Pattinson should’ve been Joker and Phoenix should’ve been Batman

DISCUSSION

There’s no denying that both actors have proven their diverse acting prowess throughout their careers, Phoenix for a while and Pattinson more recently. Both are no stranger to subversive roles, and I thought they played their parts in each movie very well.

That being said, I think that when Pattinson delves into a grotesque role such as in *The Rover* and *Good Time*, he truly shines. Given the role of Joker, I could envision a performance comparable to—though not as perfect—as Ledger’s was.

Phoenix has more of the potential for vulnerability and in my opinion, the right look to play Batman/Bruce Wayne, though he might be a little old. His hair alone is right out of the Frank Miller books. I would love to see him cast in a true adaptation of The Dark Knight Returns.

4 Comments Share Save Hide Report 13% Upvoted

About Community

r/DC_Cinematic

Your one stop for DC Films news and discussion, as well as past DC films and Vertigo adaptations!

330k	815
Metahumans	Heroes United

Created Sep 21, 2013

Join

Help About
Reddit Coins Careers
Reddit Premium Press

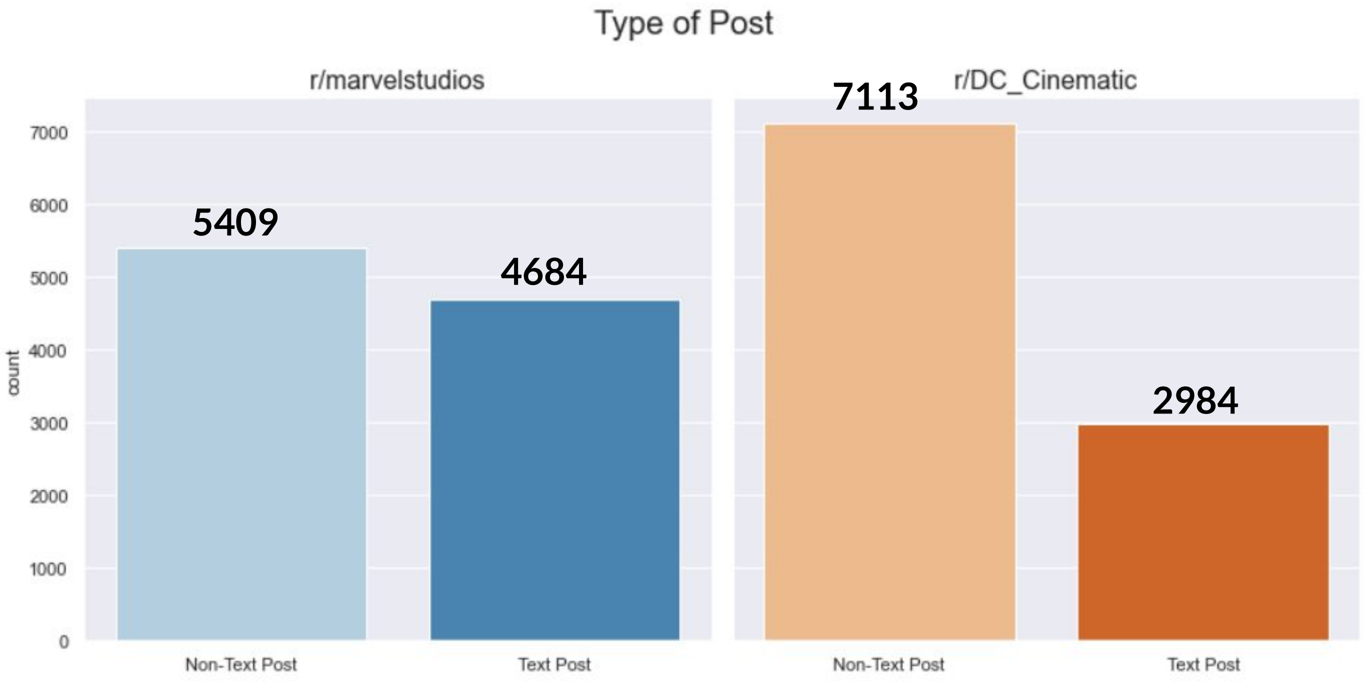
1. Total data = 20,000 rows

2. Duplicates = 6% of 20,000 post (subsequently removed)

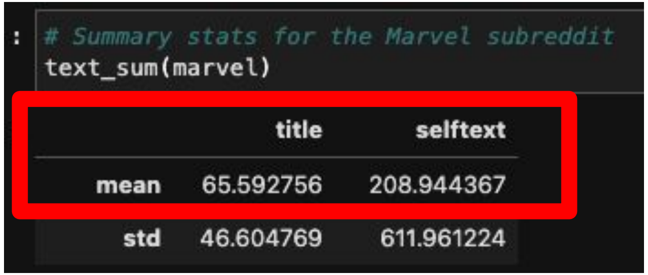
EDA : Type of Posts

Summary:

1. To retain the data classification of the post from the two subreddits, the selftext and the title of the post was combined.

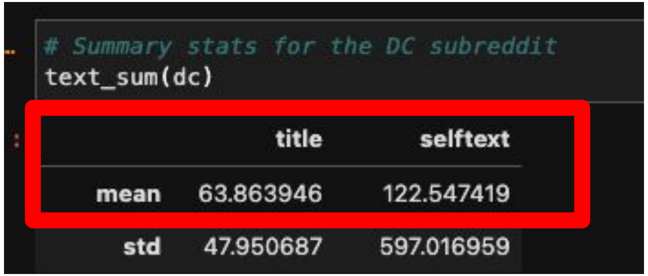


EDA: Text title and selftext of posts



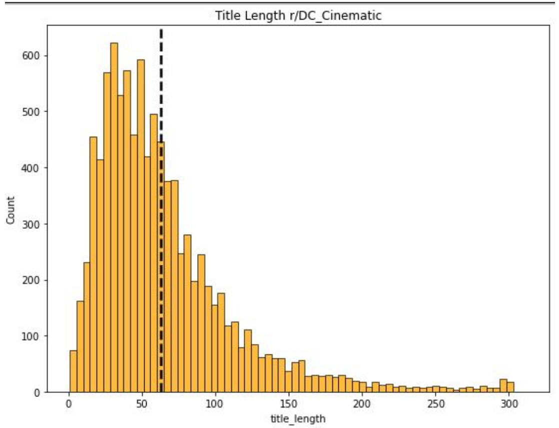
```
# Summary stats for the Marvel subreddit
text_sum(marvel)
```

	title	selftext
mean	65.592756	208.944367
std	46.604769	611.961224

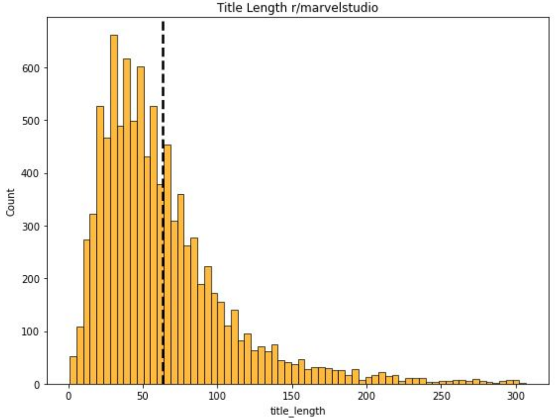


```
# Summary stats for the DC subreddit
text_sum(dc)
```

	title	selftext
mean	63.863946	122.547419
std	47.950687	597.016959



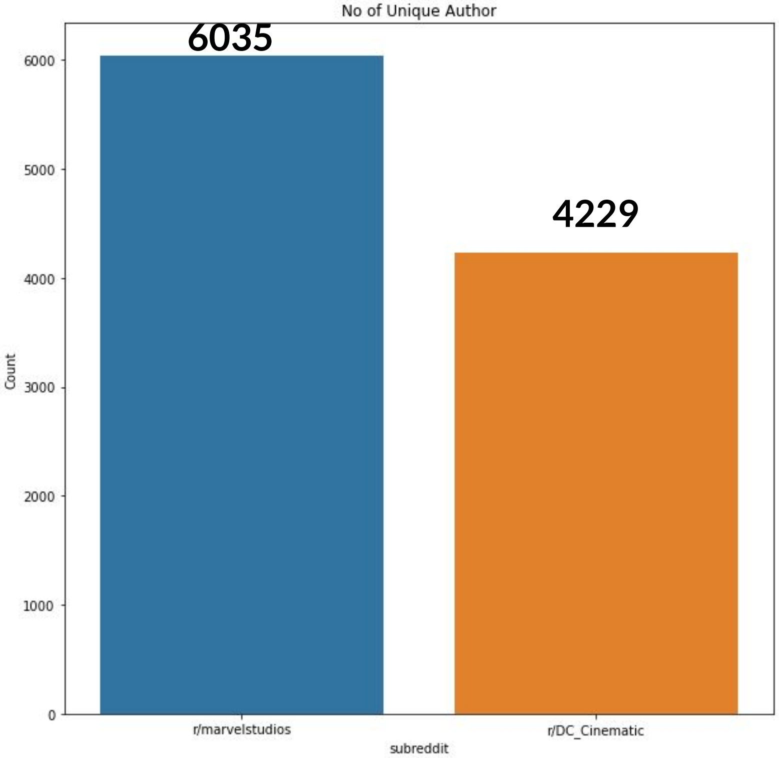
Title Length r/DC_Cinematic



Title Length r/marvelstudio

Posts in both subreddits have a similar average characters and title and selftext.

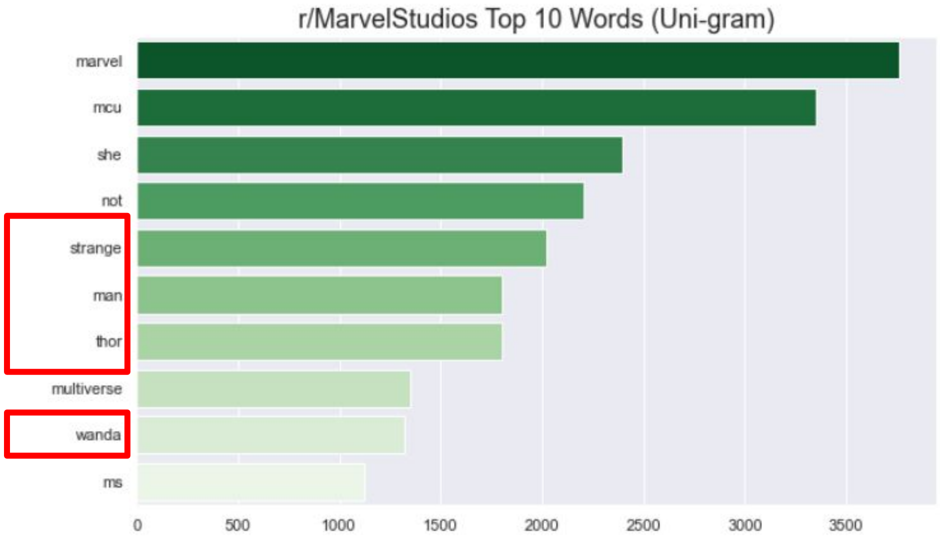
EDA: Unique redditors

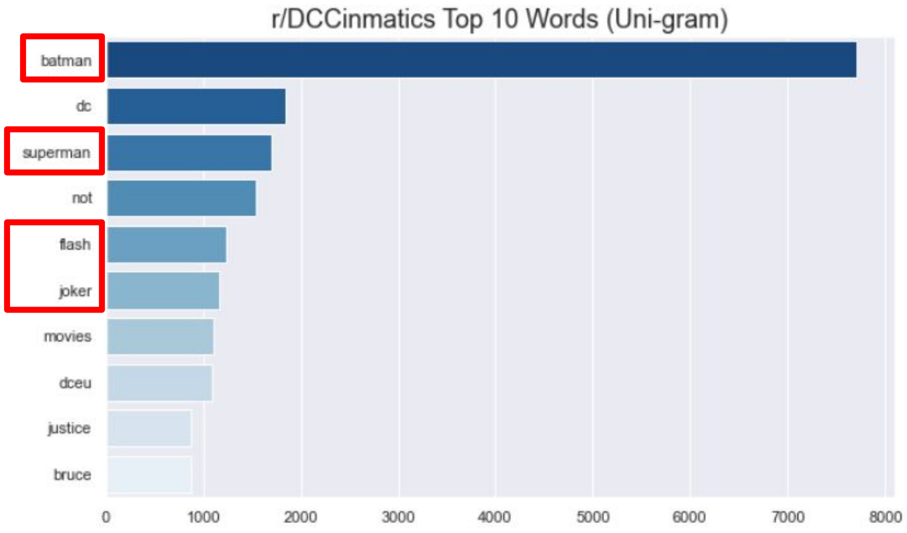


Summary:

1. r/marvelstudios has a larger active fanbase on reddit.

EDA: Uni-grams

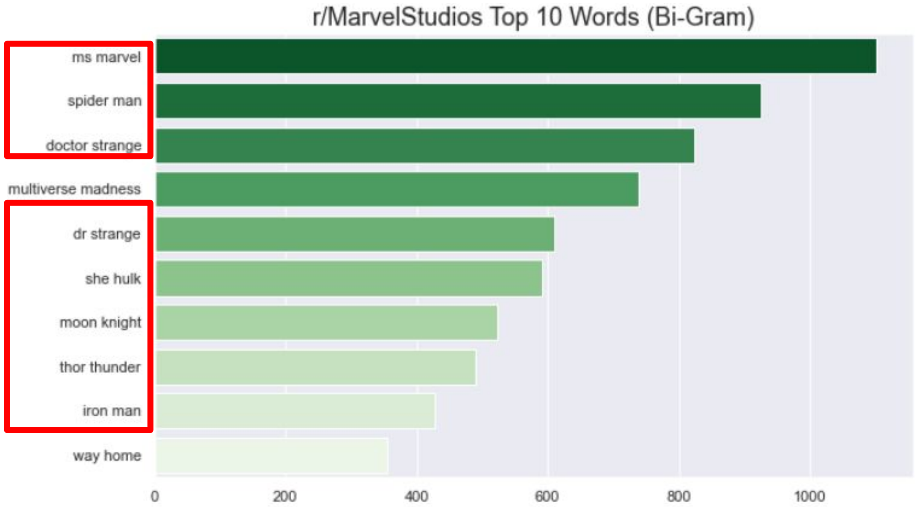


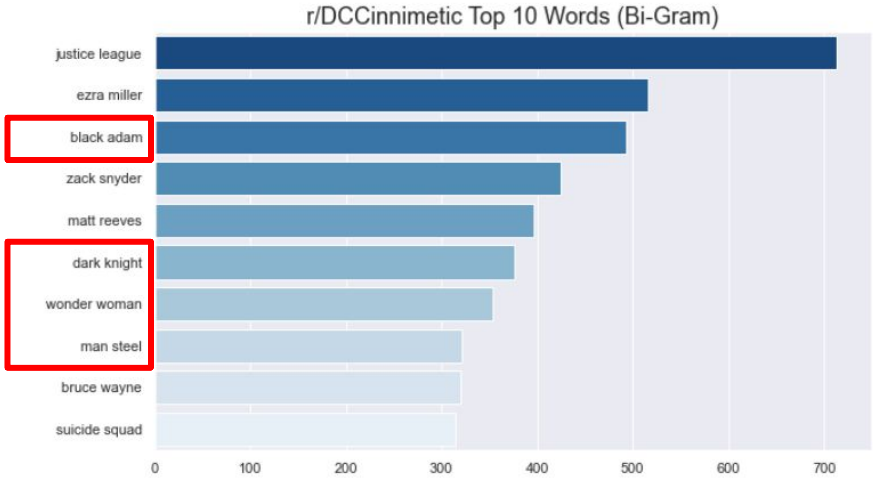


EDA Summary:

1. Marvel identified characters such as strange, thor, man, and wanda
2. DC identified characters such as batman, superman, flash, and joker

EDA: Bi-grams

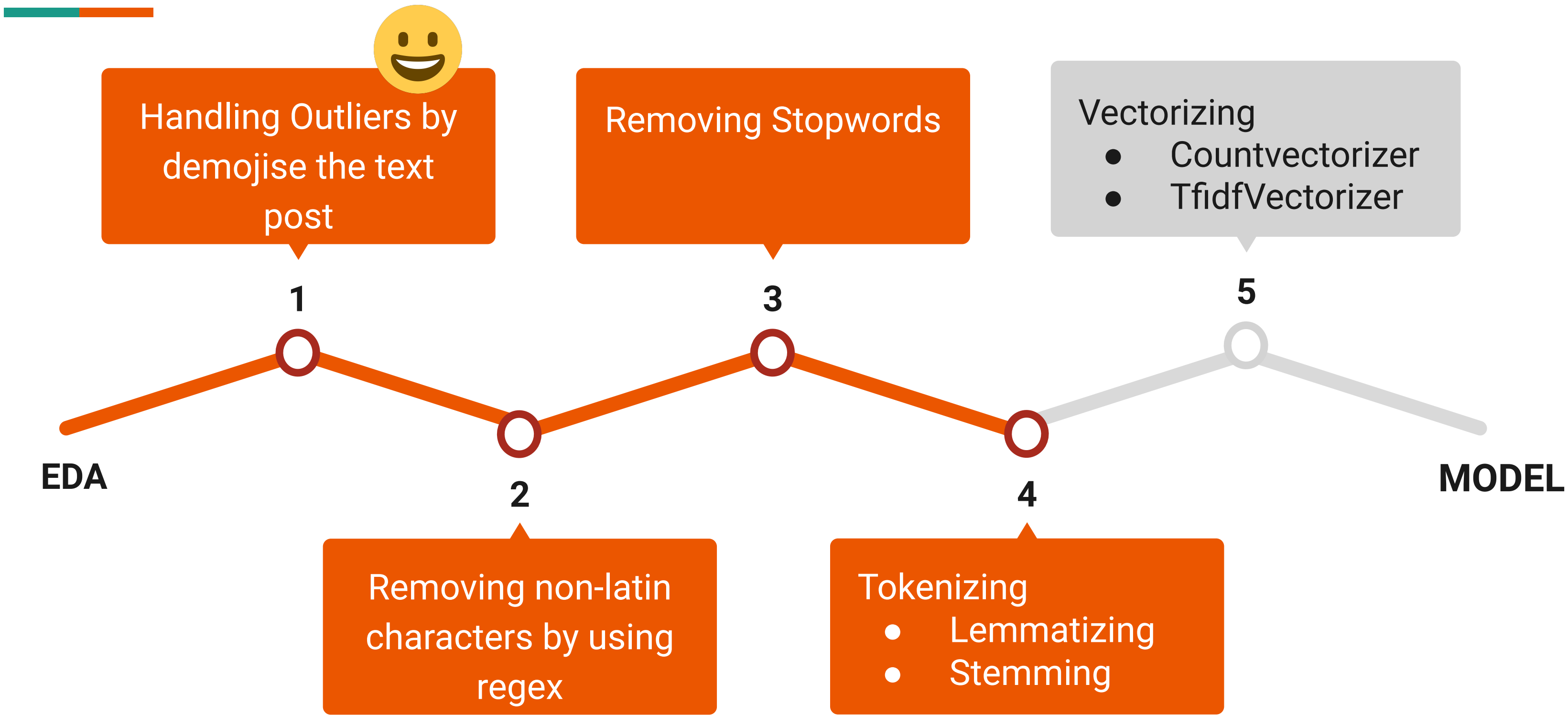




EDA Summary:

1. Marvel identified characters such as ms marvel, spiderman, dr. strange, she hulk, moon knight, thor and iron man.
2. DC identified characters such as black adam, dark knight, wonder woman, man steel.

Preprocessing & Vectorizing



EDA

1 Handling Outliers by demojise the text post

2 Removing non-latin characters by using regex

3 Removing Stopwords

4 Tokenizing
- Lemmatizing
- Stemming

5 Vectorizing
- Countvectorizer
- TfidfVectorizer

MODEL

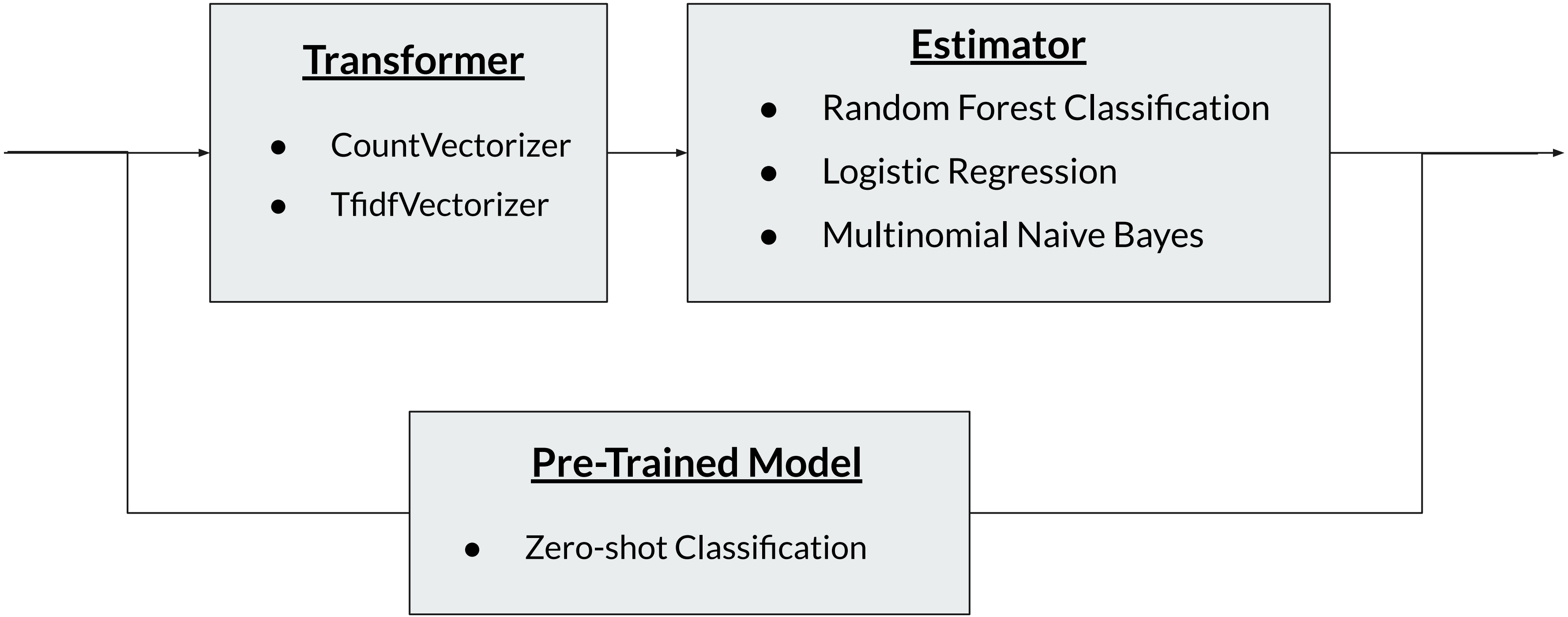
Modelling - Baseline Model

1. Normalized value count of target data set as baseline model

2. Shows 50/50% distribution between two classes

3. Dataset is balance

Modelling - Model Testing



Transformer

- CountVectorizer
- TfidfVectorizer

Estimator

- Random Forest Classification
- Logistic Regression
- Multinomial Naive Bayes

Pre-Trained Model

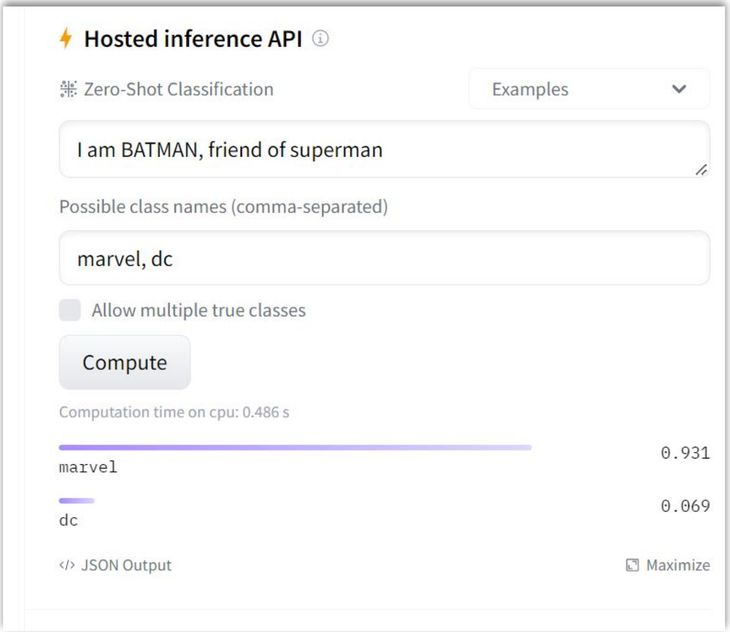
- Zero-shot Classification

Modelling - Zero shot Classification Model

1. Bad test score of around 0.49

2. Equivalent to randomly assigning a post to a class

3. The word “dc” is too short and generic



Hosted inference API

Zero-Shot Classification

I am BATMAN, friend of superman

Possible class names (comma-separated)

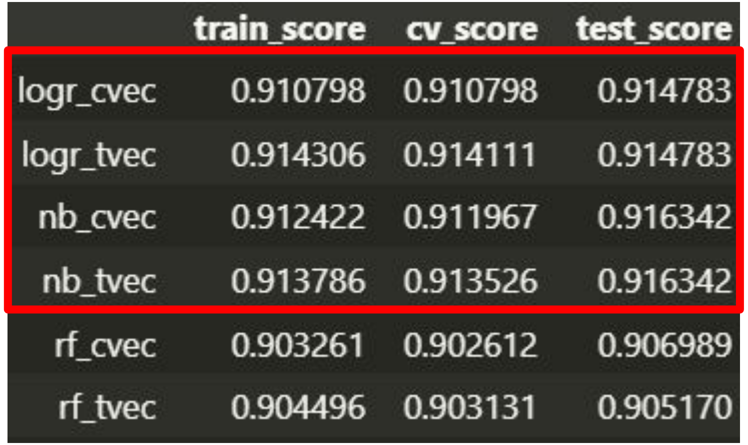
marvel, dc

□ Allow multiple true classes

Computation time on cpu: 0.486 s

```
marvel    0.931
dc        0.069
```

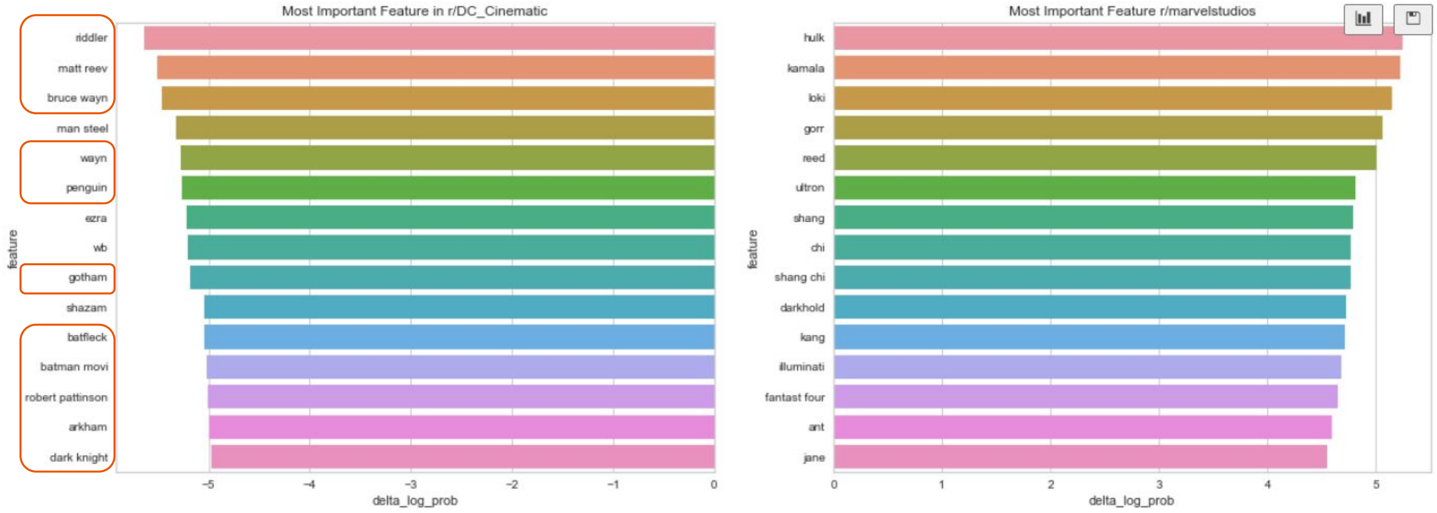
Production Model: Evaluation and Selection



	train_score	cv_score	test_score
logr_cvec	0.910798	0.910798	0.914783
logr_tvec	0.914306	0.914111	0.914783
nb_cvec	0.912422	0.911967	0.916342
nb_tvec	0.913786	0.913526	0.916342
rf_cvec	0.903261	0.902612	0.906989
rf_tvec	0.904496	0.903131	0.905170

- Both transformer perform very similarly
- The model from the GridSearchCV are well fitted
- Narrowed down to Naive Bayes and Logistic Regression model based on the scores

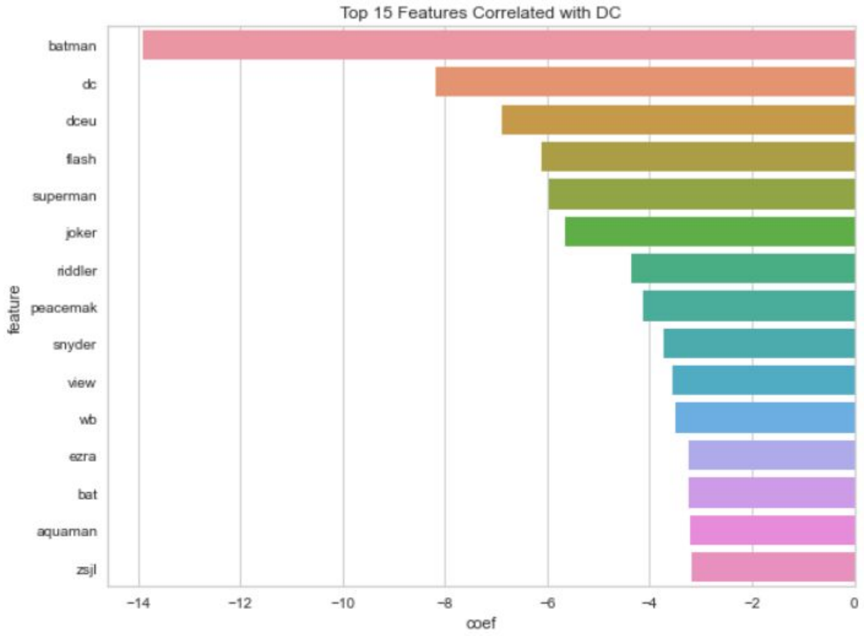
Production Model: Disadvantages of Naive Bayes

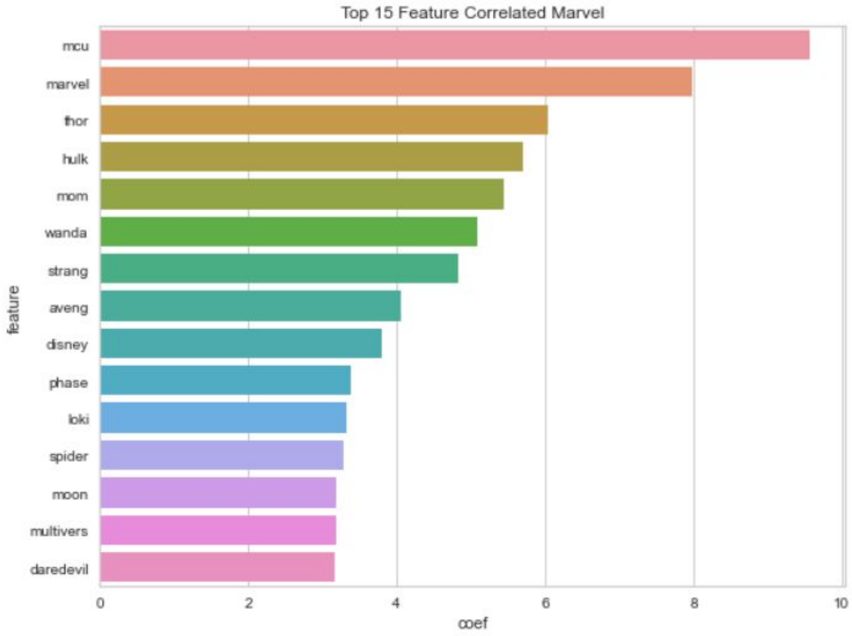


There are 2 **significant disadvantages** of Naive Bayes:

- Assumption of independence between words
- Determining feature importance requires the use of predicted probabilities which are known to be unreliable hence the top feature list from NB model may not be a representative list

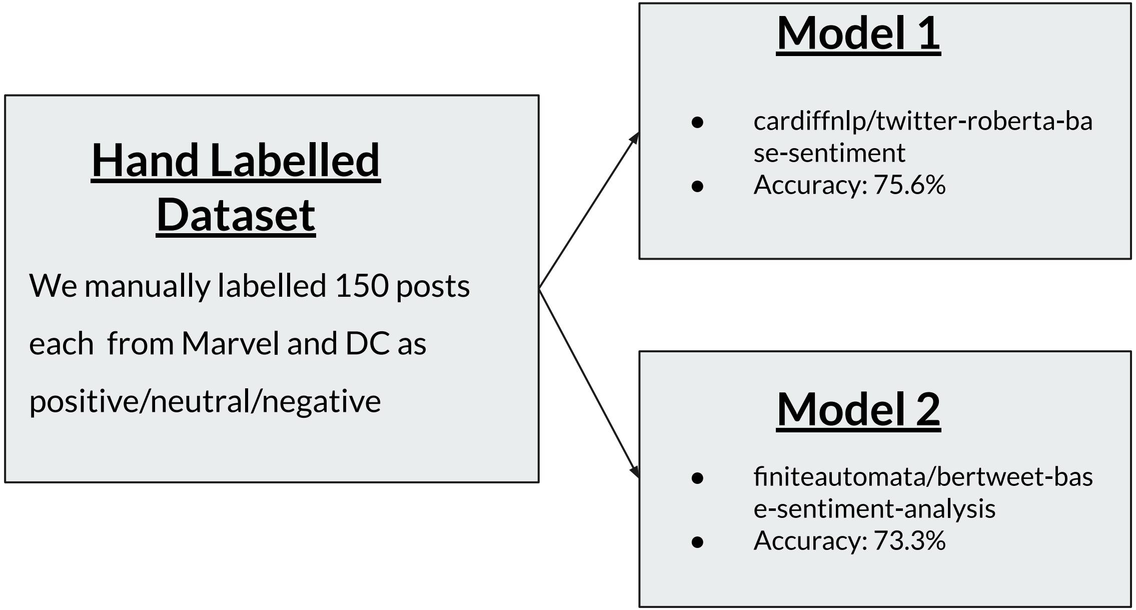
Production Model: Logistic Regression


Top 15 Features Correlated with DC


Top 15 Feature Correlated Marvel

- **Logistic regression** was chosen as the production model
    - The features are more diverse, and independent from one and another

Sentiment Analysis: Choosing a Model



Hand Labelled Dataset

We manually labelled 150 posts each from Marvel and DC as positive/neutral/negative

Model 1

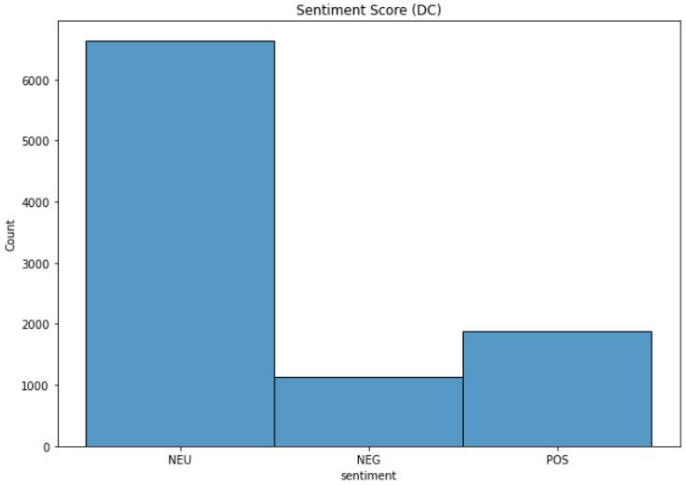
- cardiffnlp/twitter-roberta-base-sentiment
- Accuracy: 75.6%

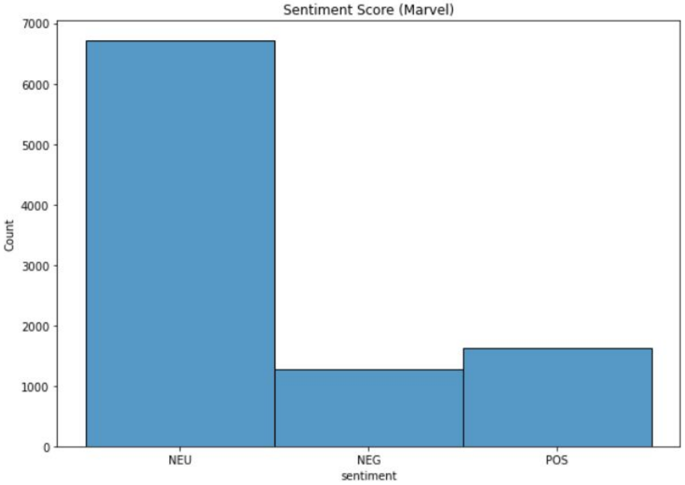
Model 2

- finiteautomata/bertweet-base-sentiment-analysis
- Accuracy: 73.3%

1. Manually labelled test dataset
    a. During this process, we noticed that many posts had neutral sentiments so we needed a model that provided neutral labels
2. Ran test data on 2 models that were able to produce POS/NEU/NEG labels
3. Select the best performing model

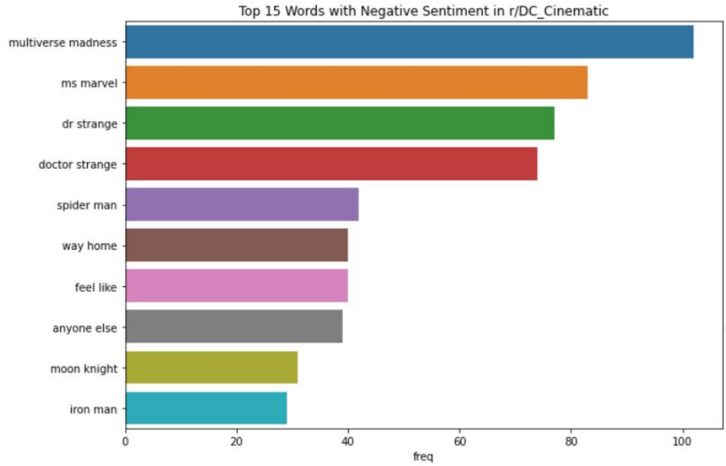
Distribution of posts by sentiment type

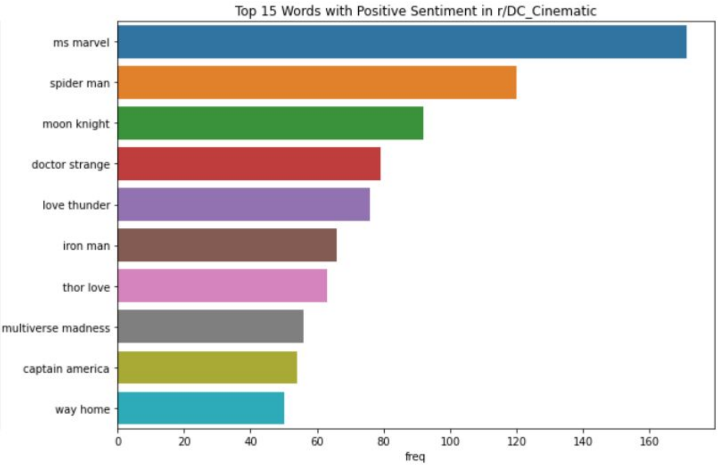




1. Majority of posts in both subreddits are neutral
    a. This is because a lot of the posts are discursive in nature e.g. “What do you think of Dr Strange?”
    b. However, these posts can still be useful for coming up with new toy/marketing ideas because it’s feedback from fans about what they want to see

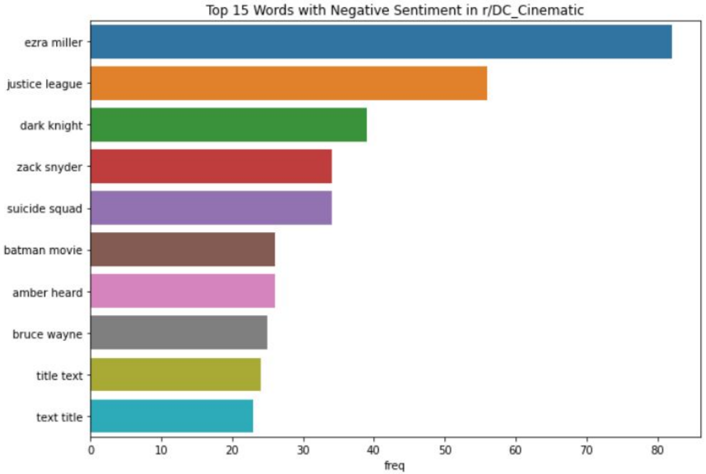
Top positive and negative bigrams for Marvel

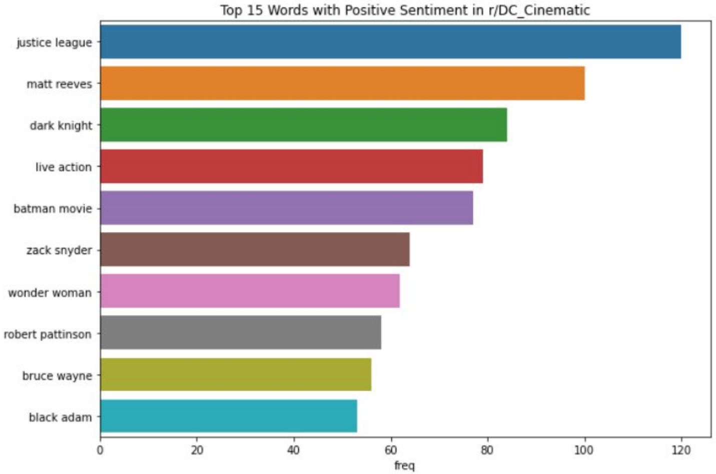




1. There's some overlap between top words with negative and positive sentiment
    a. This is likely because the most popular characters/concepts/movies are likely to have a sizable group of fans and haters

Top positive and negative bigrams for DC





1. Phrases with negative sentiments (and no positive sentiments) should be avoided for product releases/marketing
    a. Certain characters and actors under words with negative sentiments are involved in legal issues/controversy and should be avoided

Conclusion and Future Steps

1. The model is able to successfully classify Reddit posts as Marvel/DC content, possible future applications include
   a. Classify text data from other non-Reddit sources
   b. Can be used to to determine popularity/public interest in each brand
   c. Other downstream analysis e.g. sentiment analysis
2. Key findings from sentiment analysis:
   a. Characters to develop toys and marketing initiatives for
   b. Key characters to avoid