SUBREDDIT CLASSIFICATION

TOPICS OF DISCUSSION

- Problem Statement
- Process of Data Collection
- Data Cleaning and Preprocessing
- Observations
- Best Model
- Conclusions and Recommendations

PROBLEM STATEMENT

- A popular online superhero merchandise company called **Superherostuff** (https://www.superherostuff.com) wants to categorize customer posts on its online review page as either Marvel or DCComics fans.
- A data scientist has been hired to successfully develop a classification model based on posts from two popular subreddits, r/Marvel and r/DCComics and validate its accuracy. This model would then be used by the company to categorize customer reviews.

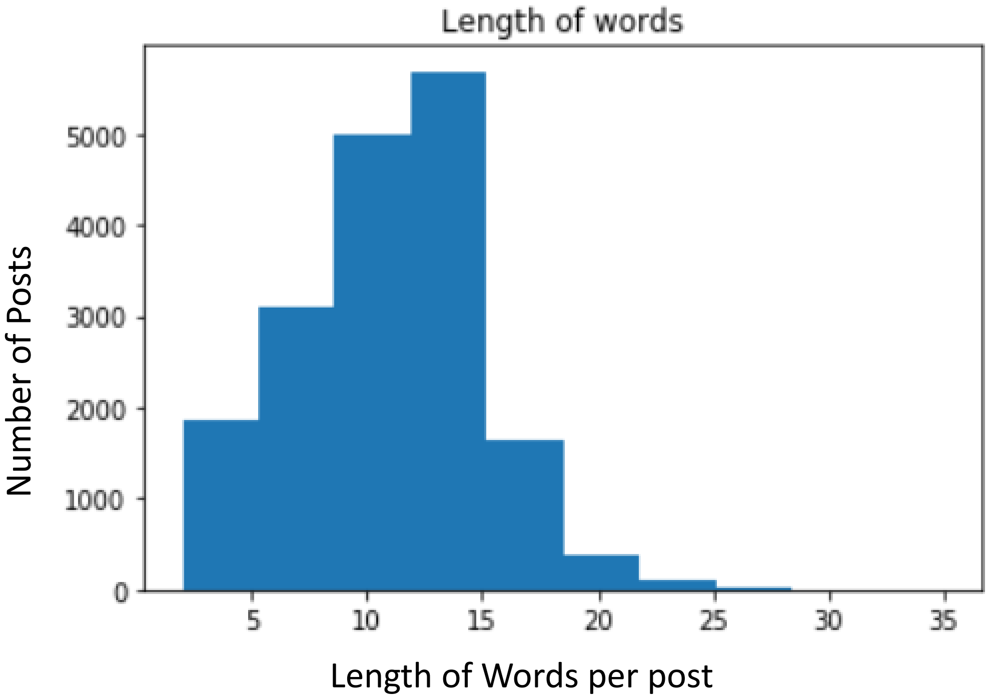
DATA COLLECTION

- Posts were scraped from two popular subreddits in Reddit ; r/Marvel and r/DCComics.
- A total of 1500 posts from each subreddit were scraped.
- Although subreddit titles and associated selftexts were collected for cleaning, selftexts had several null values and hence the column was dropped.
- Subreddit titles- The title of the submission
- Subreddit selftext- The markdown formatted content for a text submission

DATA CLEANING AND PREPROCESSING

- Subreddit titles were cleaned using Regular Expression.
- Stop words were removed
- Lemmatizing and Stemming were explored
- Certain expressions were intentionally removed because their frequency added no value , example: spoiler, art work ,fan work
- The word ‘Bot’ was further analyzed and a non-english phrase was discovered and deleted.

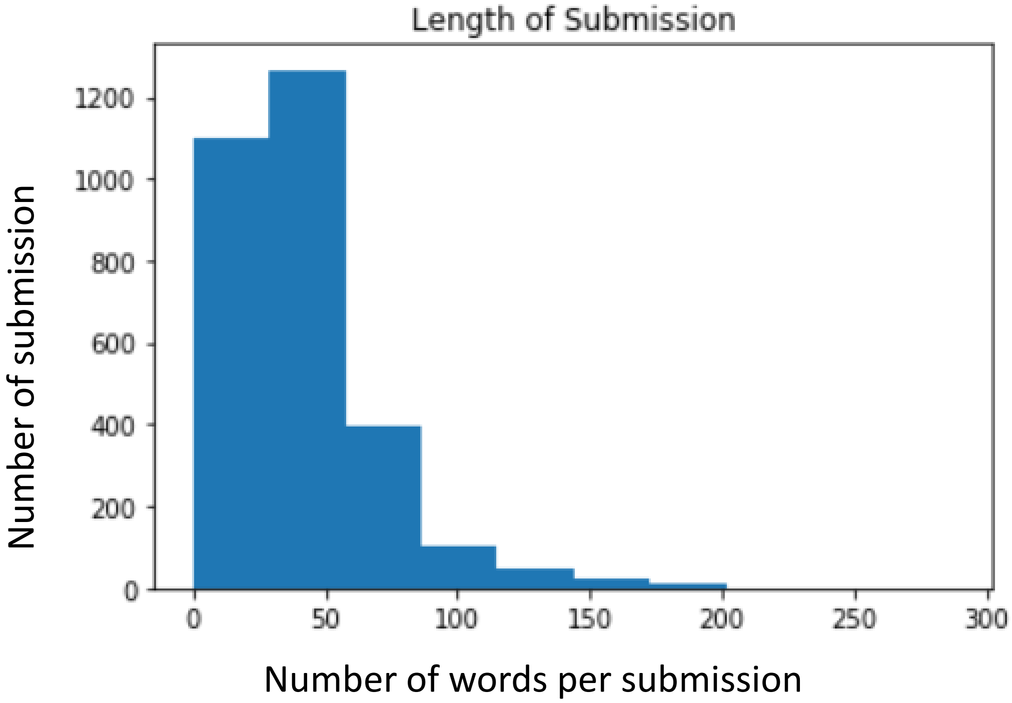
OBSERVATIONS



- The word length distribution is skewed right, which the highest word length between 10 to 15 letters.

- This shows the frequency of many two word combinations example iron man and justice league

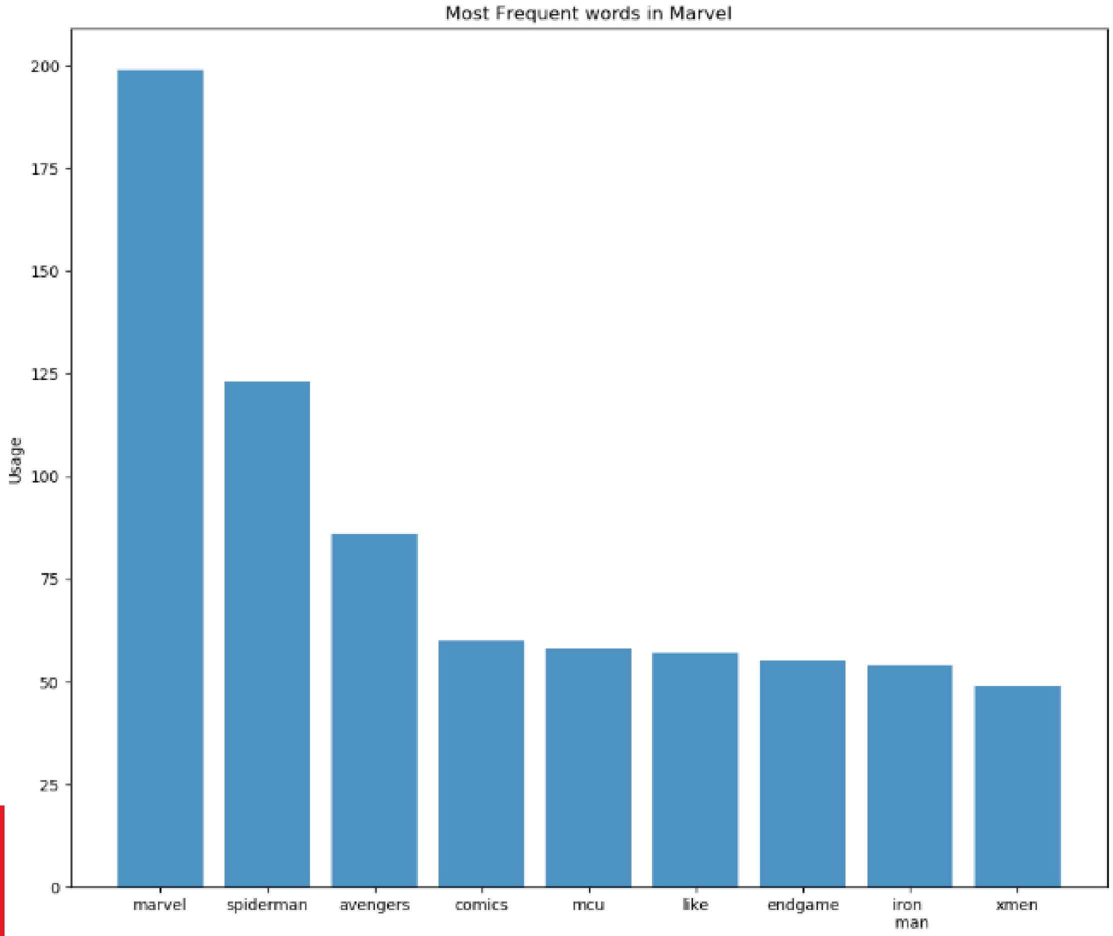
OBSERVATIONS



- Most submissions were short less than 50 words . The distribution is skewed right.

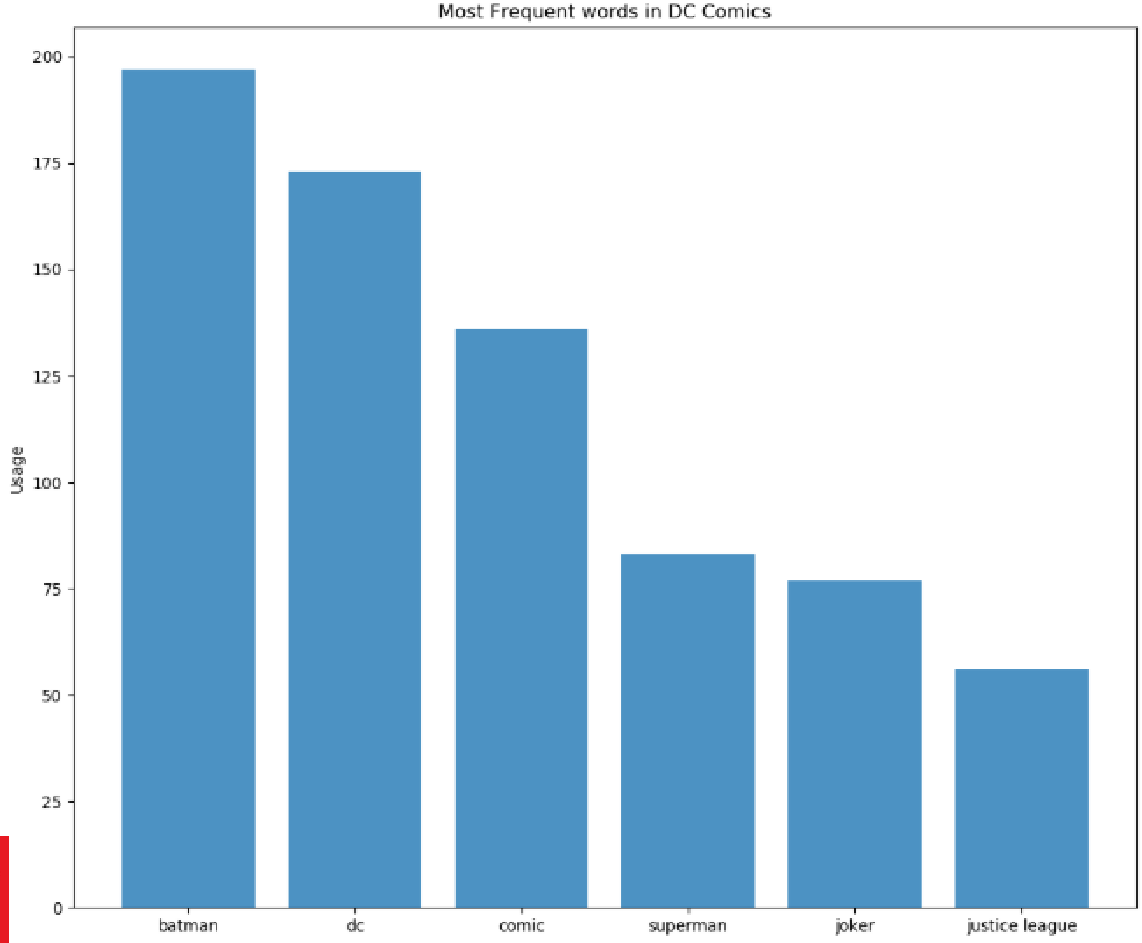
OBSERVATIONS

Most frequent words from both subreddits



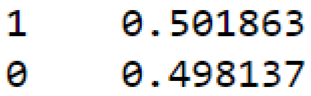
OBSERVATIONS

Most frequent words from both subreddits



BEST MODEL

- Baseline Model



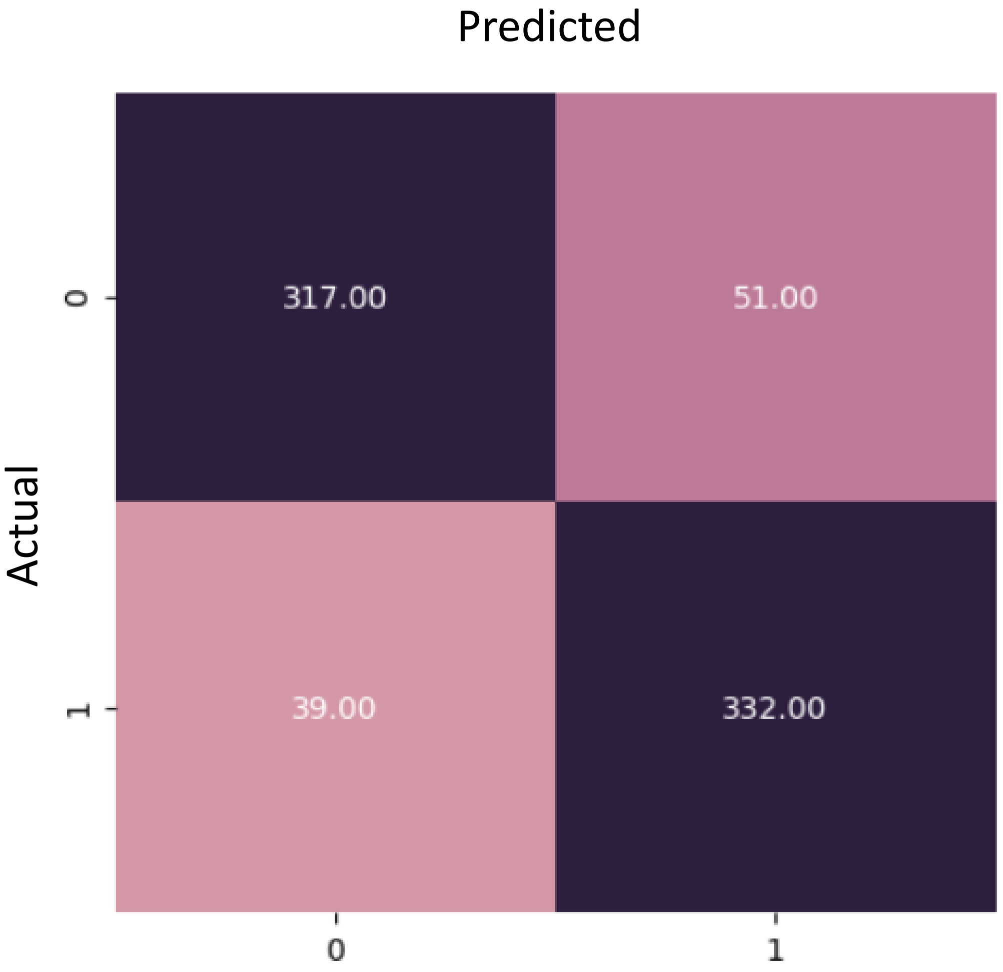
```
1     0.501863
0     0.498137
```

50.18% chance of guessing a post belongs to Marvel subreddit correctly

- Best Model

Best model was a Multinomial Naive Bayes model using Count Vectorizer with an accuracy of **94.26%**

MODEL VALIDATION



Confusion Matrix

Sensitivity: 89.5 %

Specificity: 86.1 %

Precision: 86.7 %

CONCLUSIONS AND RECOMMENDATIONS

- The model can be further improved by looking into the posts that were misclassified.
- Hyperparameters can be further explored and finetuned for various models that were analyzed.
- The total dataset consisted of only 3000 subreddit posts. Maybe more subreddit posts can be web scraped from the reddit platforms to construct a more generalized model and avoid overfitting.
- For this project, only the subreddit posts were considered. As an extension to improve the model, the comments and self text columns can be considered as well.
- Subreddit posts from a non biased subreddit like r/Comics can be classified to further validate the accuracy of the model.

QUESTIONS?