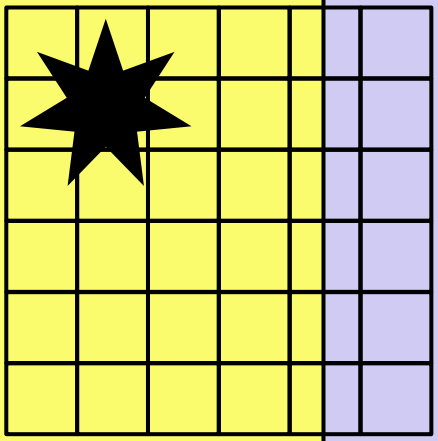
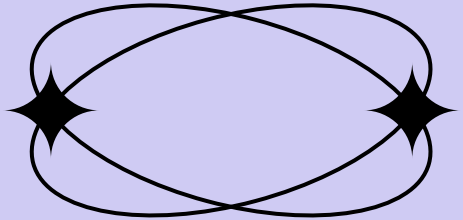




# Project 3: Web APIs & NLP

## Subreddit Comparison

By Eli Daniels

# METHODOLOGY

| | |
|---|---|
| Data Collection | Utilizing Reddit's API, data will be gathered on the number of subscribers, posts, comments, and other relevant metrics for both subreddits over a selected time period |
| Text Analysis | Natural Language Processing (NLP) techniques will be applied to analyze the sentiment of posts and comments, extract keywords, and classify content types |
| Data Visualization | o facilitate easy comparison and interpretation, data will be visualized using charts, graphs, and heatmaps. |
| Hypothesis Testing | Statistical analysis may be employed to test hypotheses regarding user engagement differences between the two subreddits |
| Content Analysis | A manual review of a subset of posts and comments will be performed to validate the results from the automated NLP analysis. |

# PROBLEM STATEMENT

**In this study, we aim to delve into the online communities of two popular subreddits, r/Marvel and r/DCcomics, to understand the unique characteristics, discussions, and trends within each subreddit outside of superhero names**

**We aim to determine which subreddit has a stronger positive sentiment among its users and explore whether certain topics or themes dominate the discussions within each.**

# CLEAN!!

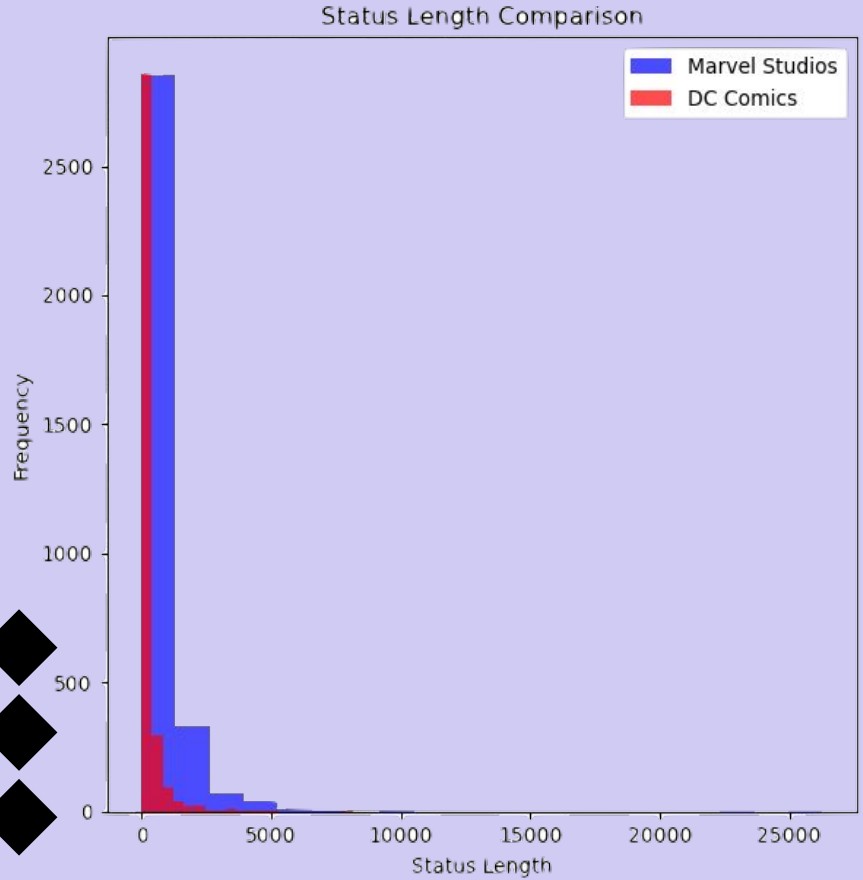
**01**
Web Scraping

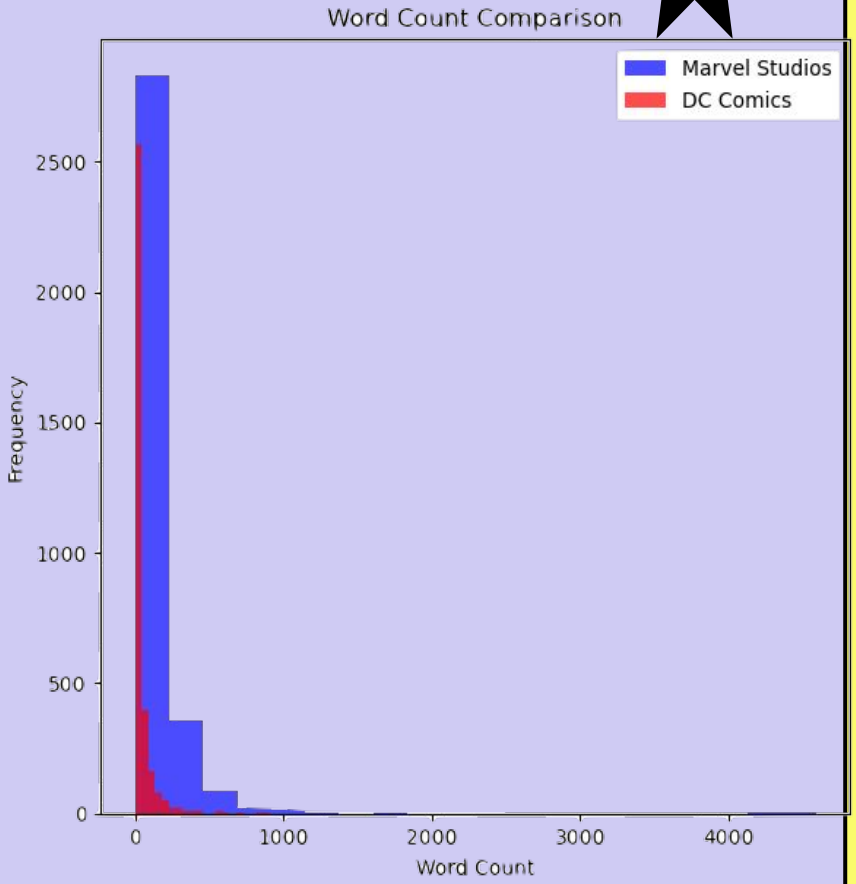
**02**
Removing Null Values

**03**
Removing Punctuations

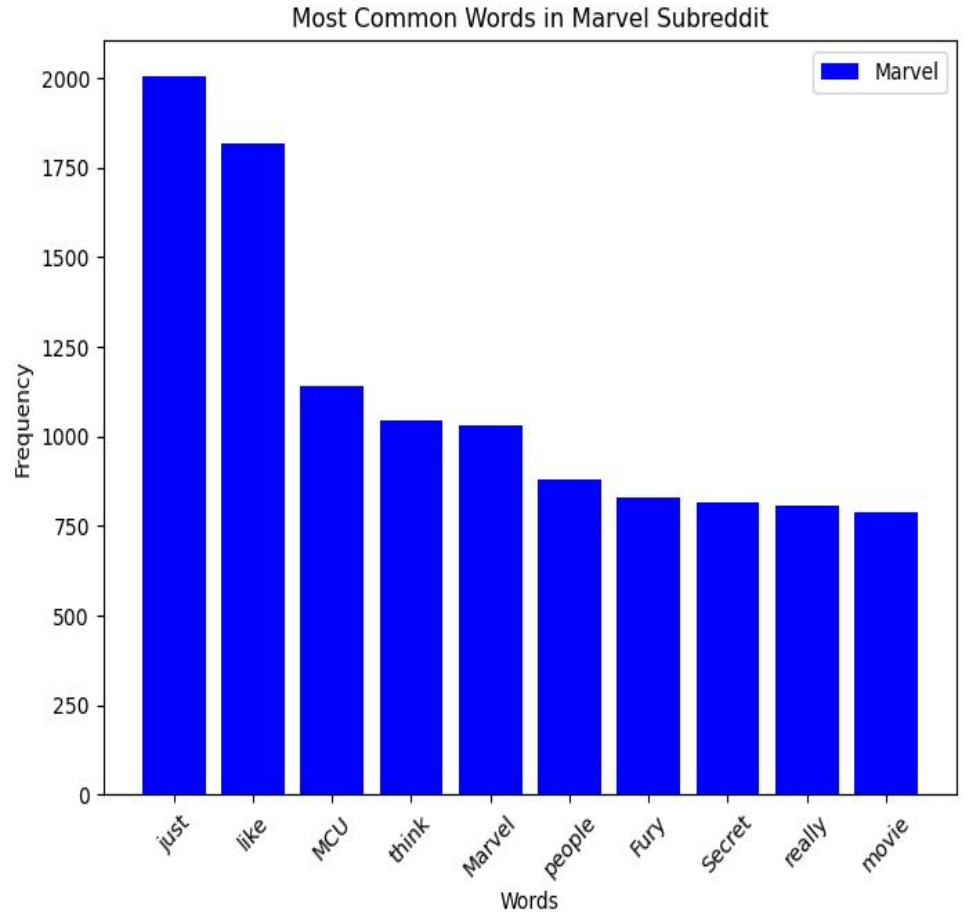
**04**
Remove HTML, Hyperlinks,whites pace, etc

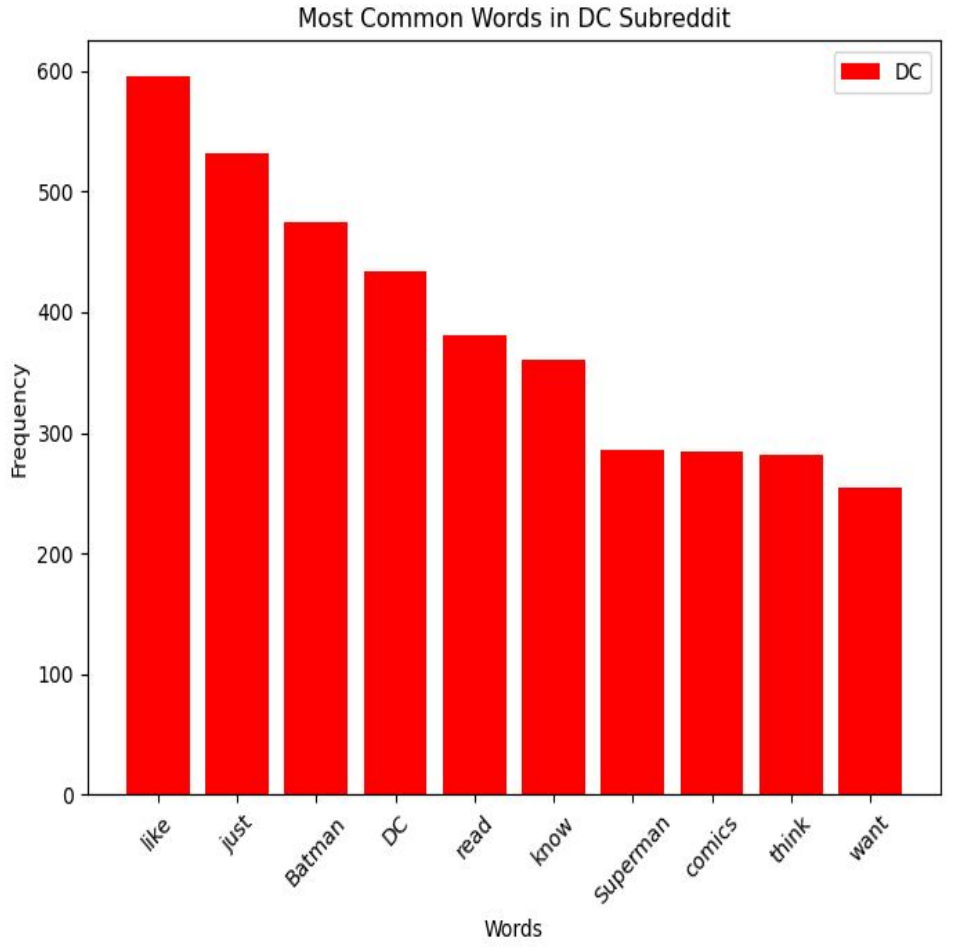
# Lets make some new columns!!



# We Can Do Better: The Count  though



Bi
Gram

**Most Common Bigrams in Marvel Subreddit:**

- Secret Invasion - 456 occurrences
- Iron Man - 157 occurrences
- feel like - 145 occurrences
- Captain America - 126 occurrences
- Captain Marvel - 115 occurrences
- Nick Fury - 108 occurrences
- Doctor Strange - 101 occurrences
- Guardians Galaxy - 90 occurrences
- Black Panther - 81 occurrences
- felt like - 79 occurrences

**Most Common Bigrams in DC Subreddit:**

- Justice League - 110 occurrences
- Green Lantern - 85 occurrences
- Wonder Woman - 56 occurrences
- justice league - 42 occurrences
- want read - 42 occurrences
- feel like - 40 occurrences
- reading comics - 34 occurrences
- League Season - 32 occurrences
- just finished - 30 occurrences
- reading order - 28 occurrences

# Building A Model

## Data Preparation

Converting text to numerical representation
Methods Used: Vectorizer, N-gram, TF-IDF

## Model's Used

Naive Bayes, Multinomial Naive Bayes,
Logistic Regression. Ensemble methods
Random Forest Grid Search

## Model Optimization

Hyperparameter Tuning

| Model | Naive Bayes | Logistic Reg | Random Forest |
|---|---|---|---|
| Baseline Accuracy: 0.50262 | **0.8787** | **0.8915** | **0.8907** |

Model 1 MNB
| Train Score: 0.8790
| Test Score: 0.8671

Model 2 Logr
| Train Score: 0.8890
| Test Score: 0.8784

Model 3 RandomForest
| Train Score: 0.8890
| Test Score: 0.8743

Model 4 KNN
| Train Score: 0.8491
| Test Score: 0.7880

Model 5 RandomForest Grid
| Parameter: {'rf__max_depth': None,
| 'rf__max_features': 0.5,
| 'rf__n_estimators': 150}
| Train Score: 0.8890
| Test Score: 0.8772

**Would the model be able to predict which Subbreddit by removing top superhero names from the equation.**

Since we know the top words and bi-grams from each subbreddit we can use those to remove

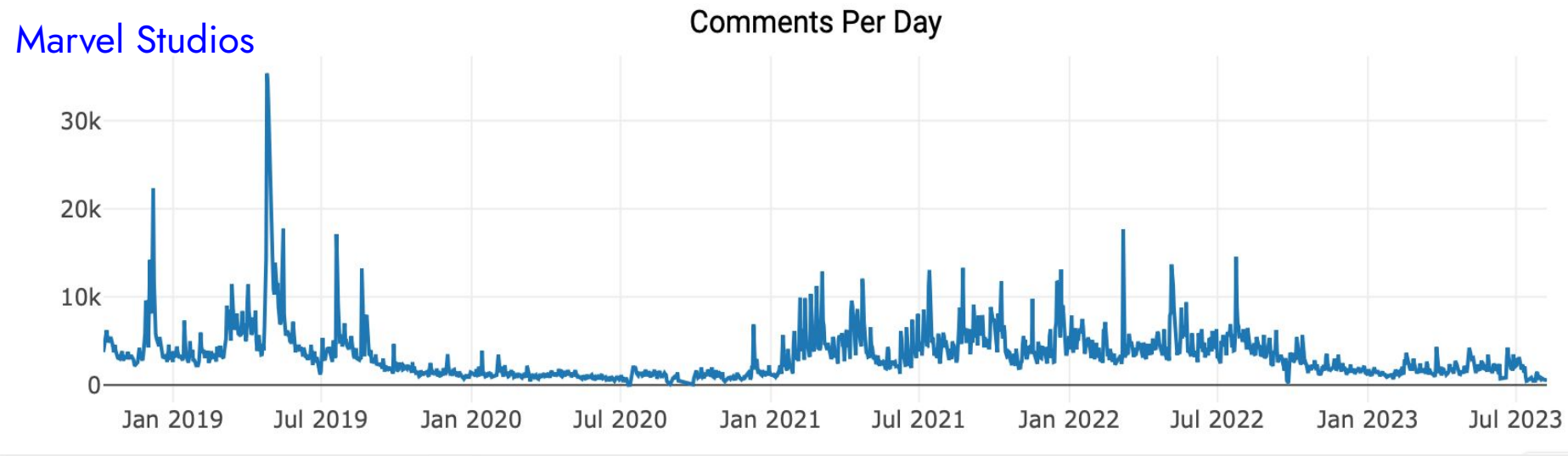
```
vectorizer = TfidfVectorizer(ngram_range=(2, 2), stop_words='english')
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)
```

```
# Remove rows with the top superhero names from the 'Selftext' column
final['Selftext'] = final['Selftext'].apply(lambda text: ' '.join(word for word in text.split() if word.lower() not in superhero_names))
```

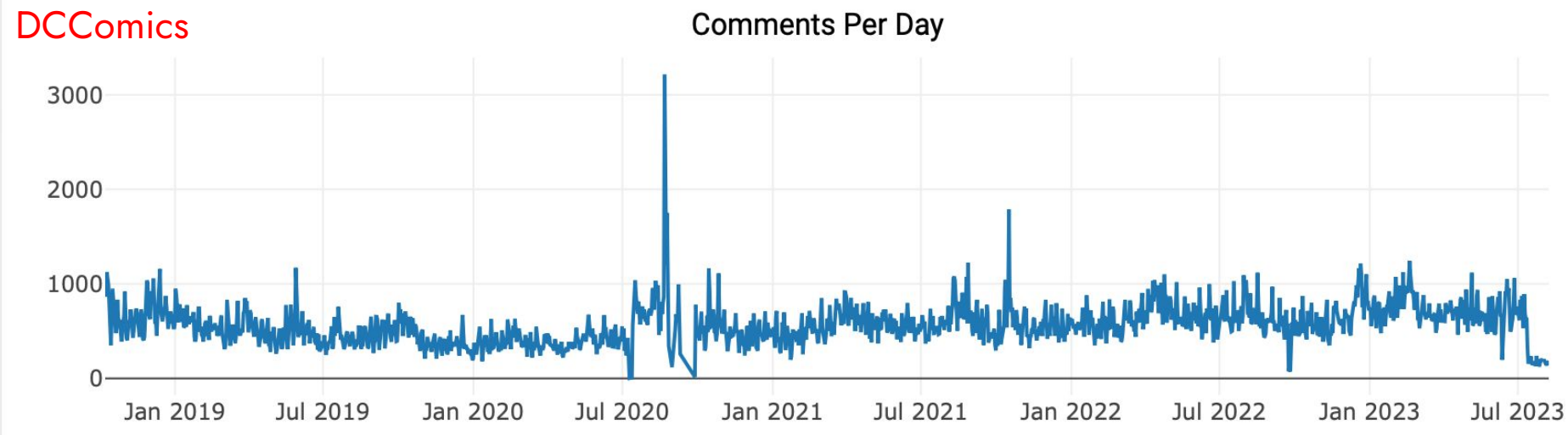
```
Accuracy: 0.8885
Classification Report:
              precision    recall  f1-score   support

           0       0.82      1.00      0.90       656
           1       1.00      0.78      0.88       680

    accuracy                           0.89      1336
   macro avg       0.91      0.89      0.89      1336
weighted avg       0.91      0.89      0.89      1336
```

Opted for TF-IDF vectorization over CountVectorizer, a technique that not only considers the word frequency within a document but also factors in its significance within the entire corpus.

Marvel Studios




DCComics



# Average Sentiment

## Marvel Reddit: 0.0817

## DC Reddit: 0.1058

**1 indicates a strongly negative sentiment**

**0 indicates a neutral sentiment**

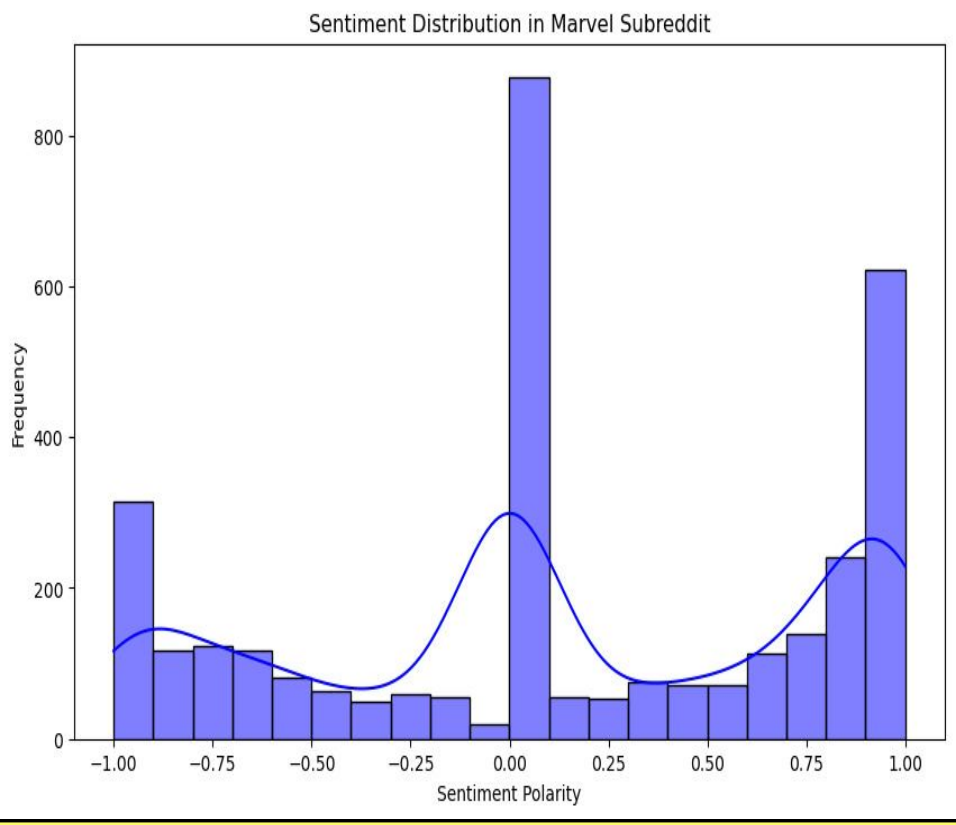
**+1 indicates a strongly positive sentiment**

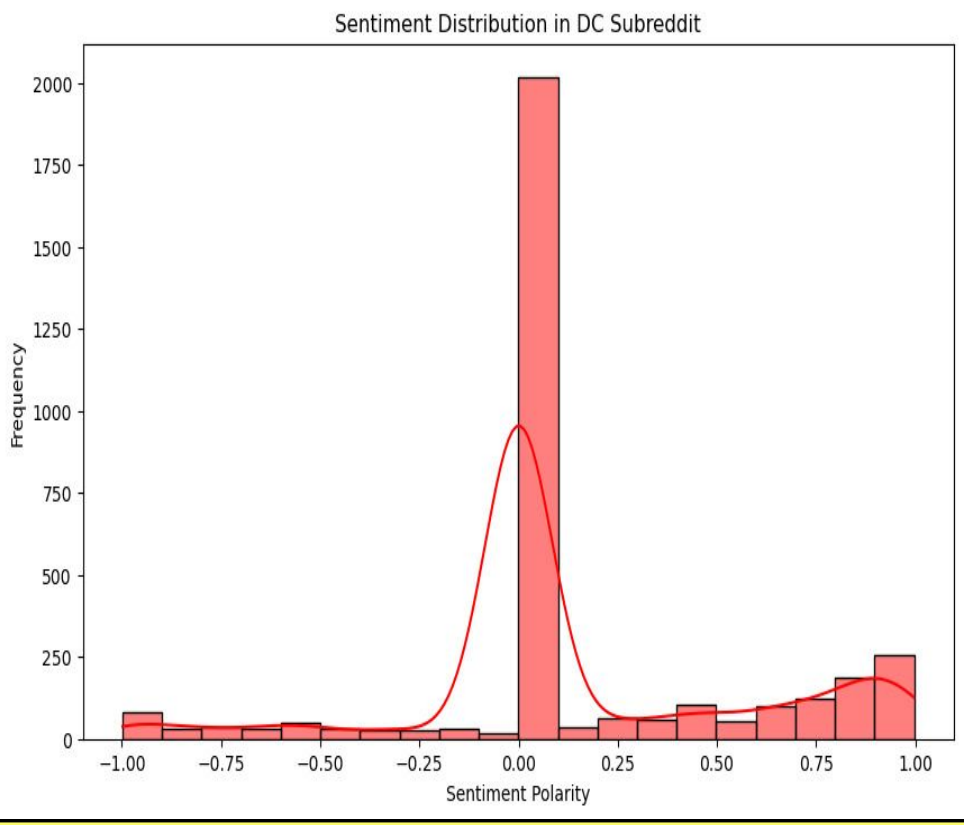
**In this case, the sentiment polarity scores are:**

**Average sentiment polarity for MS subreddit: 0.0817 (Slightly positive)**

**Average sentiment polarity for DC subreddit: 0.1058 (Slightly positive)**

**Both subreddits have slightly positive sentiment, but the DC subreddit has a slightly higher average sentiment polarity compared to the MS subreddit. Although Marvel having more engagement should be taken into consideration**




**Marvel subreddit exhibits a greater disparity in sentiment polarity scores across both positive and negative ends of the spectrum, which implies a wider range of emotional tones in the discussions.**

# Conclusion

## Marvel > DC

Marvel's iconic characters hold greater recognition than DC's except for Superman and Batman. The model's ability to maintain a high level of predictive accuracy even after eliminating dominant bi-grams and superhero names reflects its robustness in capturing meaningful patterns beyond surface-level linguistic features.

Marvel and DC could potentially leverage the insights derived from analyzing forums like subreddits to gain a deeper understanding of ongoing discussions and prevalent sentiments related to their respective shows. This approach holds promise for both companies to stay attuned to audience preferences and tailor their strategies more effectively.

# THANK YOU!