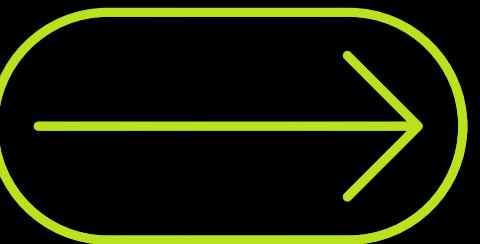


Reddit Project: Web API and NLP (Harry Potter and Marvel)

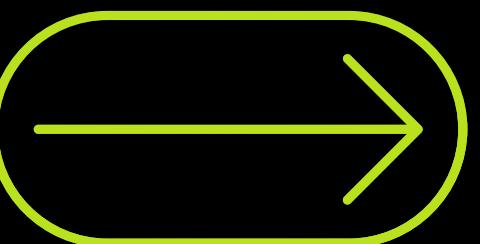
Mohona Yesmin



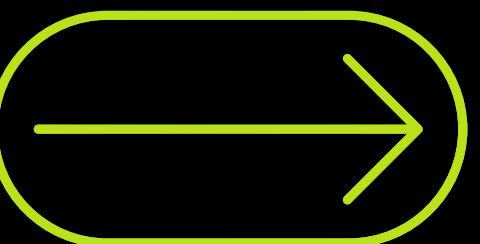
01 - Introduction



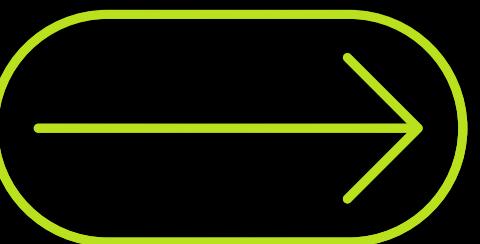
02 - EDA Data Visualization



03 - Model Analysis



04 - Conclusions





01 - *Introduction*

Problem Statement

Identify the preferences of Harry Potter and Marvel fans on social media platforms to inform targeted advertising strategies for Halloween merchandise

Marketing companies seek to better understand the preferences of Harry Potter and Marvel fans on social media to optimize targeted Halloween merchandise ads. This understanding enables more effective advertising, leading to increased engagement and conversions for Halloween sales.

01 - Data Introduction

Data set

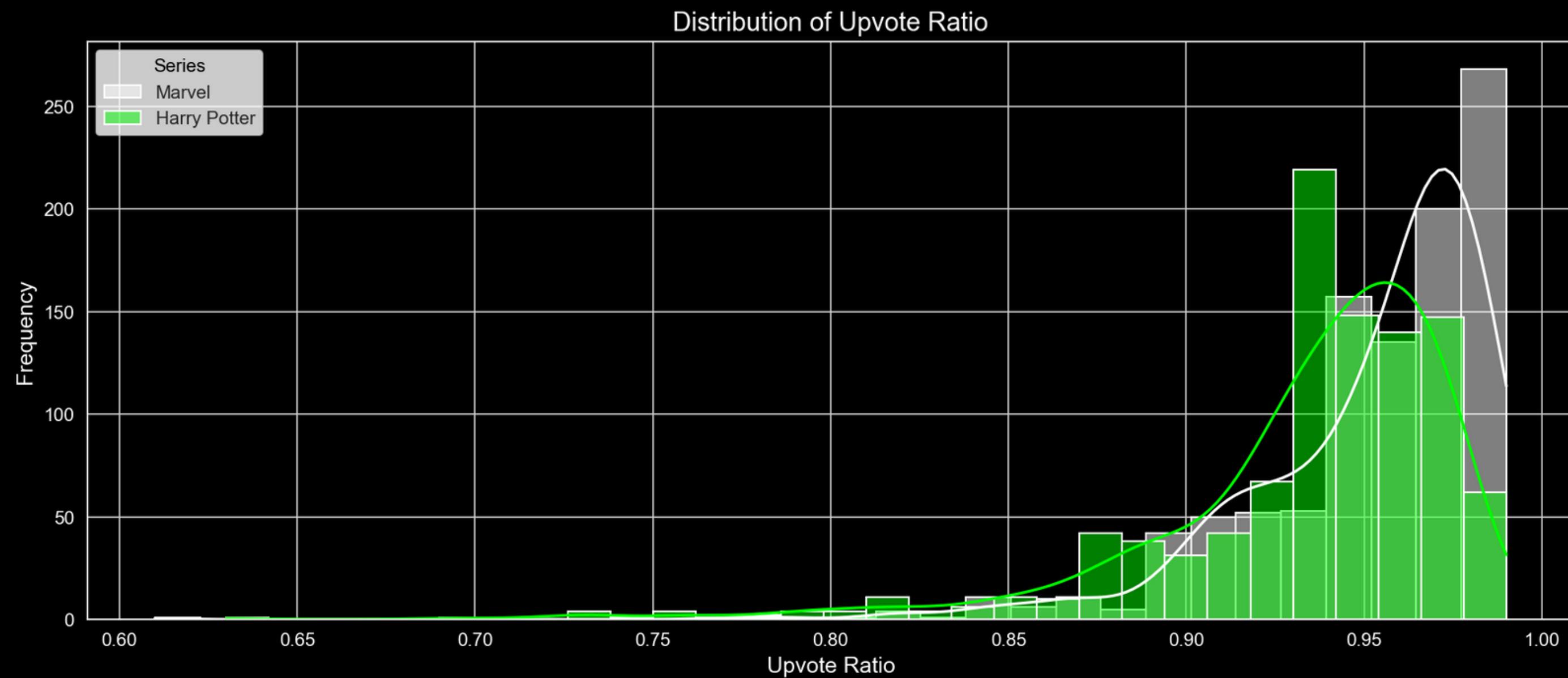
- 1. *marvel.csv*
- 2. *harrypotter.csv*

Data

- 1. *'id'*,
- 2. *'title'*,
- 3. *'content'*,
- 4. *'score'*,
- 5. *'num_comments'*,
- 6. *'author'*,
- 7. *'created_utc'*,
- 8. *'upvote_ratio'*,
- 9. *'subreddit_flair'*,
- 10. *'submission_datetime'*,
- 11. *'post_type'*,
- 12. *'entity_recognition'*,
- 13. *'text_sentiment'*

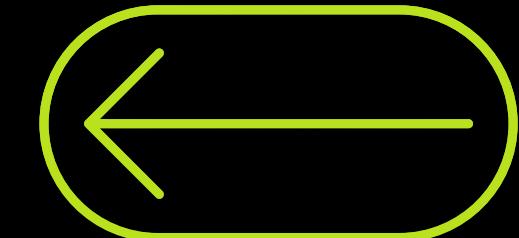
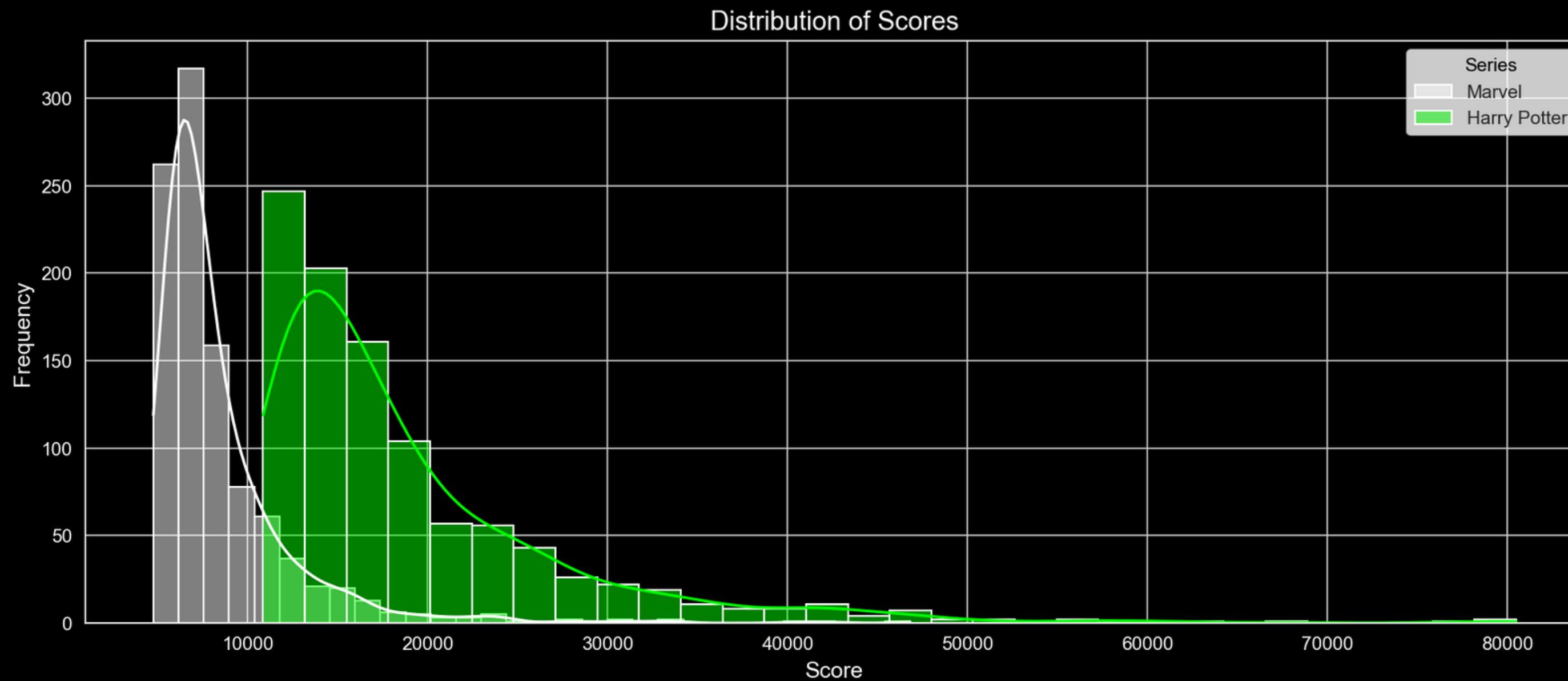
02 - Data Visualization

Distribution of Upvote Ratio - Marvel vs. Harry Potter



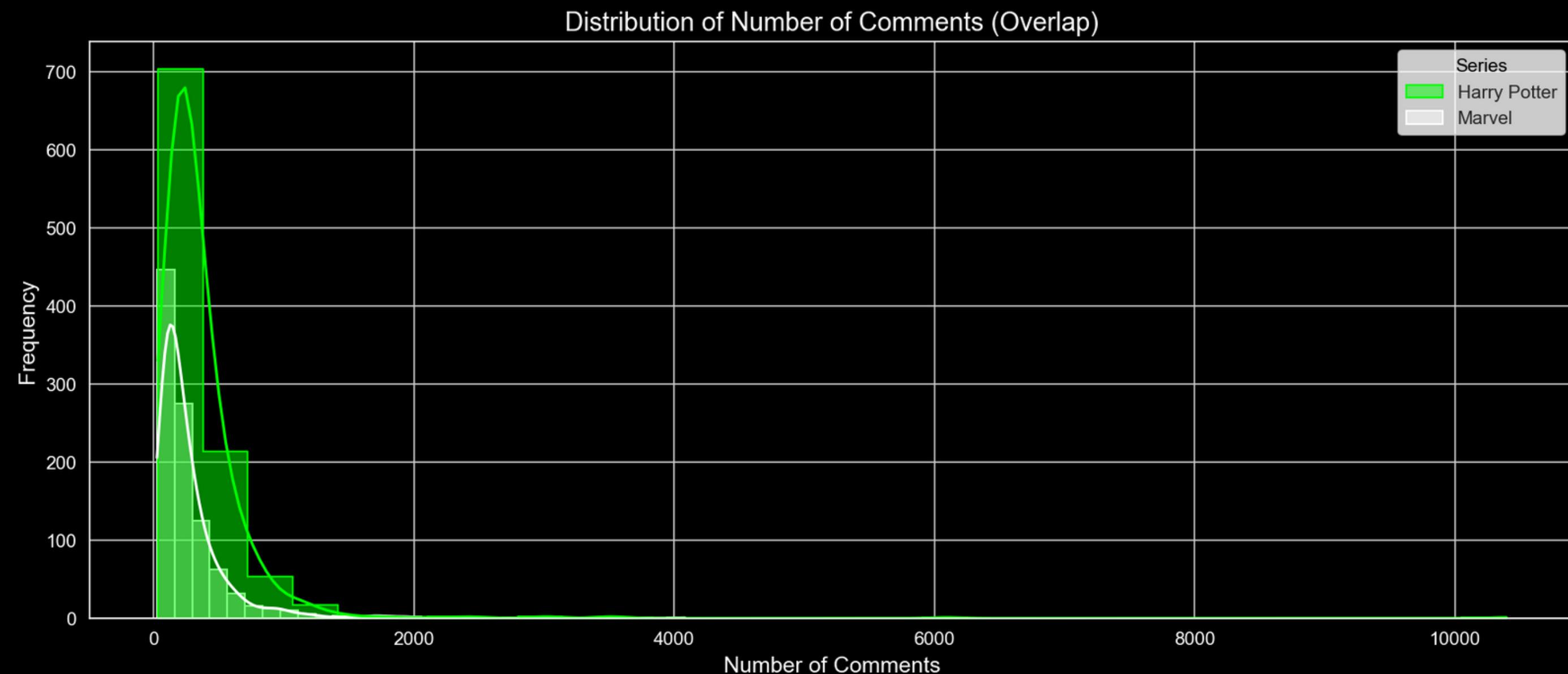
02 - Data Visualization

Distribution of Scores - Marvel vs. Harry Potter

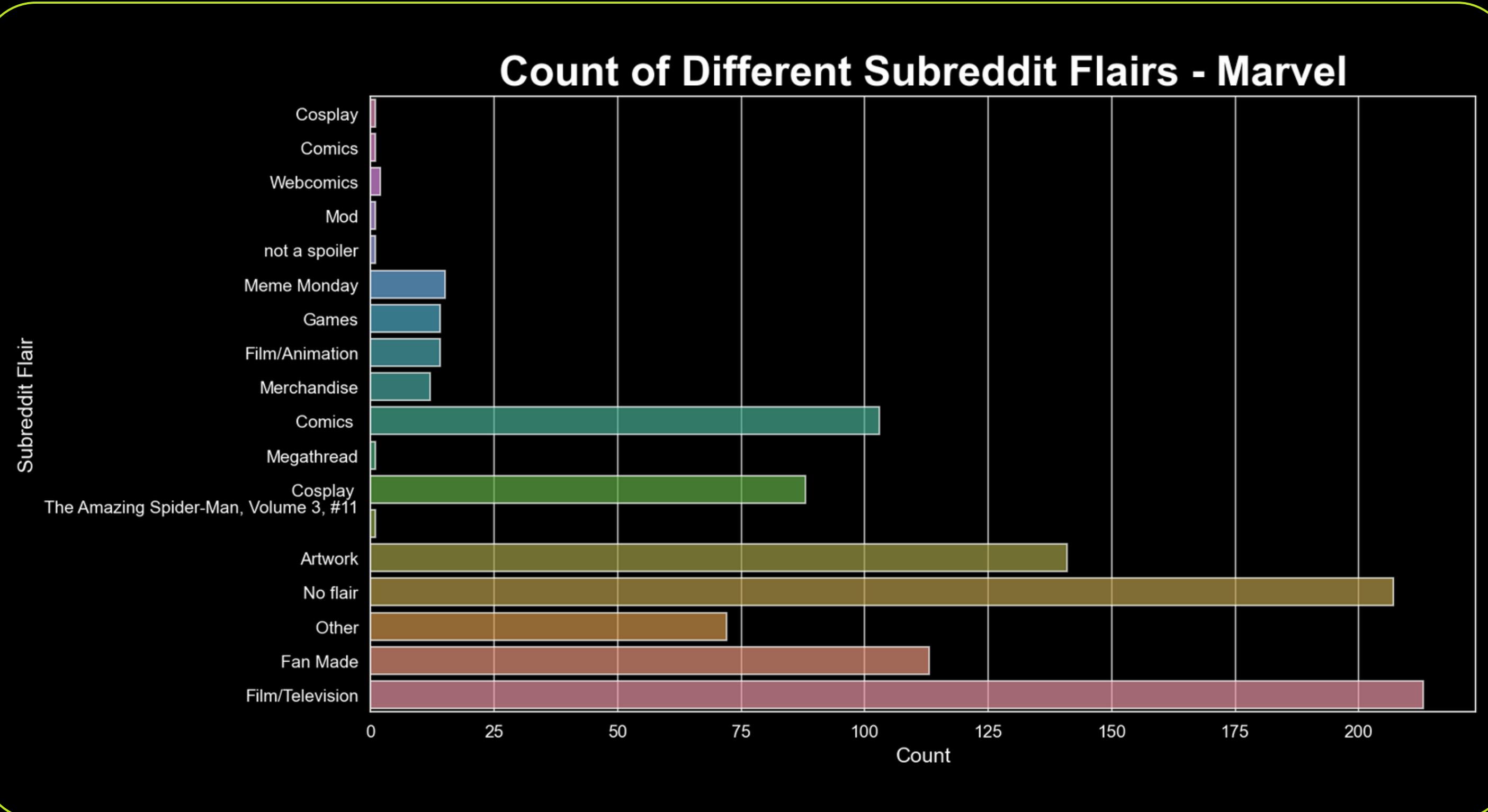


02 - Data Visualization

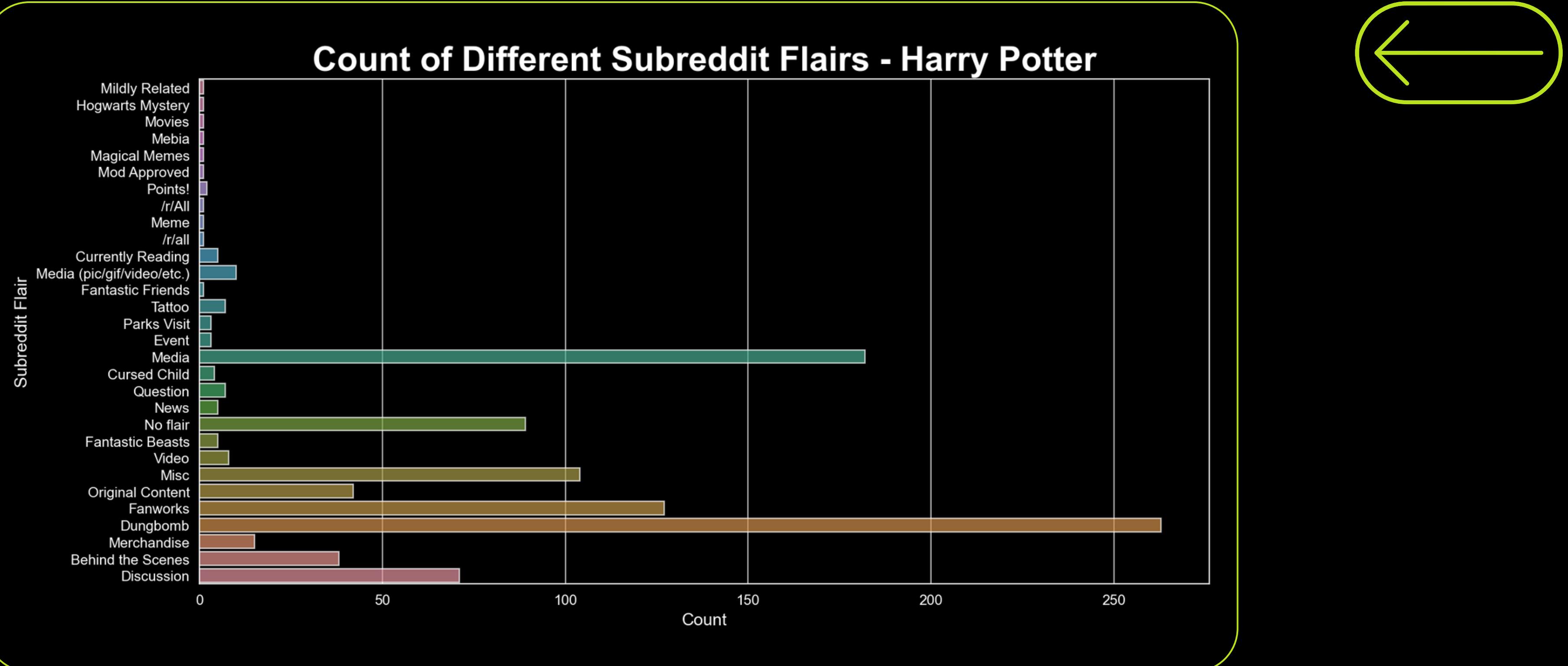
Distribution of Number of Comments - Marvel vs. Harry Potter



02 - Data Visualization

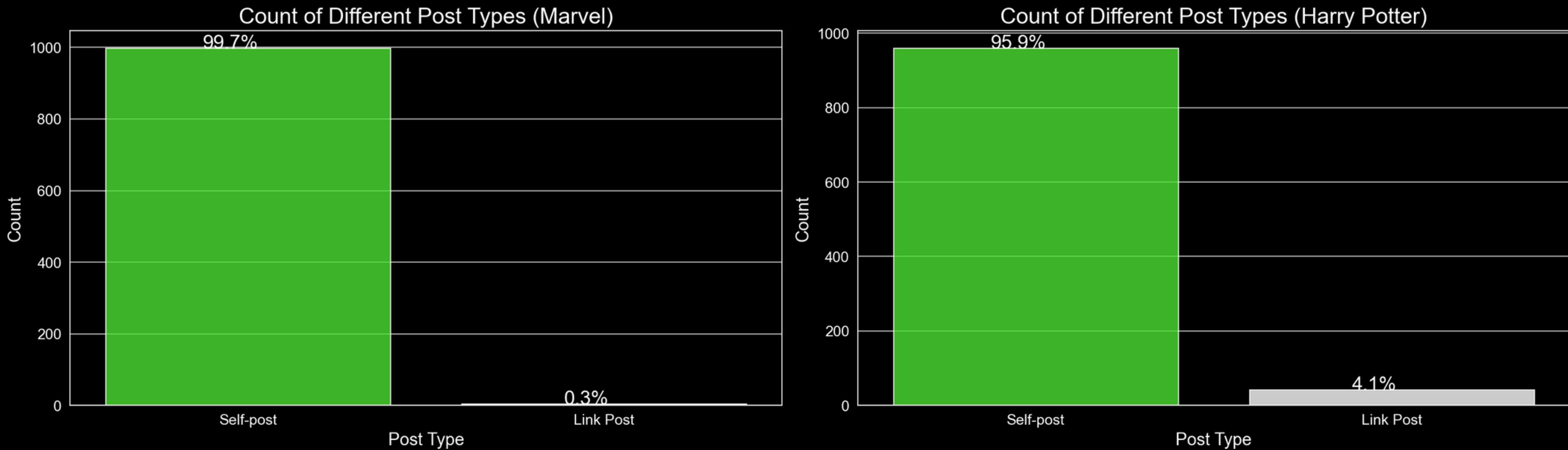


02 - Data Visualization



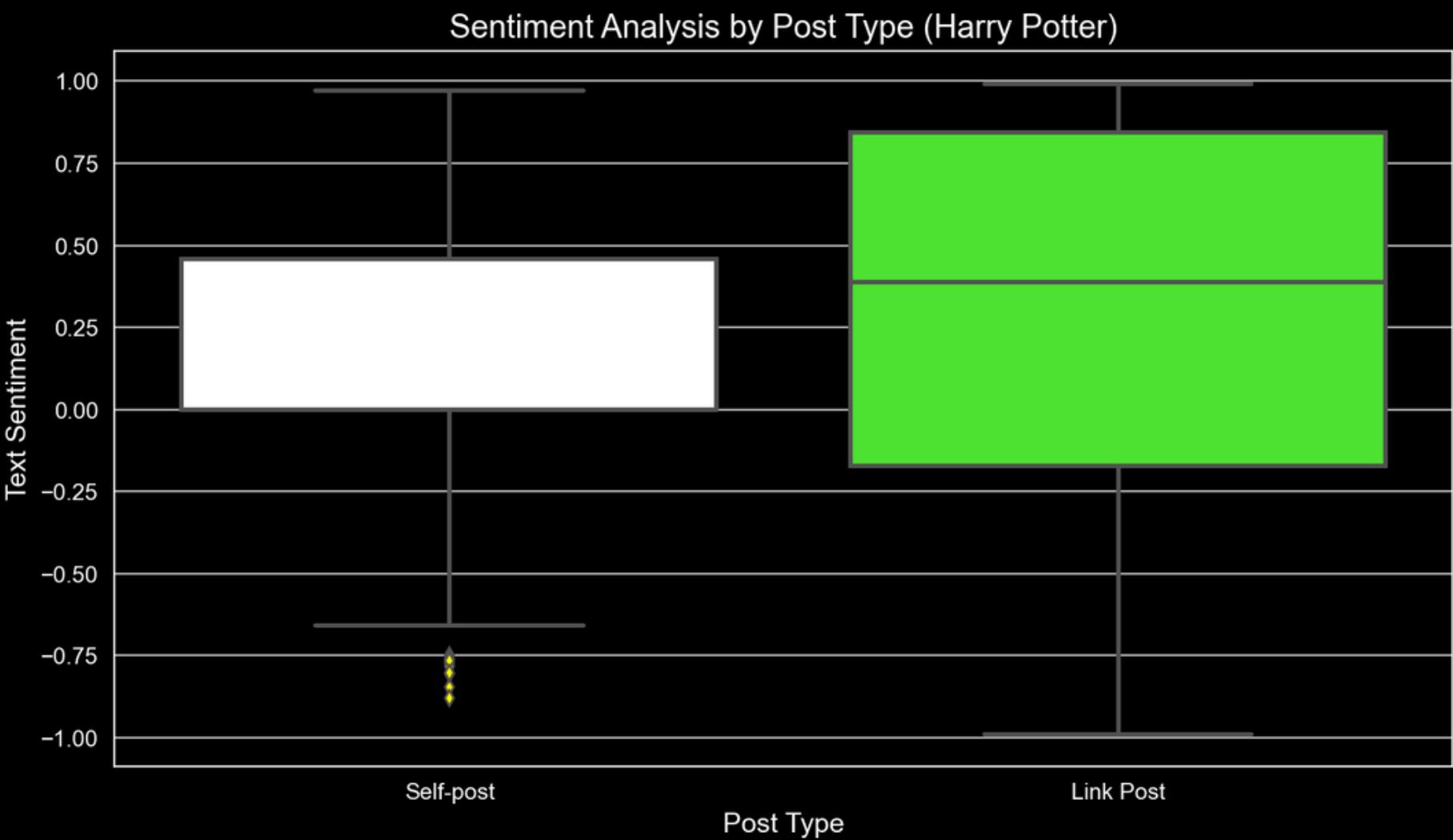
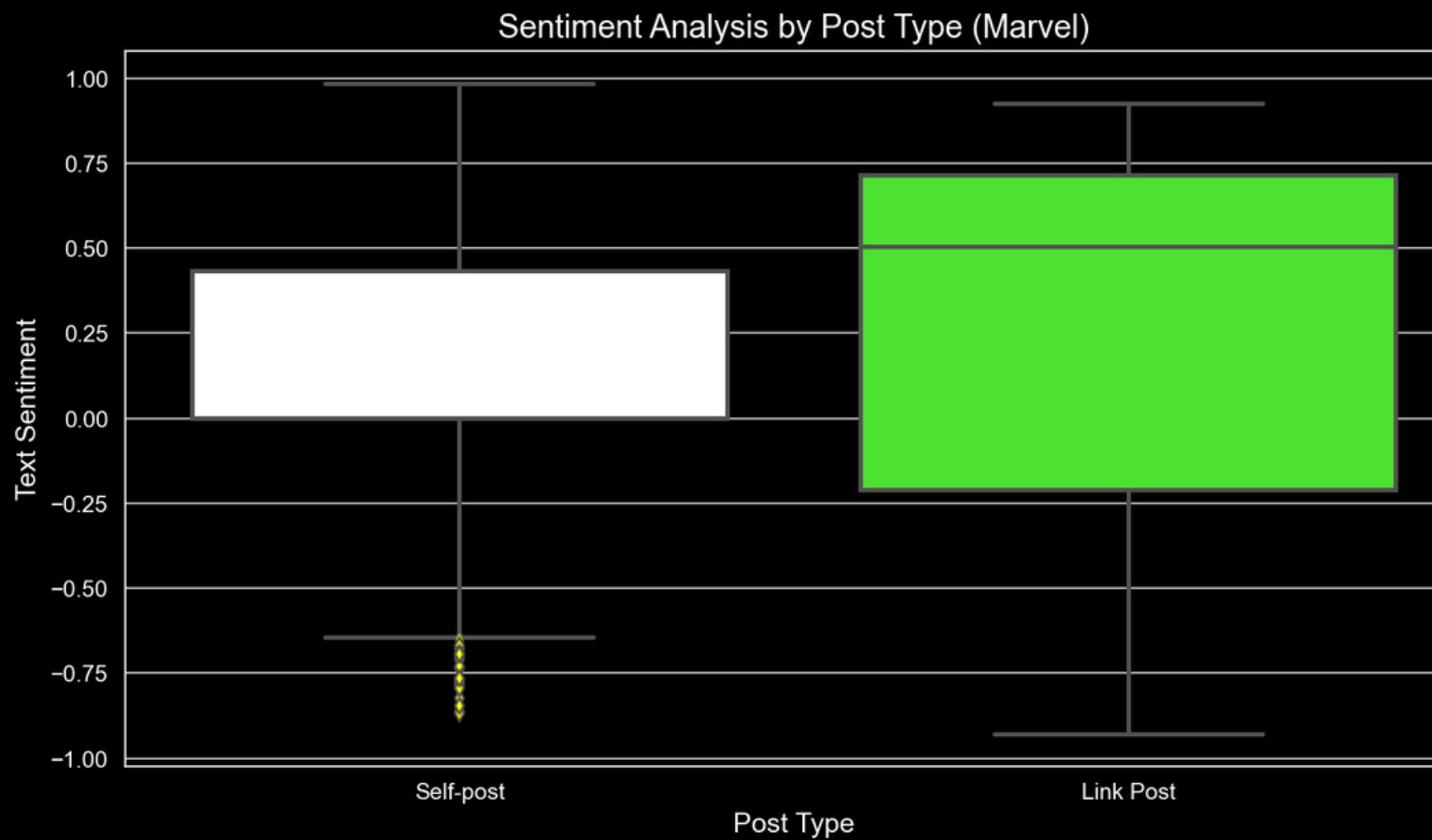
02 - Data Visualization

Count of Different Post Types - Marvel vs. Harry Potter



02 - Data Visualization

Sentiment Analysis by Post Type - Marvel vs. Harry Potter



Common Words

Harry Potter

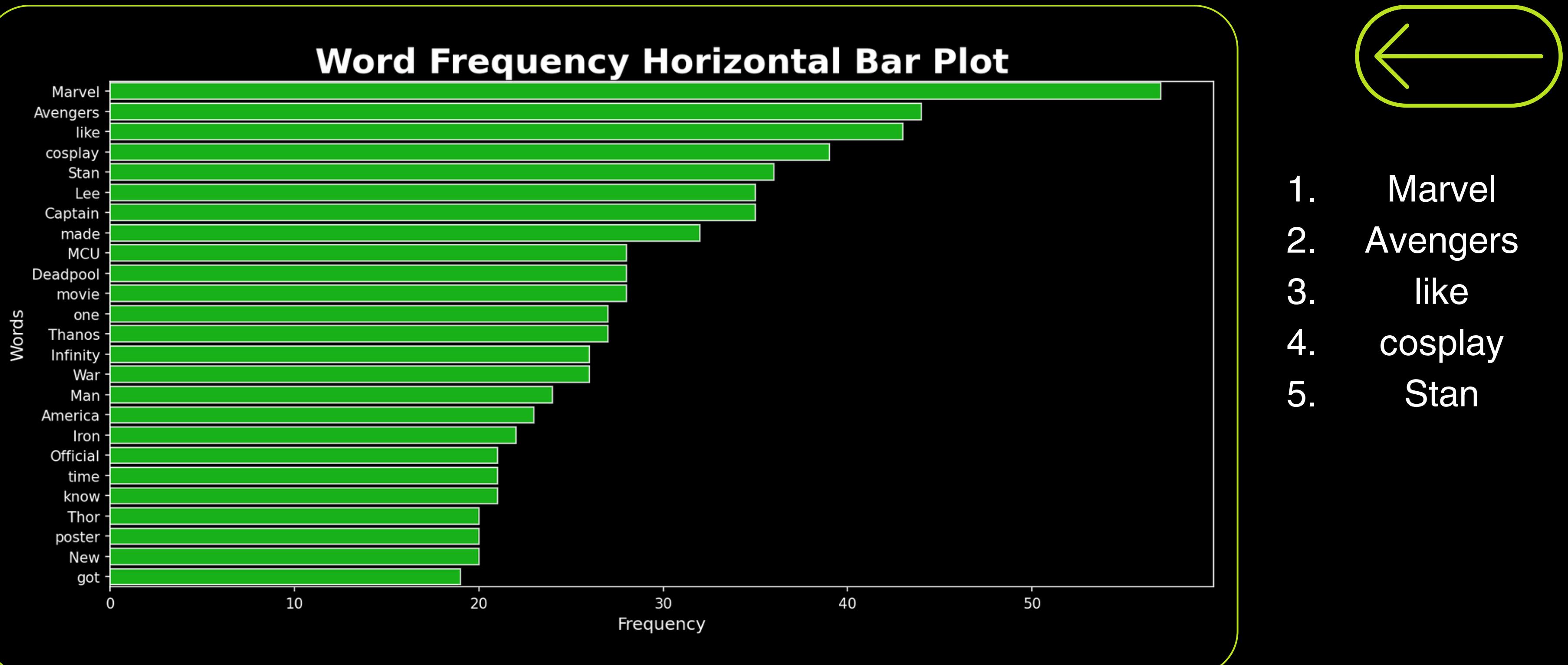
Data Columns

1. *Title*
2. *Content*

Common Words in Title and Content

Hermione thought books would
movie like Voldemort Harry series
Ron year movies
Hogwarts really made one
Weasley Potter Snape
time first Dumbledore
know

02 - Data Visualization



Common Words

Data Columns

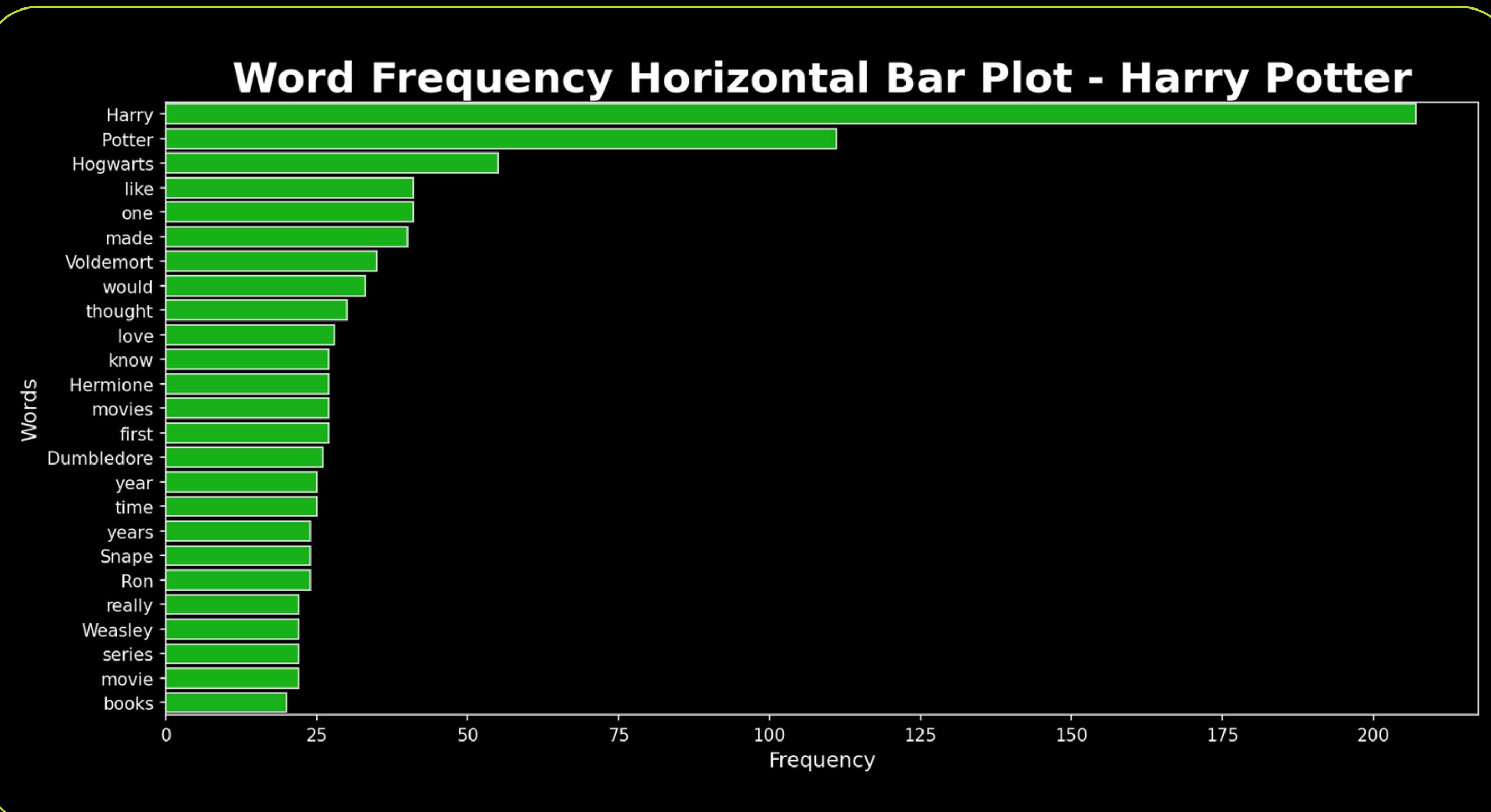
1. *Title*
2. *Content*

Marvel

Common Words in Title

The word cloud is centered around the words "Marvel" and "Avengers". "Marvel" is in large yellow font, with "War" in orange and "Official" in red positioned below it. "Avengers" is in large orange font, with "Stan" in yellow and "Captain" in red positioned below it. To the right of "Marvel", the word "like" is written vertically in yellow, with "one poster" in red above it. To the left of "Avengers", the word "cosplay" is written vertically in red, with "time" in yellow and "Man" in red below it. Other words visible include "Daredevil" (yellow), "movie" (yellow), "made" (red), "Iron" (yellow), "MCU" (red), "America" (yellow), "Infinity" (red), "got" (yellow), "Thanos" (red), and "Deadpool" (yellow).

02 - Data Visualization



1. Harry
2. Potter
3. Hogwarts
4. like
5. one

Logistic Regression Analysis

```
Cross-Validation Scores: [0.771875 0.80625 0.815625 0.815625 0.80625 ]
```

```
Mean CV Score: 0.803125
```

```
Training Score: 0.90875
```

```
Testing Score: 0.825
```

```
Accuracy: 0.825
```

	precision	recall	f1-score	support
Harry Potter	0.80	0.87	0.83	201
Marvel	0.86	0.78	0.82	199
accuracy			0.82	400
macro avg	0.83	0.82	0.82	400
weighted avg	0.83	0.82	0.82	400

Tips



Logistic Regression and TF-IDF model achieved 82.5% accuracy, demonstrating strong subreddit differentiation with balanced precision and recall, consistent cross-validation, and high training performance.



Tips

Accuracy: 0.8

	precision	recall	f1-score	support
Harry Potter	0.79	0.83	0.81	201
Marvel	0.81	0.77	0.79	199
accuracy			0.80	400
macro avg	0.80	0.80	0.80	400
weighted avg	0.80	0.80	0.80	400

Cross-Validation Scores: [0.8175 0.8125 0.84 0.8325 0.845]

Mean CV Score: 0.8295

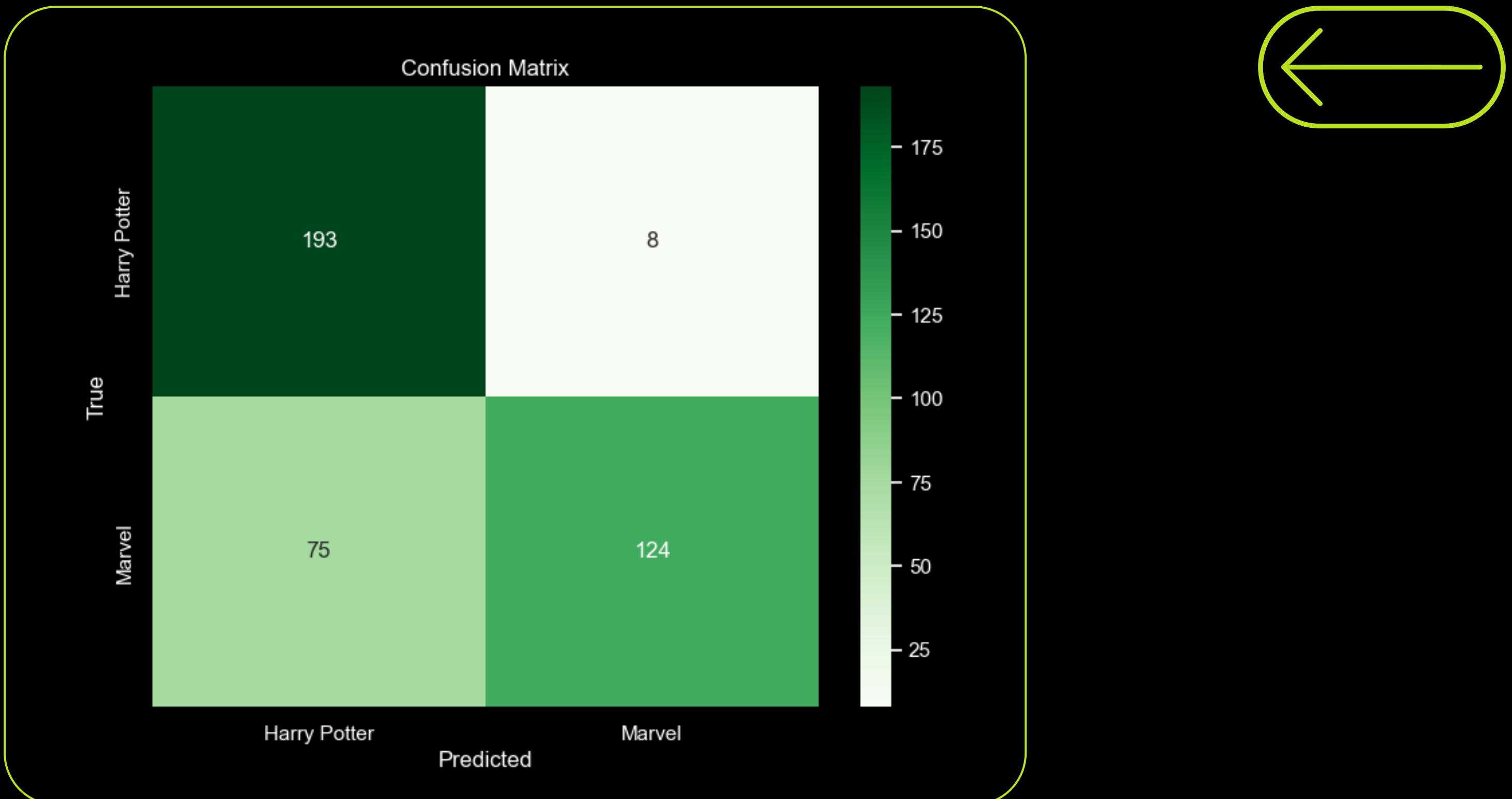
Training Score: 0.94625

Testing Score: 0.8

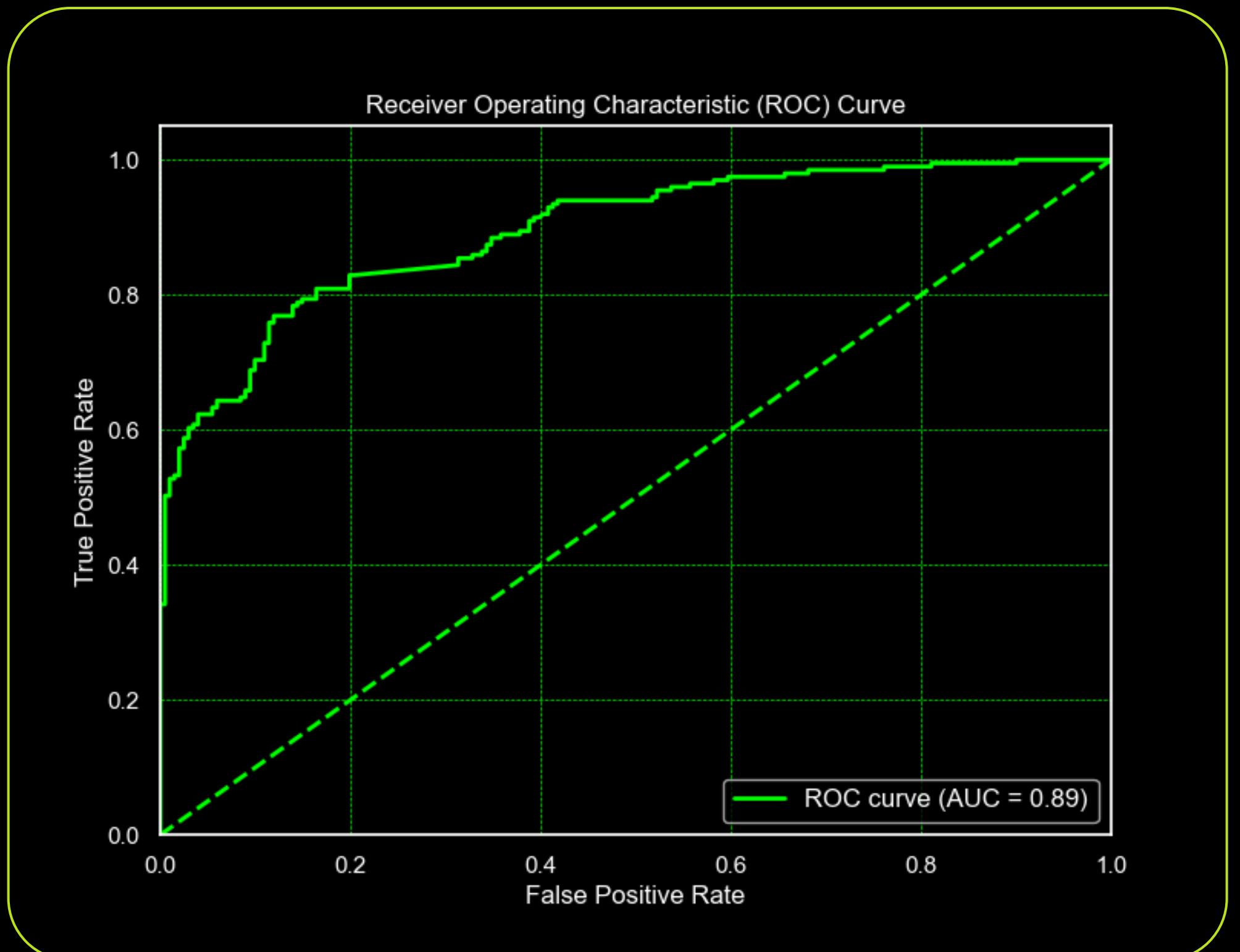
Naive Bayes and TF-IDF model achieved 80% accuracy, demonstrating strong subreddit differentiation with balanced precision and recall, consistent cross-validation, and high training performance.

Naive Bayes Analysis

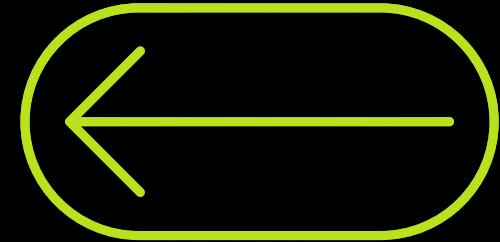
Prediction



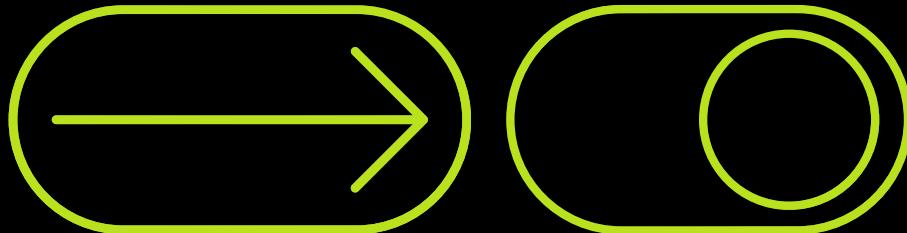
Predictions and Inference



ROC AUC 0.89



04 - Conclusions



1. ***Best Classifier Model: Logistic regression Analysis.***
 - a. ***Title and Content***
2. ***TF-IDF Vectorization (Term Frequency-Inverse Document Frequency)***
3. ***CV = 0.80, Train = 0.91, Test = 0.825***
4. ***ROC AUC = 0.89***

Thanks
