Subreddit

 vs. 

Jackie Petersen

Problem Statement

- Can I build a model that can classify the text as belonging to one subreddit or the other?

The Data

- r/harrypotter⚡ and r/Marvel 🕷
- 15,000 posts each
- Date range: December 2021 to April 2022

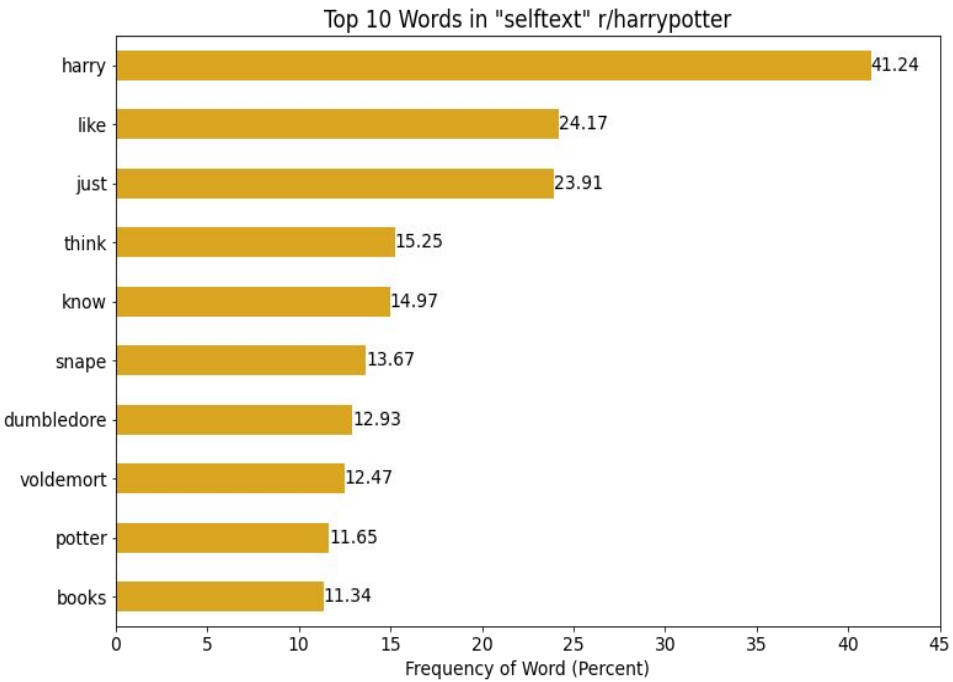
The Data Cont.

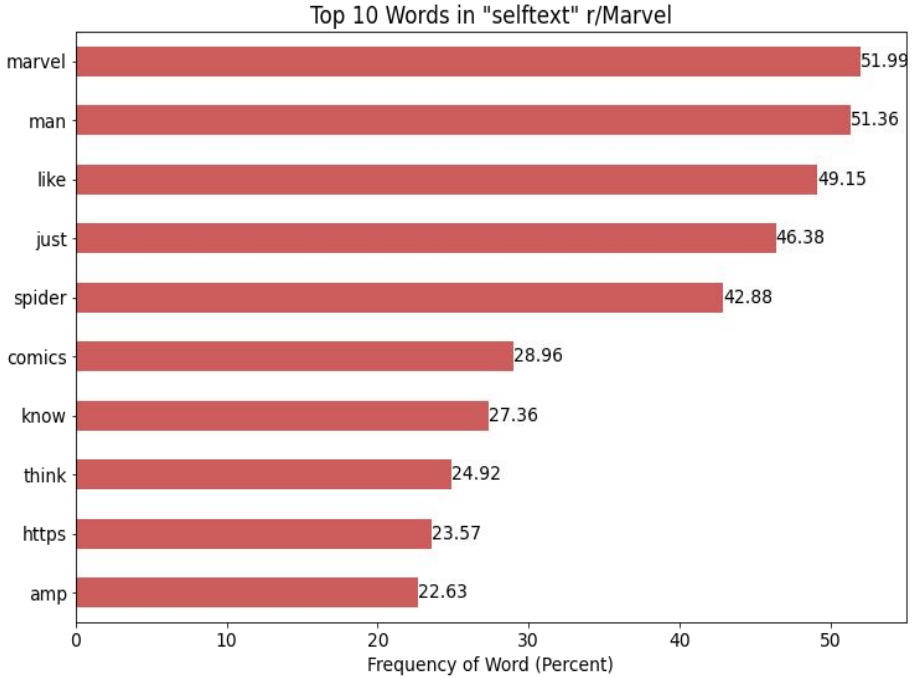
- Dropped:
  - Duplicates
  - ‘Removed’/‘deleted’ posts
  - NaN values
- Removed unnecessary columns

The Data Cont.

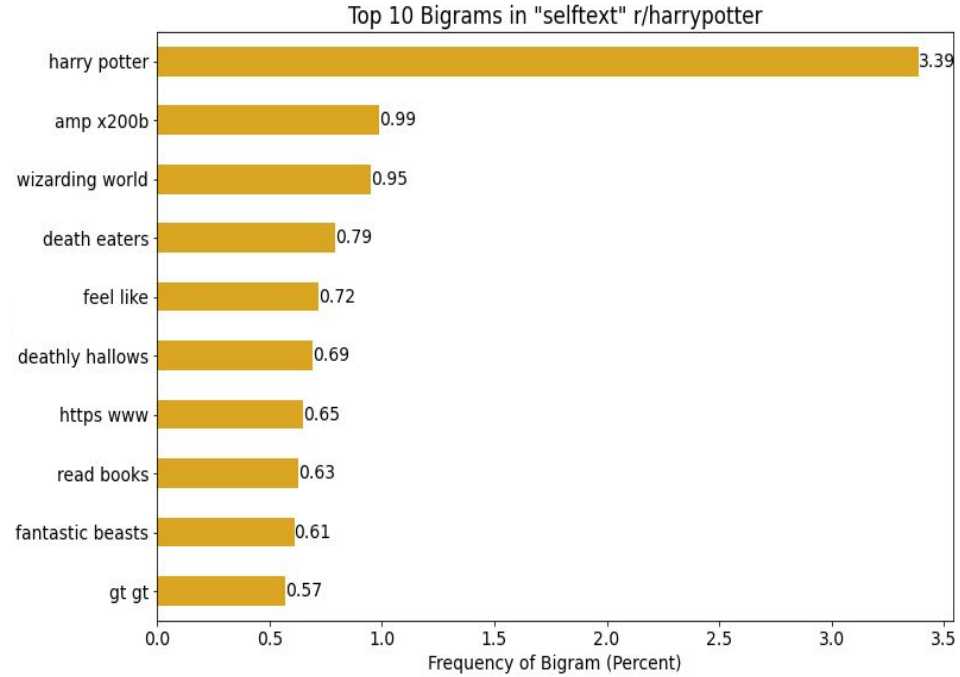
- Final count:
  - 6109 (64%) from r/harrypotter
  - 3491 (36%) from r/Marvel

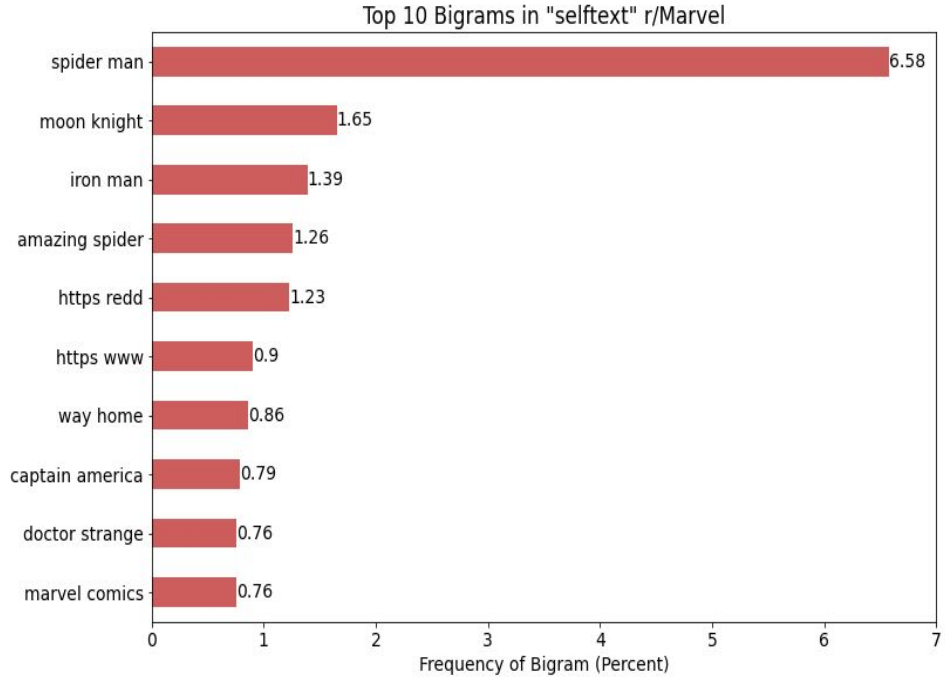
Most Common Words in ‘selftext’



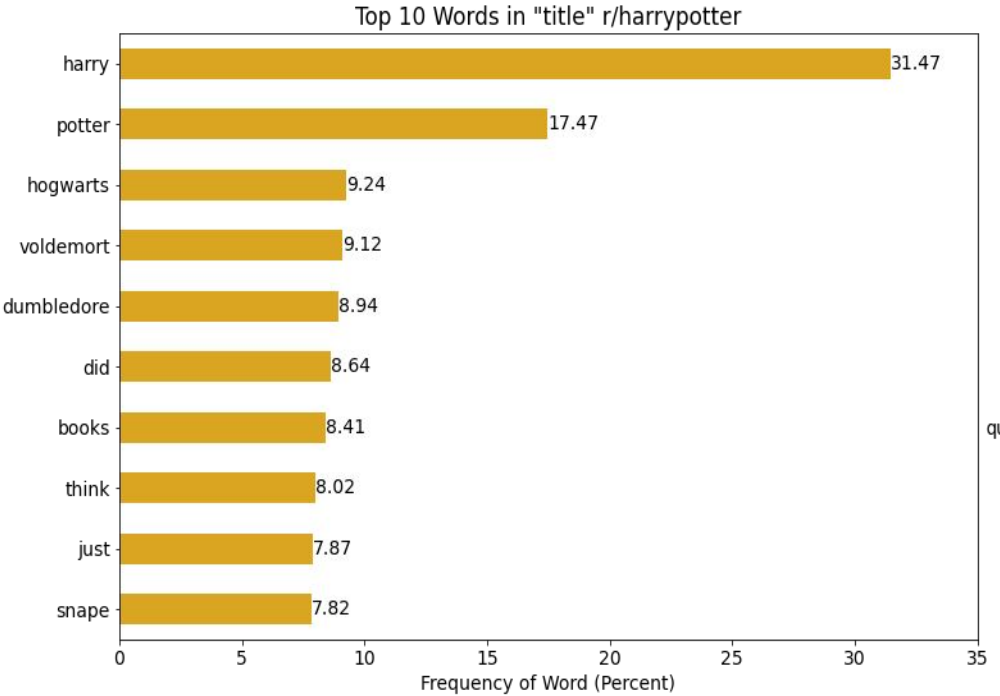


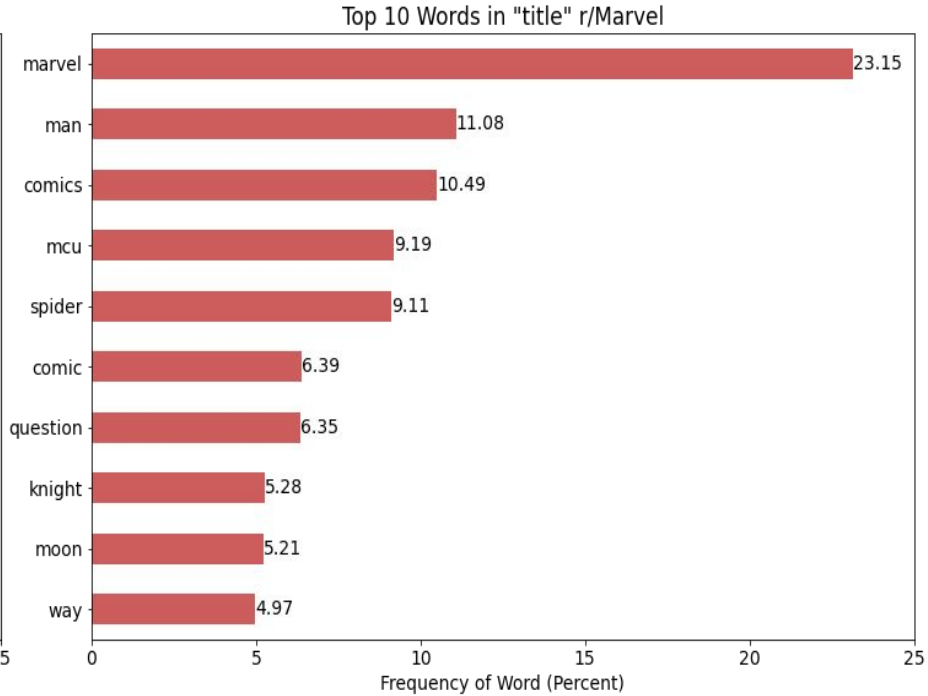
Most Bigrams in ‘selftext’



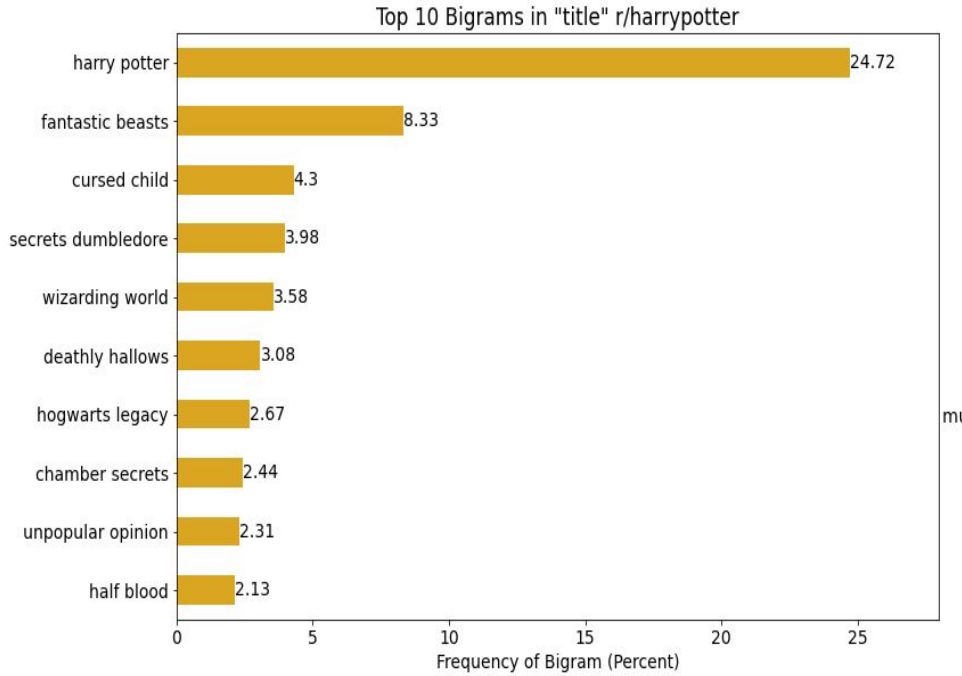


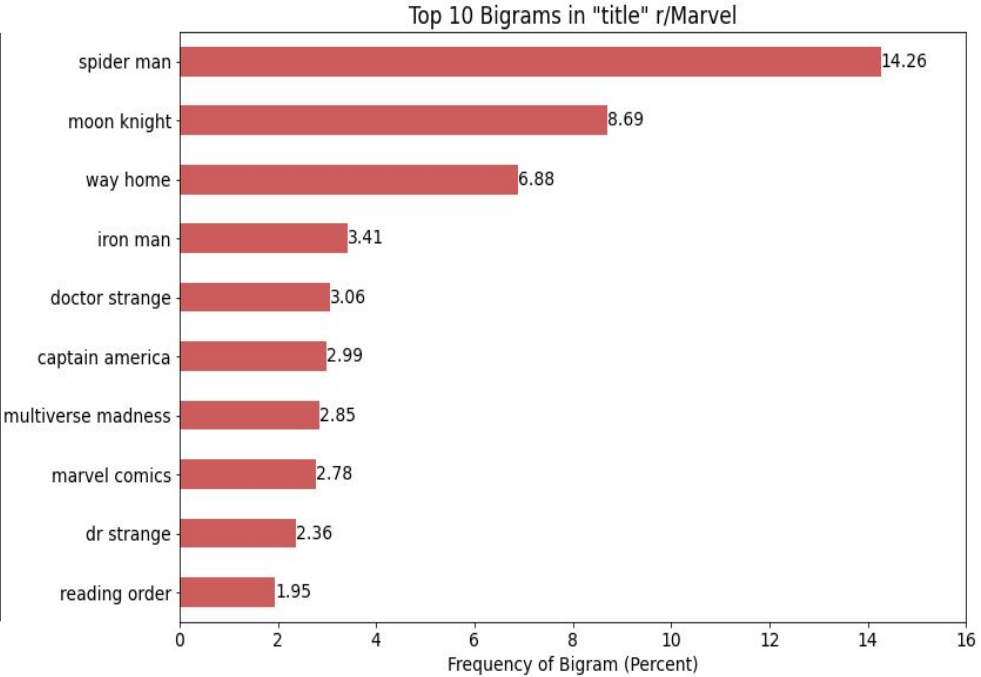
Most Common Words in ‘title’





Most Common Bigrams in ‘title’



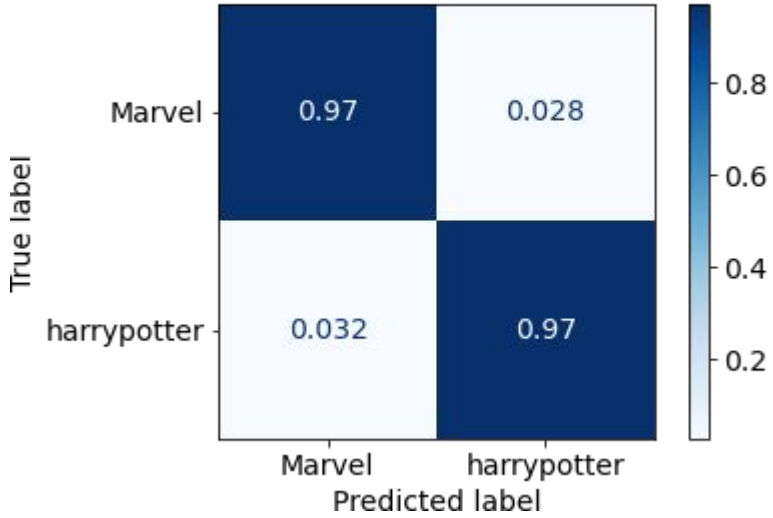


Model Process

- Baseline: 64% accuracy
- Features: ‘selftext’ and ‘title’
- Split data into ‘train’ and ‘test’ sets
- Vectorized text
  - Tried lemmatize, accuracy ⬇
- Used ‘english’ stop words
- Removed words that appeared too infrequently

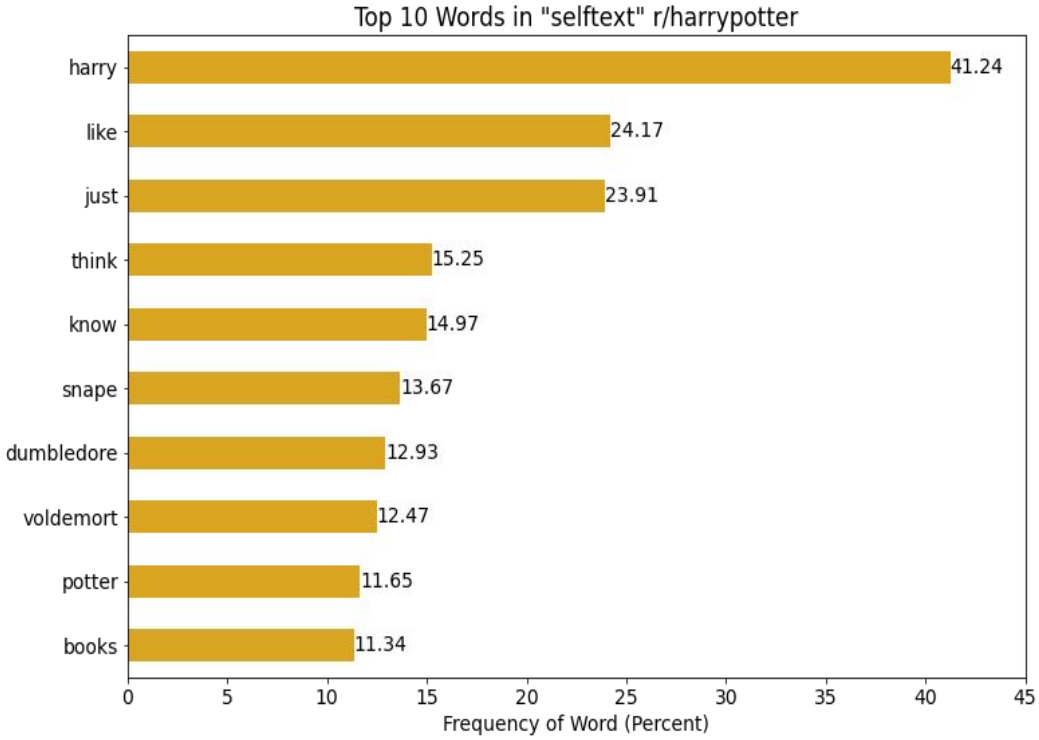
The Model

- Accuracy score of 97%



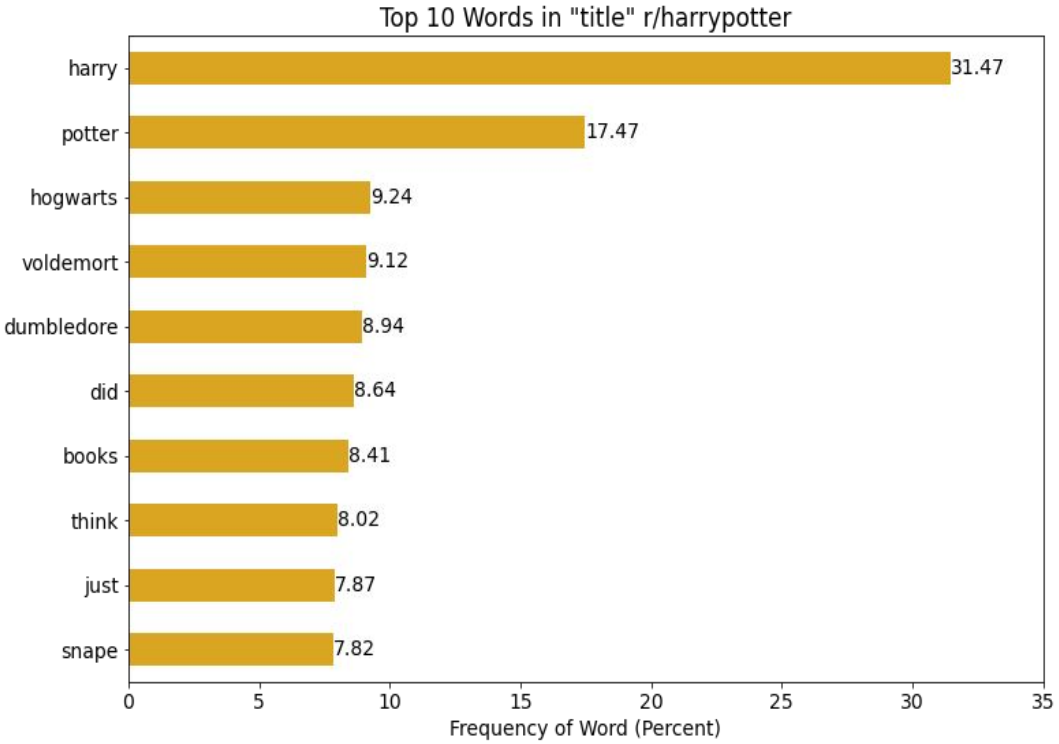
The Model

Words in ‘selftext’	More Likely to be in HP
harry	8.23
books	4.52
hogwarts	4.10
voldemort	3.37
hp	3.33
dumbledore	3.09
wand	2.94
snape	2.61
potter	2.58



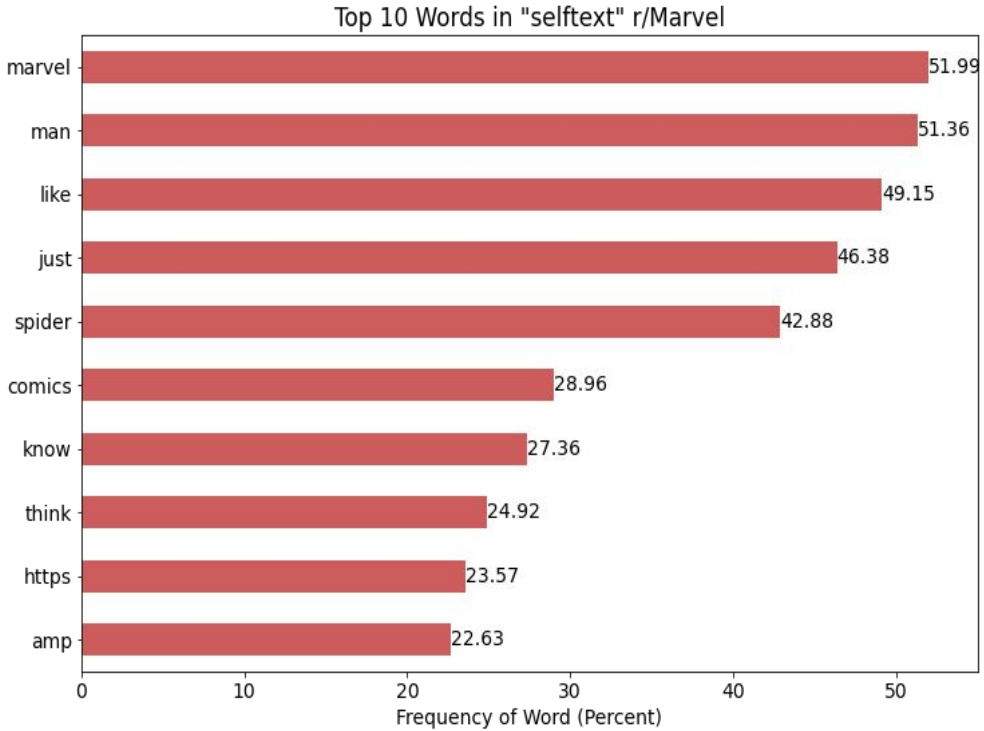
The Model

Words in ‘title’	More Likely to be in HP
harry	7.81
hogwarts	5.91
voldemort	4.34
dumbledore	4.03
hp	3.47
beasts	3.31
snape	3.13
potter	2.56
secrets	2.42



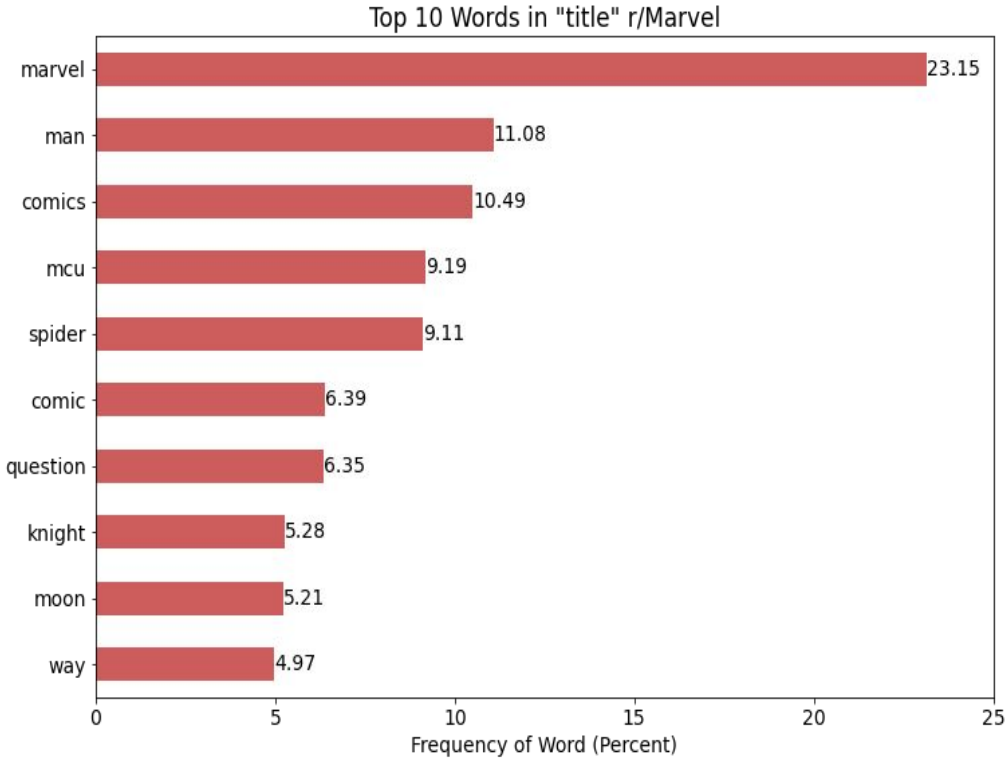
The Model

Words in ‘selftext’	Percent More Likely to be in Marvel
marvel	94%
comics	89%
comic	88%
mcu	85%
avengers	78%
wolverine	70%
spiderman	70%
thor	66%
spider	66%



The Model

Words in ‘title’	Percent More Likely to be in Marvel
marvel	94%
mcu	87%
moon	79%
morbius	78%
comic	77%
comics	77%
strange	73%
spiderman	72%
man	66%



Can the model keep performing this well moving forward?

Conclusions and Recommendations

- The words are dependent on current trends
- Continue to gather data to check on model

Thank you!

Any questions?