Reddit Classification

By Erik Mercado

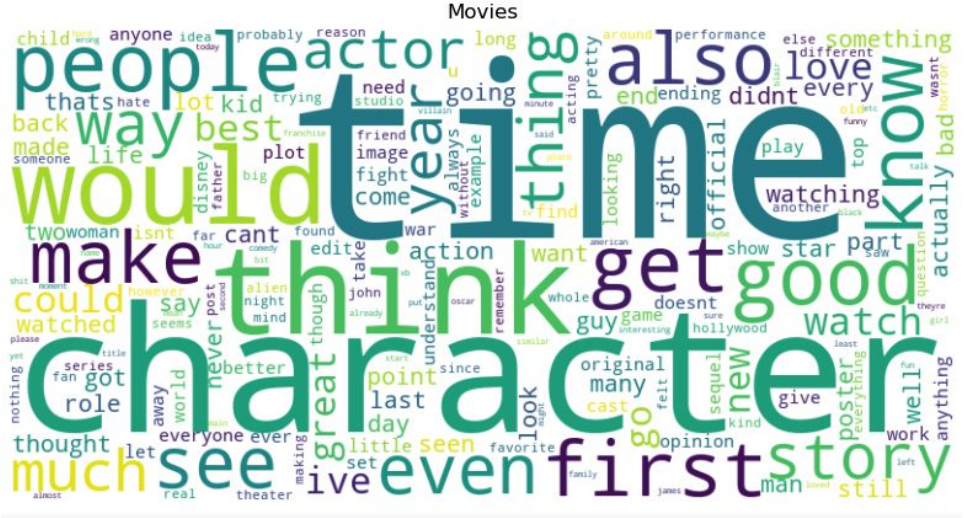
Problem Statement

We are running a ad compagine and want to learn what people are talking about in r/books and r/movies.

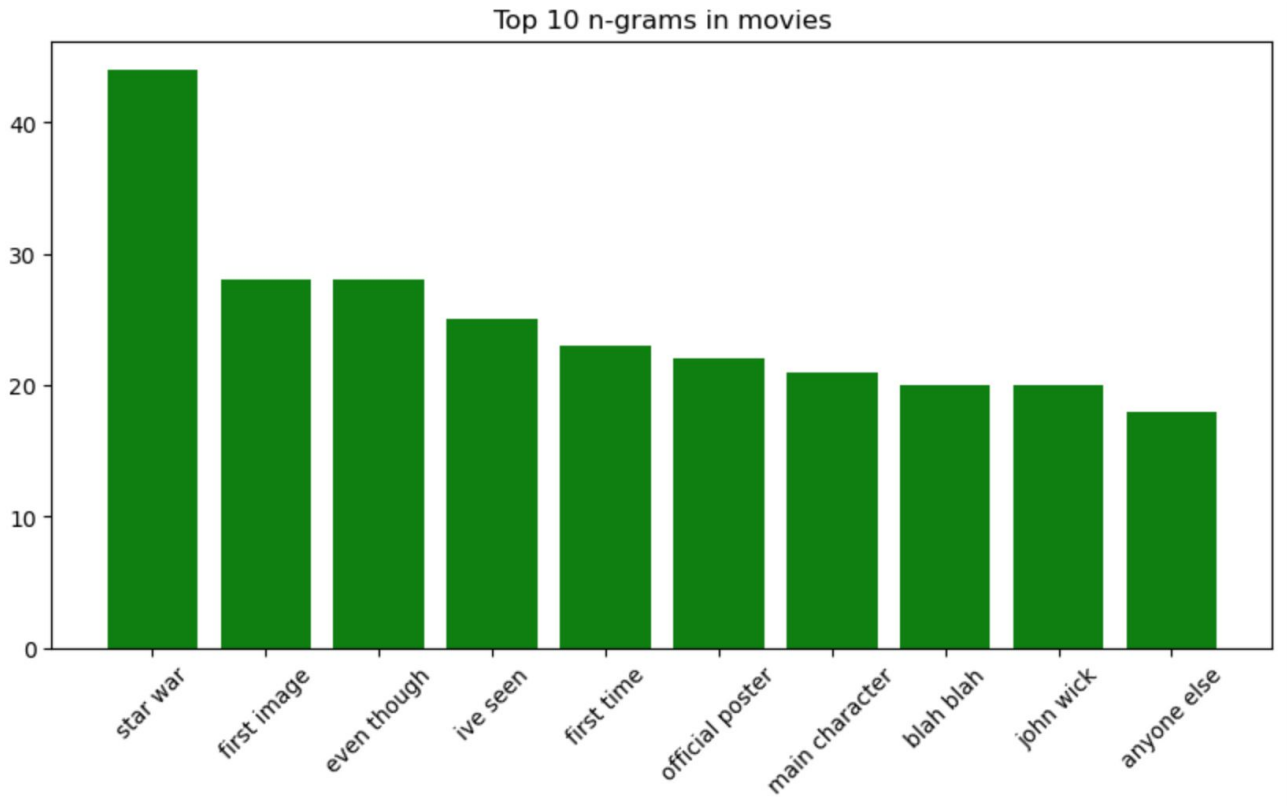
Processes

- Scrape Reddit
  - Hot, new, top, rising, controversial
  - 1227 posts from r/Movies
  - 1179 posts from r/Books
- Clean Data
  - Remove URLS, Special Chars, and Nums, Stop
  - Remove obvious words from each subreddit
    - Book, Reading, Author, Page
    - Movie, Film, Director, Scene
- Modeling
  - Random Forest
  - SVC

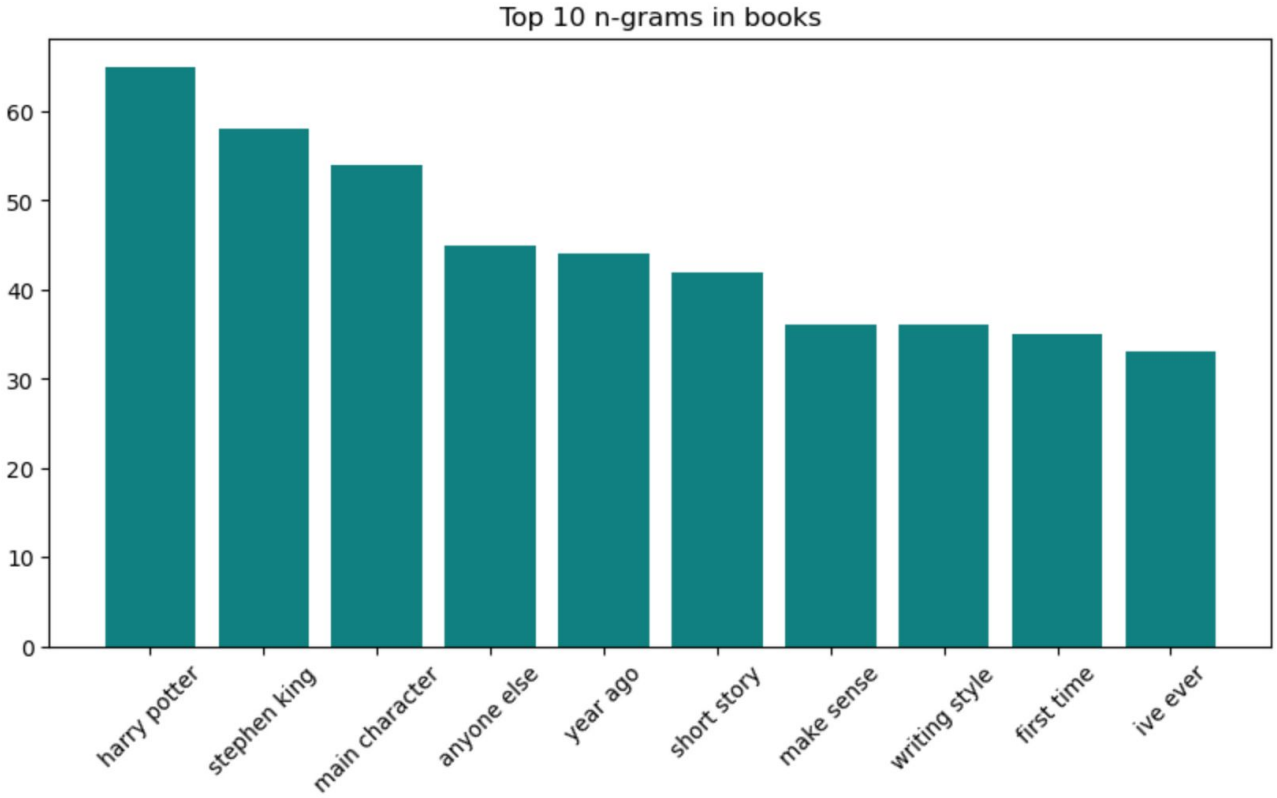
EDA


Movies
people
actor
also
time
would
character
think
make
get
good
know
watch
story
see
even
first
much
thing
year
love
way
best
life
look
Books
good
much
world
year
even
character
people
time
story
think
know
love
also
thing
novel
way
would
make
want
get
go
see
first
life
series
thought
take
give
work
reader
woman
ive

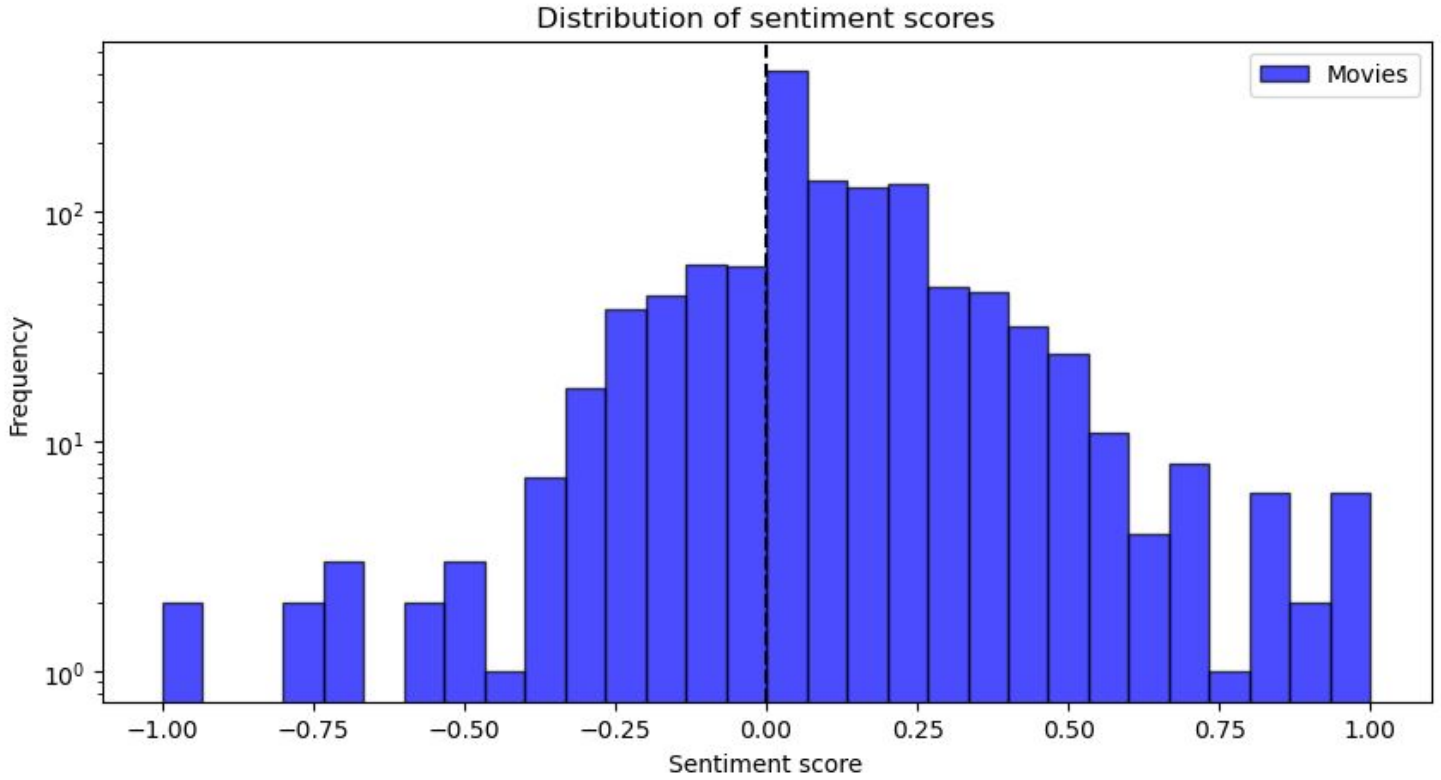
EDA


Top 10 n-grams in movies
40
30
20
10
0
star war
first image
even though
ive seen
first time
official poster
main character
blah blah
john wick
anyone else

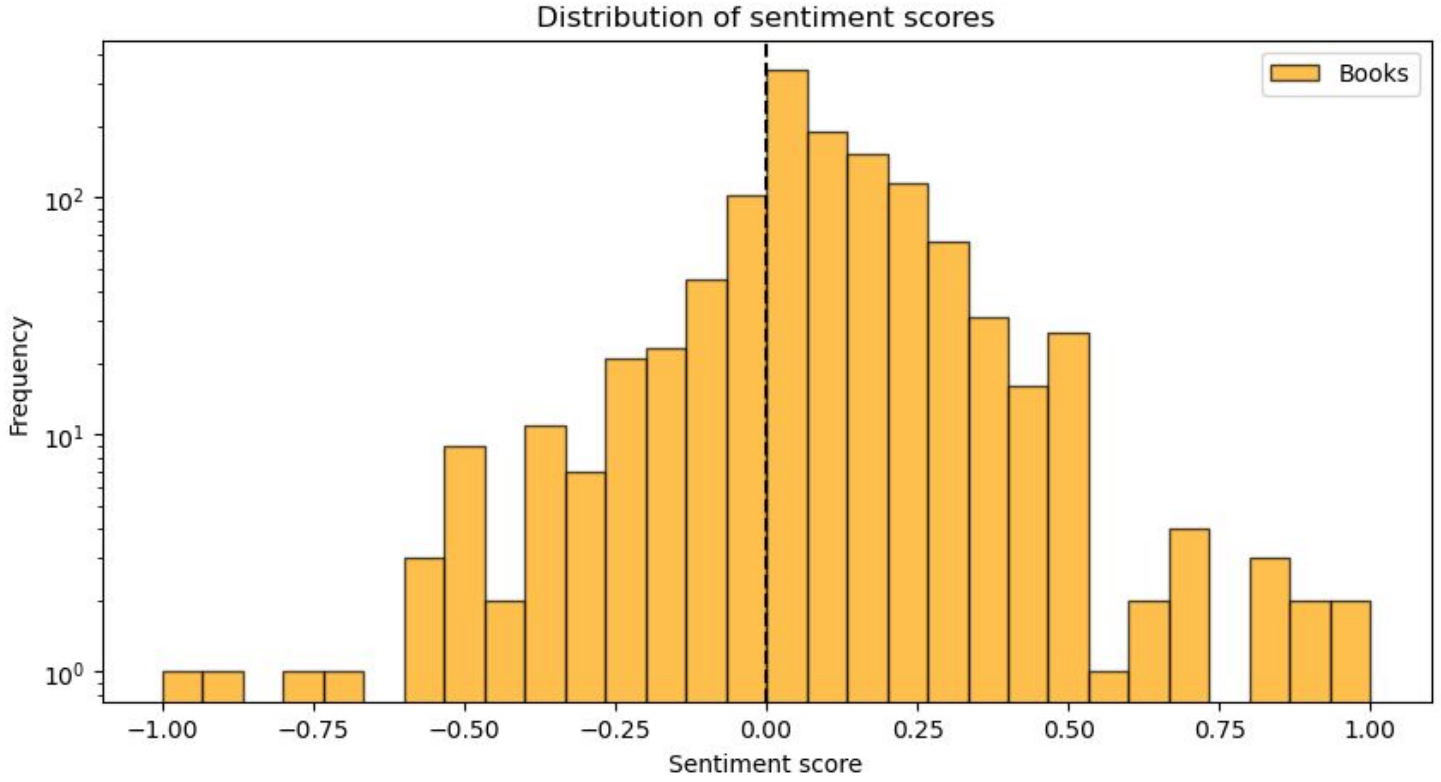
EDA


Top 10 n-grams in books
60
50
40
30
20
10
0
harry potter
stephen king
main character
anyone else
year ago
short story
make sense
writing style
first time
ive ever

EDA


Distribution of sentiment scores
Movies
Frequency
10^2
10^1
10^0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment score

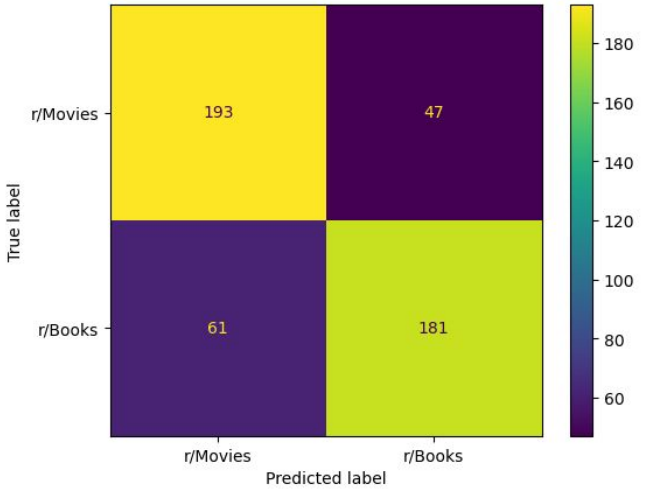
EDA


Distribution of sentiment scores
Books
Frequency
10^2
10^1
10^0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment score

Evaluations

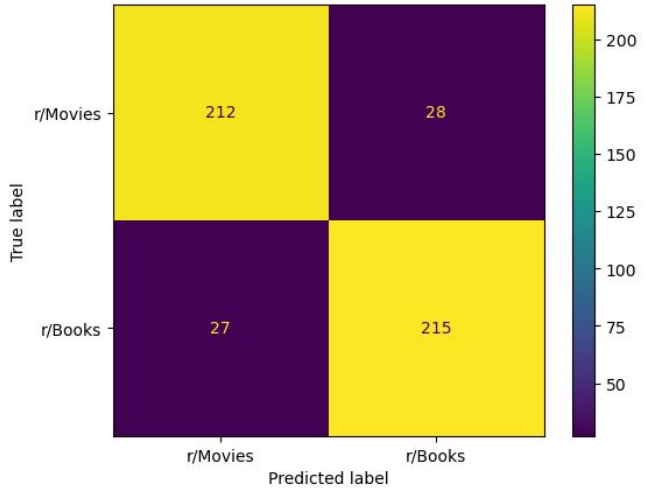
Random Forest

```
Best Accuracy: 0.8040449134199135
Test Accuracy: 0.7759336099585062
```

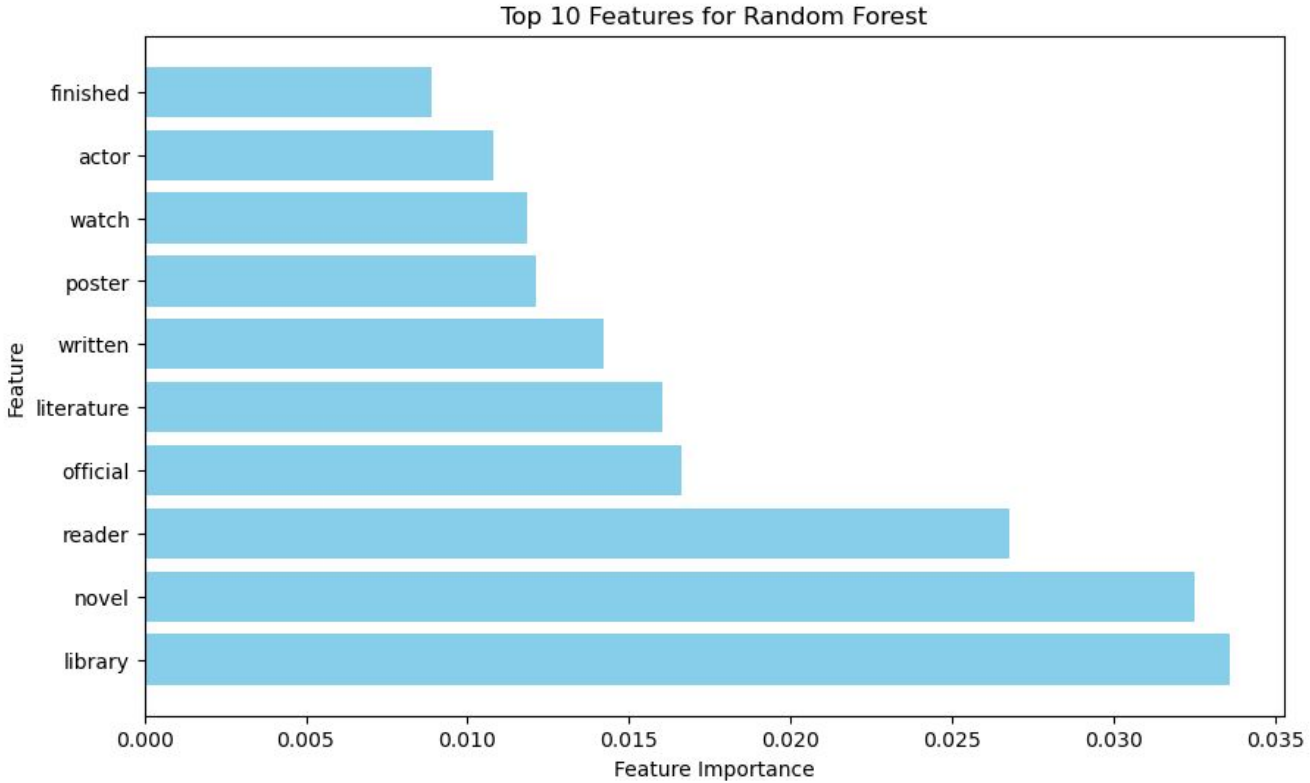


SVC

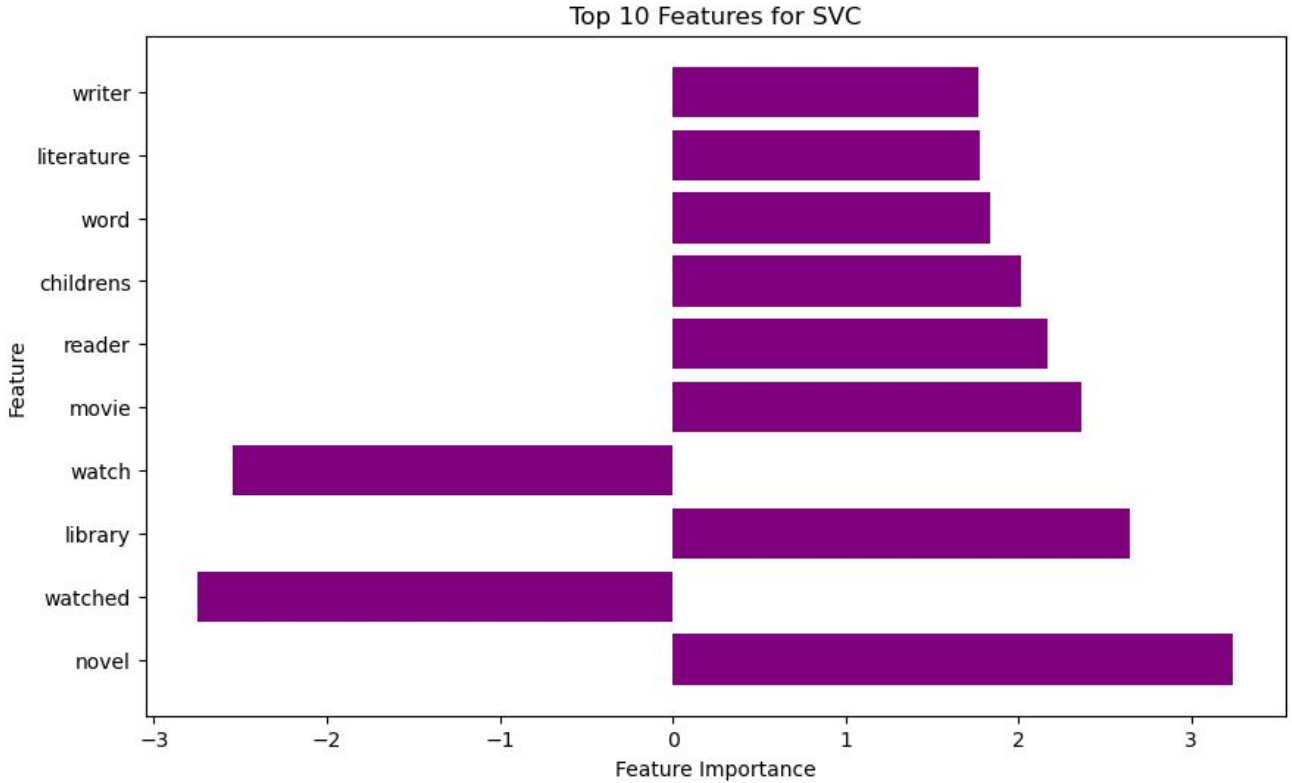
```
Best Accuracy: 0.8762987012987011
Test Accuracy: 0.8858921161825726
```



Feature Importance



Feature Importance



Thank You