PROJECT 3:
WEB APIS & CLASSIFICATION

BY
RICHA MARWAH

CONTENT

- Background
- Problem Statement
- Datasets
- Data Cleaning
- Text Cleaning
- EDA
- Preprocessing Data and Model Prep
- Model 1: Naive Bayes
- Model 2: Logistic Regression
- Conclusion

BACKGROUND

InfluencerX is a media organization that represents entertainment influencers. Its employees and influencers contribute actively to various entertainment-related subreddits, including ones for Netflix and Amazon Prime Video.

To streamline content posting, InfluencerX is looking for a model that can help label the posts as belonging to either the Netflix subreddit or the Prime Video subreddit.

PROBLEM STATEMENT

To help InfluencerX create a classification model that can accurately label which subreddit a given post belongs to, in order to improve the process of content posting.

DATASETS

- Posts from the Netflix subreddit (collected 625 posts)
- Posts from the Amazon Prime Video subreddit (collected 613 posts)
- For the purpose of the project, we only used the text columns (title, selftext) and the target (subreddit).

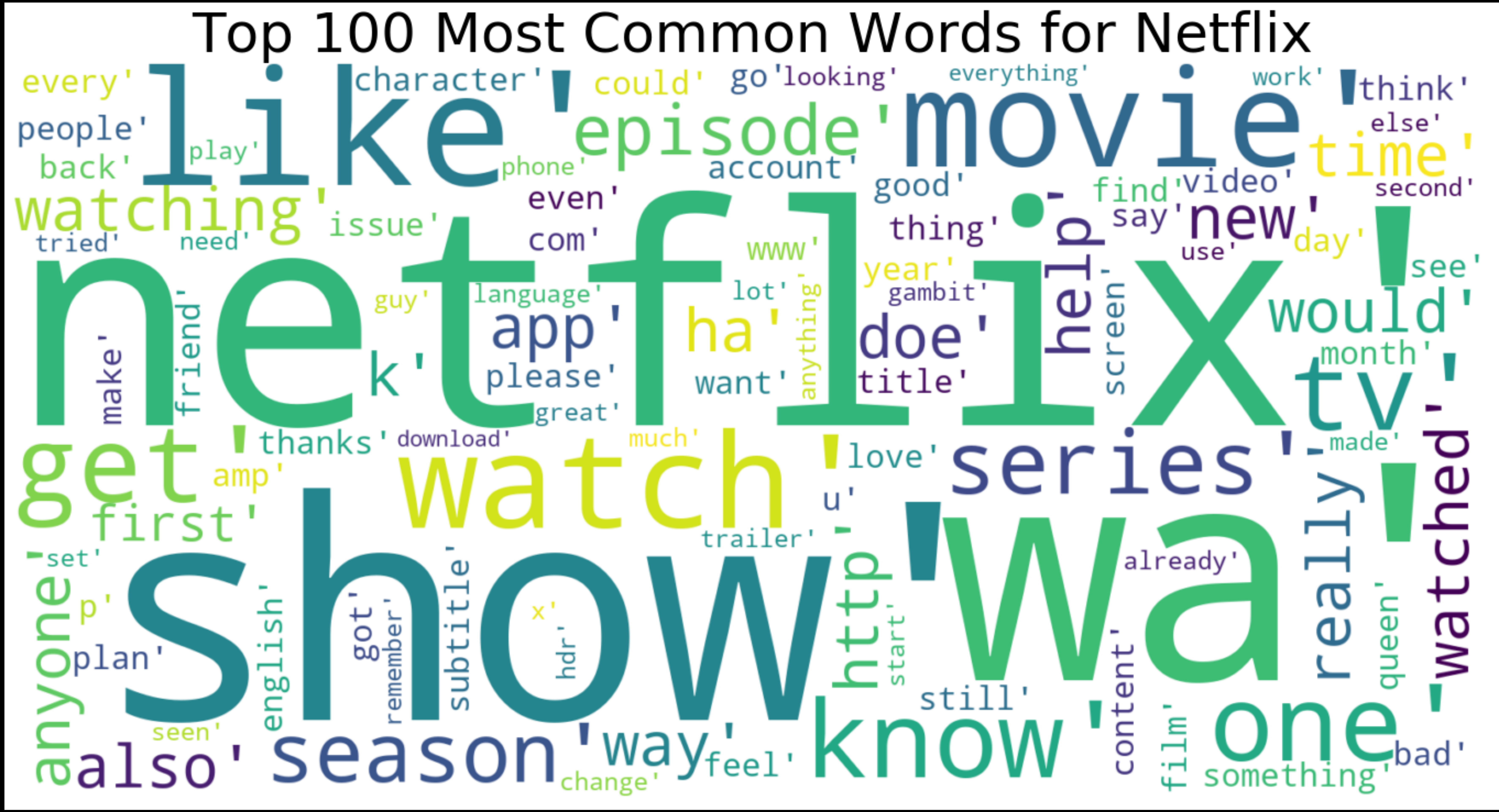
DATA CLEANING

- Filling in missing values
- Combining columns
- Dropping duplicates
- Balancing Classes

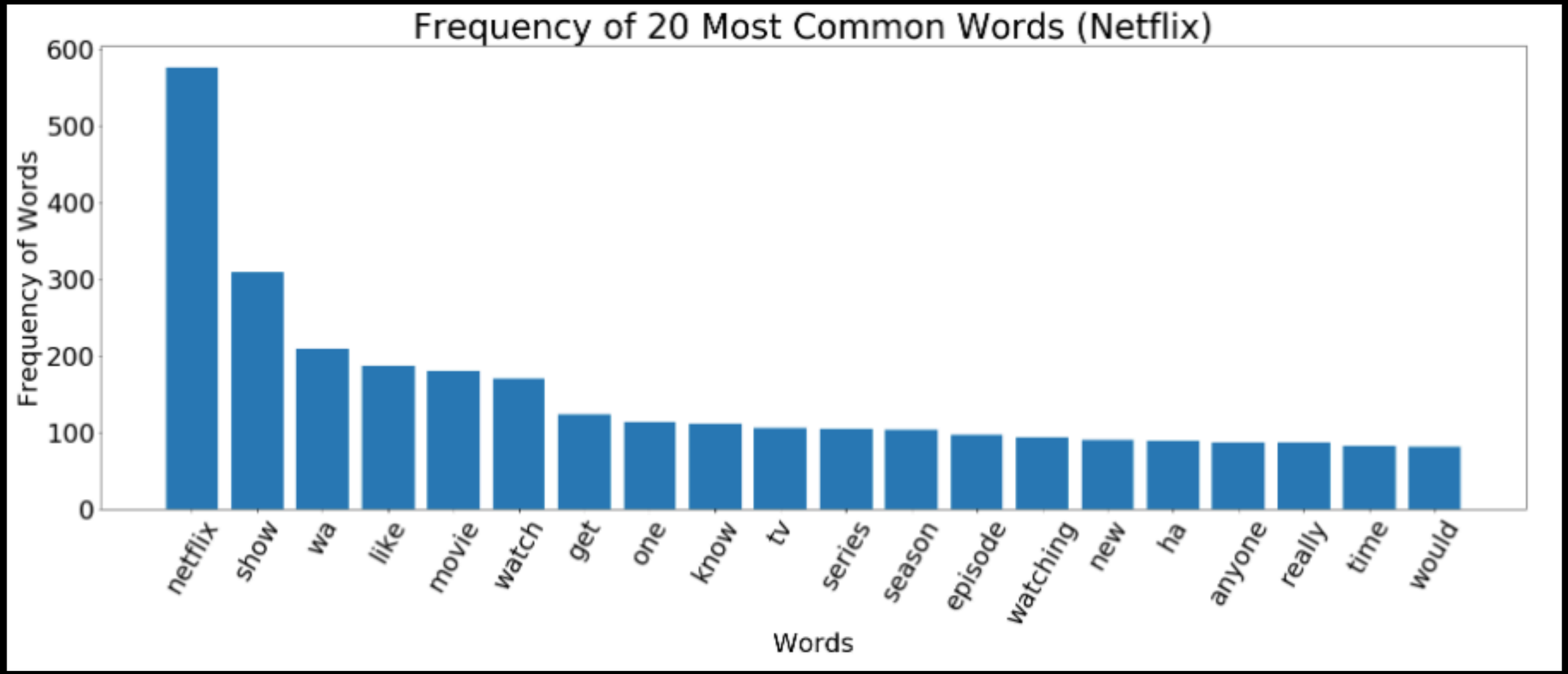
TEXT CLEANING

- Removing non-letters with regex
- Lemmatizing
- Removing stop words

EDA (NETFLIX)

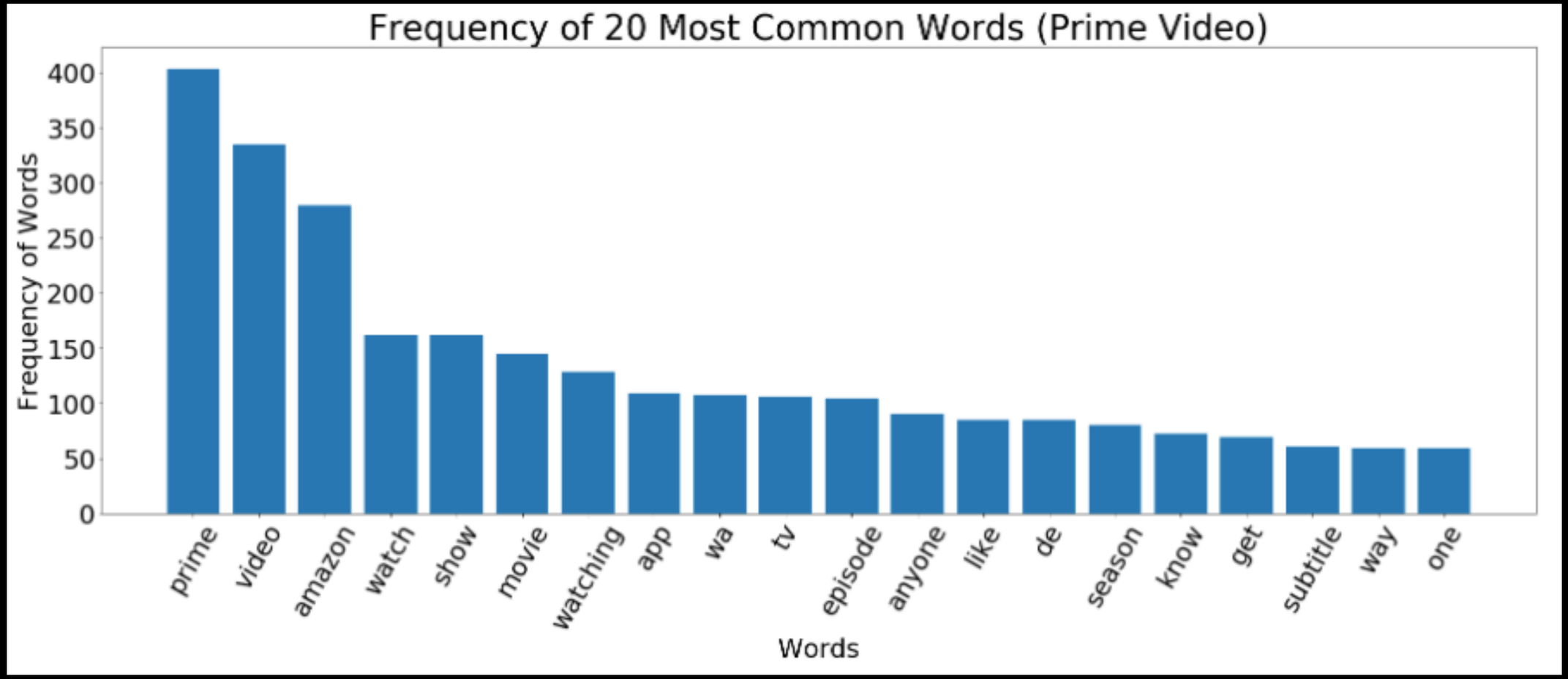

Top 100 Most Common Words for Netflix
every'
like'
character'
could'
go'
looking'
movie'
everything'
work'
think'
people'
play'
episode'
else'
back'
phone'
account'
time'
watching'
even'
good'
find'
video'
second'
issue'
say'
new'
tried'
need'
com'
thing'
day'
www'
help'
use'
year'
see'
netflix'
guy'
language'
lot'
anything'
gambit'
app'
ha'
doe'
would'
screen'
month'
k'
please'
want'
title'
tv'
make'
friend'
great'
thanks'
download'
much'
made'
get'
watch'
love'
series'
watched'
amp'
u'
first'
really'
trailer'
set'
already'
anyone'
http'
show'
wa'
p'
x'
queen'
plan'
english'
got'
remember'
subtitle'
hdr'
start'
content'
still'
one'
seen'
season'
way'
feel'
know'
film'
bad'
also'
change'
something'

EDA (NETFLIX)


Frequency of 20 Most Common Words (Netflix)
Frequency of Words
600
500
400
300
200
100
0
netflix
show
wa
like
movie
watch
get
one
know
tv
series
season
episode
watching
new
ha
anyone
really
time
would
Words

EDA (PRIME VIDEO)

EDA (PRIME VIDEO)



PREPROCESSING DATA AND MODEL PREP

- Combining dataframes
- Transforming text to structured data
  - CountVectorizer
  - TfidfVectorizer

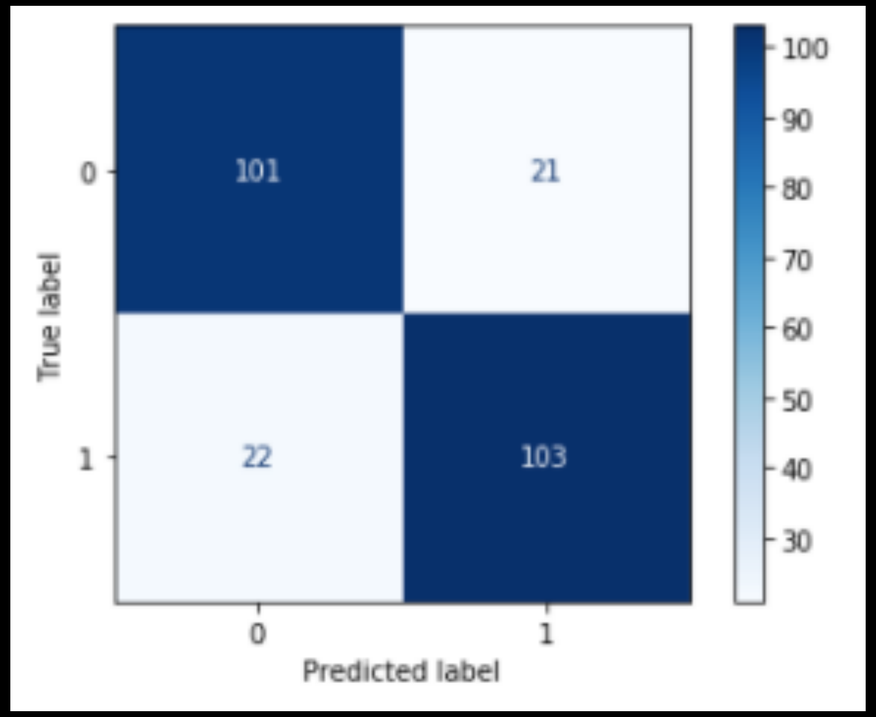
MODEL1: NAIVE BAYES

Training Score: 0.97

Testing Score: 0.83

While the model accuracy was a lot higher than the baseline accuracy, it seemed to perform better on the training dataset, suggesting some overfitting.

MODEL1: NAIVE BAYES



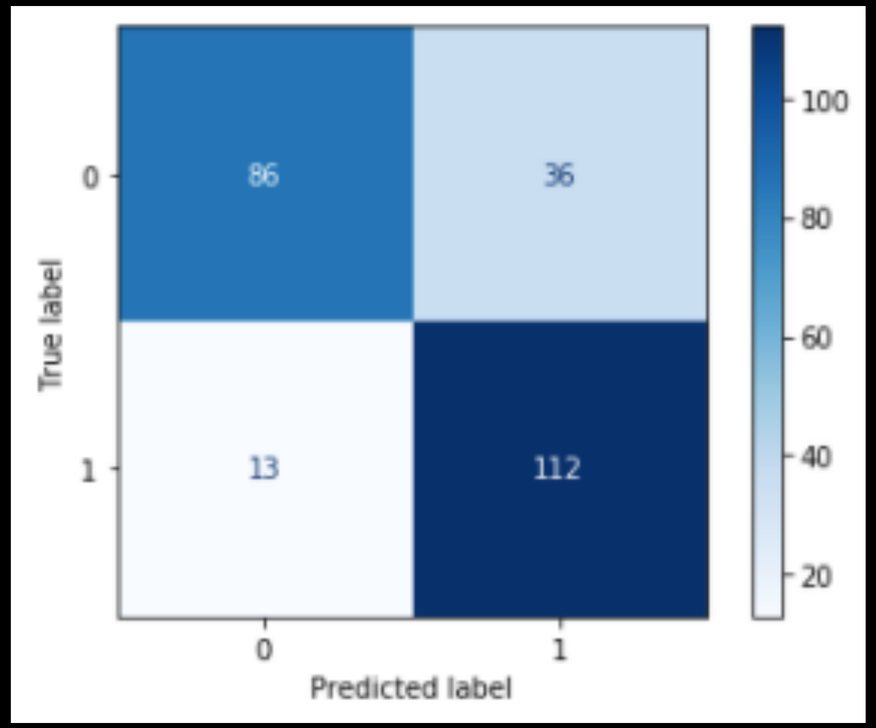
MODEL2: LOGISTIC REGRESSION

Training Score: 0.96

Testing Score: 0.80

Similar to the Naive Bayes model, while the accuracy of the Linear Regression model was a lot higher than the baseline accuracy, the model performed better on the training dataset.

MODEL2: LOGISTIC REGRESSION



MODEL EVALUATION

While the test accuracy of the Linear Regression model and the Naive Bayes model is only marginally different, the Linear Regression model seems to be more biased towards the positive class.

Overall, the Naive Bayes model is a more robust model as it is not biased towards either class and has a higher likelihood of labelling the post correctly.

CONCLUSION

By deploying the Naive Bayes classifier model, InfluencerX can label a post as belonging to either the Netflix subreddit or Prime subreddit with 82.59% accuracy.

As the model is not a 100% accurate at this point, there will likely still be a need for some human intervention before posting.

However, the model can help streamline the content posting process and reduce man-hours spent on the process.

CONCLUSION

To further improve accuracy, future projects can evaluate other models, should the project scope allow it, and even look at ensembling, which combines predictions from multiple separate models.

In the absence of resources for developing additional models, existing models can be fine-tuned by training models with more data, if available, so that the model may get better at filtering out noise that does not aid in classification.

THANK YOU