Pokemon vs digimon

Mak Wei Yip
DSIF-SG-9


VS
Pokémon
DIGITAL
DIGIMON
MONSTERS

Problem statement

We want to use word input by users to help recommend whether pokemon or digimon is more suitable for a user. This is to recommend the right merchandise to the user.

Project Goal:
Classification of comments from two subreddits

Process:

1. Data collection from pokemon and digimon subreddit and data cleaning
2. EDA
3. Models
4. Evaluation
5. Conclusions

Data Collection & Data cleaning

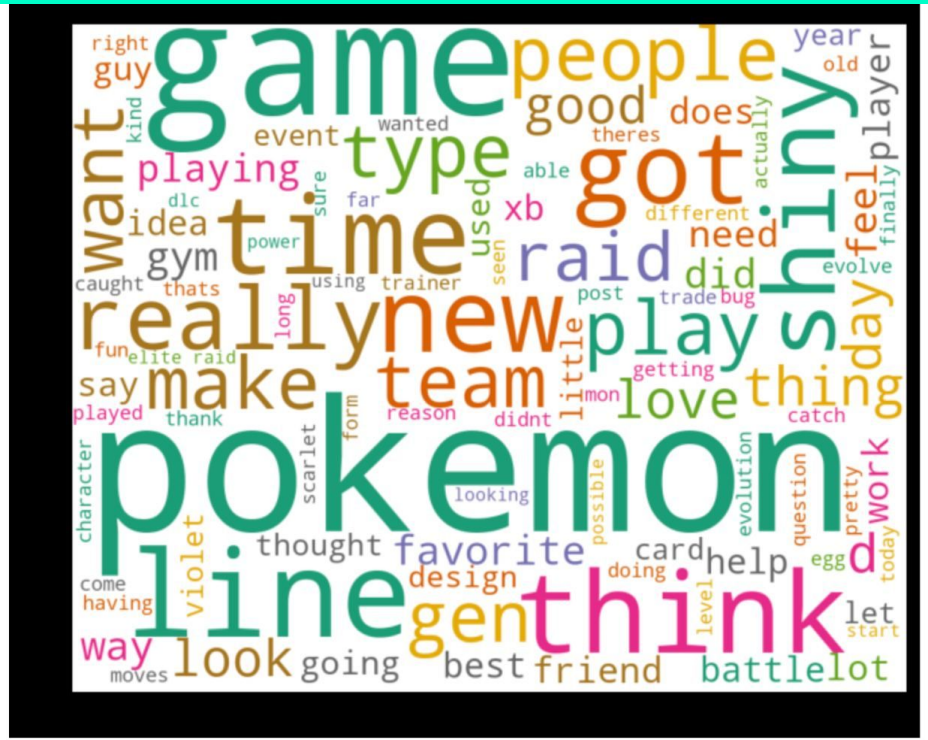
Reddit API:

1. Limited to 1000 post per subreddit
2. Downloaded title and body text from both pokemon and digimon
3. Need to do cleaning of text html, tags,etc more may required
4. Save into a csv file for modeling

Data Cleaning:

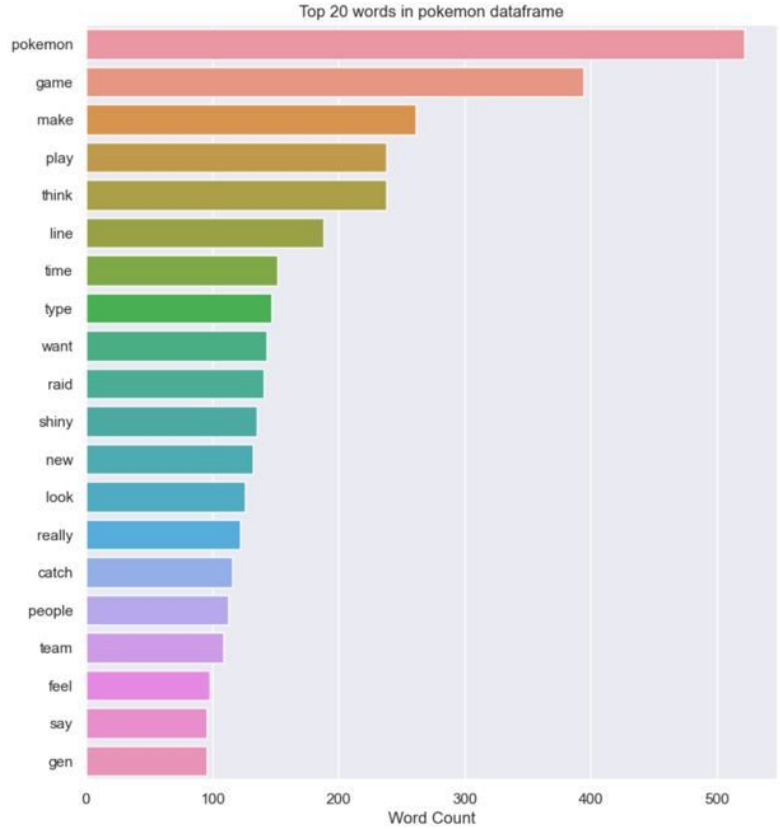
1. Remove html, hyperlinks, punctuation, word with 2 or fewer letters, whitespace and number

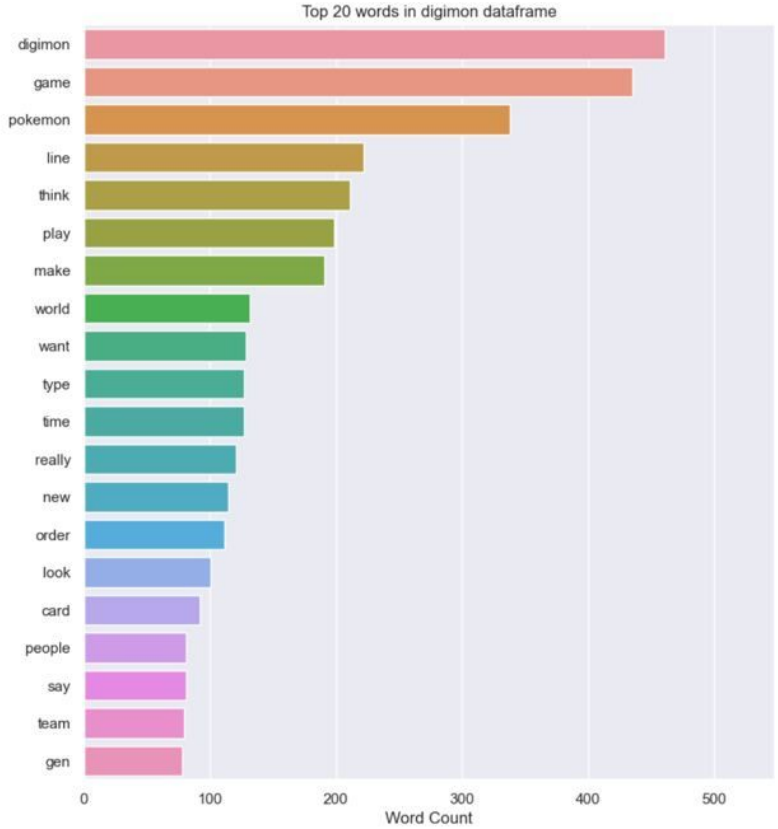
POKEMON



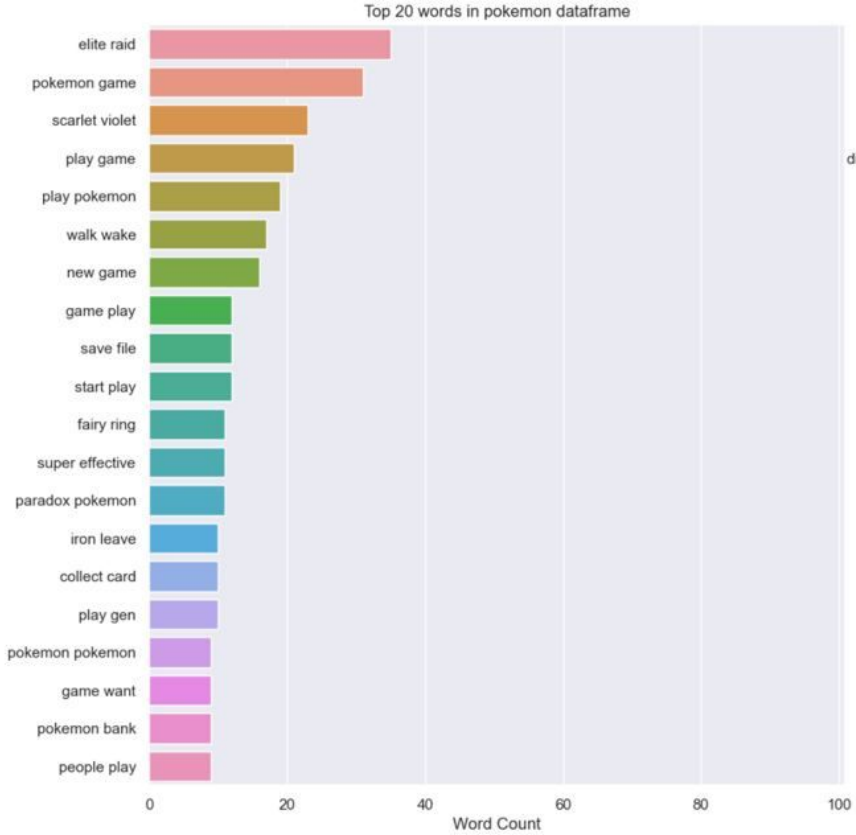
DIGIMON

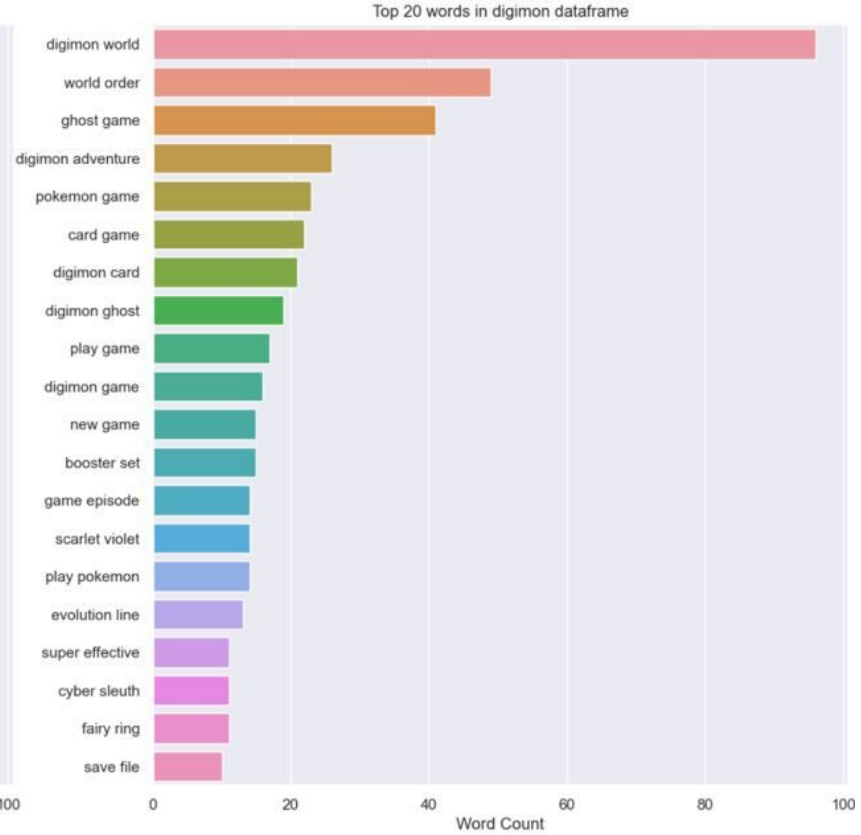
EDA on words


Top 20 words in pokemon dataframe
pokemon
game
make
play
think
line
time
type
want
raid
shiny
new
look
really
catch
people
team
feel
say
gen
0
100
200
300
400
500
Word Count


Top 20 words in digimon dataframe
digimon
game
pokemon
line
think
play
make
world
want
type
time
really
new
order
look
card
people
say
team
gen
0
100
200
300
400
500
Word Count

EDA on bi-gram





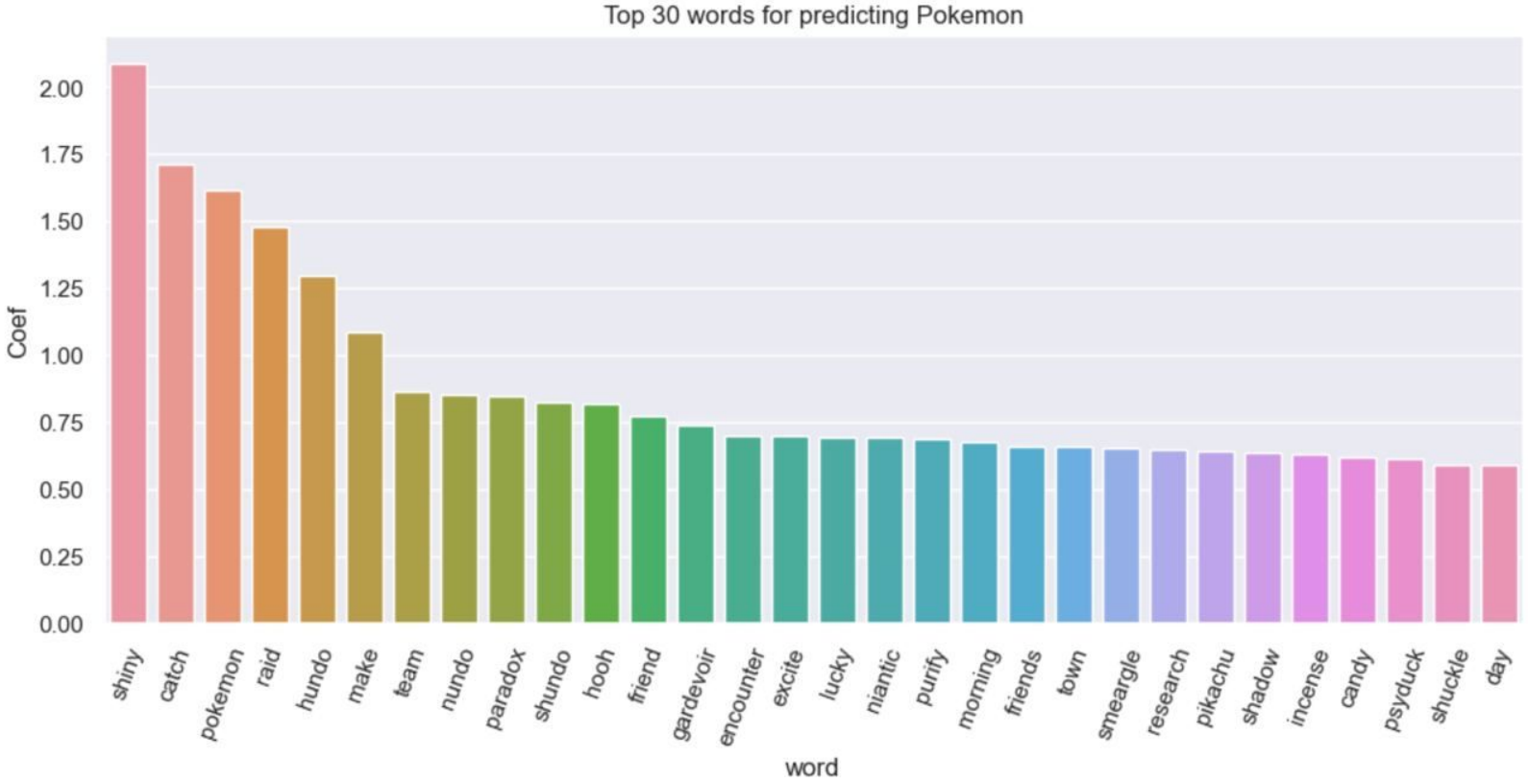
MODELS

Models:

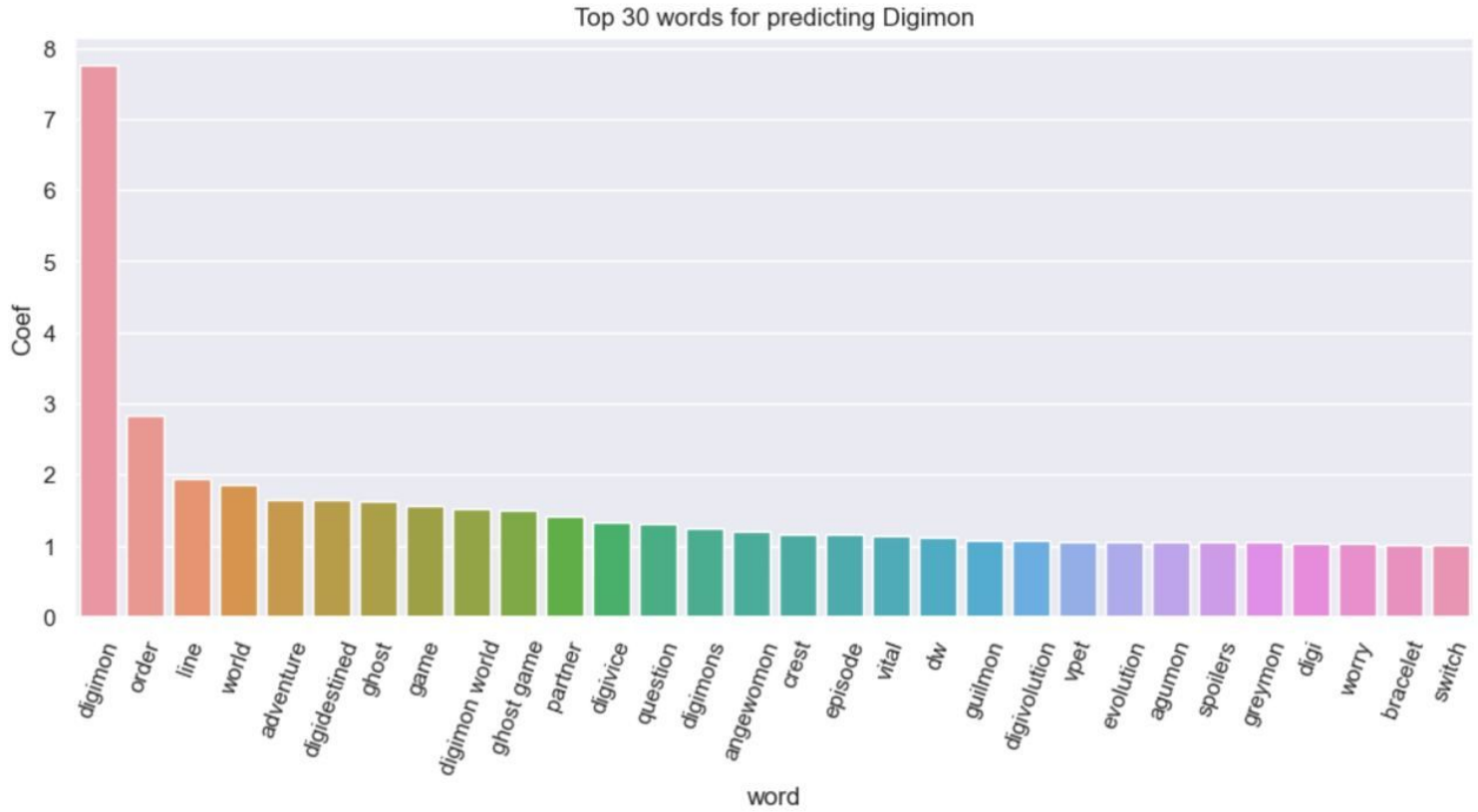
1. Baseline model will be 0.5 accuracy as there is only 50% chance this model will predict a text to be from pokemon
2. Model aim to get the best accuracy score
3. Use different model with gridsearch to get the best parameter.

models	Accuracy score(test)
Multinomial Naive Bayes	0.78
Logistic Regression	0.78
SVM	0.76
RandomForest	0.75

coefficient Logistic Regression



coefficient Logistic Regression

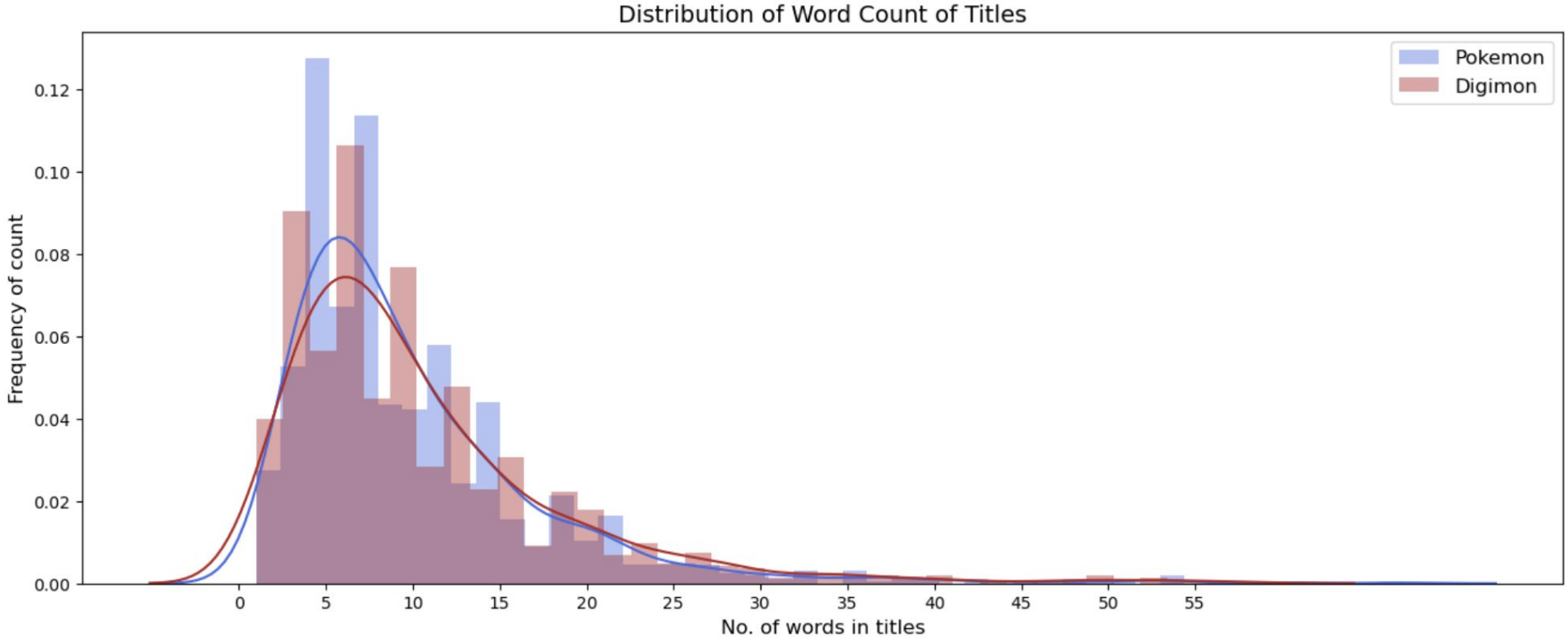


Conclusion

models	Train score	Test score	Recall score(test)	Precision score(test)	F1 score (test)
Naive Bayes	0.82	0.78	0.95	0.74	0.83
Logistic Regression	0.84	0.78	0.93	0.74	0.83
SVM	0.89	0.76	0.89	0.75	0.81
RandomForest	0.98	0.75	0.83	0.76	0.79

We will be using Naive Bayes for our model deployment as it has the highest accuracy and recall.

Thank you


Distribution of Word Count of Titles
Pokemon
Digimon
Frequency of count
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
40
45
50
55
No. of words in titles