Sentiment Analysis of Video Games on Twitter

Ryan Mitchell

The Dataset

- Each entry is a tweet categorized by its sentiment positive, negative, or neutral/irrelevant.
- Includes 12447 unique entries.
- Dataset was compiled in 2022 into a CSV file and retrieved from Kaggle.
- Goal: analyze how twitter feels about some of the most popular video games.

Categories

- The data contains 32 unique categories mostly pertaining to videogames or companies in that realm.
- Contains some of the most popular video games (LoL, WoW, CS-GO,ect.)

```
Unique Games Mentioned in the Dataset:
1. Borderlands
2. CallOfDutyBlackopsColdWar
3. Amazon
4. Overwatch
5. Xbox(Xseries)
6. NBA2K
7. Dota2
8. PlayStation5(PS5)
9. WorldOfCraft
10. CS-GO
11. Google
12. AssassinsCreed
13. ApexLegends
14. LeagueOfLegends
15. Fortnite
16. Microsoft
17. Hearthstone
18. Battlefield
19. PlayerUnknownsBattlegrounds(PUBG)
20. Verizon
21. HomeDepot
22. FIFA
23. RedDeadRedemption(RDR)
24. CallOfDuty
25. TomClancysRainbowSix
26. Facebook
27. GrandTheftAuto(GTA)
28. MaddenNFL
29. johnson&johnson
30. Cyberpunk2077
31. TomClancysGhostRecon
32. Nvidia
```

Preprocessing

- Data is cleaned by removing mentions, URL's, Punctuation, and whitespaces.
- Then using the NLTK tool stopwords are removed from the text.
- The data that is left is then put in a list where each word is an index and can be analyzed.

Words

Pre Cleaning

```
Total Words: 271903
Unique Words: 23941
Average Words per Tweet: 21.84
Total Characters: 1339723
Average Characters per Tweet: 107.63
```

Post Cleaning

```
Total Words: 126450
Unique Words: 19652
Average Words per Tweet: 10.16
Total Characters: 1167354
Average Characters per Tweet: 93.79
```

Twitter contains a lot of weird text not considered parts of speech such as (@) mentions and links that need to be cleaned resulting in a fair percentage of words from the tweets not being analyzed.

Sentiment

- Total positive tweets: 3472
- Total negative tweets: 3757
- Total neutral tweets: 5218
- Each category contains roughly 400 tweets including neutral tweets.

Sentiment Game	Positive	Negative	Total	Positive %
AssassinsCreed	241	63	304	79.28
RedDeadRedemption(RDR)	155	51	206	75.24
Cyberpunk2077	161	65	226	71.24
Borderlands	170	71	241	70.54
CS-GO	128	58	186	68.82
WorldOfCraft	123	57	180	68.33
Xbox(Xseries)	132	63	195	67.69
PlayStation5(PS5)	157	76	233	67.38
Hearthstone	139	88	227	61.23
Nvidia	136	87	223	60.99
CallOfDutyBlackopsColdWar	144	96	240	60.00
Battlefield	99	79	178	55.62
Overwatch	122	105	227	53.74
ApexLegends	107	100	207	51.69
GrandTheftAuto(GTA)	104	99	203	51.23
LeagueOfLegends	103	107	210	49.05
HomeDepot	130	150	280	46.43
Fortnite	94	117	211	44.55
Microsoft	101	129	230	43.91
Dota2	97	128	225	43.11
TomClancysGhostRecon	103	150	253	40.71
Google	60	99	159	37.74
PlayerUnknownsBattlegrounds(PUBG)	68	116	184	36.96
Amazon	52	96	148	35.14
CallOfDuty	75	149	224	33.48
Verizon	88	183	271	32.47
TomClancysRainbowSix	88	187	275	32.00
FIFA	84	196	280	30.00
johnson&johnson	45	141	186	24.19
NBA2K	71	246	317	22.40
Facebook	29	120	149	19.46
MaddenNFL	66	285	351	18.80

Positive Tweet Lexical Analysis

- A lot of positive associations like love, like, and best.
- According to the data most positive tweets contain a lot of the same words and don't overlap much with negative tweets

Top 20 Words by Lexical Dispersion in Positive Tweets:

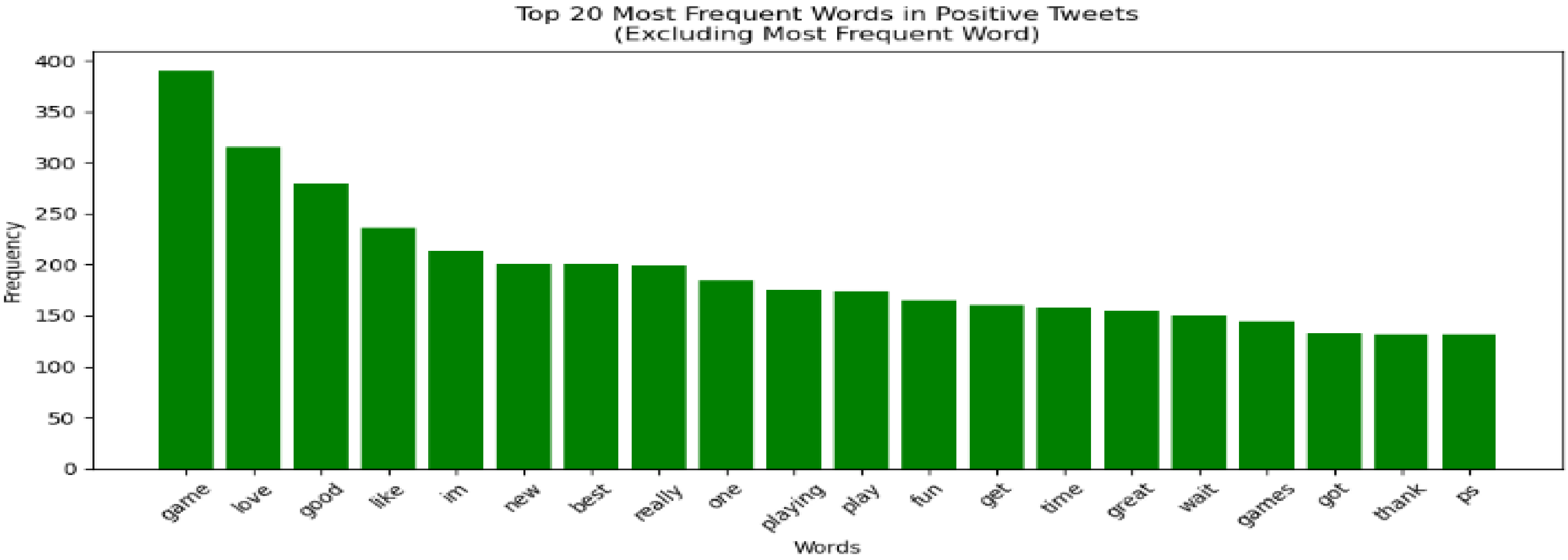
Word	Dispersion	Frequency
game	334	390
love	290	315
good	266	280
like	213	236
im	195	213
new	185	200
best	190	200
really	187	199
one	173	184
playing	166	175
play	167	174
fun	159	165
get	150	160
time	153	158
great	148	154
wait	143	150
games	133	144
got	130	133
thank	129	132
ps	115	131

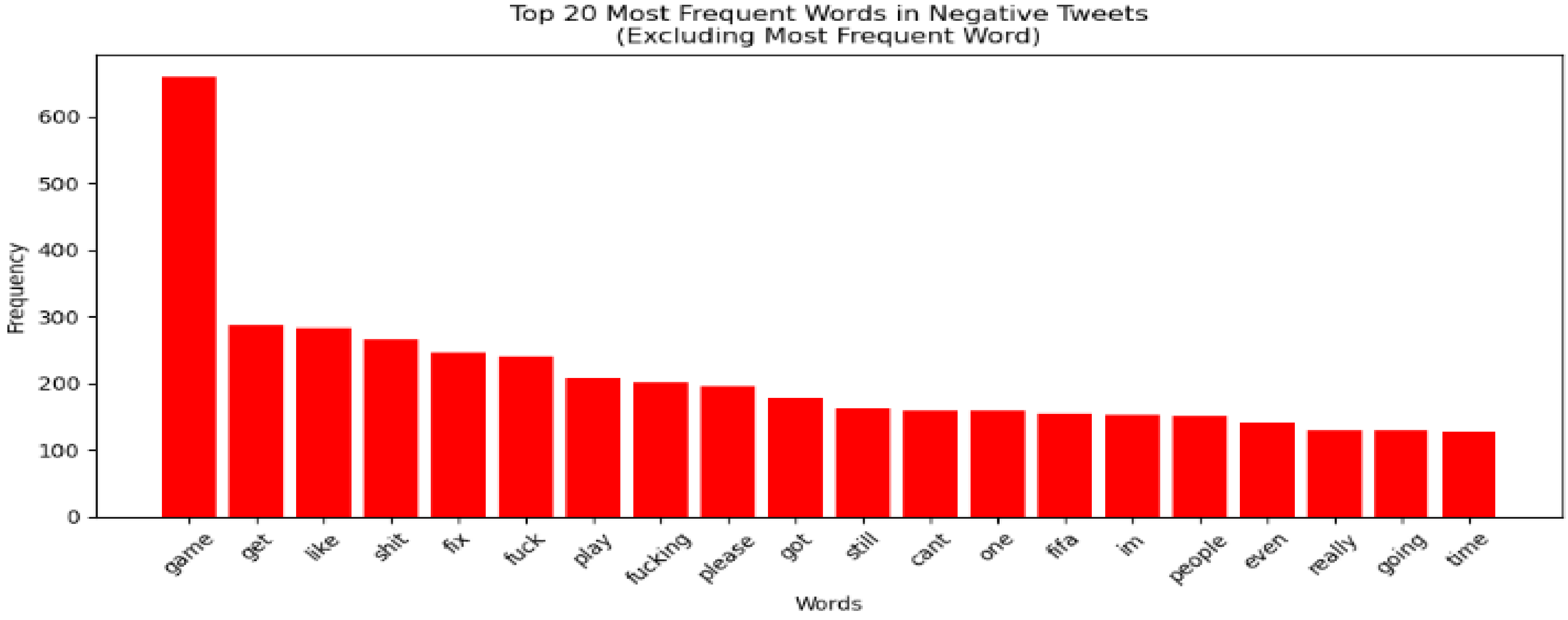
Negative Tweet Lexical Analysis

- A lot of curse words.
- FIFA which was the one of the most categories associated with negative sentiment.

Top 20 Words by Lexical Dispersion in Negative Tweets:

Word	Dispersion	Frequency
game	549	660
get	260	286
like	266	282
shit	250	265
fix	238	246
fuck	213	240
play	189	207
fucking	177	201
please	175	195
got	160	178
still	159	162
cant	151	159
one	153	158
fifa	144	155
im	141	152
people	144	150
even	139	142
going	124	129
really	117	129
time	119	127


Top 20 Most Frequent Words in Positive Tweets
(Excluding Most Frequent Word)
Frequency
0
50
100
150
200
250
300
350
400
game
love
good
like
im
new
best
really
one
playing
play
fun
get
time
great
wait
games
got
thank
ps
Words


Top 20 Most Frequent Words in Negative Tweets
(Excluding Most Frequent Word)
Frequency
0
100
200
300
400
500
600
game
get
like
shit
fix
fuck
play
fucking
please
got
still
cant
one
fifa
im
people
even
really
going
time
Words

Model Creation

- Goal: Predict the sentiment of tweets.
- Tool Native Bayes Classifier from NLTK.
  - Split used 80% training 20% testing.
- I only used positive negative tweets to make the data less ambiguous.
- This greatly reduced the amount of training data I had to train the model with.

Results

- Accuracy 79-82% which was very good most likely due to the lack of overlapping keywords between positive and negative tweets.

Classification Report:

Label	Precision	Recall	F1-Score
Positive	81.63	74.59	77.95
Negative	79.42	85.38	82.29

Average Processing Time per Tweet: 0.000030 seconds

Analysis

- The model is correct more often when predicting positive tweets shown by the higher precision score
- The model is more likely to miss label negative tweets and is overall more likely to predict a negative tweet.
- This is in-part due to the data containing more negativet weets than positive ones

Most Informative features

```
Most Informative Features
                     fix = True           Negati : Positi =     27.6 : 1.0
                 awesome = True           Positi : Negati =     24.4 : 1.0
                 excited = True           Positi : Negati =     22.4 : 1.0
               beautiful = True           Positi : Negati =     17.4 : 1.0
                 trailer = True           Positi : Negati =     17.2 : 1.0
                    dope = True           Positi : Negati =     15.0 : 1.0
                bullshit = True           Negati : Positi =     14.9 : 1.0
                  loving = True           Positi : Negati =     13.6 : 1.0
                valhalla = True           Positi : Negati =     13.6 : 1.0
                   sucks = True           Negati : Positi =     12.5 : 1.0
                 loading = True           Negati : Positi =     11.8 : 1.0
                 account = True           Negati : Positi =     11.6 : 1.0
              appreciate = True           Positi : Negati =     11.4 : 1.0
                birthday = True           Positi : Negati =     11.4 : 1.0
                    quit = True           Negati : Positi =     11.2 : 1.0
                  shitty = True           Negati : Positi =     11.2 : 1.0
                   toxic = True           Negati : Positi =     11.2 : 1.0
                  stupid = True           Negati : Positi =     11.0 : 1.0
                   worst = True           Negati : Positi =     10.6 : 1.0
                   error = True           Negati : Positi =     10.4 : 1.0
```

Most Informative POS

- According to the data, adjectives give the most insight to which sentiment the tweet will fall under.
- Awesome and beautiful are high indicators of a positive tweet.
- Toxic, worst, or stupid indicate a negative tweet.
- Verbs seem to indicate a negative tweet more often than not fix, sucks, quit, and loading all indicate a negative tweet.

Conclusion

- The data was imbalanced having roughly 400 more negative entries vs positive one.
- The neutral data ended up being junk at least for my goal and would decrease the accuracy of the model when used.
- The most informative POS that indicate sentiment at least pertaining to twitter are adjectives.