

Twitter Sentiment Analysis for Bungie and Destiny 2

Sebastian Alvis

SpringBoard Data Science Bootcamp

Overview

- What's Bungie / Destiny 2?
- Getting Data From Twitter
- Exploratory Data Analysis
- Machine Learning Analysis
- Conclusions

Overview

- What's Bungie / Destiny 2?
- Getting Data From Twitter
- Exploratory Data Analysis
- Machine Learning Analysis
- Conclusions

What's Bungie / Destiny 2?

- Bungie
 - Game Company
 - They're in Bellevue
- Destiny 2
 - Hit game by Bungie
 - Magic space ninjas with guns



Social Media Presence

- Official Twitter Accounts
 - @Bungie
 - @DestinyTheGame
- Lots of tweets
 - Announcements
 - Etc
- Use Twitter data
 - What do players think of recent updates to Destiny 2?
 - What do people think of Bungie?

Overview

- What's Bungie / Destiny 2?
- **Getting Data From Twitter**
- Exploratory Data Analysis
- Machine Learning Analysis
- Conclusions

Data Acquisition

- Twitter Standard Search API
 - Get tweets from the last 7 – 10 days
 - Not exhaustive (\$\$\$)
 - Search terms
- Search for 4 types of tweets
 - Tweets about Destiny 2 (21.4k)
 - Tweets about Bungie (24.7k)
 - @DestinyTheGame's tweets (46)
 - @Bungie's tweets (107)

Data Cleaning

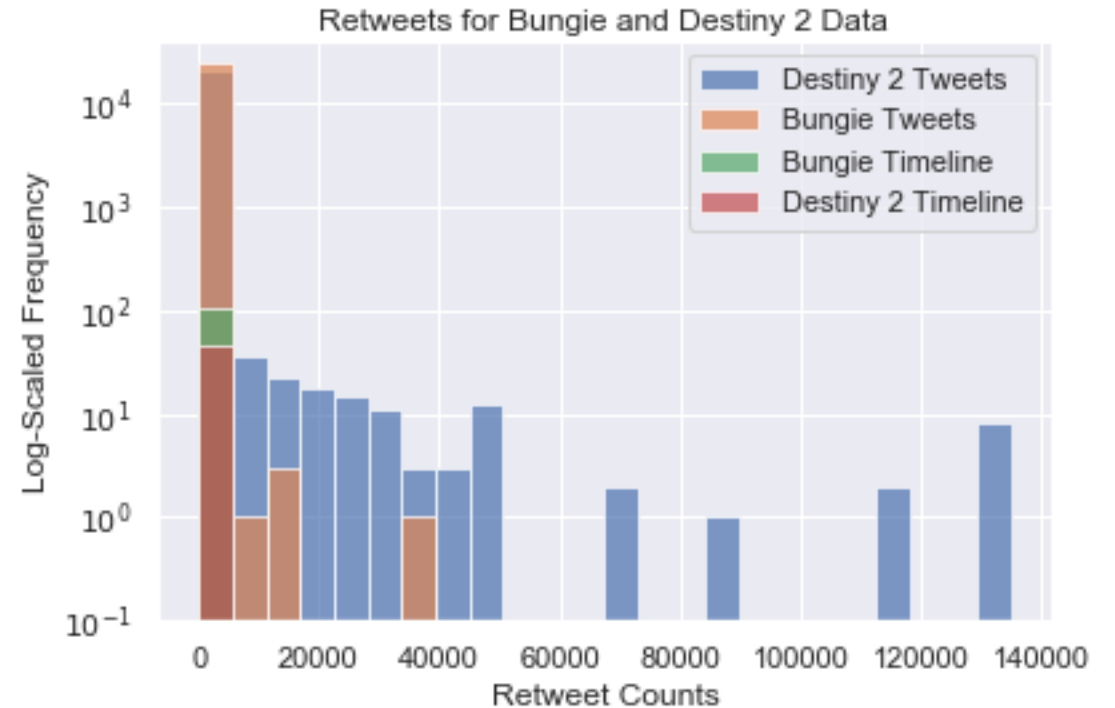
- Missing / misaligned data
- Reduced columns
 - 320 to 50
- Kept the data as 4 DataFrames
 - Concatenate them when appropriate
- Datetime columns

Overview

- What's Bungie / Destiny 2?
- Getting Data From Twitter
- **Exploratory Data Analysis**
- Machine Learning Analysis
- Conclusions

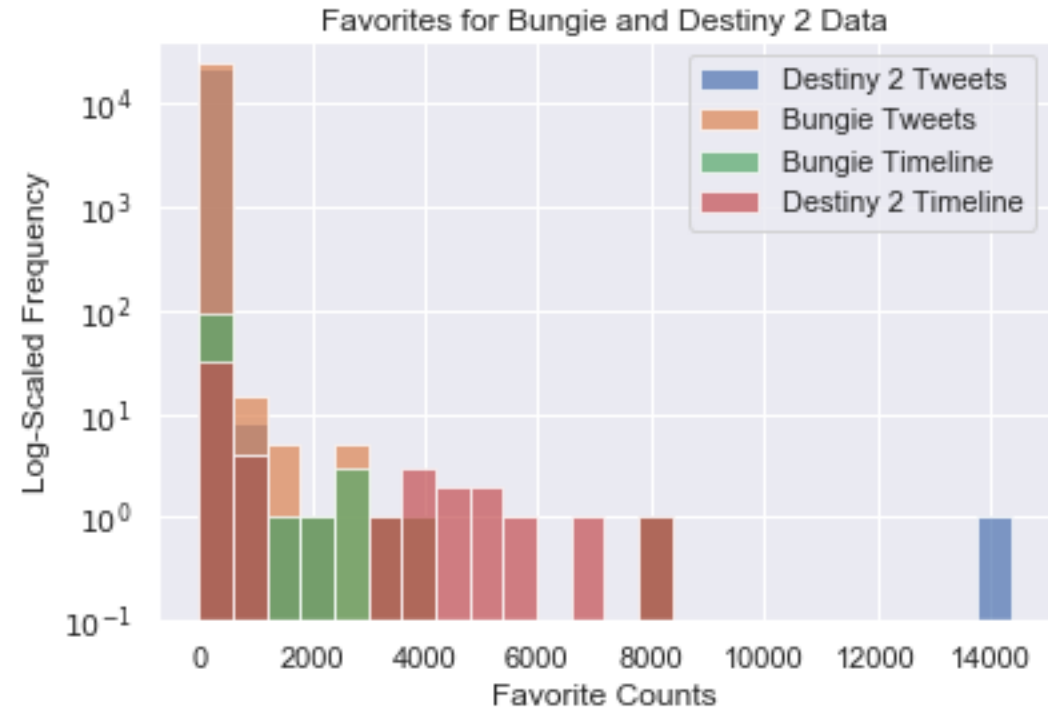
How Do I Measure Engagement?

- Retweets
 - Official accounts don't get many compared to other accounts



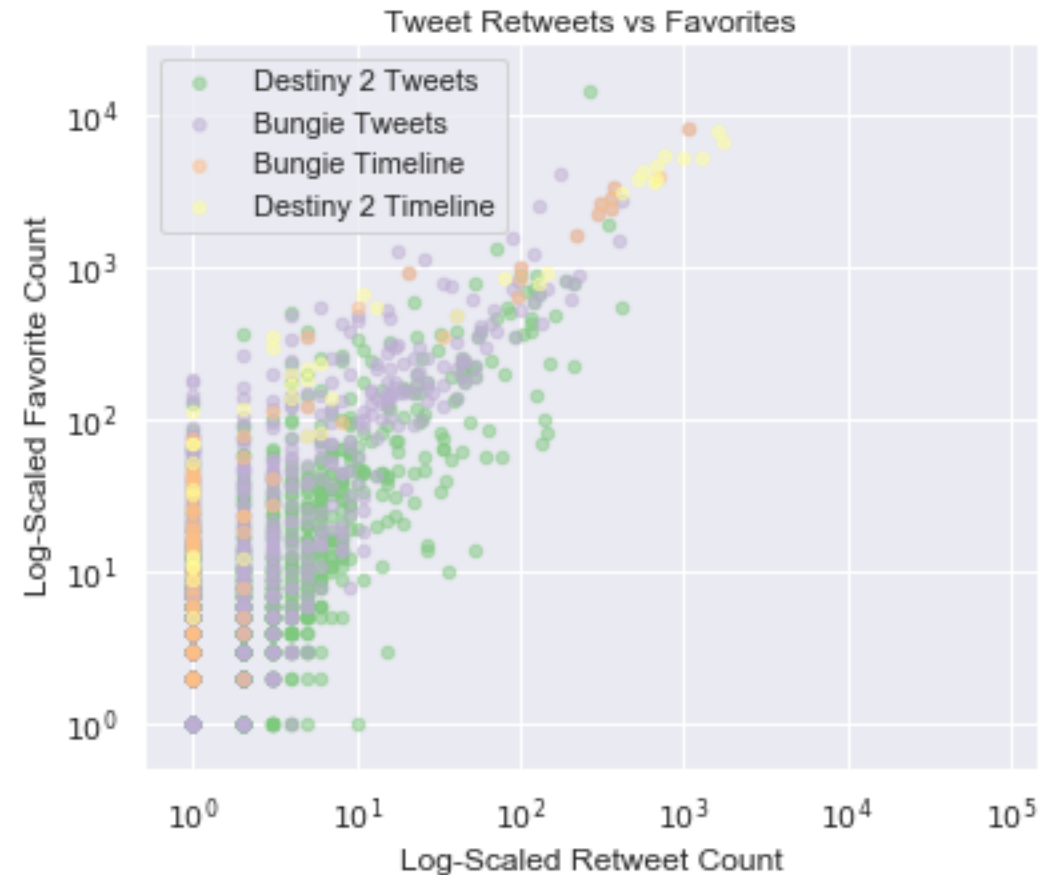
How Do I Measure Engagement?

- Retweets
 - Official accounts don't get many compared to other accounts
- Favorites
 - Official accounts do get at least as many as other accounts



Do Favorites and Retweets Correlate?

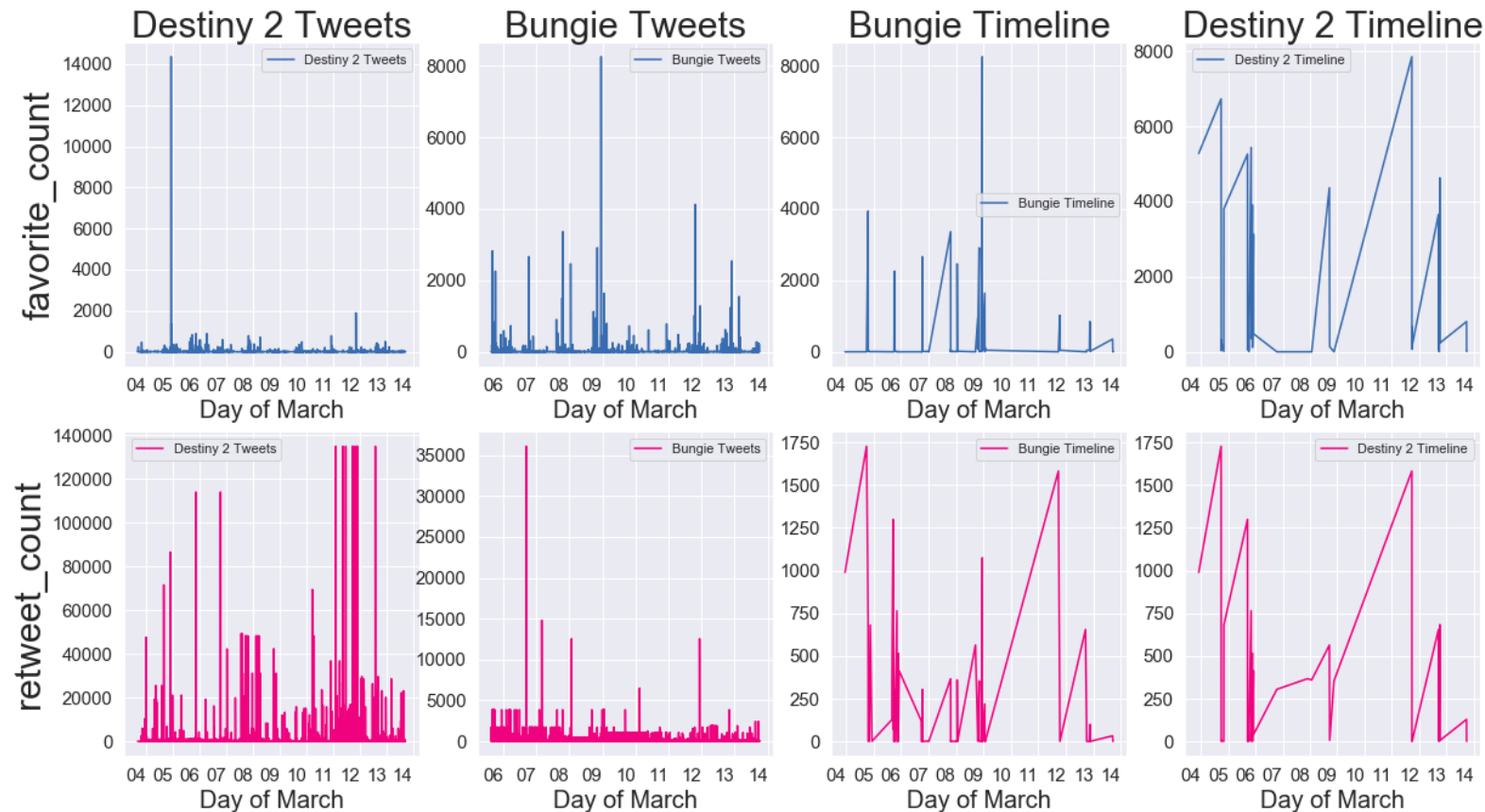
- Only if you log scale them both
 - Darn outliers
- Generally more favorites than retweets!
 - The outliers ruin the visuals



Time Correlations?

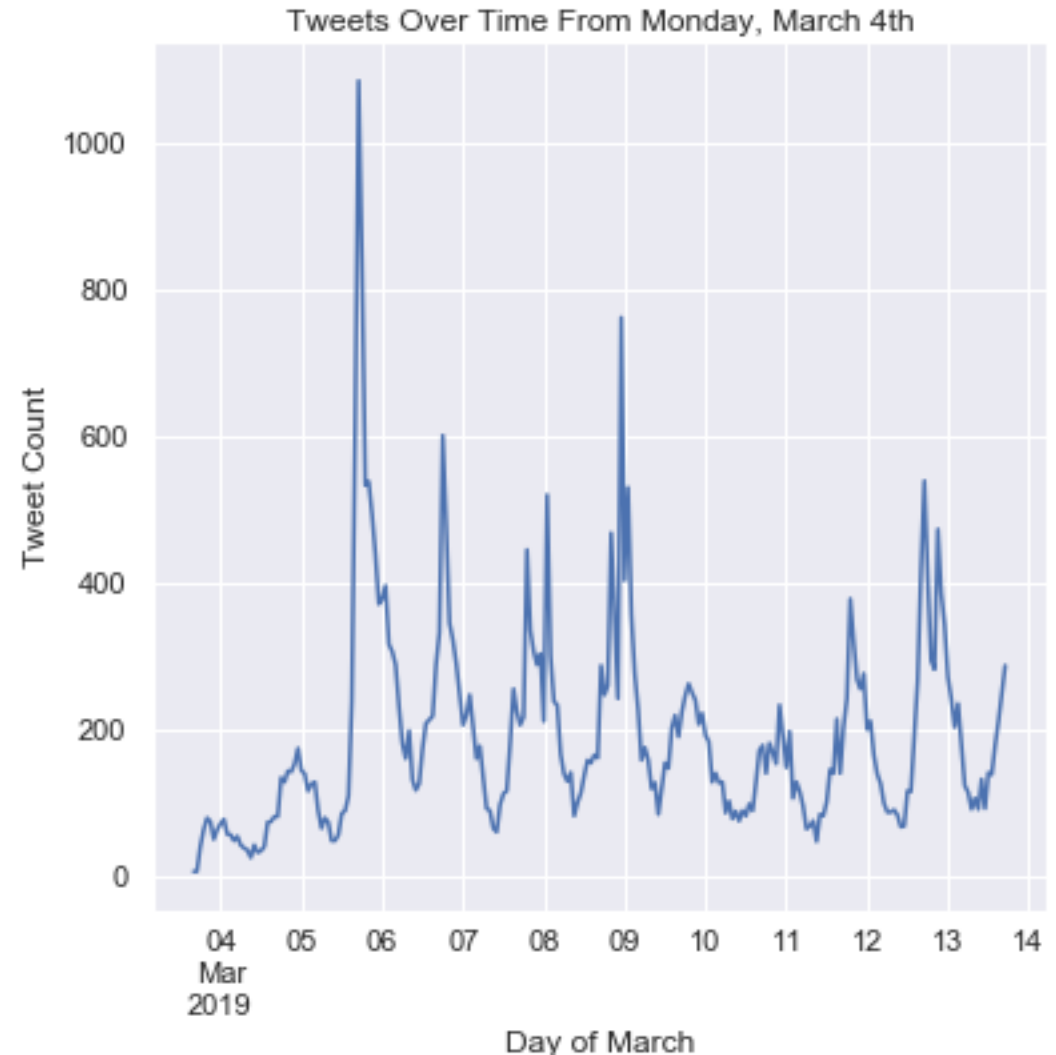
- Only on the official accounts

Engagement Metrics Across All Datasets



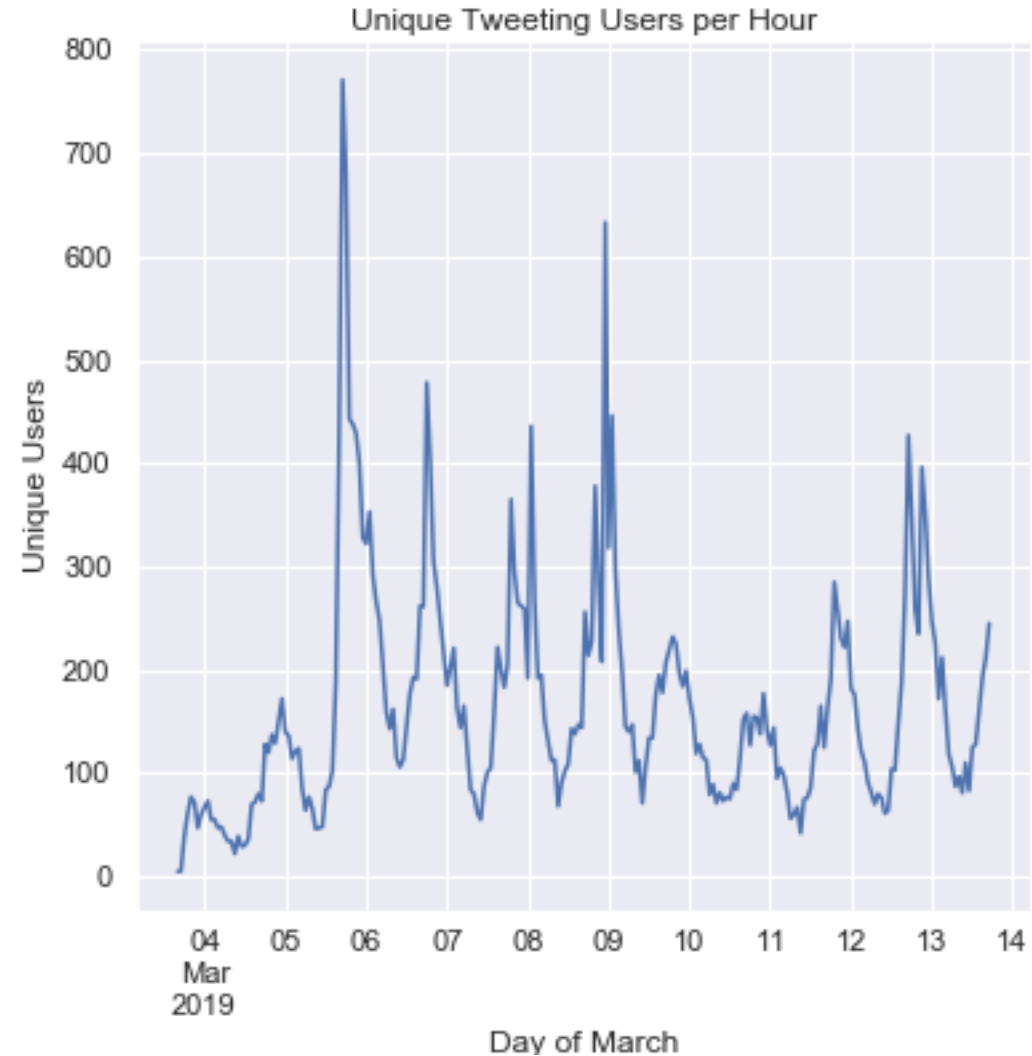
Speaking of Time...

- When are these tweets posted?
 - March 3rd – 13th
- Note:
 - The Destiny 2 tweets start on the 3rd
 - Bungie tweets start on the 5th
- Tweets posted in the evening, mostly on weekdays



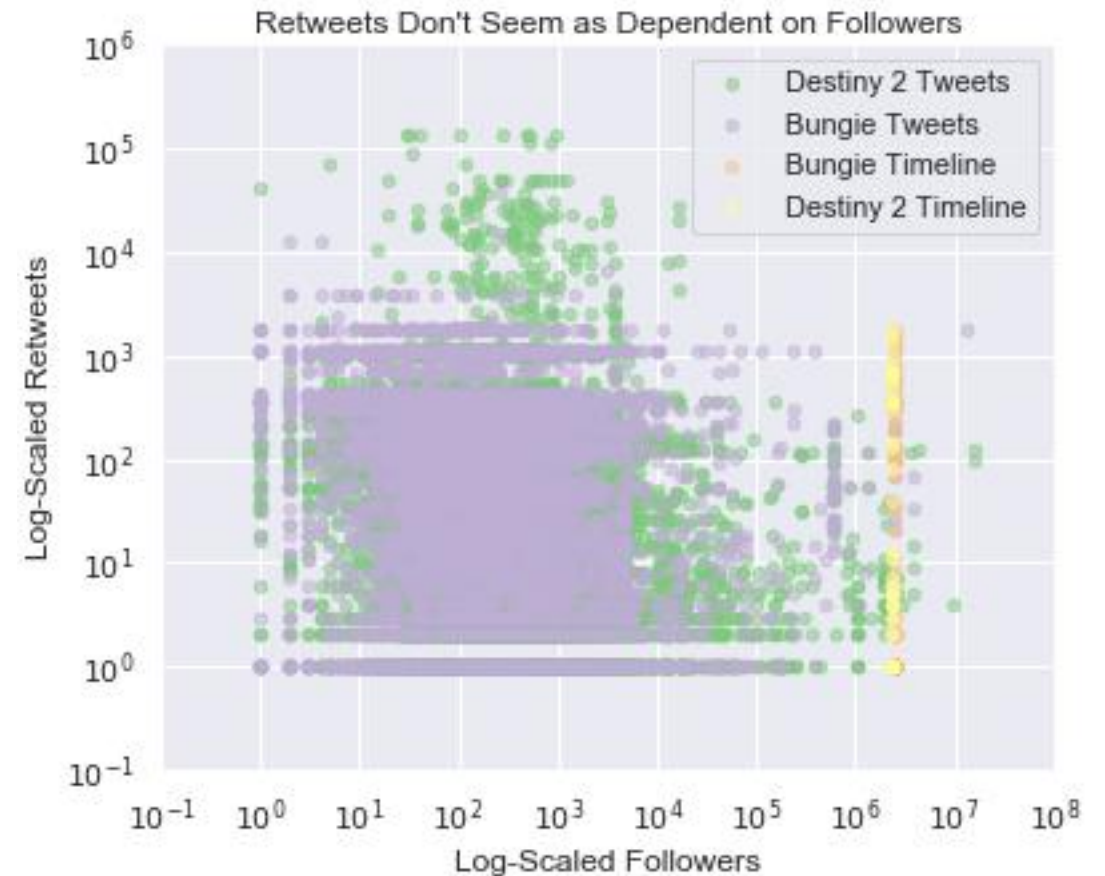
Speaking of Time...

- When are these tweets posted?
 - March 3rd – 13th
- Note:
 - The Destiny 2 tweets start on the 3rd
 - Bungie tweets start on the 5th
- Tweets posted in the evening, mostly on weekdays
- Unique users mirrors the same trend



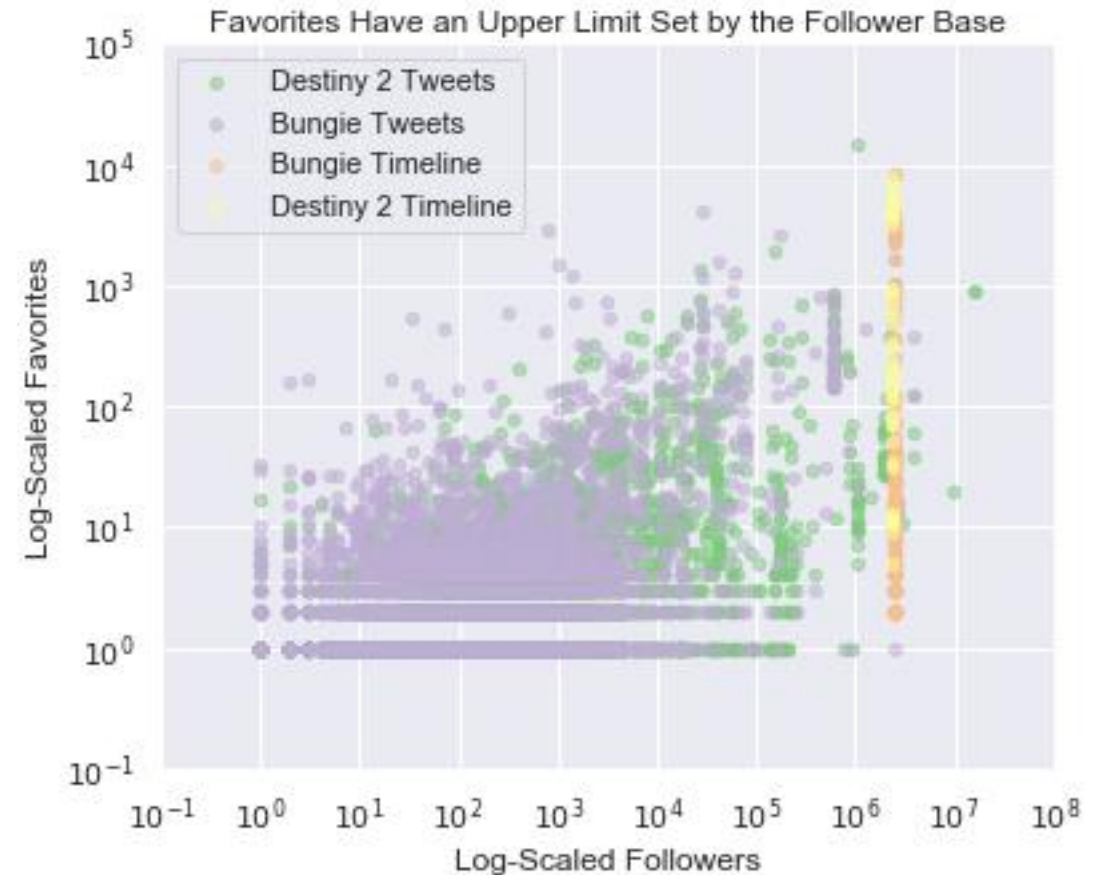
What Drives Engagement?

- First Guess: Followers
- More people will see your tweets, and then interact with it?
 - Apparently not for retweets



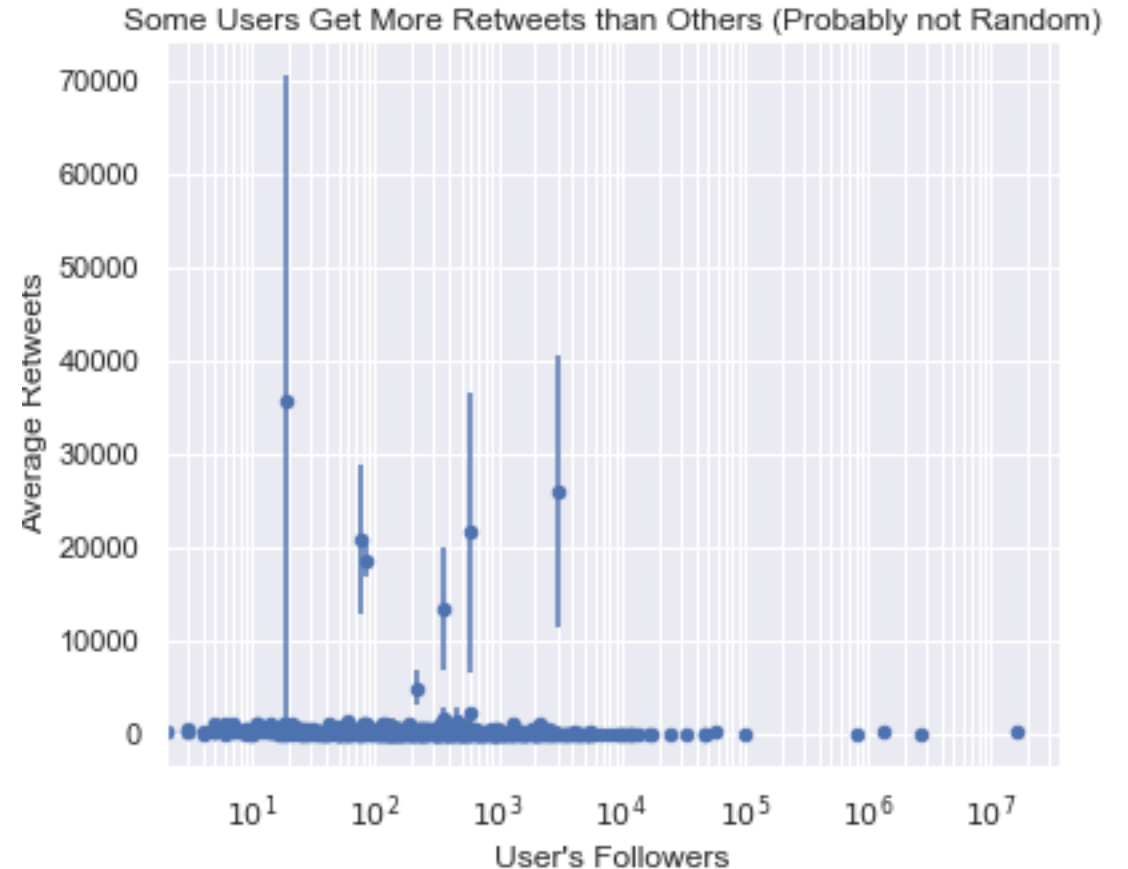
What Drives Engagement?

- First Guess: Followers
- More people will see your tweets, and then interact with it?
 - Apparently not for retweets
 - Maybe for favorites
- These are single tweets
 - Aggregate over each user and check again



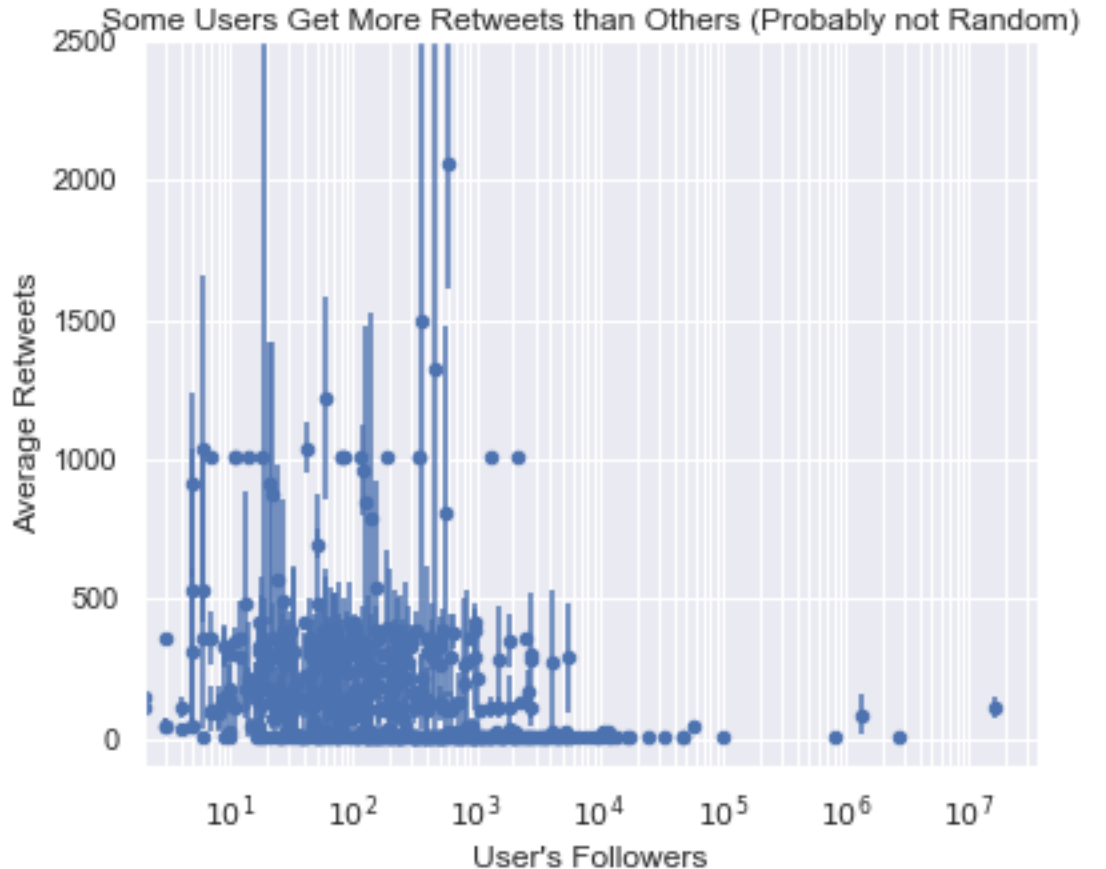
Engagement per User

- See if this gets us better results
- Can do a z-test
 - Average retweets
 - Compare to the baseline of 0
- Error bars are the 95% confidence interval for average retweets
 - Some users definitely get more retweets



Engagement per User

- See if this gets us better results
- Can do a z-test
 - Average retweets
 - Compare to the baseline of 0
- Error bars are the 95% confidence interval for average retweets
 - Some users definitely get more retweets
 - Zooming in, more evidence

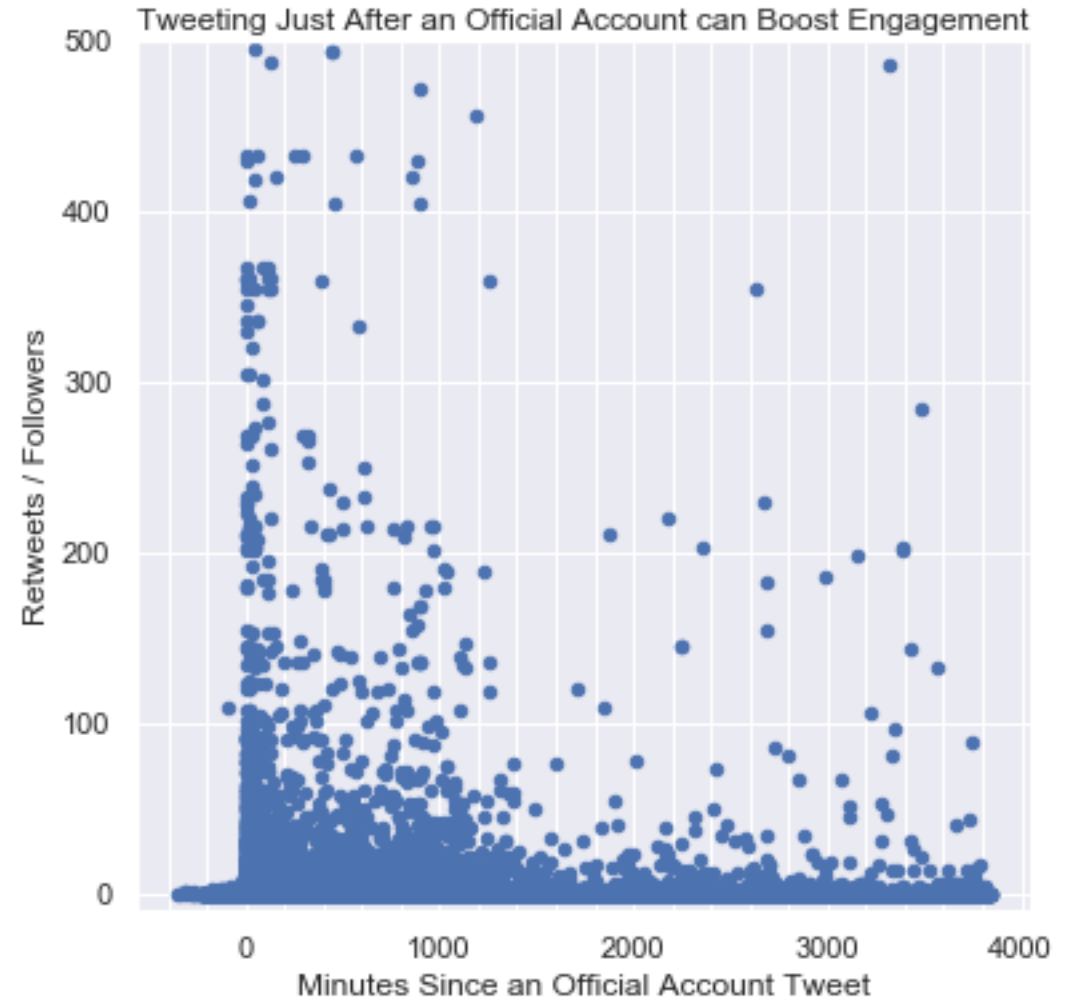


What Drives Engagement?

- Timing?
- Can check time of day
- Time since @Bungie or @DestinyTheGame tweeted?

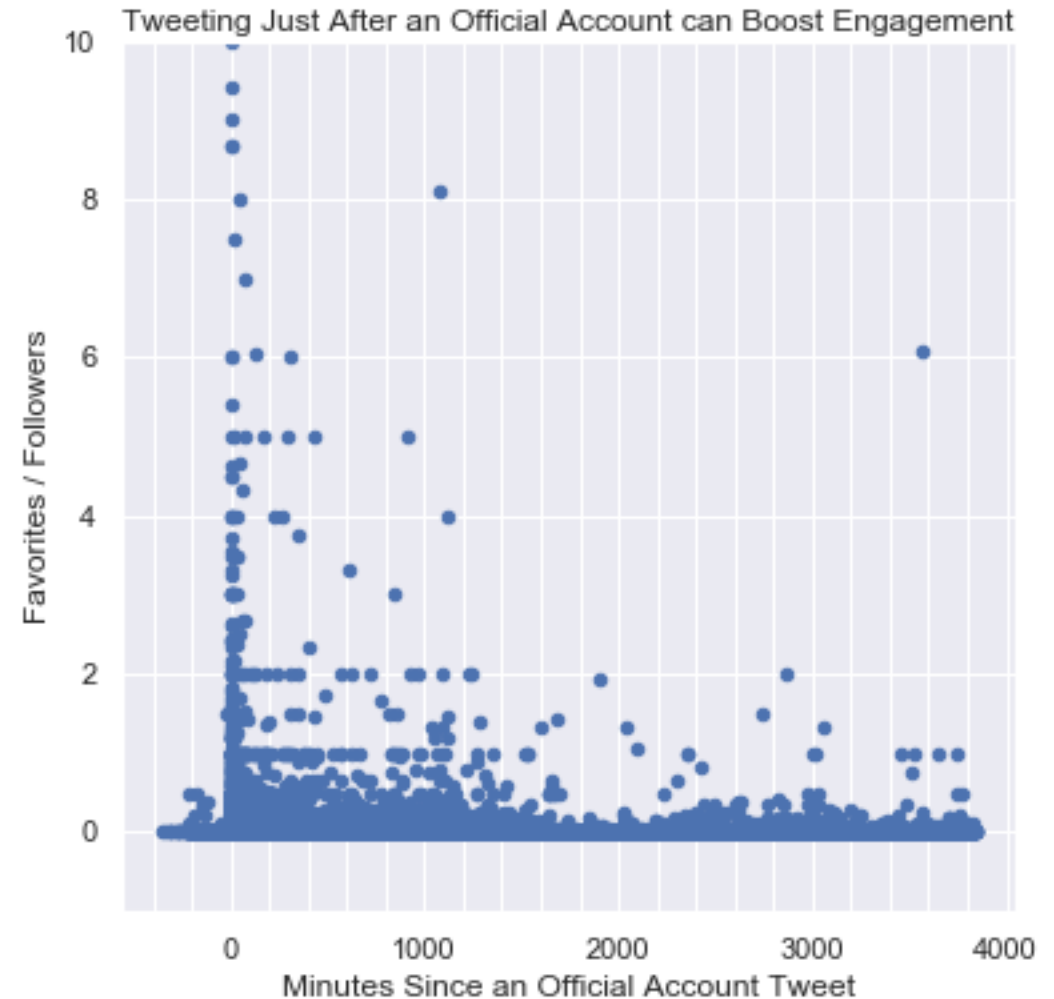
What Drives Engagement?

- Timing?
- Can check time of day
- Time since @Bungie or @DestinyTheGame tweeted?
 - Negative correlation
 - Official tweets can boost your signal



What Drives Engagement?

- Timing?
- Can check time of day
- Time since @Bungie or @DestinyTheGame tweeted?
 - Negative correlation
 - Official tweets can boost your signal



Overview

- What's Bungie / Destiny 2?
- Getting Data From Twitter
- Exploratory Data Analysis
- **Machine Learning Analysis**
- Conclusions

NLP

- Don't want to label ~48k tweets with sentiments
- Package to do that for me?
 - Feature Engineering rather than Machine Learning
- VADER
 - In NLTK

Valence Aware Dictionary and sEntiment Reasoner

- <https://github.com/cjhutto/vaderSentiment>

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
```

- Specializes in social media text
- Can interpret special text cases
 - Emojis
 - ALL CAPS
 - Excessive punctuation

VADER Scores

- Vader will score Strings
- Multivariate
 - Positive: Proportion of words in the text that are positive
 - Negative: Proportion of words in the text that are negative
 - Neutral: Proportion of words in the text that are neutral
- Univariate
 - Compound: Score between -1 and 1

VADER Score Examples

- Neg: 0.815 “@Bungie Cancer”
- Comp: -0.8552 “@Bungie It’s worst in Gambit Prime killed 3 envoys with hammerhead and we killed the prime evil only 2% increase are we serious \U0001f633”
- Pos: 0.879 “@DestinyTheGame @Bungie Yes yes yes yes YES!”
- Comp: 0.9836 “AAAAAAHHHH I forgot how fun Destiny 2 is. Gambit Prime is intense but super fun. :) :) :) :0 :)”
- Neu: 1.0 “RT @DestinyTheGame: Season of the Drifter is underway and the latest Bungie ViDoc outlines what to expect all season long. \U0001f4a0”

Feature Engineering and Processing

- All 4 VADER scores
- Options for the time since the last official account tweet
 - Linear: $-t$
 - Inverse: $\frac{1}{t}$
 - Exponential Decay: e^{-t}
- Scale the Data
 - 0 – 1
- Train test split
 - Test size is 30%

ML Model 1

- Predict number of retweets
 - Best metric for Twitter engagement
- Linear Regression
 - Vary the time column
 - Vary the scaling
 - Remove negative time values
- Performed terribly
 - All the R-squared values were very low (> 0.0007)

ML Model 1

- Exact retweet numbers
 - Lots of randomness
- Would need a hugely complex model to get a good accuracy
- Tried a random forest regressor
 - Also did not perform well
- What now?

ML Model 2

- Retweets are important
- Predict a range
- Class 0: 0 – 99 retweets
- Class 1: 100 – 9999 retweets
- Class 2: 10000+ retweets
- Classification problem
 - K Nearest Neighbors
 - Random Forests

ML Model 2

- K Nearest Neighbors Classifier
- Test Score: 0.901
- After hyperparameter tuning: 0.915
- Confusion matrix:

Predictions	Class 0	Class 1	Class 2
Actual Class 0	11020	340	3
Actual Class 1	800	1546	1
Actual Class 2	25	0	9

ML Model 2.1

- Class imbalance -> Upsample!
- About 5 times more Class 0 data than Class 1
- About 332 times more Class 0 data than Class 2
- Retry K Nearest Neighbors: 0.921
- Hyperparameter tuning: 0.951

Predictions	Class 0	Class 1	Class 2
Actual Class 0	9714	1596	52
Actual Class 1	52	11681	3
Actual Class 2	0	0	11355

ML Model 2.1

- Model works great!
- Might be biased with repeated data for “nearest neighbors”
- Interpret the model...
 - I can't
 - K Nearest Neighbors is a black box
- Try Random Forests

ML Model 3

- Normal data sampling: 0.920
- Hyperparameter tuning: 0.924
- Confusion matrix has the same trends as K Nearest Neighbors

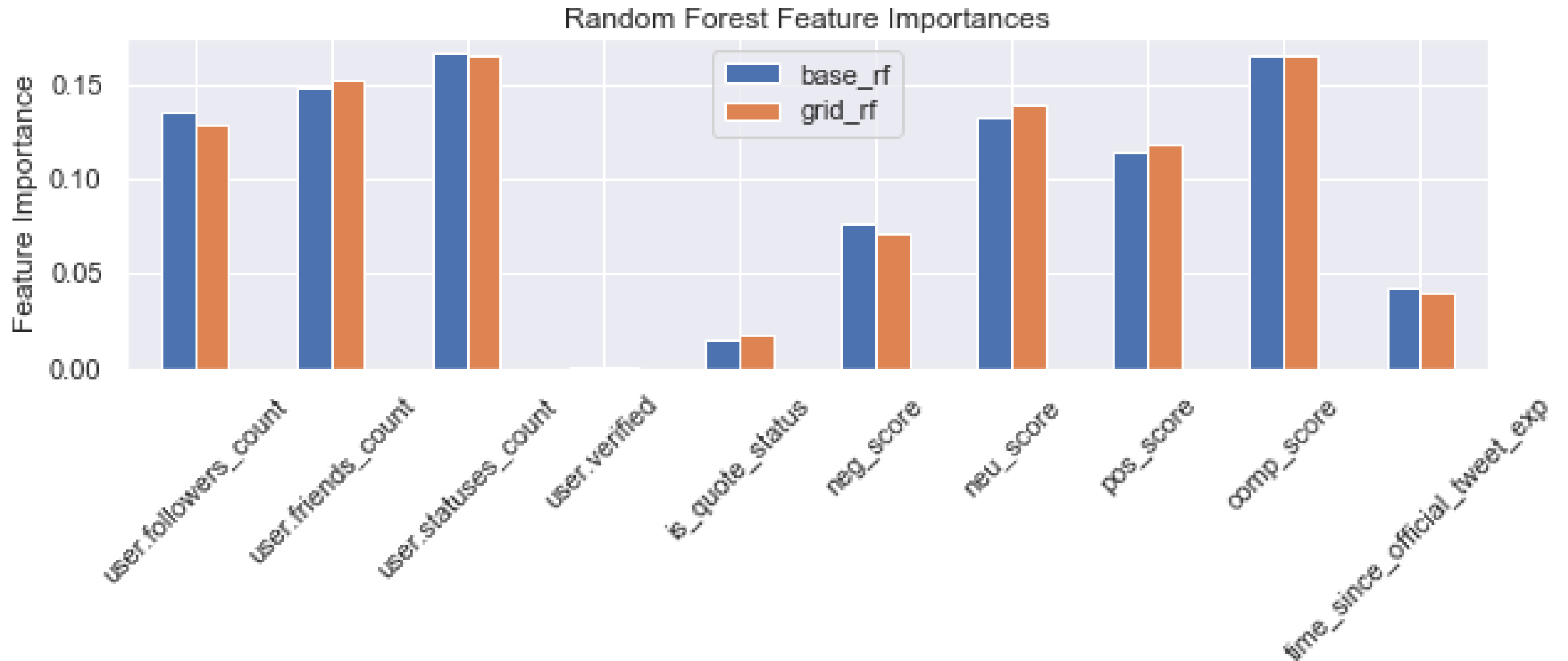
Predictions	Class 0	Class 1	Class 2
Actual Class 0	11206	157	0
Actual Class 1	860	1487	0
Actual Class 2	31	0	3

ML Model 3.1

- Upsampled data: 0.973
- Hyperparameter tuning: 0.975
- Again, same trends as K Nearest Neighbors with upsampled data
- Only this time, I can extract feature importances

Predictions	Class 0	Class 1	Class 2
Actual Class 0	10583	771	8
Actual Class 1	84	11647	5
Actual Class 2	0	0	11355

Random Forest Feature Importances

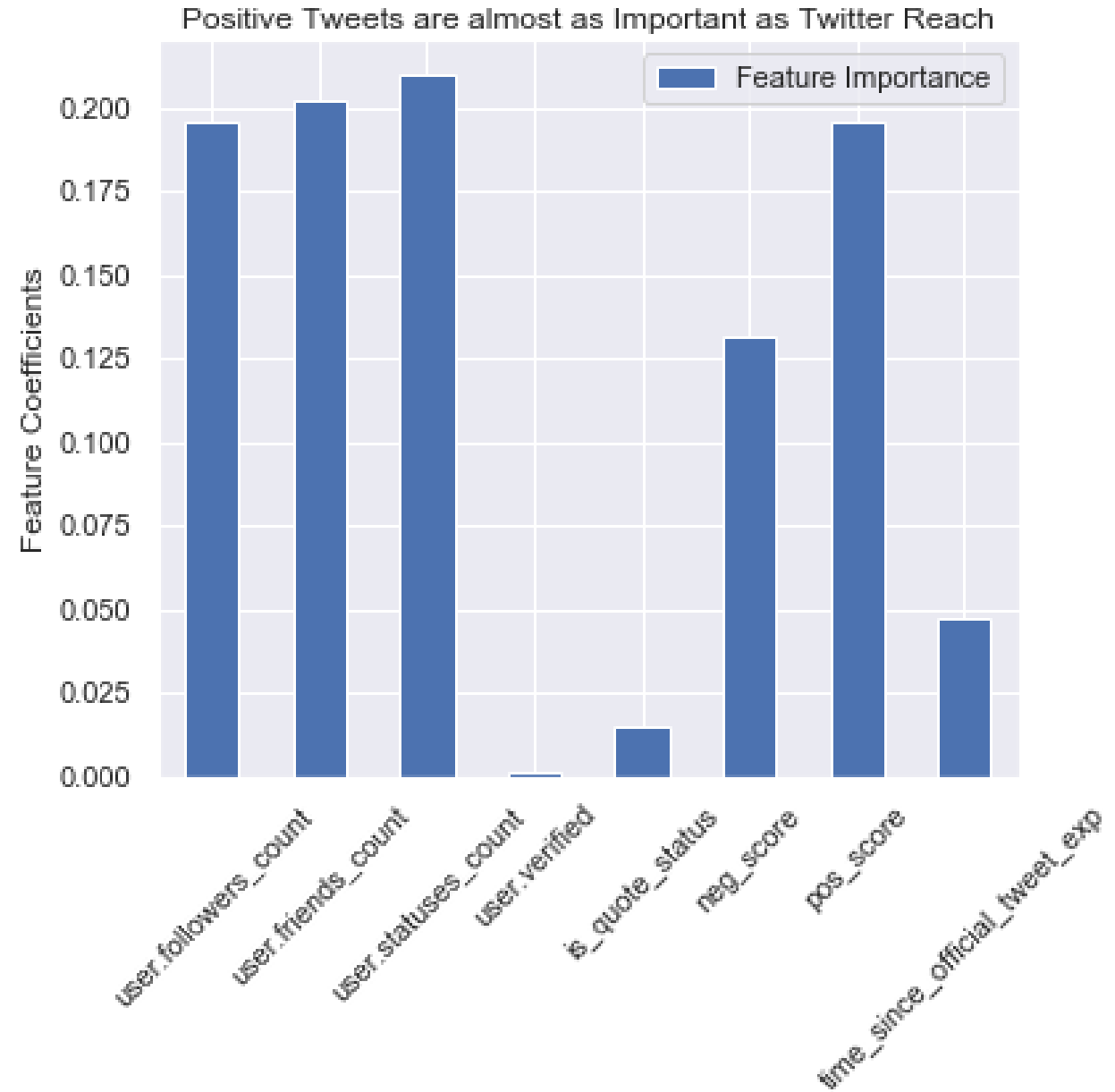


ML Model 3.1 Correlation Problems

- VADER Scores
 - Higher positive correlates with higher compound
 - Higher negative correlates with lower compound
 - Higher neutral correlates with compound closer to 0
- I want to know if positive or negative tweets get more engagement
 - Keep positive and negative scores
 - See which is more important

ML Model 3.2

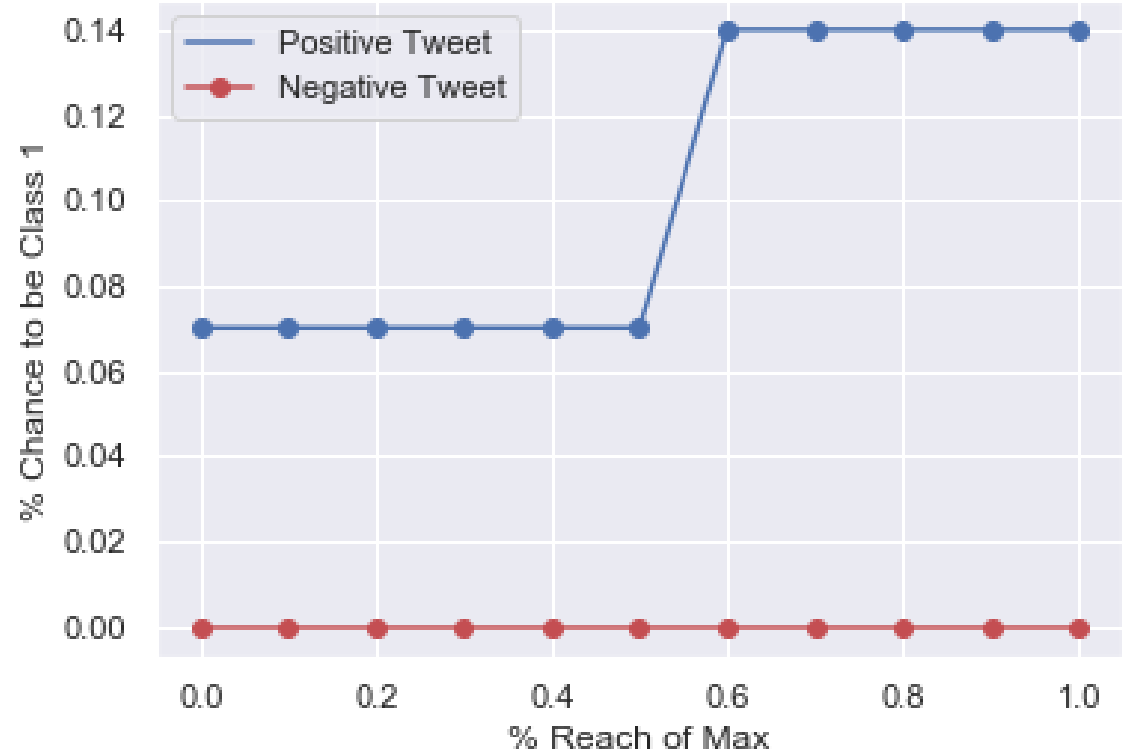
- Drop neutral and compound scores
- Use features that hyperparameter tuning found for v3.1
- Test Score: 0.971
 - Very similar
- Careful, feature importance is not necessarily positive



Positive vs Negative: Which Gets More Retweets?

- Predict class probabilities on simulated data
- One has positive sentiment, one has negative
 - Everything else controlled

Positive Tweets Always Have a Much Better Chance to get more Engagement



Overview

- What's Bungie / Destiny 2?
- Getting Data From Twitter
- Exploratory Data Analysis
- Machine Learning Analysis
- **Conclusions**

Conclusions

- Wanted to draw conclusions from social media data
- Found that:
 - People are more active on Twitter in the evening
 - People can tweet right after Bungie or Destiny to get a small engagement boost
 - ML models can predict if tweets will fall into a certain range
 - Can repurpose this to help Bungie craft tweets that should get more engagement
 - People like positive tweets about Bungie and Destiny 2 way more than negative ones!

Future Work

- Collect cleaner data
- More exhaustive exploratory analysis
- Use sentiment scores better
- More categories for the ML models
- Run a ML model on only official account tweets