Examining Tweet Performance and Virality

Long Ho

Advisor: Professor David Dobkin

Abstract

It can be quite puzzling why certain pieces of media can garner significant attention while others barely get a glance. In this paper, we begin to unravel this mystery by examining why certain tweets acquire more likes than others. We start by examining whether there are any noticeable relationships between certain tweet characteristics, such as how subjective a tweet is, and a tweet's like count. Then, we delve into tweet topic and content matter to figure out what exactly top performing tweets are *talking* about and *mentioning*, with a special focus on tweets related to covid-19 and the election. Indeed, all of our experiments are carried out on tweets created from October - December 2020, extracted using the Twitter API. From our experiments, we not only distinguish tweet characteristics that are relevant to like count, but we also pinpoint more precisely the content of viral, top-performing tweets down to specific keywords.

1. Introduction

1.1 Motivation

The field of unveiling why some things garner more attention than others has been well mined. And rightly so. At least in the perspective of companies and businesses, being able to produce content and media that reaches the eyes of many people very intuitively means an opportunity for a large consumer base and even larger profits. The political arena similarly relies on the widespread dissemination of information and rhetoric to rally support and sway voters.

Even just achieving widespread attention and virality in itself has its own appeal. I seek to tackle this mystery of what kinds of things seize attention and virality on a specific social media platform: Twitter. More specifically, my project seeks to examine why some tweets do better than others in terms of likes, considering a wide range of factors from the actual tweet content to the quantitative aspects of the tweet, like tweet length. What characteristics seem to matter the most in acquiring likes and which seem to be (perhaps surprisingly) irrelevant? What do top performing tweets uniquely talk about?

1.2 Related Research and Background

A number of papers have already investigated these preceding questions to some degree. For instance, researchers from the Hasso Plattner Institute found from their twitter dataset that longer and more subjective tweets tend to acquire more likes and retweets [1]. In another paper titled "Everyone's an Influencer: Quantifying Influence on Twitter", researchers attempted to predict viral tweets by examining tweet topic, user statistics (like number of followers and tweets), and emotional positivity. This paper discovered, for instance, that tweets related to topics such as **lifestyle** and **technology** tend to be the most popular, whereas topics like **sports** and **business/finance** do not receive as much attention [2]. However, this same paper noted that attributes relating to the content were not as predictive of tweet performance as attributes related to the user, such as user follower count and tweet count [2]. A third paper titled "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter" moreover investigated the actual *wording* of these tweets, finding for example that tweets that imitate headlines tend to do better on average [3]. One of my goals was determining the replicability of some of these relationships that have been investigated, namely with the datasets that I have been using. More particularly, I am acquiring more recent tweets from October 2020 -

December 2020, (interestingly distinctive because it is in the middle of an election season and in the midst of a pandemic), and I sought to determine whether the patterns from these papers were consistent with what I found and how some of the contradictions I found between the papers played out in *my* data. Some characteristics that past papers discussed that I sought to examine again were tweet length, user followers, tweet polarity (how positive or negative a tweet is), emotional divergence, and topic matter. One discrepancy between the two of the aforementioned papers was that it wasn't entirely clear whether more positive or more negative tweets did better (although both claimed there was indeed some sort of relationship), and I sought to determine what pattern arose in my data. Some other characteristics that I investigated that weren't touched upon in my reading of the literature were the relevance of pictures, videos, and urls; tweet readability; topic matter; early tweet performance and specific use of keywords. In my data analyses, I investigate these tweet characteristics and their relationship with tweet performance to first determine if there are any noticeable patterns and then how exactly those patterns manifest.

2. Approach

Starting with the rather general question -- why do some tweets do better than others in terms of likes? -- I broke this down into three specific sub questions that could be answered more concretely:

1. What tweet characteristics matter and how do they affect tweet performance?
2. What general tweet topics do the best?
3. What do top performing tweets uniquely talk about and mention?

2.1 Approach: What tweet characteristics matter and how do they affect tweet performance?

To answer this subquestion, I first gathered tweets using the Twitter API function that allowed me to randomly stream a sample of public tweets published in real time [4]. For a dataset of tweets published on a given night, I then sought to examine tweet characteristics that had noticeable relationships with like count. Some of these datasets were special in that I streamed tweets on debate nights as well as nights during election week, allowing me to compare patterns among these nights with patterns overall.

One thing I was interested in observing was whether a tweet's **use of media (pictures/videos)** and **use of URLs** had any relationship with how many likes it acquired. To analyze this, with each dataset on a given night, I calculated the average likes of tweets with media and compared it to the average likes of tweets without media in a given dataset, and quantified the significance of this difference (to be elaborated in *Implementation)*. This same process was done to examine use of URLs.

Furthermore, I investigated qualities of a tweet that can be assigned a *number*, namely tweet **length, polarity** (how positive or negative a tweet is), **subjectivity** (how emotionally charged a tweet is), **early tweet performance**, and **readability** as well as information regarding the user posting the tweet, such as **user follower count** and **tweet count**. To calculate the effect of these different tweet characteristics, I studied these qualities among most highly liked tweets and least liked tweets in a dataset, keeping an eye for characteristics where the difference between the two groups was most noticeable. Moreover, I looked for noticeable correlations across the *entire* collection of tweets in a given dataset.

Ultimately, my goal was to be able to pinpoint the tweet characteristics that had the most noticeable relationships with like count and to extract the nature of these relationships. Given that I was observing new, unexplored data and even examined some of the previously explored tweet characteristics with new techniques, some of my biggest tasks included figuring what portions of the literature coincided with my findings and perhaps reconciling *discrepancies* in the literature. Of course, some tweet characteristics, such as use of media, readability and early tweet performance, represented new investigations that were not as heavily explored in the literature.

2.2 Approach: What general topics do the best?

While one paper I discussed also examined the topic matter of tweets, it did so using human classifiers [2]. In my investigation, I utilized a relatively new feature of the Twitter API that annotates each tweet with a particular "context", which are essentially topics such as "sports" or "videogames". To examine topic matter, I used the Twitter API to stream tweets using a filtered stream that collected tweets of a specific context. I then compared the mean acquired likes of the tweets from each of these context-specific streams and ranked the topics from highest average acquired likes to lowest average acquired likes. Again, given this new approach to extracting the general topic matter of a tweet, I sought to determine how my findings compared to the findings of a prior paper that used human classifiers to determine topic matter.

2.3 Approach: What do top performing tweets uniquely talk about and mention? (covid and election tweets)

To examine the content matter of top performing tweets and how it differs from other tweets, I examine particular *key words* and more generally, words that are specifically mentioned in tweets that do well. To perform this analysis, I narrow my focus to tweets related to two topics most pertinent to the time period of this semester: **covid-19** and **election** related tweets. By narrowing the focus of my analysis of content matter to these two topics, I could more easily determine the types of *specific* content, words, and topics that dominate the highest performing tweets. Moreover, this allows for a unique analysis of current events that has not been replicated in previously discussed literature.

The first way I tackled this task of figuring out what top performing tweets are talking about was by analyzing another new feature of the Twitter API (notably a tool that prior papers did not have access to) that annotated each tweet with specific "entities" which are specific concepts, organizations, people, among other things, such as "Barack Obama" or "The Supreme Court."

The second way I analyzed content matter involved a machine learning and natural language processing algorithm that again was not discussed heavily in the literature: topic modeling. I extracted latent topics underlying a large collection of tweets, with each topic composed of keywords among the tweets, to create topic models. The advantage of this method is that it latches onto a wide variety of words that are actually mentioned in the tweets as opposed to "entities" which are words that are selected by Twitter itself.

3. Implementation

My investigation was a multipronged process that can be broken down into first collecting the tweets primarily through the Twitter API; pre-processing, cleaning, and filtering out the tweet data; extracting information about the tweets; and finally looking for patterns and relationships with tweet performance.

3.1 Implementation: What tweet characteristics matter and how do they affect tweet performance?

3.1.1 Tweet Collection

To randomly collect tweets, I used the Twitter API to randomly stream 1% of all public tweets that were being published in real time on nights throughout October 2020 - November

2020 [4]. I randomly collected tweets for about half an hour, and collected these same tweets again about three to four days later to record how many likes each individual tweet received in the three to four day span. The only exception to this was that to examine the relevance of early tweet performance, I examine entirely new datasets of randomly streamed tweets created in late December[1], where I streamed for a shorter time of 15 minutes to ensure tweets of the same dataset were a bit more precisely created at the same time.

The fact that the tweets in a given dataset were created roughly at the same time and were given the same time period to acquire likes allowed me to control for the time factor in tweet performance (tweets visible longer will get more likes).

3.1.2 Tweet Preprocessing, Filtering, and Information Extraction

I first excluded tweets from users with less than 10 followers. Doing so helped produce more interesting results, because tweets are first only visible to a user's followers and can only be spread to the timeline feed of other users via liking or retweeting a tweet. By excluding tweets from users with near negligible followings, I help ensure that the tweets in my dataset are not hindered merely by the negligible following of the posting user. Moreover, I excluded retweets to only focus on original postings. The reason for this is that if a retweeted post does not have text from the retweeting user it can not be liked, and even if the retweet *did* have text, the performance of that tweet would be confounded with the tweet content from the original user.

The preprocessing of tweets and extraction of tweet information varied depending on the particular metric I was trying to obtain. **Like count, user follower count, user tweet count,** and whether a tweet contained **media** were all metrics that came with each tweet in the JSON payload via the Twitter API. A tweet's **use of URLs** was determined by the presence of the

[1] This investigation involves a new dataset because I did not think of this experiment until late December, and it is impossible to retrieve a tweet's like count at a specific time in the past with the Twitter API.

string “https://” in the tweet text. To obtain **tweet length, subjectivity score**, and **polarity score**, I first had to remove the urls and emojis in each tweet. Doing so allowed me to remove extraneous information about a tweet relative to the length and the sentiment of a tweet’s *content*. To calculate sentiment scores, I used the Textblob library to extract subjectivity and polarity of a given tweet [5]. The subjectivity score ranged from 0 to 1, with 0 being the least subjective and 1 being the most subjective, while polarity score ranged from -1 to 1, with -1 indicating highly negative, 0 indicating neutral, and 1 indicating highly positive [5]. Finally, to extract information about **readability**, I used the *textstat* library to obtain the Flesch-Kincaid readability score of a given tweet, which is a score that increases the easier a text is to read; the maximum score is 121.22 and there is no theoretical minimum score [6]. This reading score is based on the average words per sentence and the average number of syllables per word in a text [7]. Before calculating this score, I had to remove non alphanumeric characters, such as extra punctuation, since this score is based primarily on the *words* of a text. Finally, I investigated the relevance of **early tweet performance**. After streaming my late-December tweet dataset, I collected the same tweets and their like count *3 hours* after and compared it to the like count acquired after *3 days*, to investigate if it matters how a tweet performs in the first few hours of being visible.

I organized this tweet information into a pandas dataframe where each row represents a tweet and each column represents a numerical quality. In the case of use of media and use of URLs, these columns were a boolean representing the presence of that feature.

3.1.3 Looking for Patterns

To analyze the impact of urls and media, I calculated the average like count of tweets with media and urls and compared it with the average like count of tweets without media and urls. I used Cohen's D effect size to gauge the significance of this difference. For instance, to calculate effect size of the use of media, the formula is as follows

$$\frac{\left(M_{media} - M_{no\ media}\right)}{Pooled\ Standard\ Deviation}$$

where M represents the mean of the groups and the difference is divided by the standard deviation of the entire dataset [8]. Moreover, I calculated the two-tailed p-value of this difference, which is the probability that the difference between these two groups was due to chance, using the *ttest_ind* function from the scipy stats library [9]. A description of the two-tailed p-value is explained in detail in an [informational article from the Eberly College of Science]() [10]. Shown below is an example of a data report for the dataset of tweets collected on 11-21-20.

```
11-21-20 Number of data points: 1592

11-21-20 Mean Likes for Tweets overall: 6.348618090452261
11-21-20 Standard Deviation for Likes for Tweets overall: 35.31894292126997
11-21-20  Mean Likes for Tweets(#1271 tweets) with no attachments: 3.5932336742722266
11-21-20  Mean Likes for Tweets(#321 tweets) with attachments: 17.258566978193148
t-score: -6.268016564874303 and p-value: 4.699341785262718e-10
Effect size: 0.38691229616873124

11-21-20 Mean Likes for Tweets overall: 6.348618090452261
11-21-20 Standard Deviation for Likes for Tweets overall: 35.31894292126997
11-21-20  Mean Likes for Tweets(#543 tweets) with urls: 10.9060773480663
11-21-20  Mean Likes for Tweets(#1049 tweets) with no urls: 3.9895138226882745
t-score: 3.7191453660905895 and p-value: 0.00020684078550744554
Effect size: 0.1958315553439934
```

To analyze more numerical qualities of a tweet, one tactic I used was calculating the average of each of these qualities among the top 10% of tweets and comparing it with the bottom 10%, such as the average readability score of the top 10% of tweets versus the average readability score of the bottom 10%. I again calculated Cohen D's effect size and two-tailed p-value, but in this case these metrics measure the difference between the top performing and bottom performing tweets.

Finally, I looked for general linear correlations across the *entire* dataset on a given night, by calculating the Pearson correlation coefficient between like count and a given numerical quality of a tweet, such as readability score, using the *corrcoef* function from the numpy library. The mathematical formula behind the Pearson correlation coefficient is described in detail here [11]. An example of the information I collected is shown below, specifically for the dataset of tweets on 11-20-21.

```
11-21-20 Correlation coefficients
Correlation coefficient for tweet length vs likes : 0.053350192764057086
Correlation coefficient for polarity vs likes: 0.035120965816882926
Correlation coefficient for subjectivity vs likes : 0.042212520871619925
Correlation coefficient for follower count vs likes : 0.1975558301626437
Correlation coefficient for tweet count vs likes : -0.038263077016974394
Correlation coefficient for readability vs likes: -0.016320053456336612
```

```
Analyzing best and worst tweets (includes outliers, n = 1594)
11-21-20 Polarity Score
Average polarity of top 10% of tweets: 0.07347860472860474
Average polarity of bottom 10% of tweets: 0.045367922475311476
t value: -1.1165729564666094 and p value: 0.26445489756997104
Effect Size: -0.09454869611562305
```

3.2 Implementation: What general topics do the best?

3.2.1 Tweet Collection and Information Extraction

I also streamed tweets using a Twitter API filter function that collected tweets based on a particular context, which represents a topic like "Sports" [4]. Because of a monthly stream limit in the Twitter API for this particular function, my datasets only come from tweets created in late November and late December. Given that many of these Twitter selected contexts were related to each other, like "Videogame Conference" and "Video Game Tournament", I narrowed the scope to topics that were markedly distinct from each other and represented a diverse list of possible topics, namely **movies, sports**, **products, video games, TV shows, politicians, interests/hobbies, journalist, holidays.** After collecting tweets and creating distinct datasets corresponding to distinct contexts, I collected these tweets again after three to four days to record how many likes they acquired, similar to the random sample stream.

3.2.2 Looking for Patterns

I calculated the average like count of each of these context-specific datasets and ranked them, and looked for similarities across the different nights on which I streamed data points based on a specific context. One example of how I ranked different contexts is shown below from one night of tweet streaming.

Dataset 1		
Topic	**Mean Likes**	**n**
Movies	19.32	228
Sports	9.77	1797
Products	9.39	790
Videogames	8.06	2167
TV Shows	7.30	11806
Politicians	6.77	10030
Interests/Hobbies	5.75	9400
Journalist	5.63	2806
Holidays	4.64	1145

3.3 Implementation: What do top performing tweets uniquely talk about and mention?

3.3.1 Tweet Collection

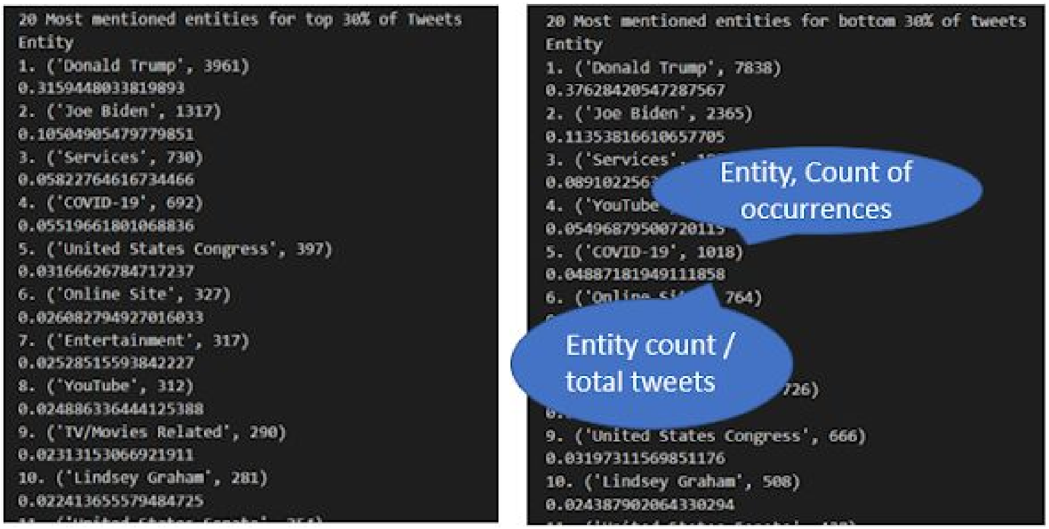
For both election and covid-19 related tweets, I used the 7 day historical search which allows me to fetch tweets that contain specific keywords and on specific days in the last 7 days of using the API [4]. I fetched tweets that had keywords "**covid**" and "**election**" and were created in **mid to late November**.

Furthermore, to expand on my analysis of election related tweets I analyzed tweets that were randomly streamed on the night of the **vice presidential debate**, the **last presidential debate**, and **election night**. A single dataset comes from a single night of either tweet collection or streaming, and I looked for patterns across these different datasets.

3.3.2 Tweet Preprocessing, Filtering, and Information Extraction

To preprocess and filter the twitter data, I built on the procedures already used on the randomly sampled tweets as described in section 3.1.2, including removing urls, emojis, non alphanumeric characters. To create the most informative topic models, I also made all characters lowercase, lemmatized words using the WordNetLemmatizer from the nltk library, and finally removed stop words, which are words that are not informative to the actual content of the tweet text [12]. These stop words came from a mixture of manually created words, the nltk.corpus library, and sklearn feature extraction library [13] [14]. These preprocessing steps ensure that the text corpus that our topic modeling algorithm will analyze primarily contains words that are informative of the content of the tweet and can more accurately pick out meaningful key words.

I created two collections of tweets from each dataset, one from the top 30% of each dataset and the bottom 30%. The first thing I analyzed was the "entity" Twitter annotations attached to each tweet, which as discussed before represents specific people, concepts, organizations, or things. For each collection of tweets, I used the Counter class from the python *collections* module to keep count of the different entities used and I generated a list of the top 20 entities, additionally keeping track of the number of times that entity appears in the collection of tweets out of the total number of tweets [15]. Shown below are two ranked lists of entities from 11-19-20 regarding covid related tweets.

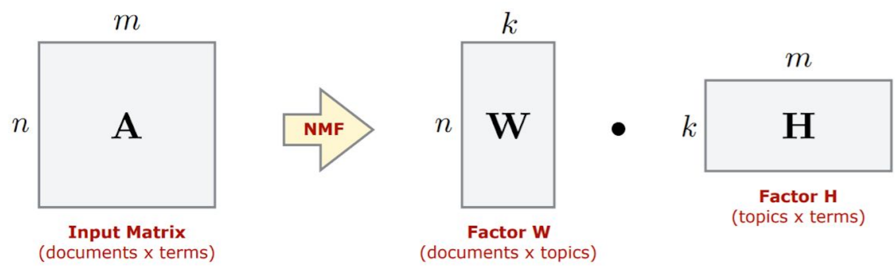


The second component of my analysis of each collection of tweets was topic modeling. For each collection of tweets, I first used the TfidfVectorizer class from the sklearn feature extraction library library to transform the collection of tweets into a tf-idf matrix, which results from multiplying the term frequency matrix of the collection of documents by the inverse document frequency vector [14]. Put more precisely,

$$tf_{m,n} = \frac{\#\ times\ word\ m\ appears\ in\ tweet\ n}{total\ \#\ of\ words\ in\ tweet\ m}$$

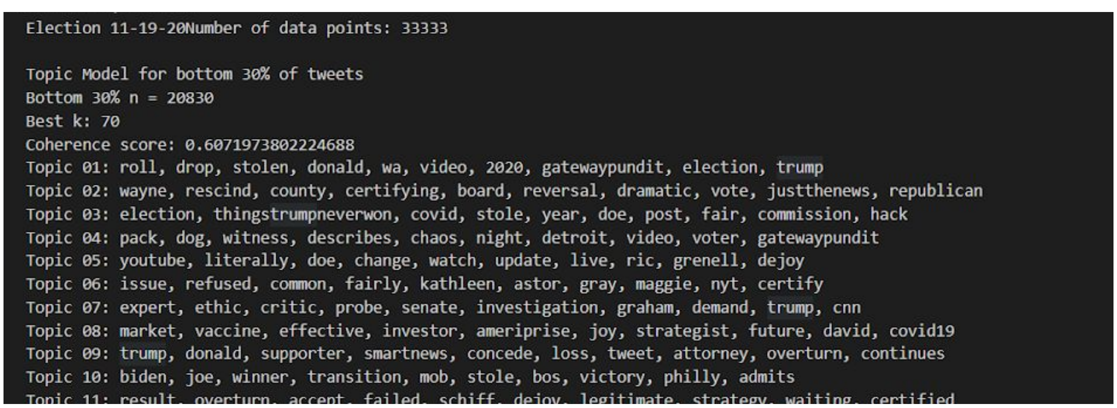
$$idf_{n} = \frac{total\ \#\ of\ tweets}{1 + total\ \#\ of\ tweets\ with\ word\ n}$$

The tf-idf matrix is created by broadcasting the tf matrix and the idf vector. The topic modeling algorithm we will use here is non-negative matrix factorization. (As a note, I chose this algorithm because it seemed to produce the most coherent topics with meaningful and well related words, as opposed to other algorithms I tried like Latent Dirichlet Allocation, which produced topics with harder to understand and less meaningful topics.) Non-negative matrix factorization (NMF) factorizes the tf-idf matrix into the product of a matrix W, where the rows are the tweets and columns represent the topics, and matrix H, where the rows are topics and the columns are particular terms that make up each topic [16] .



[16]

I performed this factorization using the NMF class from the sklearn.decomposition module [17]. To create topic models, we focus on matrix H which is a matrix of topics by term weights. For each topic (row), we sort the terms so that the terms with the highest weights appear first and then retrieve the words associated with the top 10 terms for each topic. An example topic model is shown below.



```
Election 11-19-20Number of data points: 33333

Topic Model for bottom 30% of tweets
Bottom 30% n = 20830
Best k: 70
Coherence score: 0.6071973802224688
Topic 01: roll, drop, stolen, donald, wa, video, 2020, gatewaypundit, election, trump
Topic 02: wayne, rescind, county, certifying, board, reversal, dramatic, vote, justthenews, republican
Topic 03: election, thingstrumpneverwon, covid, stole, year, doe, post, fair, commission, hack
Topic 04: pack, dog, witness, describes, chaos, night, detroit, video, voter, gatewaypundit
Topic 05: youtube, literally, doe, change, watch, update, live, ric, grenell, dejoy
Topic 06: issue, refused, common, fairly, kathleen, astor, gray, maggie, nyt, certify
Topic 07: expert, ethic, critic, probe, senate, investigation, graham, demand, trump, cnn
Topic 08: market, vaccine, effective, investor, ameriprise, joy, strategist, future, david, covid19
Topic 09: trump, donald, supporter, smartnews, concede, loss, tweet, attorney, overturn, continues
Topic 10: biden, joe, winner, transition, mob, stole, bos, victory, philly, admits
Topic 11: result, overturn, accept, failed, schiff, dejoy, legitimate, strategy, waiting, certified
```

first 10 topics of the topic model for election related tweets on 11-19-20 (bottom 30% group)

Note that k, the number of topics, is a parameter that I manually had to select. To choose the best number of topics, I used a topic model evaluation metric called *topic coherence*. To calculate this, I first converted the collection of tweets into tokens of words and created a Word2Vec model, using the gensim library, which uses word associations within this large body of text to represent each term as a vector, allowing me to determine as a specific value the semantic similarity between any two words [18]. The coherence score of a *topic* is the average semantic similarity between the pairs of words composing that topic; the coherence score of a *topic model* is the average coherence score across all its topics [16]. To roughly approximate the best number of topics, I created different topic models with different k values, iterating k from 1 to 80 by step size 3. I then selected the topic model that had the highest coherence score.

3.3.3 Looking for Patterns

To analyze the entity models, I manually picked out words from the list of the top 20 entities from the top performing tweet collection that were represented more relative to the entities from the bottom 30% of tweets. More precisely, I identified entities where the entity count divided by the number of tweets in the top 30% tweet collection was greater than that of the bottom 30% tweet collection and the percent difference was greater than 15%.

In other words, I looked for entities such that

$$t = \frac{\#\ of\ occurrences\ in\ top\ tweets_{entity\ m}}{total\ \#\ of\ tweets\ in\ top\ tweet\ collection}$$

$$l = \frac{\#\ of\ occurrences\ in\ bottom\ tweets_{entity\ m}}{total\ \#\ of\ tweets\ in\ bottom\ tweet\ collection}$$

$$\frac{t - l}{\frac{(t + l)}{2}} > 0.15$$

After identifying these keywords in a dataset, I looked for patterns *across* the datasets, entities that popped out in *multiple* datasets under a given topic.

The analysis of topic models was two-fold. For a more holistic approach, I manually read the topic models and compared which new stories, people, concepts, and things were being mentioned. However, picking out specific keywords that were represented more in the top tweet topic model required a more precise, programmatic approach. I first found key words that were present in the top 30% topic model but not in the bottom 30% topic model and identified words that were this way across *multiple* datasets under a given topic. As a note, at this stage, although we filtered out stop words in the preprocessing step, there were still stop word-like words that showed up in our topic model, so we focus on words that provide meaningful information. Moreover, similar to analyzing entities, I used the Counter class from the python collections module to keep a count of the most mentioned words in a given topic model and calculated a proportion p, the occurrence of that word divided by the total number of words in the topic model [15]. I then looked at the most occurring keywords up to those that were tied for 10th place (often meaning about 15 words in total) in the topic model of the top 30% of tweets and among these words I identified those where proportion p was greater than in the bottom 30% topic model, such that the percent difference was greater than 15%. In other words, among the top 10 key words, I looked for words m such that

$$p_{top\ tweets} = \frac{\#\ of\ occurences\ of\ \ word\ m\ in\ top\ tweet\ topic\ model}{total\ \#\ \ of\ \ words\ in\ top\ tweets\ topic\ model}$$

$$p_{bottom\ tweets} = \frac{\#\ of\ occurences\ of\ \ word\ m\ in\ bottom\ tweets\ topic\ model}{total\ \#\ \ of\ \ words\ in\ bottom\ tweets\ topic\ model}$$

$$\frac{p_{top\ tweets} - p_{bottom\ tweets}}{\left(p_{top\ tweets} + p_{bottom\ tweets}\right)/2} > 0.15$$

Looking at words represented more in a topic model is not quite an exact approach to measuring representation in the actual corpus of tweets. However, doing this still gives us some insight about words that are mentioned more because our topic modeling algorithm picks out words that appear frequently together in the collection of tweets and groups them as topics. If a word appears in many topics, it is a rough signal that that key word appears often in the actual collection of tweets.

4. Evaluation

4.1 Results and Evaluation: What tweet characteristics matter and how do they affect tweet performance?

From our random stream of tweets from October 5 - December 31, we obtained 124441 total tweets for our datasets across 23 different nights of streaming, ranging from 990 to 21985 data points. The full details are described in Appendix 1.1. Keep in mind, our datasets from late December were purely to examine early tweet performance and are distinguished in the appendix.

4.1.1 Results: Comparing Means of Different Groups and the Story Told By Effect Size & P-Value

Keeping in mind effect size and p-values as described in *Implementation* (Section 3.1), I make special note of the tweet characteristics with weighted average p-value less than 0.05 and effect size greater than 0.1 to precisely define for the purposes of this paper which tweet characteristics have noticeable relationships with like count and which do not.

Use of Media and URLs

I describe in Appendix 2.1 a comparison of the average likes of tweets with media and urls with the average likes of tweets without them for all the datasets. In summary, I noticed that with each of our datasets, our tweets that had **media** on average received **more likes** than tweets without them and tweets with **URLs** on average also received **more likes** than those without them.

Moreover, the calculated average effect size and p-values associated with these two tweet characteristics is summarized below.

Characteristic	**Weighted average p-value**	**Weighted average effect size**	**Nature of Correlation**
Use of media	6.228e-6	0.235	Use of media -> More likes
Use of URL	0.004	0.118	Use of urls -> More likes

We note that for both of these qualities, p-value is less than 0.05 and effect size is greater than 0.1, with the use of media being the stronger variable. However, though noticeable, the effect sizes of having URLs and media are around 0.1-0.2 and designate a *small* effect. These labels come from a suggested benchmark developed by statistician Jacob Cohen to put meaning to cohen d's effect size values [19]. Though it should be noted that even effect sizes as small as 0.1 can have large consequences and the purpose of these values is better suited to compare relative effects rather than to assign arbitrary labels [20].

Numerical Qualities

Our first analysis of numerical qualities involves an investigation of key differences between the top 10% and bottom 10%, where we compare the mean of different qualities between these two groups across all the datasets. In Appendices 2.2 and 2.3, I show the graphs of this analysis with tweet characteristics I designated as having the most noticeable relationships with like count and those having the least noticeable.

Our distinction of most noticeable vs. least noticeable tweet characteristics comes from an analysis of effect sizes and p-values, which quantify the significance of the difference between the top performing and bottom performing tweets.

Characteristic	Weighted Average p-value	Weighted Average Effect Size	Nature of Correlation
Polarity	0.149	0.091	positive
Subjectivity	0.035	0.150	positive
Readability	0.107	0.056	negative
Tweet Length	0.004	0.321	positive
Follower Count	1.008e-15	0.564	positive
Tweet Count	0.139	0.062	negative
Early Tweet Performance	9.62e-05	0.644	positive

Highlighted in green are the tweet characteristics with p-values less than 0.05 and effect sizes greater than 0.1, designating a characteristic that has a noticeable relationship with like count. In summary, based on these noticeable characteristics, the top performing tweets in general were **longer, more subjective, performed well early on,** and users of these tweets had **more followers** compared to bottom performing tweets. Of these noticeable characteristics, the effect size ranges from 0.15 for subjectivity, which denotes a *small* effect, to 0.64 for early tweet performance, which denotes a *medium-large* effect [19].

Though we defined polarity, readability, and tweet count to not have noticeable relationships, the reality is not as clear cut; even with these characteristics, we observe slight patterns. For instance, across all our datasets, top performing tweets on average had users who tweeted *less*, suggesting a negative correlation between user tweet count and a tweet's likes. However, an analysis of effect size and p-values lead us to not make any strong, definite conclusions about these characteristics.

4.1.2 Results: Correlation Coefficients of Numerical Qualities, A More Subdued Story

In addition to analyzing top and bottom performing tweets, I also calculated the Pearson correlation coefficient which captures linear relationships across the entire dataset. A summary of calculated coefficients is shown below.

Weighted Average Pearson Correlation Coefficients in Relation to Like Count

Tweet Length	0.024
Polarity	0.005
Subjectivity	0.004
User Follower Count	0.174
User Tweet Count	0.0005
Readability	-0.015
Early Tweet Performance	0.973

The patterns we witnessed with comparing top and bottom performing tweets also manifest with these correlation coefficients, with the exception of tweet count. Interestingly though, the majority of these correlation coefficients are rather small and almost negligible. This demonstrates a more subdued story about the conclusions we drew from comparing top and bottom performing tweets, showing that the "noticeable" relationships we defined are perhaps not as strong when broadening our focus to the entire dataset. The only exception to this pattern was with **early tweet performance** in which the correlation coefficient is remarkably high, making the relationship between a tweet's performance in its early hours and its performance days later even *more blatant*.

4.1.3 Discussion: Evaluation of Results

Upon initial analysis of top and bottom performing tweets, we picked out key tweet characteristics--follower count, tweet length, subjectivity, early tweet performance, use of media, and use of URLs -- that "mattered" in terms of how many likes a tweet received, based on our two evaluation metrics **effect size** and **p-value**. We precisely pick out these tweet characteristics on the premise that the effect they had on our dataset was both *important* --in that the effect was at least *small*-- and *statistically significant*. This requirement that these tweet characteristics have at least a small effect is admittedly far from a sweeping declaration about tweet performance. However, there was still a degree of success in our experiment in identifying characteristics that had at least *some* relevance with performance. Moreover, as noted earlier in this section, an effect labeled as *small* could still have large consequences for a tweet's performance.

However, we should also note that our additional analysis of another evaluation metric **Pearson correlation coefficient** qualifies some of the discoveries we made in our initial analysis. Across the different numerical qualities we considered, most of the correlation coefficients were very close to 0, suggesting very weak correlations with like count. This suggests that our understanding of tweet performance is still limited and requires further investigation of variables beyond what we already covered. Nonetheless, it is important to note that part of the reason for these weak correlations was that the vast majority of collected tweets simply do not do well and often do not rise above several likes. Therefore, most of our data points could be considered low performing and it was necessary to isolate top performing tweets and measure them against low performing tweets, as we did in our previous analysis. However, we note one exception in which the correlation coefficient reveals to us one of the strongest indicators of a tweet's overall performance: a tweet's like count even after several days is *significantly related* to how well the tweet performs in its early hours.

4.1.4 Discussion: Comparison with the Literature

In terms of the numerical tweet characteristics we defined to have noticeable relationships with like count, our findings are in line with what we discovered in the literature. There seems to be a relationship between tweet performance and subjectivity and tweet length according to a paper from the Hasso Plattner Institute. Moreover, very intuitively, we read from this same paper that tweets from users with more followers tend to get more attention, and that user tweet count seemed to be much less relevant [1]. Indeed, these discoveries were reflected in our data.

Moreover, it remains to address a key discrepancy in the literature: different papers had different claims about the role of polarity in tweet performance. With my own data, I found that there was a slight positive correlation with like count and that top performing tweets tended to be more positive. However, our data also showed that this relationship was not particularly strong or noticeable, something that was not reflected in the literature cited in section 1.

4.2 Results and Evaluation: What general topics do the best?

Results

In our initial analysis of context-specific tweets created in late November, it is relatively easy to pinpoint specific topics that generally performed better than others, though the difference is not particularly drastic.

I detail in Appendix 3.1 the details of the average likes by topic. In summary, based on these two datasets by themselves, I identified **sports** as having the best performance across these two datasets with **products, video games, and tv shows** also performing well, being in the top half of topics across the datasets. Meanwhile, **holidays, politicians, interests/hobbies,** and **journalists** were topics that generally stayed in the bottom half.

However, examining the datasets generated in late December[2] shows that analysis of topic matter is more complex than initially thought and sensitive to time period, as detailed in Appendix 3.2.

The first interesting difference with these December datasets is the rather significant increase in average likes across the board and the increase of the *spread* of these average likes. Moreover, two topics we previously identified to be poor performing topics, **journalist** and **politicians**, interestingly topped both the rankings for the December dataset. Conversely, two topics we previously identified to be top performing, **product and video game,** ended up in the bottom of the rankings for both datasets. Perhaps one topic that has remained in the top half across all the datasets is **sports**, though its relative performance still decreased.

<u>Discussion and Evaluation of Results</u>

We have perhaps identified **sports** as having the most consistent popular topic among those we examined, since it is in the top half of rankings for all our datasets. This actually contradicts a paper in the literature that found that sports had the worst popularity among the topics it examined, although the specific topics examined were similar to but not exactly those discussed in this paper [2]. Nonetheless, it is clear that instead of identifying topics that perform well in general, we instead have discovered top performing topics specific to particular time periods (late November vs. late December). In that sense, we were unsuccessful in determining topics that "do the best" among those that we considered. An important insight here is that even

[2] The reason I only have tweets from late November and December was a monthly cap on the filter function in the Twitter API.

with a time period difference as small as a month, the particular topics that do relatively well and the performance of tweets across the board can change drastically.

4.3 Results and Evaluation: What do top performing tweets uniquely talk about?

4.3.1 Covid-19 Related Tweets

Our analysis of covid-19 related tweets studied tweets in mid-late November that had the key word "covid" in it. We focus on three datasets that in total make up 124049 tweets, analyzing the content of the top 30% vs bottom 30% of tweets.

Results: Covid Entity Models

In terms of entities that were being mentioned in the top and low performing tweets, I should note that there was vast similarity in the entities that *existed* between the two groups. Put more concretely, a sizable majority of the top 20 entities in the top performing group also appeared in the top 20 entities of the low performing group. However, I observed subtle differences in which entities were *emphasized*.

Using the precise criteria described in *Implementation* (Section 3.3), I picked among the most popular entities those that were significantly represented more in the top performing tweets and calculated *how much* more that entity was emphasized (in full detail in Appendix 4.1).

In summary, many of the entities that were represented more so in the top performing group across *multiple* datasets are related to **food**, **drinks, community, health, medicine, wellness, home,** and **family.** Some of these specific entities include "Wellness and Health", "Common Medication", and "Home and Family". Notably, it was interesting to see the holidays **Thanksgiving** and **Christmas** were both represented more among top tweets as well. Moreover, across all the datasets, entities related to **food** had the *highest* emphasis in the top group relative to the bottom group, indicating it was one of the most highly liked entities.

Discussion and Evaluation: Covid-19 Entity Models

Collectively, the entities emphasized more among top tweets revolved around ideas of well-being, health, caring for others, as well as staying safe and socially responsible during these times. In terms of the holidays, it is likely these entities are in reference to public health experts urging people to stay home during the holidays to prevent the spread of the virus, echoing this idea of staying safe and keeping our communities safe. With our examination of entities among the covid-related tweets, we were successful in identifying a *unifying theme* that seemed to characterize *what* exactly top performing covid tweets were emphasizing relative to low performing ones and thus what tweet content performed the best.

Results: Covid-19 Topic Models

Similar to the patterns we witnessed in the entity models, I also observed that among the topic models of the top and low performing tweets, there was again vast similarity in content, but subtle differences in what words and concepts were *emphasized*. My initial analysis of the topic models relied on manually reading the different topics, and I found that there was striking similarity even in the specific people, news stories, and events that manifested in these topic models. For instance, both topic models consistently mentioned specific political figures like "Cuomo" and "Newsom" and news stories such as the outbreak at the Mayo Clinic.

Nonetheless, I pinpointed specific terms that were represented more in the topic model of top performing tweets, flushed out in detail in Appendix 4.2 along with *how much* more that term was emphasized.

To summarize, these terms included **community, love, mom, holiday, Thanksgiving,** and **travel,** which are words that suggest caring for others and keeping safe during this pandemic. For instance, “travel” likely is related to public health and safety recommendations to reduce travel.

Moreover, interestings results corroborating this pattern arose when I also identified words that were strictly mentioned in the top topic model across *all* the datasets and none of the bottom performing topic models. They included the following words: **policy, staying, community, icu, distance, distancing, waiting.**

<u>Discussion and Evaluation: Covid-19 Topic Models</u>

Looking both at the terms emphasized more and strictly mentioned in the top tweet topic model, I saw a parallel with the theme described in our analysis of highly liked tweet entities: the idea of caring for people around us with words like **community** and **mom** and the idea of being responsible and safe during this pandemic with words like **distancing** and **travel**.

The success in our experiment with covid-related tweets lay not merely in our ability to find unifying patterns across the topic models of these different covid-related tweets, but also discovering that these patterns corroborated our analysis of the tweet *entities* too. This ultimately allows us to pinpoint concretely the ideas and themes that garnered many Twitter likes among covid-19 related tweets in mid-late November.

<u>4.3.2 Election Related Tweets</u>

Similar to the covid related tweets, I collected two datasets of tweets with the keyword “election” in late November, totaling 46226 tweets. Moreover, I also analyzed the randomly streamed tweets from the vice presidential debate, the last presidential debate, and election night, as was used in section 4.1. As was the case with covid-19 related tweets, I found there was vast similarity in the content and things being mentioned between top and low performing tweets, when investigating both entity and topic models. Nonetheless, I was able to determine key words and terms that were *emphasized* more among top performing tweets.

Results: Election Entity Models

Using the same procedure as with the covid entity models, I detail in Appendix 4.3 the *entities* that were represented more in the top topic models.

To summarize recurring patterns, one interesting entity that appeared to be represented more among top tweets across all the datasets was **Twitter**. Moreover, across the streamed datasets (debate and election nights), entities related to **food** and **drinks** appeared often. I note that this was also a pattern seen with the covid-19 related tweets. With the datasets collected in late November, the entity **Covid-19** also seemed to be highly liked.

However, perhaps more informative than trying to find entities that related *all* the datasets, some of the more interesting discoveries were specific to a single dataset. For instance, **Abraham Lincoln** had one of the highest emphases among top tweets of the Last Presidential Debate dataset (indicating that it was one of the most well-liked entities). This is reflective of a specific line spoken during the debate that night. Moreover, **Kanye West** was one of the most well-liked entities among election night tweets and is perhaps indicative of his being on the ballot that night.

Discussion and Evaluation: Election Entity Models

It was more difficult to find a unifying pattern among well-liked entities among election related tweets, at least compared to our analysis of covid-related tweets. This was partly expected because I was working with datasets more spread out in time. Nonetheless, I still made interesting discoveries about the datasets of specific nights or timeframes. Though we encountered limited success in identifying any common themes among well-liked entities, our most fruitful experiments lay in the examination of specific datasets and how our discoveries related to the specific *context* of that dataset, i.e. the last debate or election night.

Results: Election Topic Models

I use the same process from our covid-19 tweet analysis to pick out words represented more in the top performing topic model, as shown in Appendix 4.4.

To summarize my main discoveries, as with the entity models, more interesting patterns emerge when looking at specific single datasets or datasets that are closer together in time. Looking at the election tweets in late November, many of the key words like **integrity, Pennsylvania, lawsuit, concede, lose, fair, honest** revolve around known current events, especially those involving attempts to overturn the election results and sentiments questioning the integrity of the election. Moreover, looking at the debate night datasets, **hate** from the last debate dataset could perhaps be tied to **racism** and **racist** from the VP debate dataset to capture some of the discussion those nights involving divisiveness and racial tensions.

Discussion and Evaluation: Election Topic Models

Unlike our analysis of covid-19 tweets, it was difficult to identify common patterns about top performing tweet content between our entity and topic models. Nonetheless, these topic models still provide us additional clues about the topics, content, and ideas that were well liked among election tweets. As seen with our initial analysis of election tweet entities, the success of our ability to determine what exactly was well-liked content was dependent on *specific* events and contexts instead of general unifying themes as seen with the covid-19 datasets.

5. Conclusion

In our investigation of the subquestion "What tweet characteristics matter and how do they affect tweet performance?", we picked out six characteristics that had noticeable relationships with a tweet's like count: user follower count, subjectivity, tweet length, early tweet performance, and use of media/urls. Namely, top performing tweets tend to be **more subjective, longer**, **perform well early on**, and come from **users with more followers**. They also tend to **include media**(pictures/videos) and **URLs.**

While tackling the question "What general topics garner the most attention?", I discovered **sports** was a topic that generally does well, but more importantly, my expectation that certain topics can consistently do well across time is not entirely true. Instead, we discover that high performing topics can fluctuate rapidly, even within a time span of a month.

Finally, our investigation of the question "What do top performing tweets uniquely talk about?" discovered across both covid-19 and election related tweets that top performing and low performing tweets had great similarity in content but subtle differences in which terms and ideas were emphasized more. In our analysis of covid related tweets, we discovered that top performing tweets emphasized themes of well-being, of caring for oneself and others, and keeping safe during the pandemic. It was much more difficult to find common themes across all our election-related tweets, partly because of their greater spread across time. Nonetheless, we discovered interesting patterns related to specific contexts and time frames, such as the distinction of "Abraham Lincoln" as a well liked entity on the night of the last debate.

6. Future Work

In terms of immediate next steps for this project, one thing I would like to investigate is how patterns change in terms of what tweet characteristics matter when we examine specific kinds of tweets, such as tweets under a particular topic, tweets by specific users, or tweets from users with small followings. Since I was a bit hindered by the monthly API cap on streaming filtered tweets, another immediate next step would be to further attain more data regarding the popularity of general topics and study topics beyond those we examine in this study.

In terms of longer term investigations, one interesting thing to consider is figuring out what *specific* media and urls garner more attention. I discovered that tweets with media and urls tend to get more likes than tweets without them. With media, do *pictures* or *videos* perform better? What kinds of media and media content work best? I predict that simply including media can't boost one's like count; rather, there are types of content that can garner more attention that easily manifests in pictures and videos. But what *is* this content? Likewise, this same idea can be applied to urls: what are patterns about the *kinds* of links among top performing tweets? Moreover, there are numerous other tweet characteristics to explore: the use of slang, quantity of hashtags, text formality, topic models of random samples of tweets. Do these tweet characteristics have any relationship to tweet performance too, and if so, what is the strength of these relationships relative to those already studied in this paper?

8. Acknowledgements

I would like to thank the spectacular support and guidance of Professor Dobkin and the COS IW 07 Seminar as I worked on this project.

9. Honor Code

I pledge my honor that I have not violated the Honor Code

Long Ho

10. Appendix: Figures from Evaluation

1. Figures from Evaluation: Data

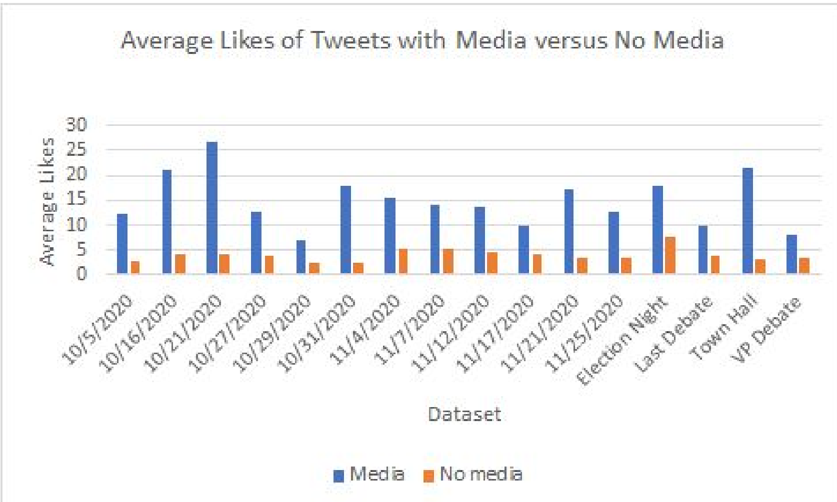
1.1 Summary of Tweets Collected Via Random Stream

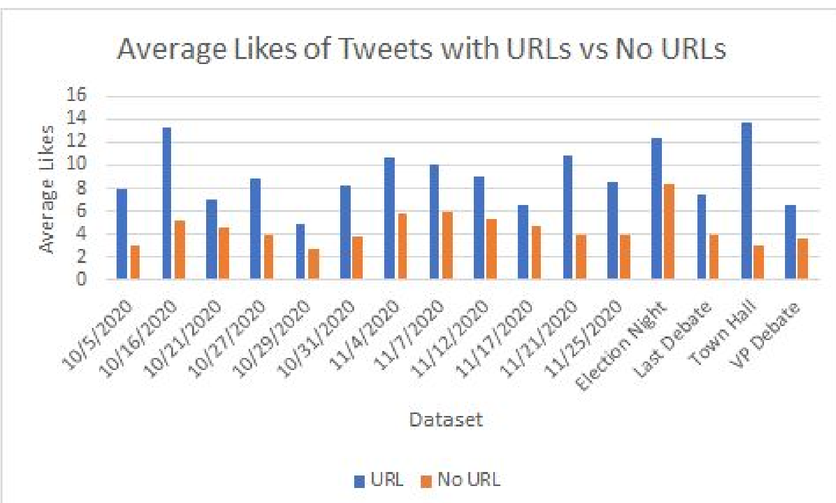
Dataset	# of Datapoints	Dataset[3]	# of Datapoints
10/5/2020	1175	12/25/20	1650
10/16/2020	4092	12/26/20	1924
10/21/2020	1222	12/27/20	1998
10/27/2020	5961	12/28/20	2873
10/29/2020	5033	12/29/20	2501
10/31/2020	1046	12/30/20	2333
11/4/2020	5631	12/31/20	1169
11/7/2020	8654		
11/12/2020	4264		
11/17/2020	1465		
11/21/2020	1594		
11/25/2020	21819		
Election Night	21895		
Last Pres Debate	12376		
Town Hall	990		
VP Debate	12776		

[3] Analysis of early tweet performance comes only from datasets in late December

2. <u>Figures From Evaluation: What tweet characteristics matter and how do they affect tweet performance?</u>

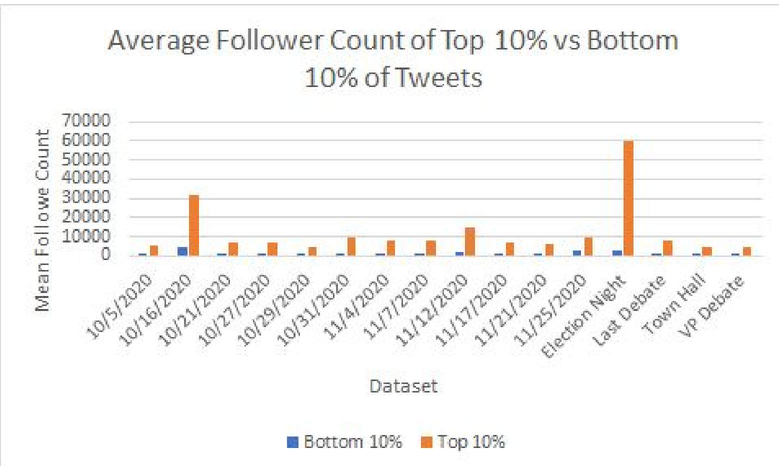
2.1 Average Likes of Tweets with Media URLs vs Tweets without them



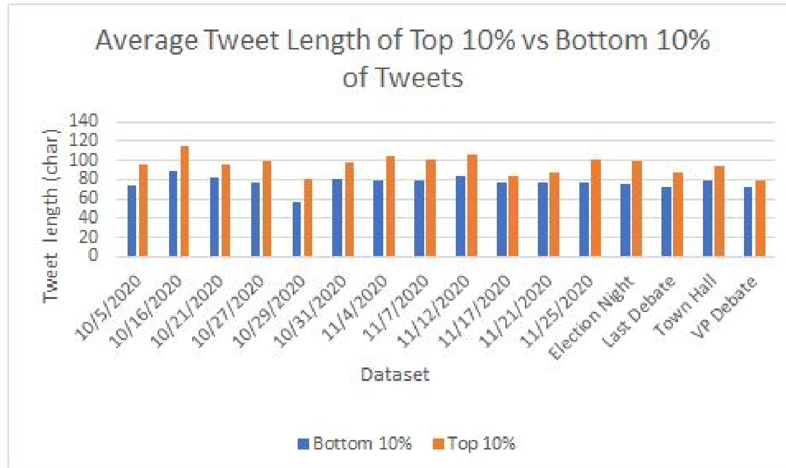


2.2 Tweet characteristics with most notable relationship with like count: Comparing Top 10% of Tweets with Bottom 10% (with regard to like count)

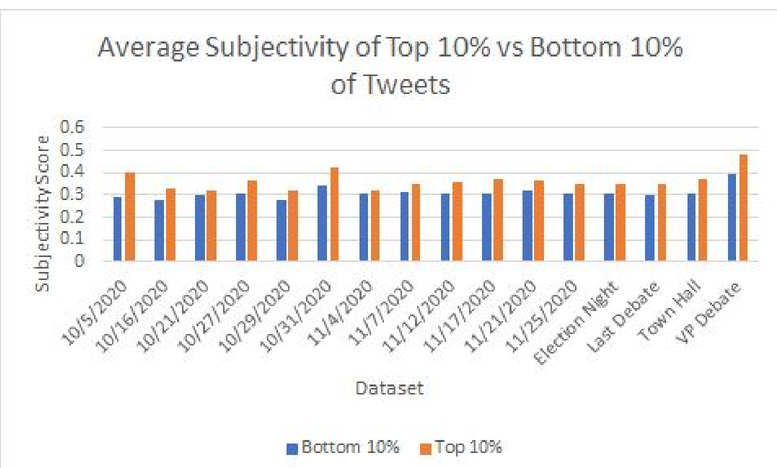
Follower Count



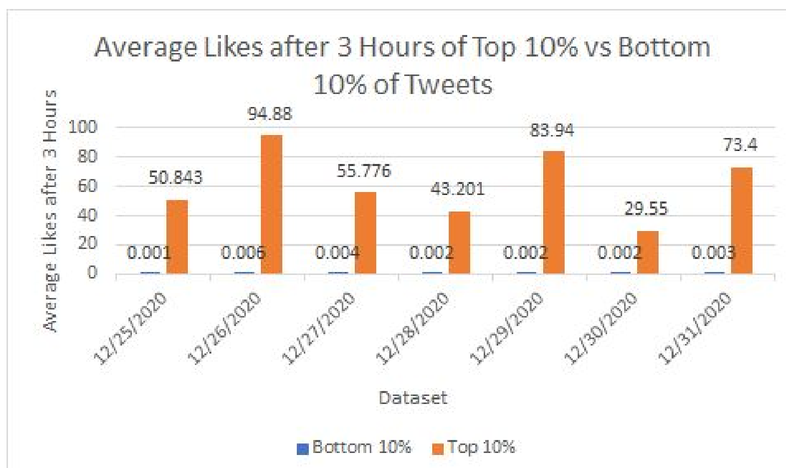
Tweet Length



Subjectivity

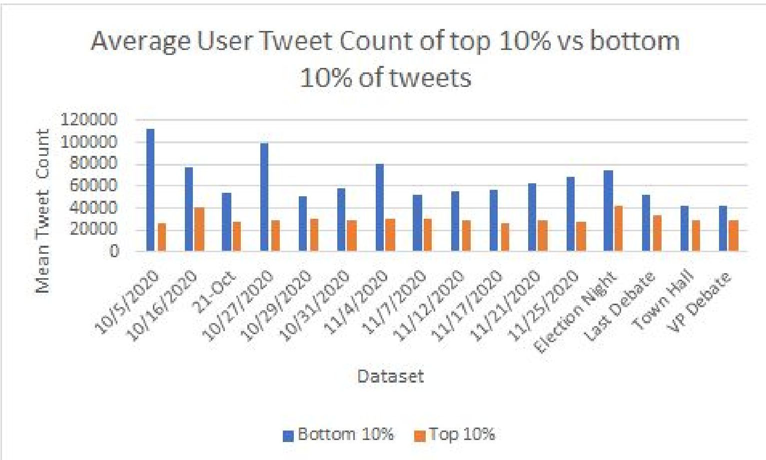


Early Tweet Performance

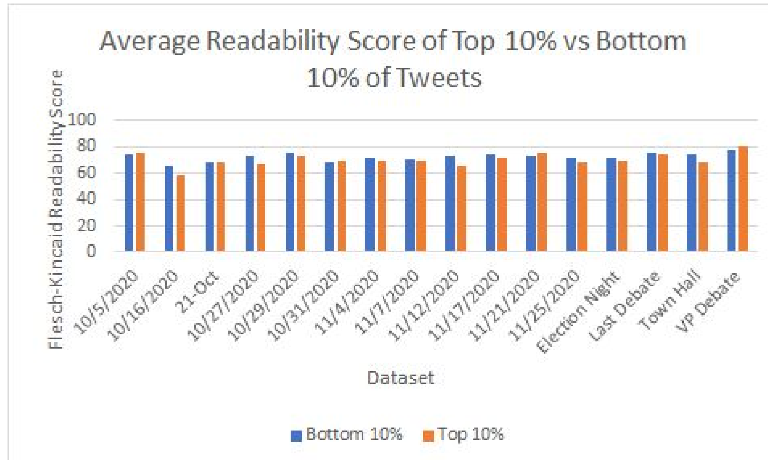


2. 3 Tweet Characteristics with least notable relationships with like count: Comparing Top 10% of Tweets with Bottom 10% (with regard to like count)

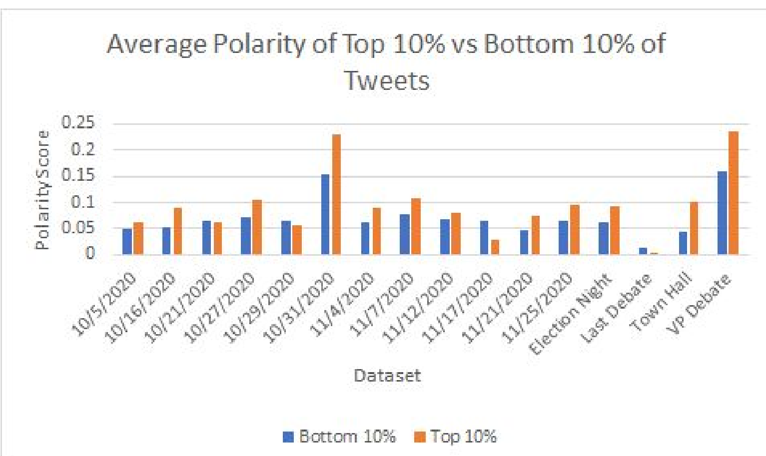
User Tweet Count



Readability



Polarity



3. Figures From Evaluation: What general topics perform the best?

3.1 Average likes by Topic of Tweets Collected By Topic/Context - Late November

Late November - Dataset 1 **Late November - Dataset 2**

Topic	Mean Likes	# of Datapoints	Topic	Mean Likes	# of Datapoints
Movies	19.32	228	Sports	15.74	1679
Sports	9.77	1797	Video games	11.64	2889
Products	9.39	790	TV Shows	8.28	3969
Video games	8.06	2167	Products	7.57	1977
TV Shows	7.30	11806	Holidays	7.20	2059
Politicians	6.77	10030	Politicians	5.10	3955
Interests/Hobbies	5.75	9400	Interests/Hobbies	4.97	14055
Journalist	5.63	2806	Movies	2.91	271
Holidays	4.64	1145	Journalist	2.16	160

3.2 Average likes by Topic of Tweets Collected By Topic/Context - Late December

Late December - Dataset 1 **Late December - Dataset 2**

Topic	Mean Likes	# of Datapoints	Topic	Mean Likes	# of Datapoints
Journalist	64.77	85	Journalist	108.72	155
Politicians	25.42	2484	Politicians	35.66	3391
Interests/Hobbies	21.08	11492	Sports	34.39	2300
Sports	17.98	3841	TV Shows	19.54	5513
Holiday	16.07	18898	Holiday	17.10	4340
Movie	15.77	498	Movie	14.42	3055
TV Shows	15.25	6021	Interests/Hobbies	12.06	15162
Video games	7.36	2587	Videogame	9.173	4612
Product	7.25	1519	Product	6.541	1951

4. Figures From Evaluation: What do top performing tweets uniquely talk about?

4.1 Entities Represented More in Top Performing Covid-19 Tweets

Datasets from late November 2020	Entity (Percent Difference *p*)[4]
Dataset 1	Food (59.03%) Community health (37.04%) Common medication (35.82%) Wellness and health (30.30%) Generic food (24.86%) Thanksgiving (20.62%) Home and family (19.35%) Christmas (15.38%)
Dataset 2	Food (61%) Generic food (46.91%) Christmas (42.42%) Thanksgiving day (36.73%) Wellness and health (32.25%) Home and family (31.68%) Common medication (30.76%) Community health (27.71%) General travel (18.17%)
Dataset 3	Drinks (50%) Food (49.76%) Wellness and health (25.23%) Generic food (20.41%) Thanksgiving day (17.28%)

[4] In parentheses is the **percent difference** of proportion *p* between the top and low performing group, where *p* is the number of occurrences of the entity divided by the total number of tweets in the group. In other words, it is a measure of *how much* more that entity was emphasized in the top performing group.

4.2. Words Emphasized in the Topic Model of Top Performing Covid Tweets

Datasets from late November 2020	Term (Percent Difference *p'*)[5]
Dataset 1	community (only in top topic model) mom (141.67%) love (141.67%) Pfizer (141.67%) Thanksgiving (129.56%) holiday (60.00%)
Dataset 2	died (only in top topic model) uk (85.00%) travel (21.35%) high (21.35%)
Dataset 3	died (134.12%) India (134.12%) flu (121.13%) money (121.13%) holiday (121.13%)

[5] In parentheses is either the percent difference of proportion *p'* for that term between the top and bottom topic models, where *p'* is the number of occurrences of the term in the topic model divided by the total number of terms in the topic model, or an indication that the term showed up *only* in the top topic model.

4.3 Entities Represented More in Top Performing Election Tweets

	Entity (% Difference *p*)[6]
Late November - Dataset 1	Twitter (19.1304%) Covid-19 (15.3077%)
Late November - Dataset 2	Insurance (59.63%) State Farm (59.56%) Business and finance (55.77%) Barack Obama (28.96%) Covid-19: (19.159%) Twitter (17.14%)
VP Debate Night (10/7/20)	Twitter (55.44%) Drinks (43.47%) Beauty (31.57%) Generic drinks (26.95%)
Last Presidential Debate Night (10/22/20)	Twitter (69.76%) Abraham Lincoln (61.82%) Drinks (42.20%) Food (28.57%) Services (28.02%) Generic food (26.67%) Insurance (19.35%)
Election Night (11/3/2020)	Generic Drinks (68.24%) Kanye West (65.08%) Drinks (53.39%) Joe Biden (41.69%) Twitter (36.36%) Consumer Packaged Goods (35.00%) Donald trump (28.67%) Food (24.51%) Generic Food (20.20%)

[6] p is as described in Appendix 4.1

4.4 Words emphasized in the top topic models of election-related tweets

	Term (% Difference *p'*)[7]
Late November - Dataset 1	integrity (148.028%) reelection (121.13%) landslide (68.23%) Detroit (68.23%) party (39.43%) pennsylvania (30.30%) lawsuit (30.30%) republican (28.00%)
Late November - Dataset 2	joebiden - only in top topic model count- only in top topic model lie- only in top topic model leader-only in top topic model lose - only in top topic model fair (32.56%) honest (32.56%) presidentelect (32.56%) covid (32.56%) concede (18%) transition (24.103%)
VP Debate Night (10/7/20)	moderator (170.984%) minute (100.84%) racism (92.00%) racist (92.00%) vicepresidentialdebate2020 (35.57%)
Last Presidential Debate Night (10/22/20)	realdonaldtrump (159.00%) running (159.00%) hate (159.00%) american (39.63%) woman (33.64%)
Election Night (11/3/2020)	result (166.38%) ill (72.23%) american (49.85%)

[7] p' is as described in 4.2

11. Bibliography

[1] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*. 2013, doi: 10.1145/2487788.2488017.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer," *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*. 2011, doi: 10.1145/1935826.1935845.

[3] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, doi: 10.3115/v1/p14-1017.

[4] "Twitter API." https://developer.twitter.com/en/docs/twitter-api (accessed Jan. 05, 2021).

[5] "API Reference — TextBlob 0.16.0 documentation." https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment (accessed Jan. 04, 2021).

[6] "textstat." https://pypi.org/project/textstat/ (accessed Jan. 04, 2021).

[7] "THE FLESCH READING EASE READABILITY FORMULA." https://readabilityformulas.com/flesch-reading-ease-readability-formula.php (accessed Jan. 04, 2021).

[8] "Cohen's D: Definition, Examples, Formulas - Statistics How To," Oct. 15, 2016. https://www.statisticshowto.com/cohens-d/ (accessed Jan. 04, 2021).

[9] "scipy.stats.ttest_ind — SciPy v1.6.0 Reference Guide." https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html (accessed Jan. 04, 2021).

[10] "S.3.2 Hypothesis Testing (P-Value Approach)." https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/hypothesis-testing/p-value-approach (accessed Jan. 04, 2021).

[11] "18.1 - Pearson Correlation Coefficient." https://online.stat.psu.edu/stat509/node/156/ (accessed Jan. 04, 2021).

[12] "WordNet Interface." https://www.nltk.org/howto/wordnet.html (accessed Jan. 04, 2021).

[13] "nltk.corpus package — NLTK 3.5 documentation." https://www.nltk.org/api/nltk.corpus.html (accessed Jan. 04, 2021).

[14] "sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn 0.24.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (accessed Jan. 04, 2021).

[15] "collections — Container datatypes — Python 3.9.1 documentation." https://docs.python.org/3/library/collections.html (accessed Jan. 04, 2021).

[16] D. Greene, "Topic Modelling With Scikit Learn," [Online]. Available: http://derekgreene.com/slides/topic-modelling-with-scikitlearn.pdf.

[17] "sklearn.decomposition.NMF — scikit-learn 0.24.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html (accessed Jan. 04, 2021).

[18] "Gensim: topic modelling for humans." https://radimrehurek.com/gensim/models/word2vec.html (accessed Jan. 04, 2021).

[19] J. Cohen, "Statistical Power Analysis for the Behavioral Sciences." 2013, doi: 10.4324/9780203771587.

[20] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Frontiers in Psychology*, vol. 4. 2013, doi: 10.3389/fpsyg.2013.00863.