Analysis of Twitter's Trending Topics in Comparison to Top News Headlines

University of California Santa Cruz

Surya Keswani
Jack Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, California
sukeswan@ucsc.edu

Zion Calvo
Jack Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, California
zcalvo@ucsc.edu

Donald Stewart
Jack Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, California
dolstewa@ucsc.edu

ABSTRACT

With the reliance on social media in the world today we explore the types of content that people consume on social media. This presents the question: are people who use social media as a primary source of news fed credible information by these platforms? Are the content people consume worthy of being "trending," or are their feeds primarily pop culture references? Is pop culture news important news for users to be receiving? This article addresses the trends, and the type of content users consume on Twitter.

MOTIVATION/OBJECTIVE

In 2006, Twitter launched as a microblogging social network. Today, Twitter services millions of people with content that ranges from news to memes. In 2018, 71% of Twitter's users sourced most of their news from the platform [1]. The platform's explosive growth over the last decade and a half has caught the attention of many researchers. They ask: does Twitter censor its trending topics list, and should it be allowed to continue this practice? Controversies have emerged when important and newsworthy topics like the elections or Ebola do not become trending, but Justin Bieber is frequently trending on the platform [2]. This project aims to analyze what Twitter flags as a trending topic and what factors contribute to a topic receiving the trending label. As more and more people source their news headlines from Twitter, we compare the trending topics to the top news headlines put out by some of the largest news sources in the world. This project sheds light on the ambiguous algorithm that classifies topics as trending on the platform.

ETHICAL ISSUES

As discussed in the previous section, Twitter has been accused of censoring its Trending list on the platform. This censoring has been criticized because millions of people use the platform as their primary news source. Censoring trending topics to favor pop culture over vital news reports raises concerns about what the company is promoting and whether these unethical practices should continue. This analysis aims to compare the trending topics Twitter aggregates with a curated list of the top news headlines. The results in this paper reveal some of the factors taken into account with the Trending algorithm and what Twitter classifies as "news-worthy" material.

DOMAIN/DATASET

The "Twitter Trending Topics" dataset was utilized for this project. It contains trending topics from 03/01/2011 to 03/08/2011. This dataset originally consisted of five fields: Md5 Hash, Date, Topic, and Type. We also append three new fields: number of Tweets, Unique Users, and Average Traffic, to create additional numeric features to investigate further the algorithm Twitter used (at the date specified above) to determine what factors contribute to a topic becoming trending.

Md5 Hash: Each of the 1,036 trending topics has a hash assigned to it. There are 1,036 files with tweet

information. Each file is titled with an Md5 hash and corresponds to the trending topic hash.

Date: Each topic is assigned the date of when it became trending. The dates in the original dataset are formatted as yyyyMMdd. The read_data() function reformats the dates as Day MM/DD/YYYY.

Topic: The trending Twitter topics, ranging from news stories to people to hashtags. The topics were classified by the following:

1. **News:** a trending topic can be categorized as *news* produced by a newsworthy event that major news outlets either had reported by the time the trend popped up or will report it soon after it broke on Twitter [5]
2. **Ongoing events:** a community of users tweeting about an ongoing event as it unfolds. These events have a wide range that includes soccer games, a keynote presentation by Apple, a music festival, and conferences. [5]
3. **Memes:** viral ideas initiated by either an individual or an organization, usually popular enough to spread something widely. Without being newsworthy or a mainstream event that a large audience is following, the event makes it to a large community of users for being funny or attractive to them. It can be from a funny message by a teen heartthrob such as Justin Bieber to a protest leader's request to spread a message in support of a plea or complaint. [5]
4. **Commemoratives:** when users congratulate a celebrity for their birthday, celebrate the anniversary of a certain event or person, or it is a memorial day. [5]

Type: Each of the 1,036 topics has been analyzed by RMIT and manually categorized. The manual annotation consists of one of the four classes in taxonomy: news, ongoing-event, meme, and commemorative.

Tweets: The number of times a topic was tweeted about during the dataset collection period. The number of tweets has been extracted from the tweets folder, looking at each topic's corresponding Md5 hash file.

Unique Users: The number of unique users that tweeted about the corresponding trending topic during the dataset collection period. The unique users have been extracted from the tweets folder, looking at each topic's corresponding Md5 hash file.

Average Traffic: The average of the # Tweets and Unique Users. This score serves as another way to measure the growing trends.

	Md5 Hash	Date	Topic	Type	# Tweets	Unique Users	Average Traffic
0	e86dba234c7429d9aea70c5f104dbebc	Sunday 3/6/11	#mileyonsnl	ongoing-event	1358	536	947.0
1	fe2dca2bdfca21e6c83567a531d6534e	Sunday 3/6/11	Leap Year	ongoing-event	1496	1335	1415.5
2	cf90d03111f55e55ff9c7b5ef079fa30	Saturday 3/5/11	Sober Valley Lodge	meme	1490	1376	1433.0
3	748089790952e9dbfb174b7ef39adf2b	Thursday 3/3/11	Howard Davies	news	557	452	504.5
4	3b7dd11b6654dd5f35745285ededdbd9	Thursday 3/3/11	Sky News	news	1483	1203	1343.0
...	...	...	...	...	...	...	...
1031	05d739a093794343728b66d5802db10b	Wednesday 3/2/11	Antonio Damasio	ongoing-event	197	166	181.5
1032	8669b32af04ae65b88b38df5662545ea	Saturday 3/5/11	Sheila Mello	ongoing-event	190	170	180.0
1033	4e49ec9f0f036f91a5b34278d15484f3	Friday 3/4/11	Gareca	ongoing-event	284	222	253.0
1034	d93d84ae006f530cbf2b33ebf824d48e	Monday 3/7/11	Patrícia Poeta	ongoing-event	309	283	296.0
1035	98d79b7652e365b4a4b260ca2bb5583d	Thursday 3/3/11	PIB	news	1497	1176	1336.5

1036 rows × 7 columns

Figure 1: **An example of our DataFrame**

Creating comparison data to have a context of pressing events throughout the time period, we web scraped headlines from five new sources: The New York Times, Wall Street Journal, Fox News, Washington Post, and CNN. This data set consisted of the Source, Date, and Headline.

MODELS/ALGORITHMS

To analyze Twitter's machine learning algorithm required data regarding their most prevalent subjects and data on the most popular headlines. The "Twitter Trending Topics" dataset was modified, a headline dataset was created via web scraping, and visualizations were created through Google's facets library.

Dataset Modification: By iterating through each set of hashed data, the number of tweets, unique users, and average traffic was obtained and added to each trending topic in the data set.

Web scraping: This method was essential for obtaining all the relevant headlines. "Urllib," "html5lib", and "BeautifulSoup" libraries were utilized for opening, parsing, and recording the HTML from each web page, respectively. Thus, adding a link from each news source provided all the information necessary to find headlines and record them into a CSV file. This file would then be used to compare the modified dataset.

Facets: This library focuses on providing interactive visualizations to the user. This library was especially useful for the analysis of the dataset. By providing this

library with a “Pandas” dataframe, the Facets library manipulates it into a graph that can instantly alter its axis with any field of the dataset. It helped reveal the most common tendencies within the most popular topics.

RESULTS AND ANALYSIS

The figure below shows the top 5 trending topics classified by Average Traffic, # of Unique Users, and # of Tweets per topic. This snapshot paints a clear picture of what is classified as Trending on the social media platform. During the week this dataset was collected, March 1, 2011 - March 8, 2011, a civil war broke out in Libya. Based on the three popularity metrics measured (Average Traffic, # of Unique Users, and # of Tweets), none of the top 5 charts cover the Libyan Civil War as trending. The Libyan Civil War is not mentioned in the top 35 Trending topics for any popularity measure. Of the top 35 news headlines we have curated from a diverse pool of news sources, 16 of the 35 headlines cover the Civil War. The trending topics seem to focus more on popular culture as memes dictate the top trending topics.

Top 5 Trending Topics based on Average Traffic Below:

	Md5 Hash	Date	Topic	Type	# Tweets	Unique Users	Average Traf
607	f6c5c0ae6c636b204045d21267608be5	Saturday 3/5/11	YOURMAN	meme	1492	1479	148
711	6401813085837e9f1f8b27eae98ac4f3	Saturday 3/5/11	Mountain Dew	meme	1493	1454	147
1011	5eb1c412235828a21e3767fc12ab9975	Friday 3/4/11	Fast Times	ongoing-event	1498	1441	146
10	e8611c0126305e7a3af7ebe8317dc699	Tuesday 3/1/11	RIP Jane Russell	news	1497	1441	146
408	257c941753478852a2436136bc72ab61	Wednesday 3/2/11	TOP10 Profile STALKERS	meme	1496	1439	146

Top 5 Trending Topics based on Unique Users Below:

	Md5 Hash	Date	Topic	Type	# Tweets	Unique Users	Average Traffic
607	f6c5c0ae6c636b204045d21267608be5	Saturday 3/5/11	YOURMAN	meme	1492	1479	1485.5
711	6401813085837e9f1f8b27eae98ac4f3	Saturday 3/5/11	Mountain Dew	meme	1493	1454	1473.5
877	463636c775a2ba72afdfd29d240b4a19	Wednesday 3/2/11	Nancy Grace	meme	1464	1445	1454.5
736	c80e68f4191a8150a52cf5cf43a6bdaf	Wednesday 3/2/11	Still Winning	meme	1462	1442	1452.0
10	e8611c0126305e7a3af7ebe8317dc699	Tuesday 3/1/11	RIP Jane Russell	news	1497	1441	1469.0

Top 5 Trending Topics based on # of Tweets Below:

	Md5 Hash	Date	Topic	Type	# Tweets	Unique Users	Average Traffic
70	050feefcdd7d85023c61753474848a48	Monday 3/7/11	Rosenmontag	meme	1500	1298	1399.0
114	4a4de7aa4553a1bd4fea73042d479f6b	Tuesday 3/1/11	Robot Chicken	ongoing-event	1500	1279	1389.5
460	6a5c42fc43e58b8ed67ab5da8d965c51	Wednesday 3/2/11	Wild Thing	ongoing-event	1500	1429	1464.5
602	6de9a283570262f6978fdbbab2c5bc41	Wednesday 3/2/11	Kalla	news	1500	1203	1351.5
809	9ff084e11fff0932581f87e9a5e19c1a	Thursday 3/3/11	Wellington Paulista	ongoing-event	1500	987	1243.5

Figure 2: **Top 5 Trending Topics based on Average Traffic, Unique Users, and # of Tweets**

The next figure below is a snippet of some of the top news headlines curated. The following figure highlights the main headlines the New York Times put the week this Twitter dataset was created.

14	NY Times	Tuesday 3/1/11	In U.S.-Libya Nuclear Deal, a Qaddafi Threat F...
15	NY Times	Wednesday 3/2/11	Arab Unrest Puts Their Lobbyists in Uneasy Spot
16	NY Times	Thursday 3/3/11	Justices Rule for Protesters at Military Funerals
17	NY Times	Friday 3/4/11	Obama Tells Qaddafi to Quit and Authorizes Ref...
18	NY Times	Saturday 3/5/11	Qaddafi Militia Storms Key Town Held by Libyan...
19	NY Times	Sunday 3/6/11	Business Side of Egypt's Army Blurs Lines of A...
20	NY Times	Monday 3/7/11	A Libyan Leader at War With Rebels, and Reality

Figure 3: **Top Headlines from New York Times**

The dataset also reveals that Twitter follows a pre-set distribution when classifying topics as Trending. The following figure shows that approximately 55% of the Trending topics are ongoing events, 20% are memes, 20% are news, and 5% are commemorative. This distribution fluctuates approximately ± 5%, as seen from the day-to-day data. Furthermore, memes hold many of the top spots for trending topics despite being a minority of the trending topics.

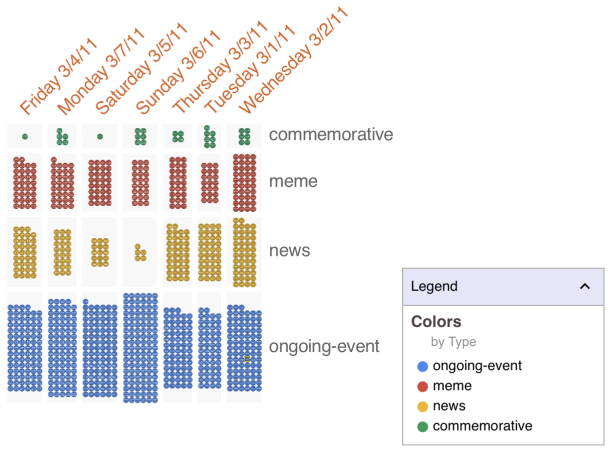

Figure 4: **Tweets organized by type of Tweet**

The largest factor contributing to topics holding many of the top trending spots is the number of times a topic is tweeted about. The figure below illustrates that approximately 50% of trending topics have 1400 to 1600 tweets. These tweets are not necessarily unique, and a single user can tweet multiple times about a topic. The volume of tweets is an important factor in determining if a topic becomes a trending label. The unique user count is not as important and negatively affects the average

traffic measure.

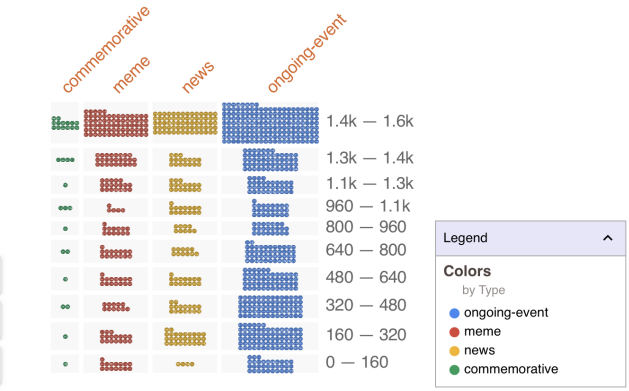

Figure 5: **Tweets organized by type of tweet and how many tweets for that trending topic**

Some of the controversies sparked around Twitter's Trending feature is user manipulation. There have been instances in which users have spammed the platform with tweets to create a large volume of tweets about certain topics artificially. This influx of misleading content forced Twitter engineers to alter the trending algorithm to account for users attempting to "game the system"[2]. It is important to note that it is unclear whether this dataset was collected before or after altering the Trending algorithm. Twitter does not reveal how the algorithm is altered, only that it is updated and optimized as the platform continues to grow and develop.

CONTRIBUTION

Zion Calvo: Cleaned dataset and scrapped the tweet files for information such as the number of tweets per topic and number of unique users per topic.

Surya Keswani: Found the dataset and cleaned the data. Processed into a pandas data frame and built a notebook to run data through Facets library.

Donald Stewart: Created the comparison dataset. Started by web scraping headlines throughout the time period, the "Twitter Trending Topics" dataset was captured.

FUTURE WORK

For future work, we would like to see a comparison among different dates/years for trending tweet types. Because our data was from one week in 2011, we feel that it would be interesting to compare and explore the following questions/ideas:

1. How consistent are the Twitter trending types throughout the given year? Does Twitter modify its trending tweet algorithm during certain parts of the year?

    a. We can measure this by analyzing different distributions of weeks/months during the year. Would the modification be used to show more memes, news, or pop culture?

2. How consistent are the Twitter trending types throughout history?

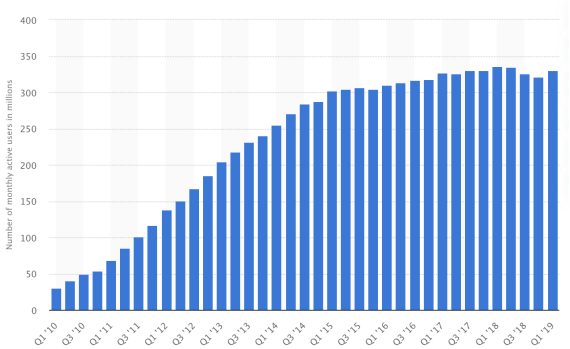

Figure 6: **Twitter user growth from Q1 of 2010 to Q1 2019**

    a. Twitter was created in 2006. With Twitter growing up to 330 million users in 2019 versus the 117 million users in our 2011 dataset, we feel that Twitter must have changed its trending algorithm a few times during its growth [3].

    b. We would like to compare the year-by-year trending topics to see different distributions of tweet types from our dataset.

We would like to find a way to classify tweets in our dataset more easily with many more users and many more tweets per day.

Our data set was individually classified into their types. One idea for future work would be to automate this

process with a high success rate so that we can further analyze tweet trending data. This would require natural language processing and retrieving each trending tweet text to check if the tweet is a meme, news, or commemorative event.

We would also like to see the overall trending attitude. Are the trending topics grief or happiness? Criticism or praise? Having that information would be interesting to analyze as we can explore more of the attitude people have when they send out tweets. We would want to web scrape each tweet's content and use IBM's Watson Natural Language Classifier [4] to process the overall sentiment and emotion of the tweet. We think we would find interesting results that we could analyze using this.

Another idea that we had while gathering data for our project was to see trending topics based on location. We found that some trending topics were in different languages. We would like to see if places with different languages have different trending topics and distributions of tweet types. If we did find a difference, it would be beneficial to further research the above questions/ideas in different languages. This, again, would require natural language processing for different languages to analyze the tweets in various types.

If possible, we would also want to compare trending topics via Twitter versus the likes of Facebook, Instagram, Tumblr, etc. We would like to see if different social media platforms have different distributions or if different social media platforms have different content for the same event. We think exploring the differences in trending topics in different social media will have a large number of future works that can be explored.

REFERENCES

[1] Elisa Shearer and Katerina Eva Matsa. 2019. News Use Across Social Media Platforms 2018. (December 2019). Retrieved March 5, 2020, from https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[2] Tarleton Gillespie. 2011. Can an algorithm be wrong? Twitter Trends, the specter of censorship, and our faith in the algorithms around us. (October 2011). Retrieved March 5, 2020, from https://culturedigitally.org/2011/10/can-an-algorithm-be-wrong/

[3] Lahle Wolfe. 2019. Twitter Statistics: How Many People Use Twitter. (October 2019). Retrieved March 5, 2020 from https://www.thebalancecareers.com/twitter-statistics-2008-2009-2010-2011-3515899

[4] Anon.Retrieved March 5, 2020 from https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

[5] A. Zubiaga, D. Spina, V. Fresno, R. Martínez. *Real-Time Classification of Twitter Trends* Journal of the American Society for Information Science and Technology (JASIST). In Press.