The Topics that Influencers Emphasize More on Twitter

- Husnul Abid

1. Introduction

People all around the world are now spent huge amount of time on various social platform. Twitter is one of the largest social media among them. Twitter is a wonderful tool for both communication and broadcasting which allows business to grow, impact other people. Twitter influencers are immensely powerful and distinctive because, at the end of the day, their followers value their opinions or personalities regarding current events/cultural moments. One study shows that 40% of Twitter users admit to being directly influenced by Tweets.

As the influencers are leaders of the digital world, it will help us to understand how they are shaping the world by influencing people on various issues. After finding the theme of the tweets and analyzing the topics, we can also gather knowledge on the social impacts. The purpose of this analysis is to identify in which topics top influencers discuss more on twitter. Topic extraction from a tweet is the prime goal of this project. Also, finding the relationship among the influencers are also a part of the project.

2. Methodology

We have used public API data from Twitter to get the tweets of top 50-100 influencers. Also, we fetched the 'following list' that helped us to use social network analysis by observing the connectivity of the top influencers. After scraping, we analyzed these data using various NLP techniques to identify trends and patterns. By accumulating the tweets, we used Latent semantic analysis (LSA) to extract most frequent topics (three topics) discussed by the influencers. We also demonstrated top five topics that were discussed on each year by Bill Gates.

We have used python as a programming language. Scikit-learn library is used to extract topics and Networkx library has been used to perform network analysis. Pandas has been used for processing and cleaning the data. We used Matplotlib for visualization.

3. Data Scraping

Twitter data is easily accessible as it supports API to fetch data. The APIs are very well-documented which helped us to get data. To access Twitter API v2, we had to apply through it's developer portal which took around a day to receive the consumer key and secret. They have also provided API token which can be used in any application. We have used this token using python to scrape following list of top influencers. The reason of fetching following list is to perform network analysis between high-profile figures.

We are defining top influencers as the most-followed Twitter accounts. The list is publicly available in Wikipedia for top 50 accounts. In this list, there are few accounts that are managed by a group of people such as NASA, Real Madrid football club and some accounts are news broadcasting channel such as BBC, CNN. We excluded these accounts as it would give us random news and specific tweets regarding niche fields. After pruning, we have 34 influencers in out list.

We collected as many tweets as we could from each of the 34 influencers. The tool that helped us to extract tweets is named Twint. This is a sophisticated Twitter scraping application built in Python that enables Tweets to be extracted from Twitter profiles without utilizing Twitter's API. Twint is also an open-source python library. We were able to collect 405,498 tweets from all influencers in 2 hours using this tool. Twint provided us tweets as a csv file of each individual account and we used python pandas to merge all the tweets in a single csv file.

Using API token of twitter endpoints, we fetched the following list of the influencers. The goal is to analyze which influencers are followed by other influencers. So, it would be efficient to just checking whether account A is being followed by account B. However, there is no direct API for such query. So, we had to extract all the accounts that an influencer follows which is very time and resource consuming which took almost a day.

4. Data Processing

As a result of the wide variety of tweets, data processing is unavoidable. There are lots of tweets that contain urls. This can be very distractive to analyze the sole purpose of the tweets. So, we have removed all the urls from all tweets at the first stage. Additionally, we deleted all the tweets that is not in English as we performed our analysis just with the English words. At this stage, we had around 270 thousand tweets including their published date from 34 influencers as our raw tweet dataset. The tweet dataset looked like the following figure 1 (a).

	account	date	tweet
2	akshaykumar	29/10/2022	Khush raho Aqsa beta. God bless you.
5	akshaykumar	26/10/2022	Thank you for celebrating your Diwali with us ...
11	akshaykumar	26/10/2022	Thank you
17	akshaykumar	24/10/2022	Iss saal Diwali hogi Action aur Adventure se b...
20	akshaykumar	22/10/2022	All set? Because the biggest adventure of the ...

Figure 1 (a): Partial tweet dataset

	following_account	account
590252	BTS_twt	justinbieber
590266	katyperry	justinbieber
619102	elonmusk	justinbieber
753481	narendramodi	justinbieber
763898	Harry_Styles	justinbieber

Figure 1 (b): Partial following dataset

We have also fetched following data for each influencer to perform network analysis on that. Unfortunately, as we needed to all the following data, we cleaned the dataset and kept just the data regarding the influencers account. We have around 291 rows in this dataset which looked like figure 1 (b).

5. Network Analysis on Influencer

Visualizing how well the influencers are connected to each other is the main purpose of this section. We defined each account as nodes. And if one account (suppose M) follows another account (suppose N), then we draw a directional edge from M to N. Directional edge will distinguish the relationship more than a non-directional edge. If both accounts are followed by each other, we will have two directional edge between themselves. We first created these edges along with the nodes using networkx library and plotted the graph with matplotlib library. We have used circular layout for the networkx as the number of nodes and edges are large amount. The whole network is shown in figure 2.

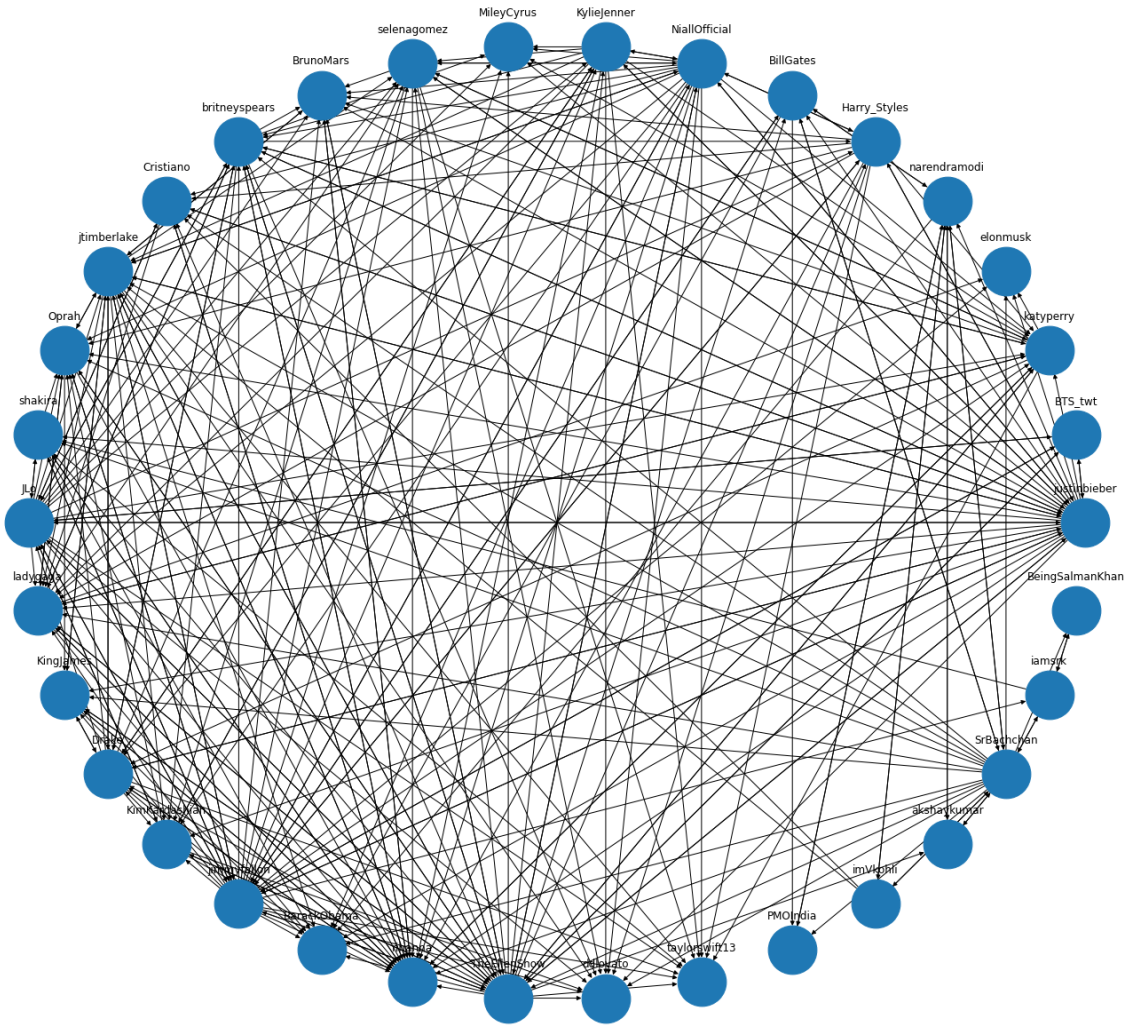

Figure 2: Top influencers followed by each other

The network is very dense on left side of the circle rather than the right side as the nodes that has on left has more degree. In graph theory, the degree of a node is the number of edges that are connected to the node. So, if an account is being followed by many other accounts, or the account follows many other accounts, the degree of the account will increase. We have also calculated the degree of each node and sorted top 10 accounts according to degree. 'TheEllenShow' has the most degree among all which means either this account follows or followed by many other accounts. We have also performed network analysis of a single account on whom they follow or followed by which accounts. By changing the account name in the code, this will help us visualizing the following/follower graph of an account easily. The graph for following/follower of 'Rihanna' has shown at figure 3.

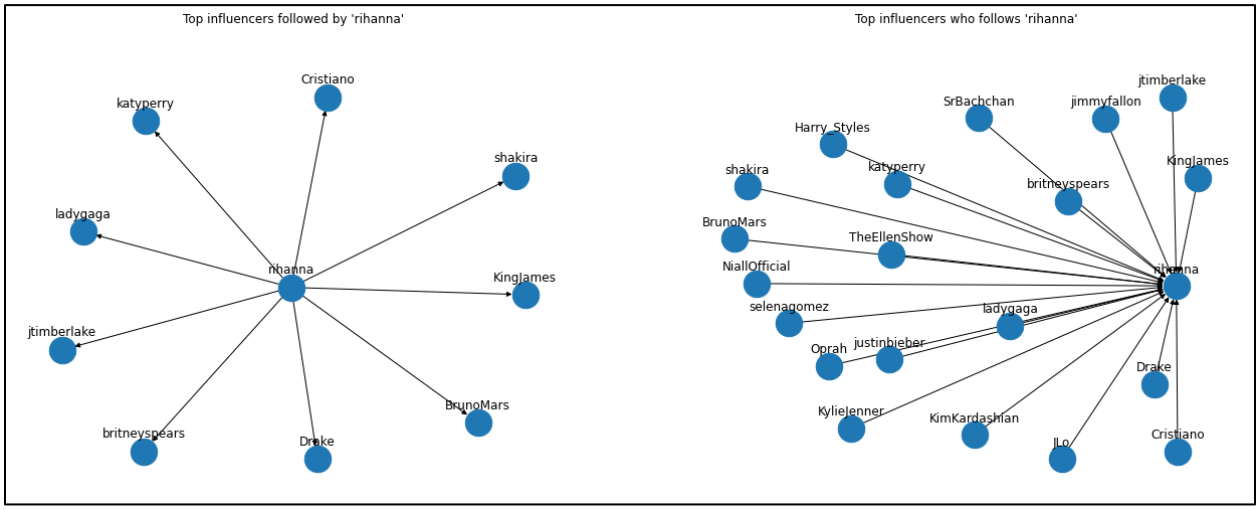

Figure 3: Following/follower graph of 'Rihanna' by only top influencers

6. Topic Extraction

As we wanted to find out which topic the influencers discuss about more often, topic extraction is the most important goal of this project. To have an overall idea, we plotted a graph with the most frequent words in all tweets which is shown in figure 4.

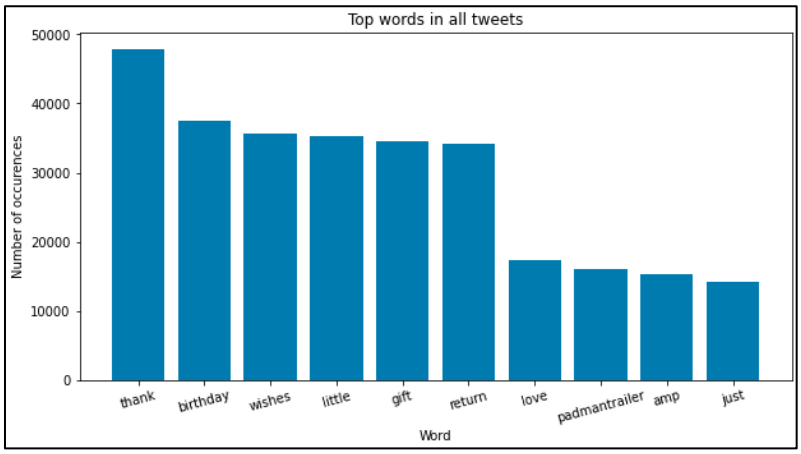

Figure 4: Top words in all tweets

We can see that, most frequent words are thank, birthday, wishes, gift, love and so on. From the word type we can come to conclusion that, influencer post many tweets thanking others and birthday wish is very common phenomenon. Another interesting word (or hashtag) is 'padmantrailer'. Pad Man is a comedy-drama by Akshay Kuman who is a top influencer in twitter. So, it seemed like a marketing strategy with this hashtag by him.

Frequency of words don't convey any strong messages about the topic of the tweet. We have used Latent Semantic Analysis (LSA) to extract the main theme of the texts. LSA entails converting a group of unstructured texts into structured data. The words we use while writing anything akin to text are not randomly selected from a lexicon. Instead, we consider a theme (or issue) before selecting words that will help us communicate our ideas to others in a more meaningful way. Typically, this theme or subject is regarded as a latent dimension. The main purpose of LSA is to find out this theme using latent dimension.

We need to feed vectorized text sample to LSA by using CountVectorizer. It breaks down a sentence or any text into words by performing preprocessing tasks like converting all words to lowercase, thus removing special characters. After that, we reduce dimension and fit into LSA model which gives us topic category and LSA count of that topic. Also, we have analyzed tweets of the Bill gates yearly with the same LSA method.

7. Result

We have extracted 5 topics for each influencer and each topic had 3 keywords. We get some idea about the topics of the tweets by influencers. Some keywords of the topics might seem little meaningless, but overall it narrates that, most influencers tweets substantially in their area. Elon Musk discuss mostly about Tesla and SpaceX. On other hand, Cristiano Ronaldo tweets about football and Real Madrid. Most common extracted topics from few influencers are show in figure 5.

```
BarackObama  -> Topic 1:  president obama watch
BarackObama  -> Topic 2:  health care americans
BarackObama  -> Topic 3:  change actonclimate climate
BarackObama  -> Topic 4:  climate change health
BarackObama  -> Topic 5:  senate doyourjob leaders


BillGates  -> Topic 1:  world great polio
BillGates  -> Topic 2:  world know did
BillGates  -> Topic 3:  energy climate need
BillGates  -> Topic 4:  health global countries
BillGates  -> Topic 5:  people live africa


elonmusk  -> Topic 1:  amp tesla spacex
elonmusk  -> Topic 2:  tesla model car
elonmusk  -> Topic 3:  spacex erdayastronaut ppathole
elonmusk  -> Topic 4:  yes teslaownerssv flcnhvy
elonmusk  -> Topic 5:  good just great
```

Figure 5: Extracted topics from tweets

After extracting topics from all influencers tweets we dig deep with just one account and analyze the tweets per year. Also, we can analyze any other account by changing the account name. In our analysis we chose Bill Gates. Bill Gates became philanthropist on 2006 and contributes in public health, vaccination of diseases, awareness about climate change, and so on. As he has interest in vast areas, we chose his account to analyze further.

Extracted topics from Bill Gates tweets shows that how different global crisis and diseases has changed over time and he tweet on the most concerning facts in each year. In figure 6, we can see that, he tweets

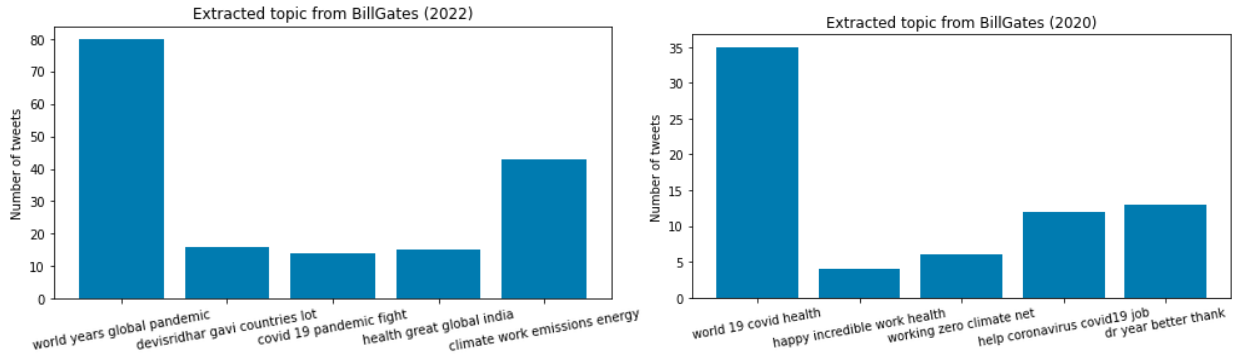

Figure 6: Extracted topics from recent tweets of Bill Gates (2022, 2020)

about covid-19 pandemic and the health issue in recent years. In 2020, all the topics was related to coronavirus. In 2020, the pandemic topic remains also highest but the concern of climate change and energy crisis also rises. So, we can get the recent world situation from his tweets.

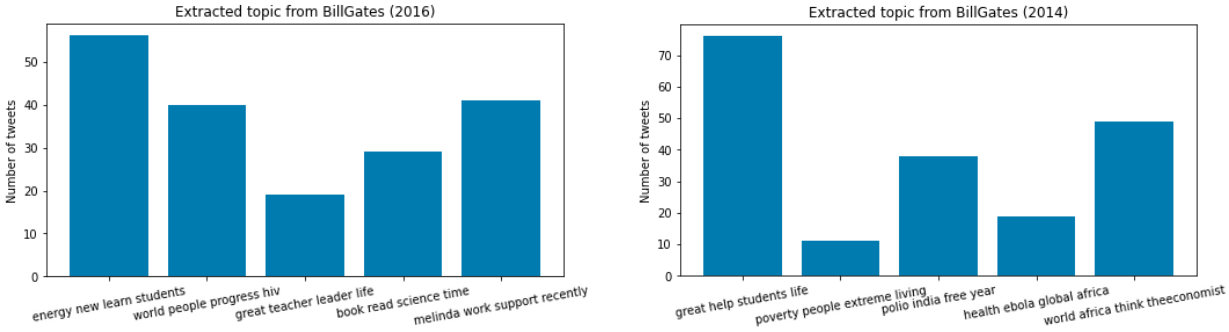

Figure 7: Extracted topics from tweets of Bill Gates (2016, 2014)

Some exciting topics are also visible from 2016 and 2014 at figure 7. We can see tweets related to HIV progress on 2016. Polio free India in 2014 which aligns with the reality. Also, we all know about the deadly virus disease named Ebola which started on 2014. Tweets from Bill Gates shows he was discussing about the disease on the same timeline.

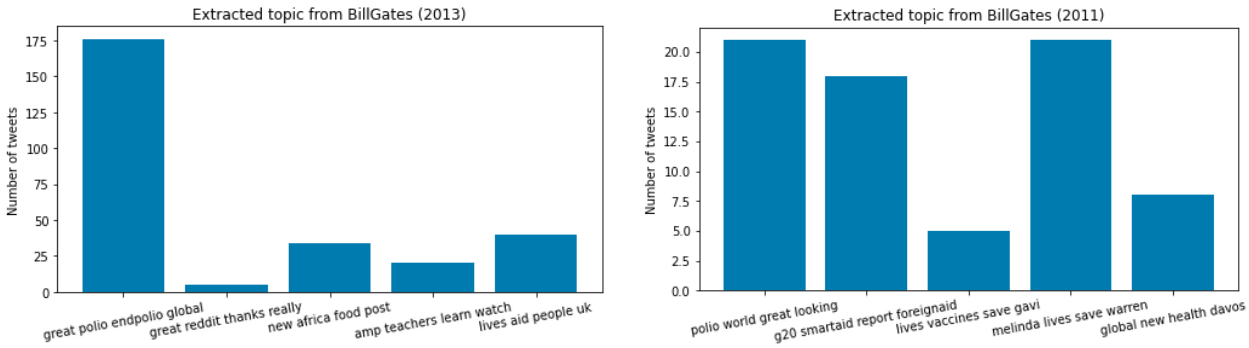

Figure 8: Extracted topics from tweets of Bill Gates (2013, 2011)

As mentioned earlier that India passed their first polio free year on 2014. If we look carefully at figure 8, we can see that Bill gates was discussing about polio in his tweets. In 2013, around 175 tweets were about ending polio and we see the result just after the next year. So, the influencers successfully influence other people to spread awareness and work on their specialized field who are making the world a better place to live for future generations.

8. Conclusion

Through this project, we analyzed tweets of influencers and how it affects us. Our prime finding is, even if many influencers use twitter for just entertainment purpose or increasing their fame and business, some influencers discuss about serious problem that we are facing and try to spread the awareness in the society. These influencers are shaping the world by broadcasting their views and raising voice. Analyzing categorized influencers' (such as Politicians, Philanthropist, Nobelist, Scientist, Economist) tweet can be done as future work which will help us understanding the influence of twitter in our daily life.