Twitter analysis

Exploring K-pop fandom - From data to model

Author: Chenxi Liu

December 17, 2019

Contents

1 Introduction

The main aim of this study is to explore the fandom culture of BTS, data for this study were collected from twitter and streaming time was from 31st of October at 22:00 UTC time till 1st of November at around 9:00, 1,004,261 tweets had been collected. It has previously been observed from the last report those fans are mainly young people based on the fact that they like using tik tok and are highly engaged with voting for their idol. Based on the number of hashtags and highlighted topic keywords, the most popular members in this band are: JinMin and Jungkook. Images, videos and emojis are the main contents of their tweets, texts length are normally quite short. Besides, they retweet a lot with tons of hashtags and the contents are mostly about concerts or new songs.
The report has been organised into two parts in order to build more models based on fans' properties. The first deals with location, mainly to find out the difference and similarities between different regions; the second part is focusing on exploring the network relations between centralized users. And due to practical constraints, this report cannot provide a review and conclusion of age predicting, though in the previous study I had given a deduction based on fans' preferences, the result couldn't been proved by real data in current state.

2 User Model

2.1 Location

Understanding who are the fans and where are they is really worthy on data analysis, the flaw of fetched tweets is the lack of "coordinate" information which can provide an accurate geographical information, so only "location" can be used to do the investigation.
First of all, I filtered tweets which only contain "location" and important user information such as "user_ID","source", ect, in order to do crossover selection later. Among the filtered 558,674 tweets (55.6% among total tweets), there are 138,051 unique users have tweeted have been selected. From the example table below it is apparent that locations can not indicate useful information sometimes, some users do not really fill the correct location.

Example Location Table	**Source**	**Hashtags**
Pekalongan Utara, Indonesia	Twitter for Android	NaN
Paraguay	Twitter for Android	NaN
namjoons dimples	Twitter for Android	NaN
tmi.tid.tvd.va.tog.BTS	Twitter Web App	['BTS', 'MAMAVOTE']
Jawa Barat, Indonesia	Twitter for Android	NaN

Table 1: **Location**: Only select 3 columns "Location", "Source", "Hashtags" for representing examples

The Table 2 below compares 2 figures, the above one shows the 15 most common locations of filtered tweets, the above one is the 15 most common locations of the unique 138,051 users. There are three parts of Mojibake, the first two are Thai, they

are presented two locations: Thailand and Bangkok, Thailand; the last Mojibake is an emoji represents the country flag of Philippines.

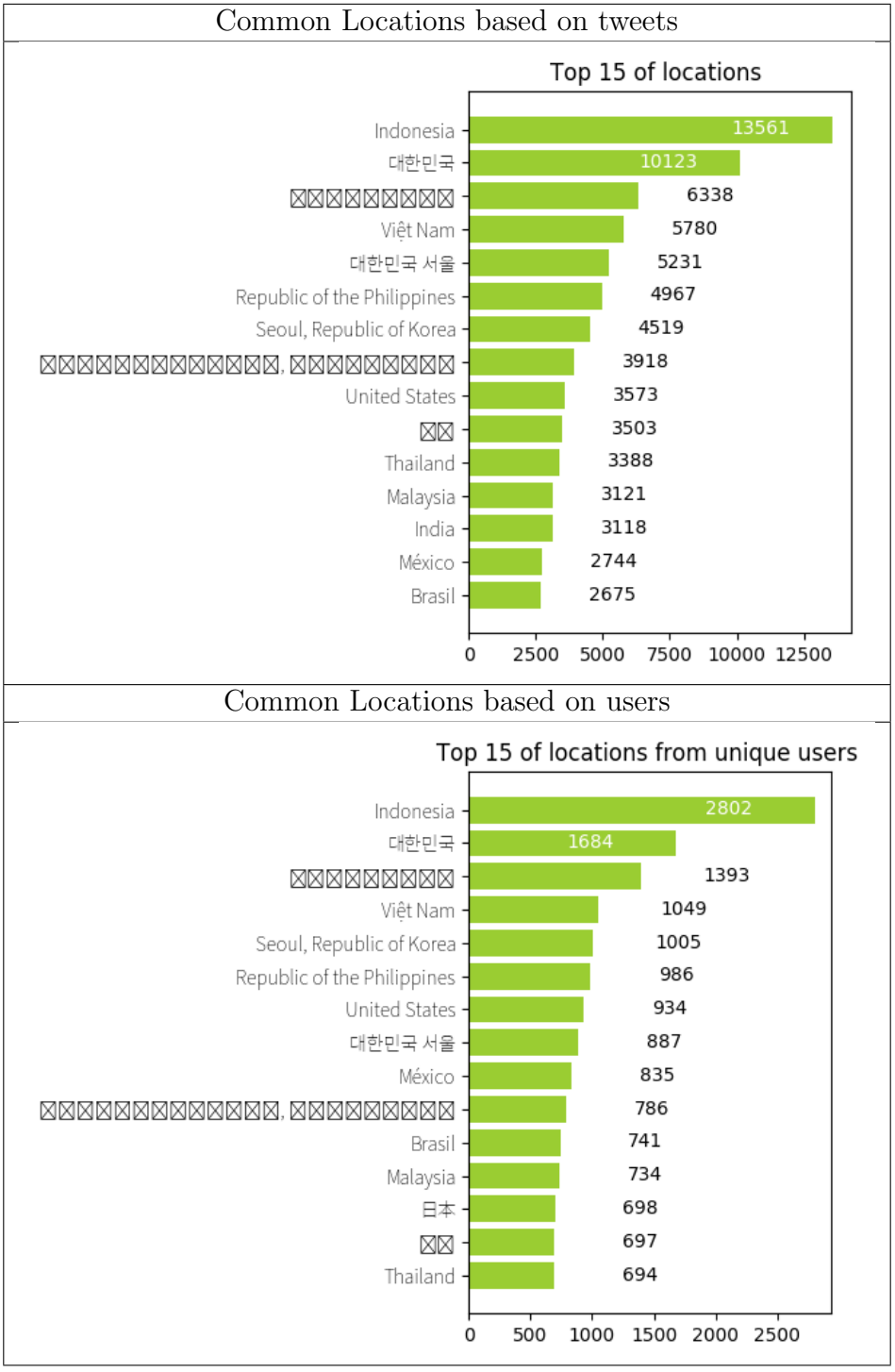

Table 2: Location

The results obtained from the two figures are quite interesting, first of all most locations are Asian countries, this can be treated as a "bias" since I fetched data in the night, at that time Asia was at daytime; secondly there is litter change among those regions,Indonesia, Korea, Thailand and Vietnam are the top four countries

which have generated most of the traffic, other countries only have changed order except India and Japan, since Japan has replaced India at the second figure we can indicate that Indian users are highly activated; Japan is just the reverse, it has more users than Indian but generate the less contents.

Considering the messy output of locations, in order to simplify the results as well as for further work, I will use country as the only standard. So among the 15 locations, there are actually 10 countries, which are: Indonesia, Korea, Thailand, Vietnam, Philippines, United States, Mexico, Brasil, Malaysia, Japan.

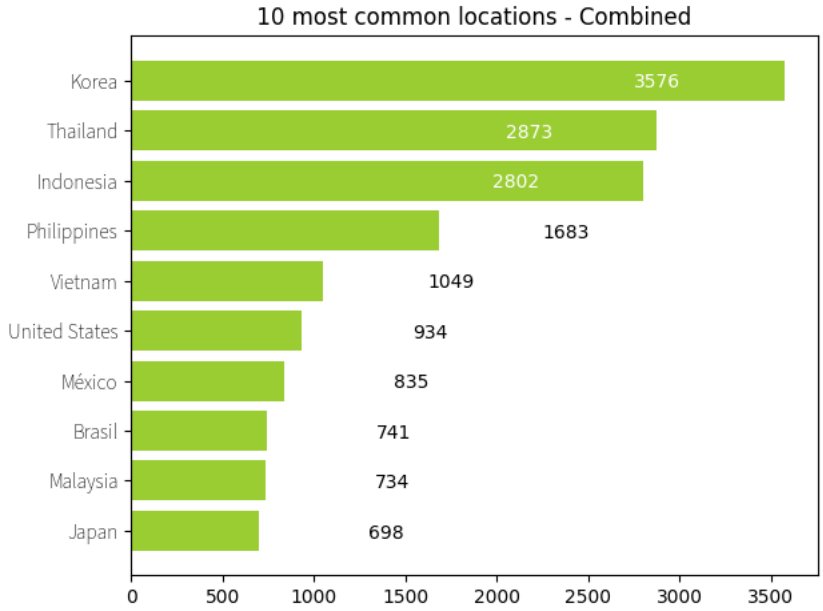

Figure 1: Filtered location

After combining pieces of location into countries, we can see Korea now becomes the top 1 country which has the most users, followed by Thailand and Indonesia.

Mobile devices are the primary platform among fans where contains a high percentage of Android users, except for Japan and Unite States, more users prefer iPhone than Android. In a way, sometimes this is important for commercial campaign.

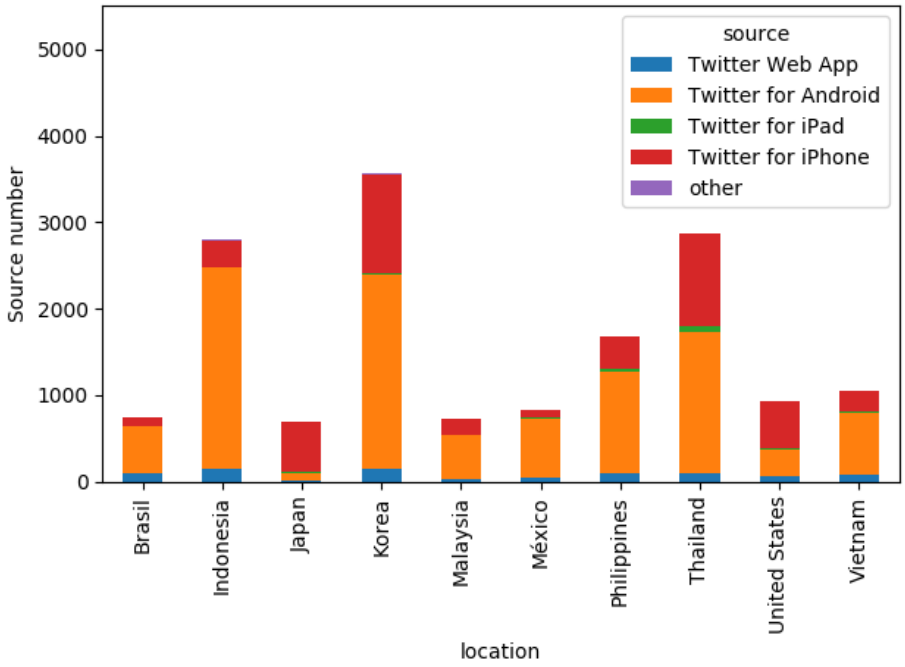

Figure 2: Source Percentage among countries

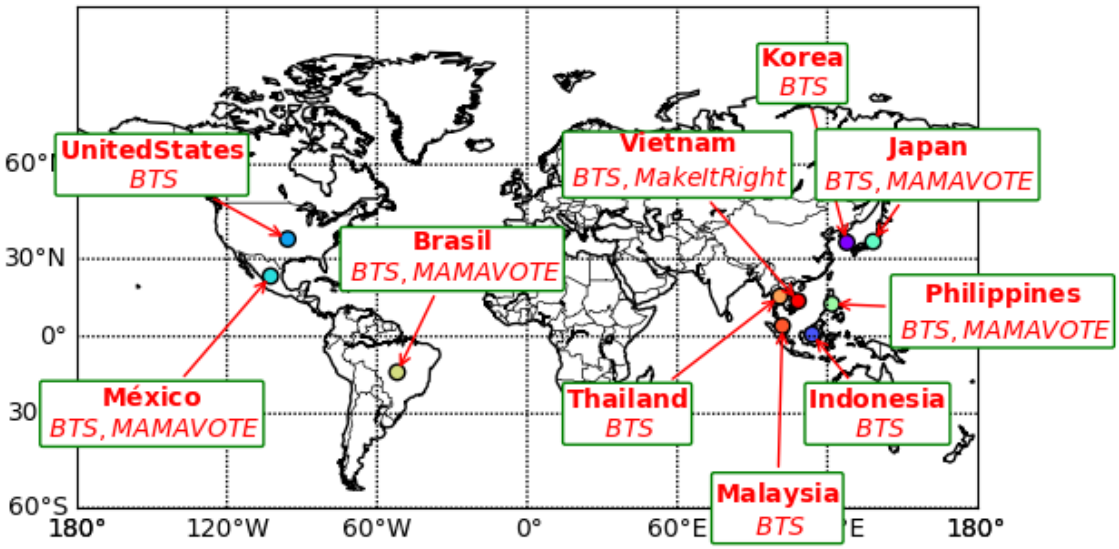

Figure 3: Map

I have also plotted the most common hashtags among those countries. Since the dataset is based on fetching hashtag "BTS", it is not surprise that it is the most common hashtag among all countries; because the time when I collected data, it was the voting season for a award among k-pop fans, which is so called "MAMAVOTE", it was a like competition among different fans groups too. Japan, Philippines, Brasil and Mexico had contributed a lot traffic for this vote; the song "Make it right" is one of the most common hashtag in Vietnam since it was really a hit during that time.

2.2 Network

Analyzing network on social media is an essential tool to understand the relationships between uses, if we can dig more, we would capture with communities which group similar users together based on the phenomenon of homophily.
For network analysis, I use networkx library. In NetworkX, a single network object, as known as graph, contains a set of nodes and edges, based on the graph we will understand how nodes and edges are related. In this case, it is meaningless to dig all fans' following list and find relationship between them. My strategy is to dig relations between retweet users and been retweeted users, and based on this relationship graph to select most important N number nodes (users), then explore the relationship between them.
After dropping duplicated information, the total number of pairs is 708,776. Table 3 provides essential graph information, from the numbers we can tell this graph has a extreme high density, and there exist quite a lot of nodes have really few connections, which means nodes connection and distribution of this graph is imbalanced.

Number of nodes	**Number of edge**	**Average degree**	**Density**
239703	707774	5.9054	2.46e-05

Table 3: Graph information

One thing we want to learn from network data is who are the most important users, and what are their inter-connected structures. Since the aim of this exercise is to understand more about BTS's fandom culture, so my next step is to find the most important nodes in this graph. Here I user degree as standard to find important nodes(users), then among those users I try to see if there exists friendship between each other, if returns true then these two users will be added as edges to the new graph.
Because of the twitter streaming time limitation, I only have selected 30 most important users as an example, the number of nodes only has 28 because there are two users have been suspended. The average degree is still around 6 which indicated the centralized users have a sort of close relations between each other.

Number of nodes	**Number of edge**	**Average degree**	**Density**
28	79	5.6429	0.2

Table 4: Graph information - 30 nodes

From Figure 4 we can see most being followed users are center on right part, the particular important node/user is BTS_twt, this is the official account of the BTS band, BigHitEnt is the second important user since it is the official account of the band's agency. What is interesting about the data is that some users whose tweets have been retweeted a lot and they had been ranked at top 10 users who had been retweeted most, for example fan accounts "BTSNewsBrasil" or "ThrowbacksBTS", but they do not have much relationship with other 27 nodes as I imaged; on the contrary,"modooborahae" is also a fan account and it has highly engagement among those users. This might be explained from account details, from the previous report I had fetched their information, "modooborahae" average tweets are as high as 150, "BTSNewsBrasil" and "ThrowbacksBTS" only tweet 48 and 20 respectively.

Moreover, UNICEF(The United Nations Children's Fund) Korea actually has a lot of connections with other fan accounts though it is not a fan page nor an offical page of BTS. This is because in 2017, BTS had joined a campaign with UNICEF Korea with an effort to raise awareness about violence towards children and young people around the world, and in 2018 there was another campaign had brought them together, and till now this account still has connection with other fan pages, also its tweets have driven a lot of traffic too though most of the contents are about funding for children. From here we can see the huge impact of this idol group.

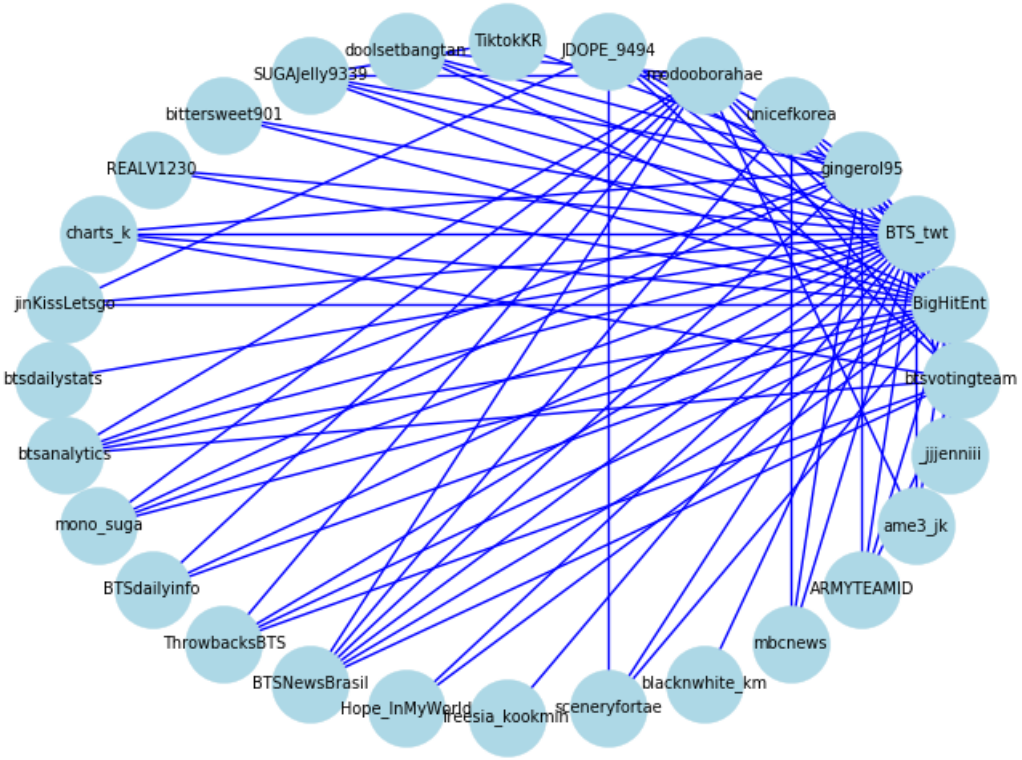

Figure 4: Networkx Graph - 30 users

3 Conclusion

This project was undertaken to explore the fandom culture of BTS, and this report has shown that Asian countries are highly active on voting, especially on Southeast Asian; Mobile devices are the main devices to publish tweets, and in general Android device is the most used. The network graph has show a high closed relations among users; and due to the impact of BTS, some unrelated tweet accounts are followed and interacted by fans; also how close between two fan pages may be determined by the account's tweet frequency.
Considerably more work will need to be done to determine the age range of fans, some work has been applied but not gain a convincing result. Some train files with a certain number of users are need but in practical it is hard to obtain.