Analysis on the Social Media Data Collection

Ansh Khandelwal

December 22, 2020

1 Data Collection

The data was collected on $18^{th} December 2020$ from the top trending hashtag "ISRO" in Hyderabad.
Exactly 10671 tweets were scraped from twitter which is sufficiently large for the analysis. The data consisted of the unique tweets around the "ISRO" hashtag and hence the data set has no retweets.
The dataset was collected using the popular "twint" tool, which is an advanced Twitter scraping OSINT tool written in Python that doesn't use Twitter's API, allows to scrape a user's followers, following, Tweets and more while evading most API limitations.

2 Basic Commands of the Twint Tool Used

- **twint -s #ISRO**: Collect every Tweet containing #ISRO from everyone's Tweets.
- **twint -u username**: Scrape all the Tweets of a user (doesn't include retweets but includes replies).
- **twint -u username –followers**: Scrape a Twitter user's followers.
- **twint -u username –following**: Scrape who a Twitter user follows.
- **twint -u username –following –user-full**: Collect full user information a person follows.
- **twint -u username -o file.json –json**: Scrape Tweets and save as a json file.

The data set consists of the tweet objects. Each tweet object has following attributes:
id, conversation_id, created_at, date stamp, timestamp, timezone, user_id, username, name, place, tweet, mentions, urls, photos, replies_count, retweets_count, likes_count, hashtags, cashtags, link, retweet, quote_url, video, user_rt_id, near, geo, source, retweet_date

Using the user_id from the tweet object we can extract the User objects. Each User object has the following attributes:

id, name, username, bio, location, url, join_date, join_time, tweets, following, followers, like, media, private, verified, avatar, background_image.

Using all the above mentioned commands and attributes, a data set was collected and the analysis has been done.

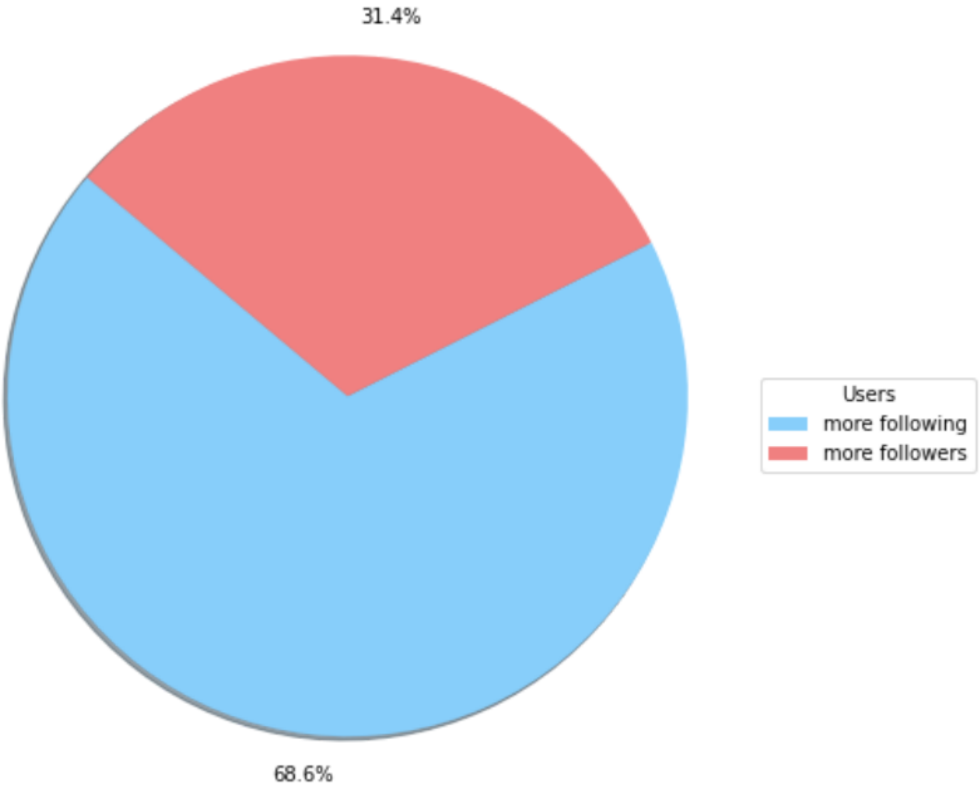
3 Analysis on Users

In this section, we try and analyse the personal traits of the users, like how many users have more followers than following. This trend with help us understand the common social dilemma behind the socials. Say, the idea is having the account as natural as possible. If you have 3000 followers but you are following 5000 people, you will generate suspicions. People will know that you follow them out of pure interest. That you follow them only to increase your followers.

3.1 Number of followers vs Number of following

The follower/following ratio is a metric some users use to judge the quality of your account. Those with low follower/following ratio are typically low-quality accounts that depend only on the follow/unfollow method to gain followers, whereas accounts with high follower/following ratio are influencers and celebrities.

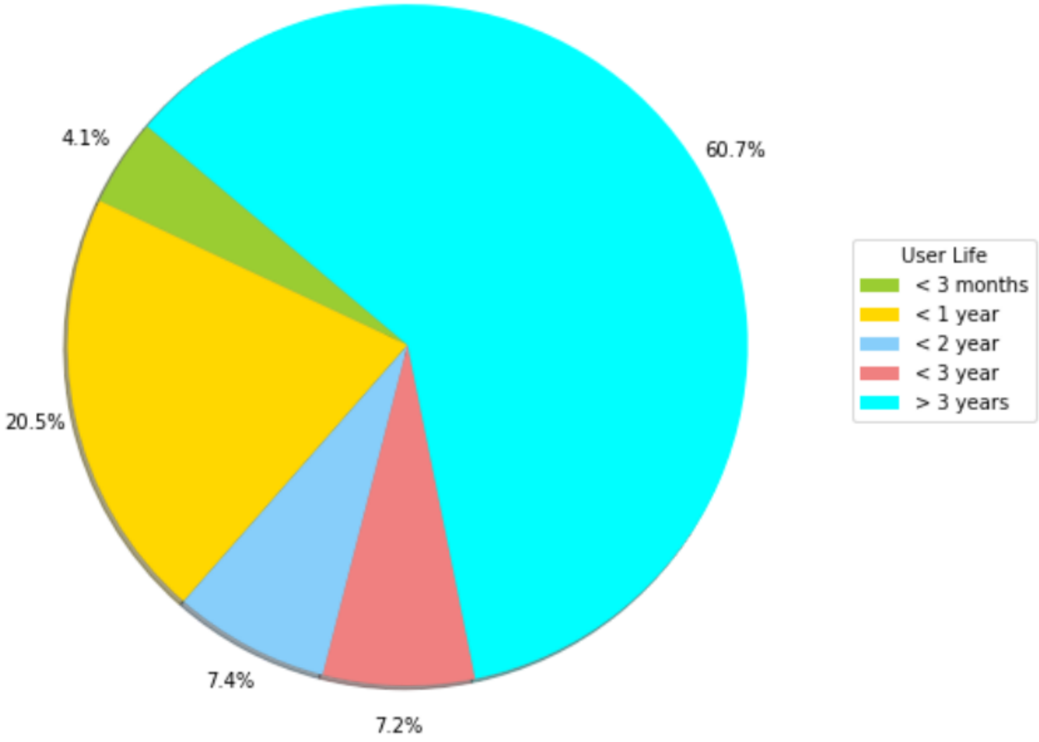
Hence this ratio is important to analyse.



From the data set we see that around 31 percent of the users have more following than followers.

3.2 Age of User's account

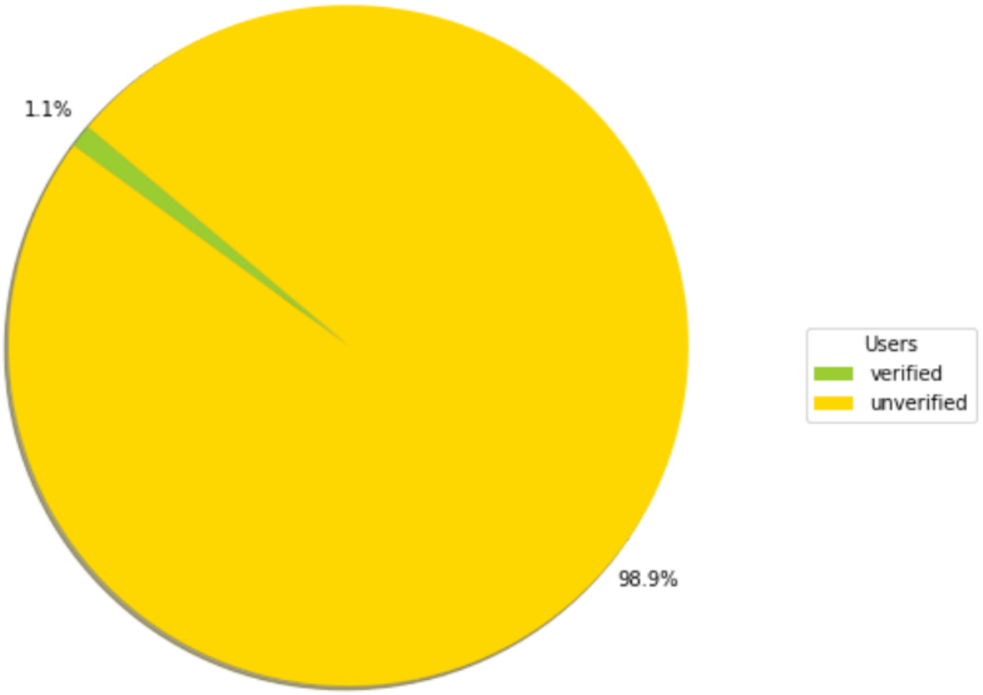
We can also talk about how long have the users been on the twitter platform.



From this graph we can that there are around 4 percent new users, who have recently joined twitter. Around 20 percent of the users have joined twitter in 2020, 7.4 percent of the users have joined in 2019, 7.2 percent in 2017. And around 60 percent of the users have been on the platform for more than 3 years!
Majority of the users are the "old/experienced twitter" users.

3.3 Verified vs Non Verified Users

Verified Twitter accounts have a blue badge with a check mark next to the profile name. The badge appears in searches and when the account owner comments on a post. This badge recognizes the account as the official profile for the person or business. Twitter verification prevents impersonation.
The blue verified badge on Twitter lets people know that an account of public interest is authentic. To receive the blue badge, your account must be authentic, notable, and active.

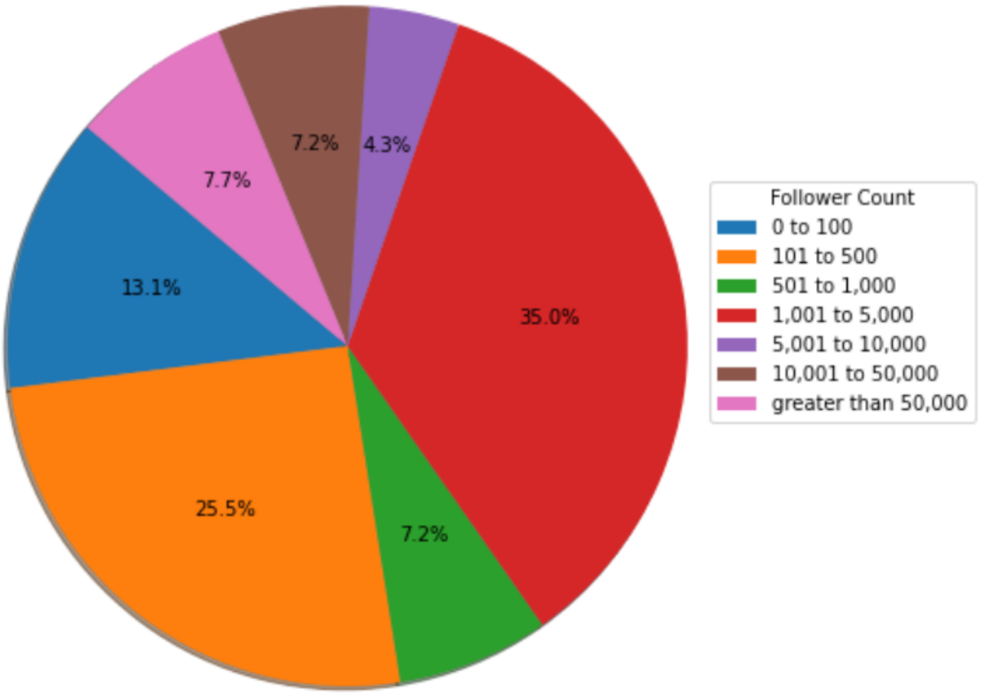


Here we see that only around 1.1 percent of the users who have tweeted the trending ”ISRO” hastag are verified.

4 Analysis on Follower and Following count

4.1 User Follower Count Analysis

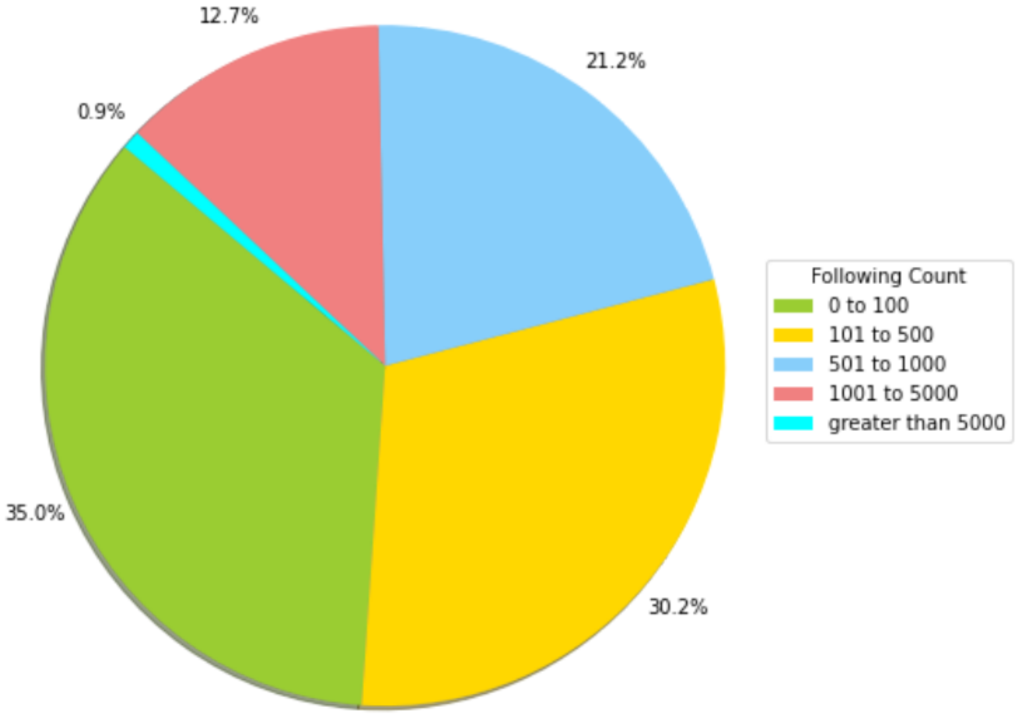
This graph divides the users on the number of followers.



We see that around 7.7 percent of the users have more followers than 50 thousand followers, around 7.2 percent of users have max 100 followers. We see that maximum of the users have around 1000-5000 followers.

4.2 User Following Count Analysis

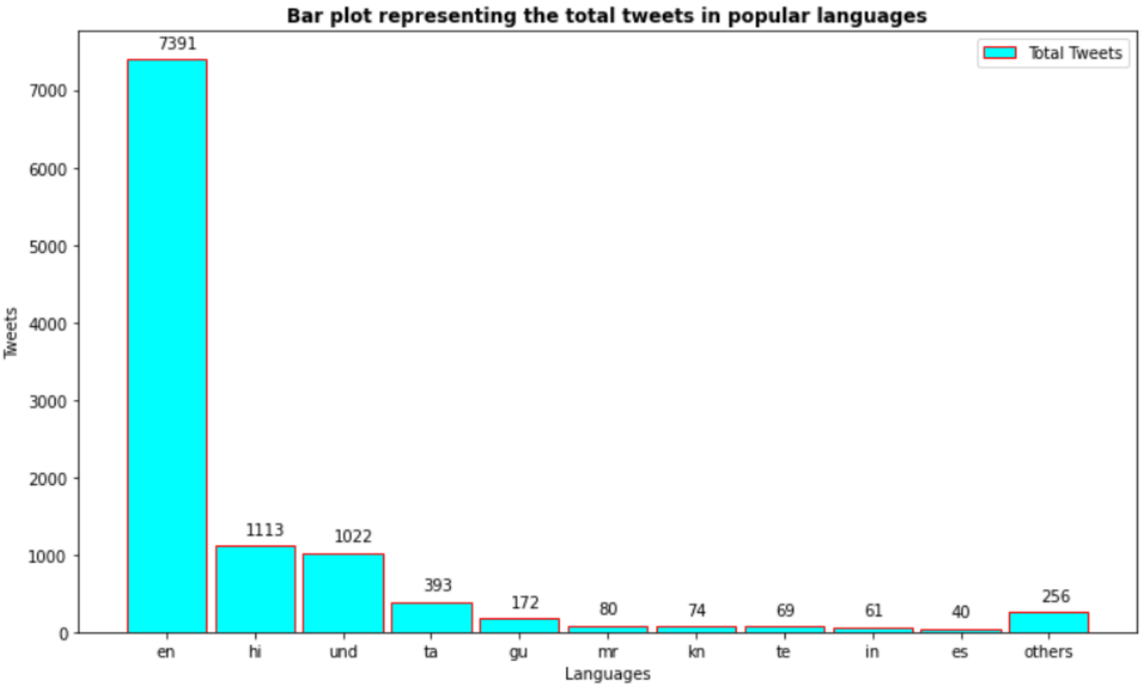
This graph divides the users on the number of following.

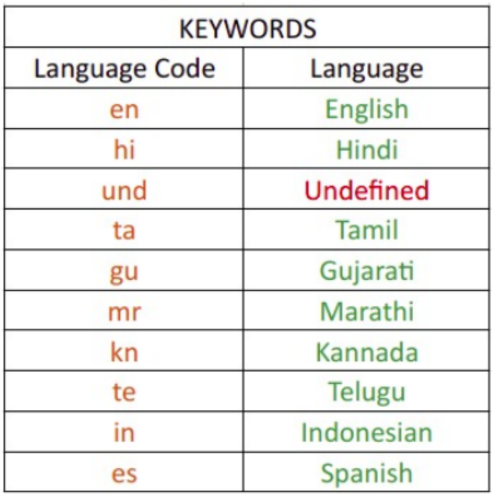


As compared to the follower count, following count is a lot lesser. Here we see that only 0.9 percent of the users follow more that 5000 twitter accounts, we see that around 65 percent of the users follow 100 to 1000 accounts, which is a majority of the users.

5 Language analysis of Tweets

This bar plot portrays the tweets and divides them on the basis of languages used. Around 34 different languages were used in the tweets, here I have plotted top 10 languages used in the tweets.
We see that English is the top language used for the tweets and Hindi is the next most popular language used for the tweets in Hyderabad.





KEYWORDS	
Language Code	Language
en	English
hi	Hindi
und	Undefined
ta	Tamil
gu	Gujarati
mr	Marathi
kn	Kannada
te	Telugu
in	Indonesian
es	Spanish