Exploratory Data Analysis

1st Abdullah Naveed
Computer Science
23100239

2nd Karan Kumar
Computer Science
23110043

3rd Ahmer Jamil
Computer Science
23100186

4th Esha Fatima
Computer Science
23100201

I. Overview of Dataset

A. Fake labelled tweets:

```
===> Number of Rows:  100200
===> Number of Columns:  12

===> Number of Null/NaN in each Columns
user_id                              0
tweet_id                             0
text                                 0
followers                            0
following                            0
location                         28905
verified                             0
statuses_count                       0
profile_background_tile              0
profile_use_background_image         0
favorite_count                       0
label                                0
dtype: int64
```

Fig. 1. Dataset

```
===> Number of unique entries in user_id                          46844
===> Number of unique entries in tweet_id                         60
===> Number of unique entries in text                             98620
===> Number of unique entries in followers                        13301
===> Number of unique entries in following                        7756
===> Number of unique entries in location                         22346
===> Number of unique entries in verified                         2
===> Number of unique entries in statuses_count                   40419
===> Number of unique entries in profile_background_tile          2
===> Number of unique entries in profile_use_background_image     2
===> Number of unique entries in favorite_count                   577
===> Number of unique entries in label                            1
```

Fig. 2. Unique Entries

B. Real labelled tweets:

C. Merged Dataset:

D. Data Summary:

The individual dataframe for fake and real tweets are available, only the location column has a significant number of missing values, 49986 combined missing values which we replace during the data transformation.

```
===> Number of Rows:  101320
===> Number of Columns:  12

===> Number of Null/NaN in each Columns
user_id                              0
tweet_id                             0
text                                 0
followers                            0
following                            0
location                         21081
verified                             0
statuses_count                       0
profile_background_tile              0
profile_use_background_image         0
favorite_count                       0
label                                0
dtype: int64
```

Fig. 3. Dataset

```
===> Number of unique entries in user_id                          14027
===> Number of unique entries in tweet_id                         21
===> Number of unique entries in text                             100962
===> Number of unique entries in followers                        8242
===> Number of unique entries in following                        5318
===> Number of unique entries in location                         6937
===> Number of unique entries in verified                         2
===> Number of unique entries in statuses_count                   20386
===> Number of unique entries in profile_background_tile          2
===> Number of unique entries in profile_use_background_image     2
===> Number of unique entries in favorite_count                   405
===> Number of unique entries in label                            1
```

Fig. 4. Unique Entries

II. Data Transformation

1) 1 The corpora of labelled fake tweets and real tweets were merged and loaded into a single data frame.
2) 2 The attributes indicating Boolean values were transformed such that False values were indicated as 0 and real were indicated with a 1. Such attributes were: Verified, Profile background tile, and profile use background image.
3) 3 The label column indicating that the tweet was fake or real is as coded as follows: Fake Tweets - 0 and Real Tweets - 1

user_id	0
tweet_id	0
text	0
followers	0
following	0
location	49986
verified	0
statuses_count	0
profile_background_tile	0
profile_use_background_image	0
favorite_count	0
label	0
dtype: int64	

Fig. 5. Number of NA entries in the entire Dataset

Number of unique entries in user_id	57945
Number of unique entries in tweet_id	72
Number of unique entries in text	198713
Number of unique entries in followers	17366
Number of unique entries in following	9113
Number of unique entries in location	26500
Number of unique entries in verified	2
Number of unique entries in statuses_count	51453
Number of unique entries in profile_background_tile	2
Number of unique entries in profile_use_background_image	2
Number of unique entries in favorite_count	711
Number of unique entries in label	2

Fig. 6. Number of unique entries for every Attribute

4) 4 The tweet ids and user ids are transformed and mapped onto unique integers via a Label Encoder.
5) 5 The missing location values are replaced with a temporary location called “not available”.

A. The transformed data set appears to be the following:

III. Hypothesis

1) H1) Some words may be found more commonly in fake news than real news and vice versa
2) H2) Verified accounts would share more real tweets than fake tweets
3) H3) Accounts with lesser followers and following would share more fake news than accounts with a greater followers list.
4) H4) Some locations would have a greater contribution to fake news than other locations
5) H5) Accounts with a profile picture are likely to be trusted accounts that share more real tweets
6) H6) Profiles with a tile as a background would be counted the same as accounts without an image and tiled vs non tiled would not influence the share of real or fake news
7) H7) we hypothesize for every tweet id it would be either fake tweet or real tweet or have a clear majority.
8) H8) there would be some correlation between the attributes.

Fig. 7. Summary

	user_id	tweet_id	followers	following	verified	statuses_count	profile_background_tile	profile_use_background_image	favorite_count	label
count	201520.000000	201520.00000	2.015200e+05	2.015200e+05	201520.000000	2.015200e+05	201520.00000	201520.000000	201520.000000	201520.000000
mean	41545.213031	66.09143	6.186774e+04	2.266402e+03	0.044184	8.064791e+04	0.15398	0.829253	5.763428	0.502779
std	18888.283189	1.56231	8.147951e+05	1.453970e+04	0.205505	1.757239e+05	0.36093	0.376289	172.766398	0.499994
min	0.000000	0.00000	0.000000e+00	0.000000e+00	0.000000	1.000000e+00	0.00000	0.000000	0.000000	0.000000
25%	26845.000000	65.00000	2.100000e+01	5.100000e+01	0.000000	6.857000e+03	0.00000	1.000000	0.000000	0.000000
50%	47513.500000	66.00000	3.290000e+02	3.430000e+02	0.000000	1.595300e+04	0.00000	1.000000	0.000000	1.000000
75%	57777.000000	67.00000	1.923000e+03	1.665000e+03	0.000000	7.446600e+04	0.00000	1.000000	0.000000	1.000000
max	57944.000000	71.00000	7.703821e+07	1.209255e+06	1.000000	9.482791e+06	1.00000	1.000000	38675.000000	1.000000

	user_id	tweet_id	text	followers	following	location	verified	statuses_count	profile_background_tile	profile_use_background_image	favorite_count	label
0	57877	64	Ellen DeGeneres Scares BTS With a Little Help ...	72.0	59.0	3Oland	0	63467.0	0	0	0.0	1.0
1	57723	64	New post (17 Royal Rules Meghan Markle Has to ...	572.0	2454.0	Nigeria	0	2155.0	0	0	0.0	1.0
2	47328	64	Shailene Woodley Is Aluminum Foil-Chic and Mor...	7831.0	2963.0	INDONESIA	0	308597.0	0	1	0.0	1.0
3	47328	64	Ellen DeGeneres Scares BTS With a Little Help ...	10021.0	2968.0	INDONESIA	0	307765.0	0	1	0.0	1.0
4	47328	64	They Didn't Mean It Like That: Actors Defendin...	7831.0	2963.0	INDONESIA	0	308597.0	0	1	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
100195	6352	68	@itchybones imagine getting dragged so thoroug...	54.0	312.0	3Oland	0	875.0	1	1	0.0	0.0
100196	57865	68	@Nomswrites @RawStory @megrutsky @GretchenCar...	3049.0	2484.0	New Jersey, USA	0	23371.0	0	1	0.0	0.0
100197	39949	68	@RedHourBen With love and good intentions, we ...	102.0	300.0	Syracuse, NY	0	672.0	0	1	0.0	0.0
100198	52363	68	@kisanye 😂😂😂😂 Don't Mind Me	829.0	836.0	3Oland	0	156493.0	0	1	0.0	0.0
100199	57769	68	I liked a @YouTube video https://t.co/WrH4beM0...	84.0	158.0	Oregon, USA	0	2735.0	0	1	0.0	0.0

Fig. 8. Transformed Data

Data Analysis

Assessing the number of real and fake tweets in the context of account

The twitter accounts can be largely split into bins based on number of followers and can be assessed as to the amount of real or fake tweets each follower bin contributes towards.

1) Number of real tweets against account followers

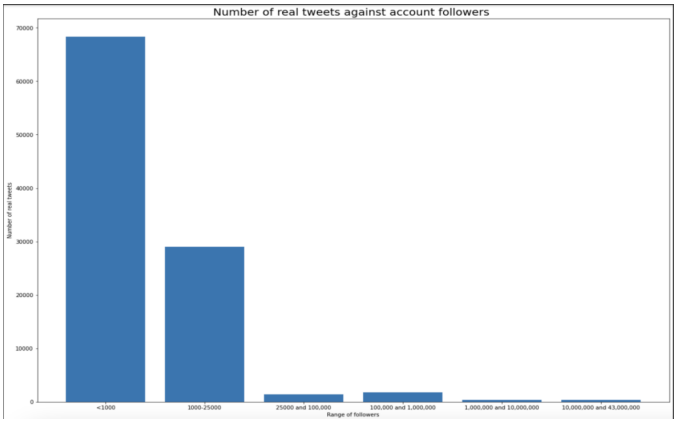


Fig. 9. This graph depicts that as the number of followers increase, the number of real tweets decreases

2) Number of fake tweets against account followers

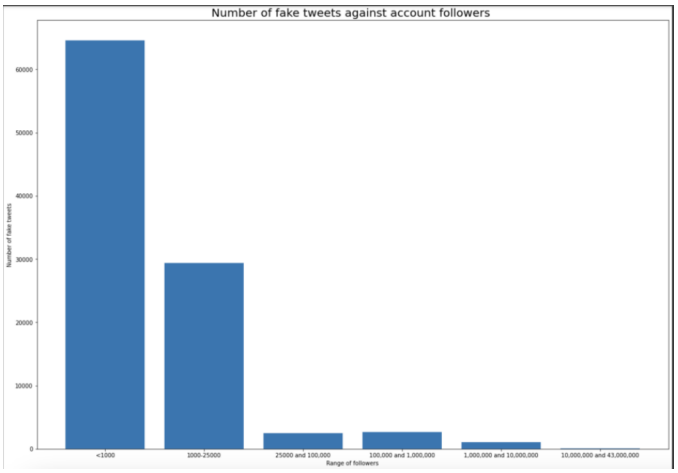


Fig. 10. This graph depicts that as the number of following increases, the number of fake tweets decreases

3) Number of tweets against account following

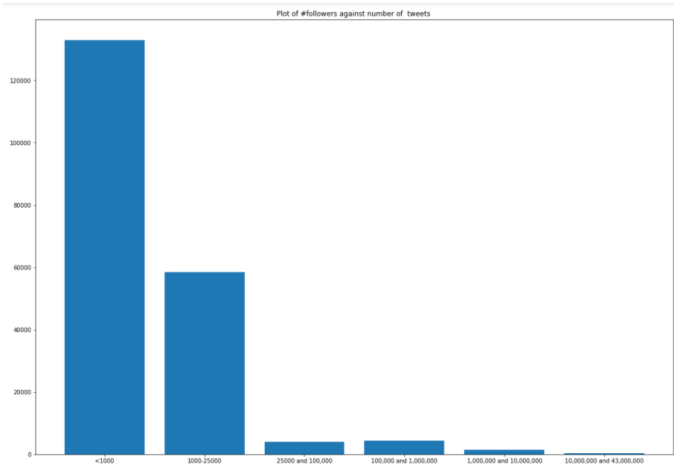


Fig. 11. This graph depicts that as the number of following increases, the total number of tweets decreases

Hence, due to the total decrease in number of tweets, a clear relationship between fake and real tweets cannot be seen.
Hence H3 cannot be proven from this as the general trend shows a decrease in number of tweets as the followers increase.

Analysing average status count for real and fake tweets in the context of account followers

1) Average status count against account followers for real tweets
2) Average status count against account followers for fake tweets This depicts that with an increase in followers, the average status count for fake tweets increases.

Analysing average status count for real and fake tweets in the context of account following

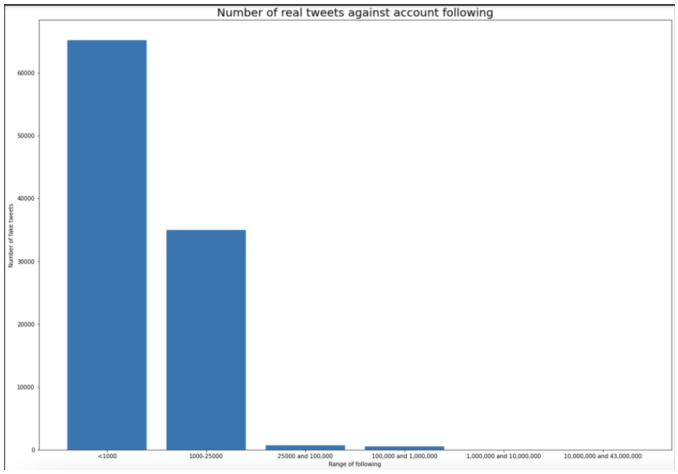


Fig. 12. This graph depicts that as the average status count remains in the same range for real tweets unless the followers are less than a thousand

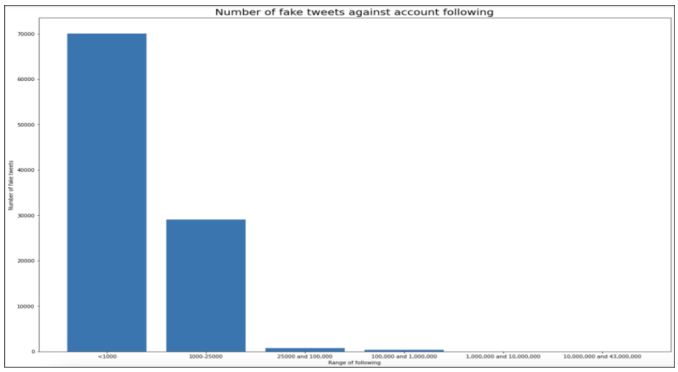


Fig. 13. This graph depicts that as the number of followers increases, the average status count increases for fake tweets

1) Average status count against account following for real tweets
2) Average status count against account following for fake tweets When comparing both the graphs with each other, users following less than 25,000 accounts are more likely to share real news, whereas users following more than these followers are likely to share more fake news.

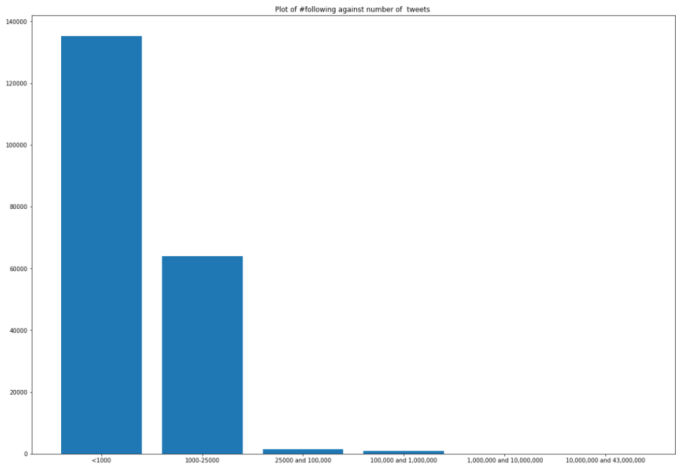


Fig. 14. This graph depicts that the distribution follows a normal distribution and users who follow 25000 to 10000 accounts are most likely to share real news

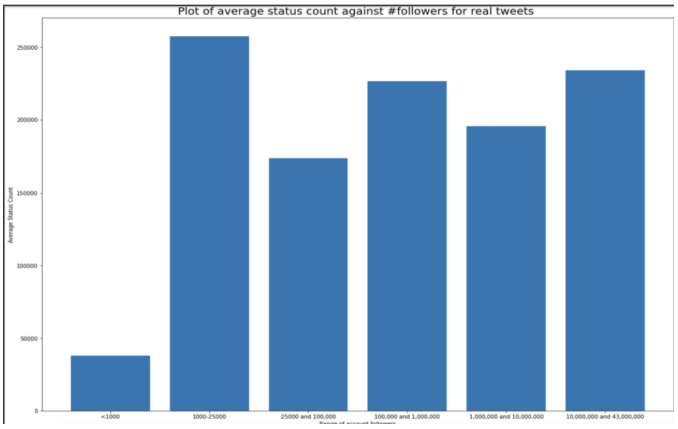


Fig. 15. This graph depicts distribution is left skewed with users following 100,000 to 1,000,000 accounts being most likely to share fake news

Computing the correlation between numerical attributes
To identify potentially related attributes within the data set, it is essential that we compute the correlation matrix for the numeric attributes.

1) Correlation matrix

	statuses_count	followers	user_id	following	favorite_count
statuses_count	1.000000	0.057845	-0.228977	0.147307	-0.000711
followers	0.057845	1.000000	-0.145459	0.215283	0.239842
user_id	-0.228977	-0.145459	1.000000	-0.133865	-0.030373
following	0.147307	0.215283	-0.133865	1.000000	0.031702
favorite_count	-0.000711	0.239842	-0.030373	0.031702	1.000000

Fig. 16. Matrix

2) The heatmap for correlations The correlation between any two attributes can be interpreted as the following:
i) Values closer to 0 indicate no correlation between the particular attributes
ii) Values closer to -1 or +1 indicate strong negative and positive correlation between the attributes respectively.

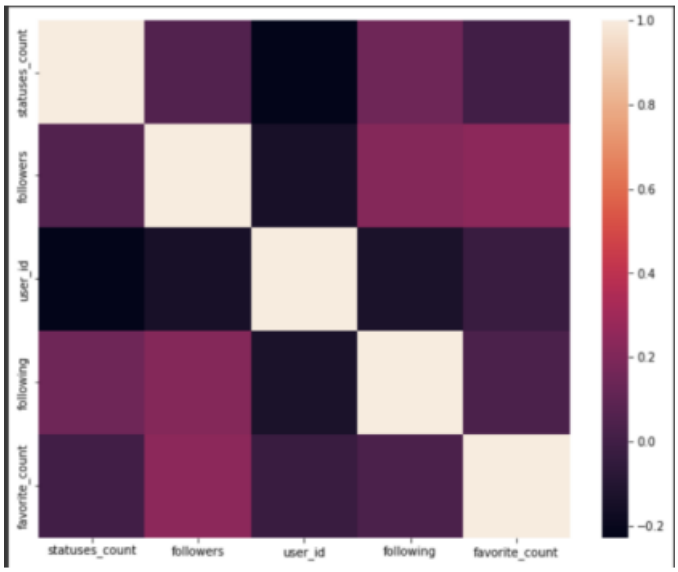


Fig. 17. Heatmap

As per the correlation values produced, there seems to be no apparently highly correlated pair of variables that has a high correlation such that one of the attributes need to be removed. **Hence, rejects H8**

Analysing total count of fake and real tweets in the context of verified accounts

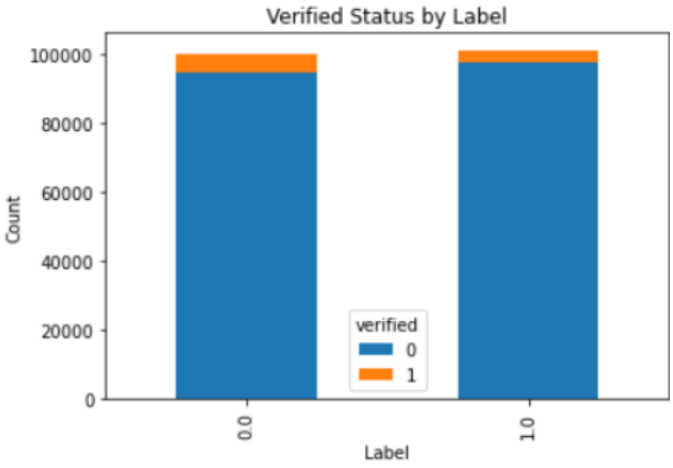


Fig. 18. Plotting the total number of verified accounts among the real (1) and fake (0) tweets

This graph depicts that for both the real as well as the fake tweets, there are more unverified users than verified users. There is a difference in the number of verified and non-verified users for both the real and fake tweets with fake tweets having a greater percentage (5.4 percent) of verified users than real news (3.37 percent).
Hence this shows that a verified user is more likely to share fake news than real news.

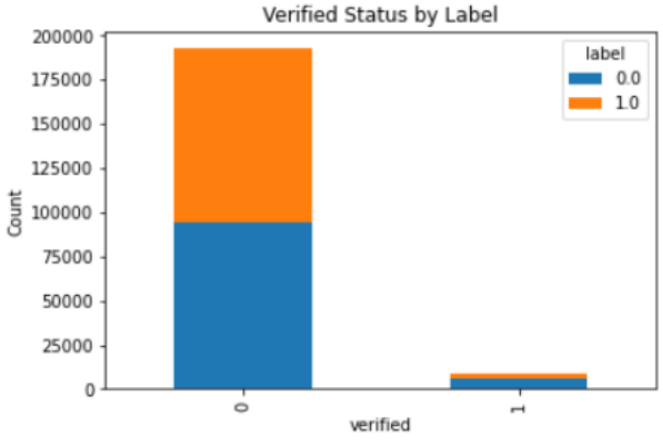


Fig. 19. Plotting the total number of real and fake tweets among the verified and non

This graph depicts that there are more unverified users than there are verified users and the difference in the number of verified and non-verified users is extreme.
However, the distribution of real and fake news among unverified users is almost the same as visible in the graph, whereas for verified users, there is a greater percentage of fake news than real news.
Hence this shows that a verified user is more likely to share fake news than real news. This disproves H2 and shows the opposite trend.

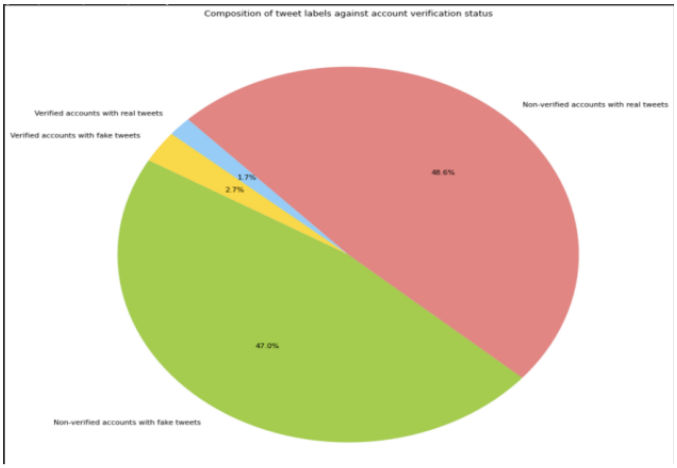


Fig. 20. Composition of tweet labels against account verification status

The pie chart shows the distribution of verified and nonverified accounts with their labels. It shows that majority of the accounts are nonverified with almost the same proportions of real and fake tweets. The verified accounts only account for 4.4 percent of the total tweets but there is a greater percentage of fake news (61.36 percent) among the verified users than real news.

This pie chart shows that there is a higher percentage of users without a tile as a background image (54.6 percent) than users with a tile as background image.
However, among the users with a tile as background image, 14 percent are likely to share real tweets while the rest are more likely to share fake tweets. Users without a tile as background image (these may include user who do not have a background image) 80 percent are likely to share real news. **Hence, for users without a tile as a background image, there is a higher possibility of sharing real news than fake news. This disproves H6 as a relationship does exist between background tile and possibility of sharing real news.**

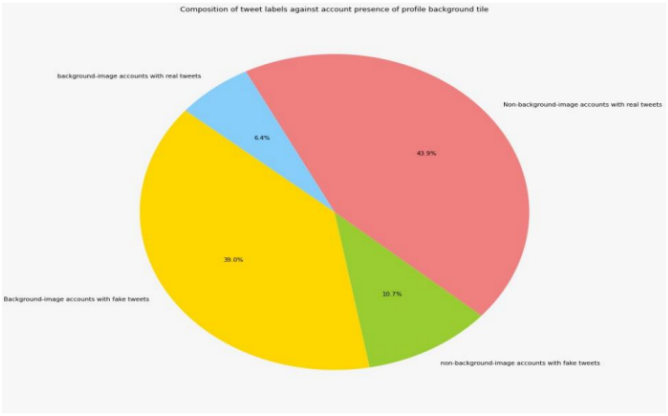


Fig. 21. Analysing the ratio of tweet labels based on the presence of account background tile

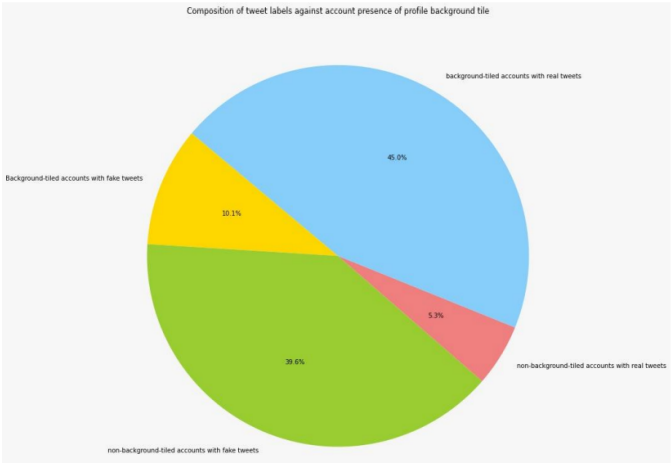
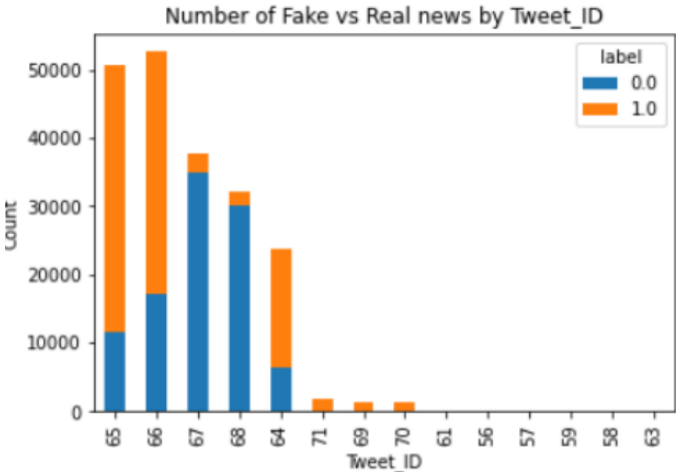


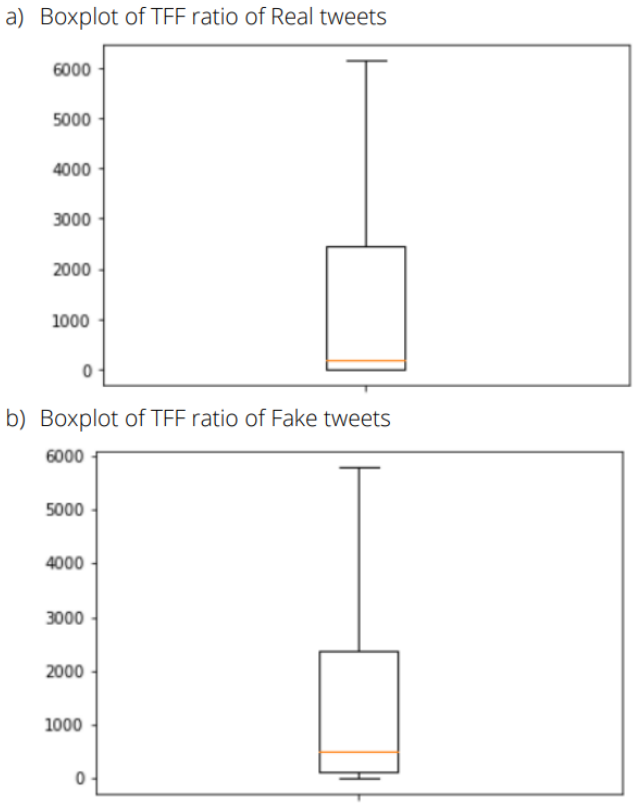
Fig. 22. Analysing the ratio of tweet labels based on the presence of profile use background image

This pie chart shows that there is a higher percentage of profiles with a background image than there are without a background image. Among the profiles with an image, there is an 80 percent likelihood of sharing real news whereas with profiles without an image, there is 11 percent likelihood of sharing real news. Hence, for accounts with a profile picture, there is a greater probability of sharing real news. This proves H5.



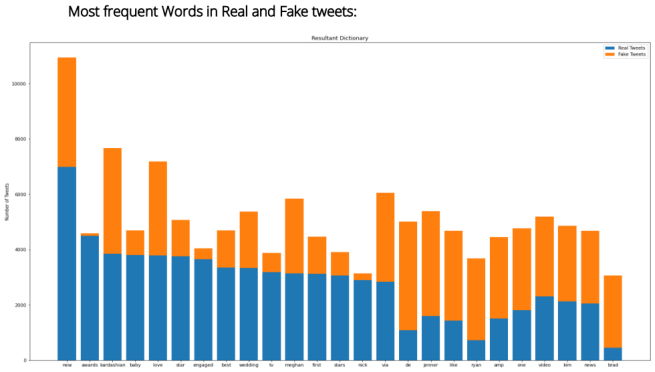
This graph shows that there are certain tweet Ids that have a greater count than other tweet ids. Among these, some tweet ids are more likely to share real news than fake news.
This confirms H7 however, each Tweet id corresponds to multiple user ids.

Fig. 24. Analysing boxplots for real and fake tweets based on TFF.



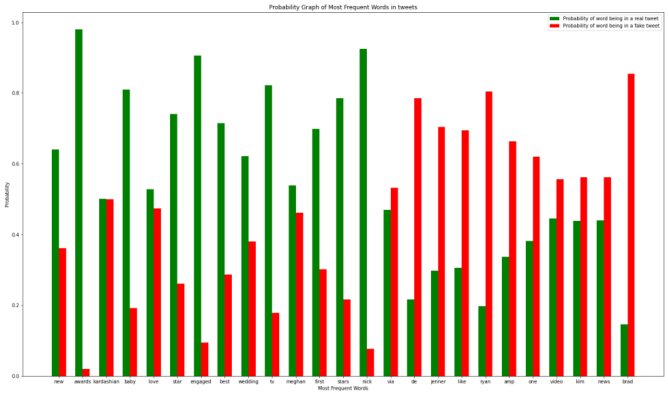
The graph above shows some frequent words in the text and their occurrences among fake and real tweets. The graph clearly depicts that some words are more likely of occurring in fake news while some occur more frequently in real tweets.

Fig. 25.



This concept of bag of words could be further extended to find phrases or commonly sequences of words among real and fake tweets.

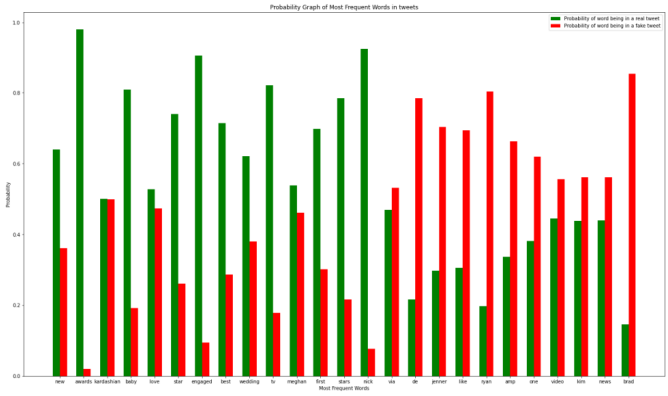
Fig. 26.



Graph shows the probabilities of certain frequently repeated words occurring in real and fake tweets. The difference in red and green bars depicts that there is a difference in the probabilities of certain words occurring in fake vs real tweets. **Hence, this confirms the hypothesis that certain words have a greater probability of occurring in fake tweets than real tweets and vice versa. This proves H1 and shows that a word might be more probable to be found in real news than fake and vice versa.**

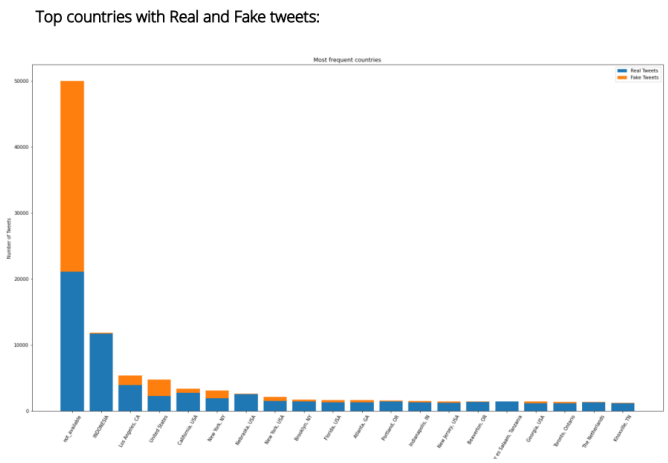
The graph shows that the greatest count of instances does not have any location associated with them. Among these tweets which do not have a location, there is a higher possibility of being a fake tweet than a real tweet. Among the tweets with a location associated, there is a higher possibility of them being a real tweet rather than a fake tweet. There is a limitation in this method as there are many locations that mention cities within USA instead of USA. However,

Fig. 27. Top countries with Real and Fake tweets



If all these are accumulated then there would be a greater percentage of real tweets than fake tweets.

Fig. 28.



The graph above shows the distribution of real and fake news more clearly for instances with a location associated with them. **This proves H4. It also depicts a relationship of instances with no location that we did not hypothesize**

DATA LIMITATIONS

The format for locations is different which creates a limitation, however, considering the most frequently mentioned locations and adding the cities in the count for the countries, we can see a general trend of more real tweets for instances that have a location.