Unraveling the Web of Fake News: An Analysis of Twitter's Information Landscape

Muhammad Hassan
23100199@lums.edu.pk

Aleem Malik
23100179@lums.edu.pk

Jawad Azhar
23100321@lums.edu.pk

Muneeb Munir
23110327@lums.edu.pk

***Index Terms*—news, tweet**

I. Introduction

In recent years, the rapid expansion of social media platforms has resulted in an unprecedented increase in the circulation of both genuine and false information. The ease with which people can disseminate information globally has created a complex information landscape, making it challenging to differentiate between facts and falsehoods. This trend has significant ramifications for various domains, including journalism, politics, and public health.

The purpose of this research paper is to investigate the differences between fake and real news in the context of social media, with a specific focus on Twitter as a primary source of data. We will analyze the features of both types of news and examine the sources of fake news to gain a better understanding of the phenomenon.

Our analysis will involve scrutinizing the language utilized in tweets, identifying patterns, and examining engagement metrics. We aim to shed light on the characteristics of fake news and its dissemination on social media to contribute to a better understanding of social media's role in shaping public discourse. We hope that by providing insights into the spread of misinformation and the features of fake news, we can develop strategies for combating the propagation of false information and promoting accurate information dissemination.

II. About the dataset

Our dataset consists of two csv files: "fake news" and "real news", both of which contain news tweets posted by various users. Each dataset includes 12 attributes:

These attributes provide a range of information about each tweet and the user who posted it. The "text" attribute gives the content of the tweet, while "followers" and "following" provide insight into the user's social network. The "verified" attribute indicates whether the user is a trusted source, and "location" can be used to identify where the news originated from. Additionally, the "statuses count" and "favorite count" attributes can be used to gauge the user's activity on Twitter. Overall, this dataset can be used to train machine learning models to distinguish between fake and real news, and to gain insights into the behavior of Twitter users who share news content.

A. Fake and real news

Our data sets include 2 csv files, fake news and real news, both of which contain news tweets posted by various users.

TABLE I
Number of Rows in Dataset

Category	Number of Rows
Fake News	645,528
Real News	995,432

B. Information about the datasets

The data provided includes various attributes related to a tweet posted by a user. The attributes are as follows: user id, which is a unique integer identifier for the user who posted the tweet; tweet id, which is the ID of the tweet sharing the news; text, which is the actual tweet shared by the user; followers, which represents the number of followers the user's account currently has; following, which represents the number of users the user's account follows; location, which is the location specified in the user's account profile; verified, a boolean value indicating whether the user's account is verified or not; statuses count, which is the number of tweets (including retweets) issued by the user; profile background tile, which is a boolean value indicating whether the user's profile background image is tiled or not; profile use background image, which is a boolean value indicating whether the user's account has a profile background image or not; favorite count, which is the number of tweets the user has liked; and label, which is a binary variable indicating whether the tweet is real news (labeled 1) or fake news (labeled 0).

III. Preprocessing techniques used

For both the Fake News and Real News datasets, some common preprocessing steps were followed. Firstly, boolean variables such as "verified", "profile background tile", and "profile use background image" were converted into 0 and 1. This was done to standardize the representation of these variables across the datasets.

IV. Data Cleaning

Irrelevant variables such as User id and Tweet id were removed from both datasets. These variables did not contain any useful information for the analysis and hence were removed. Next, any row that contained all NULL values was removed

from both datasets. This was done to ensure that the analysis was based on valid and complete data only.

V. Graphs

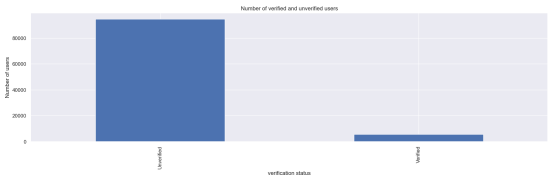

Fig. 1. Number of verified and unverified users (Fake News)

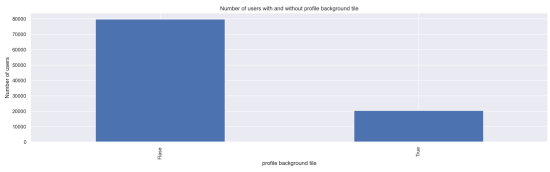

Fig. 2. Number of users with and without background tile (Fake News)

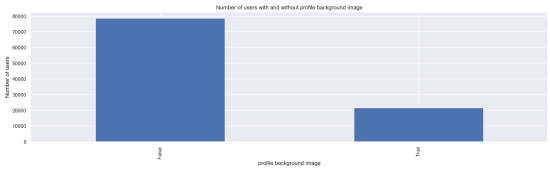

Fig. 3. Number of users with and without background image (Fake News)

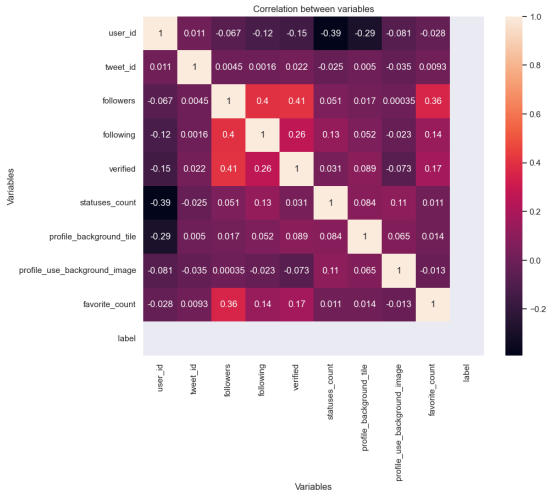

Fig. 4. Correlation between variables (Fake News)

A. Analysis

According to the pie charts showing the top countries for Real and Fake news, we can see that the Fake news is mostly originating from USA and UK. Surprisingly, most of the Real news is being originated from Indonesia. USA and Tanzania

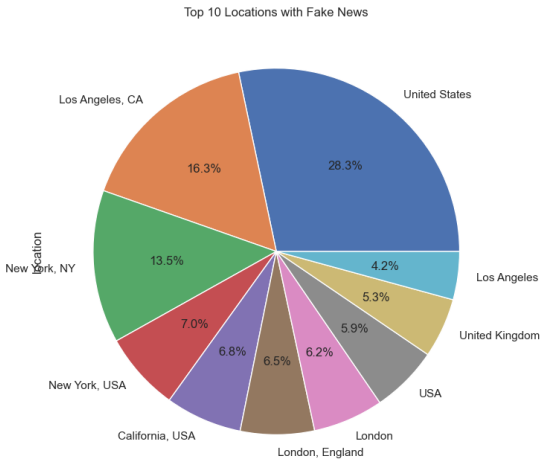

Fig. 5. Top 10 locations with Fake News

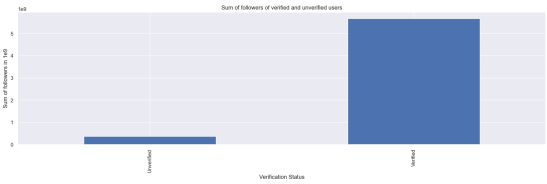

Fig. 6. Sum of followers of verified and unverified users (Fake News)

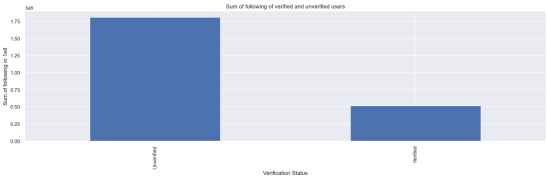

Fig. 7. Sum of following of verified and unverified users (Fake News)

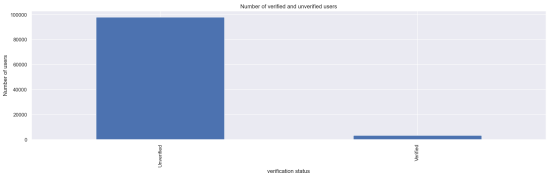

Fig. 8. Number of verified and unverified users (Real News)

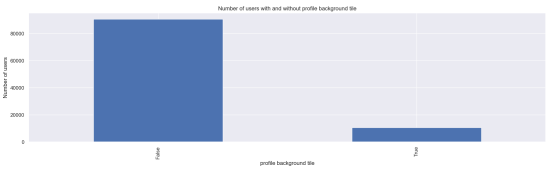

Fig. 9. Number of users with and without background tile (Real News)

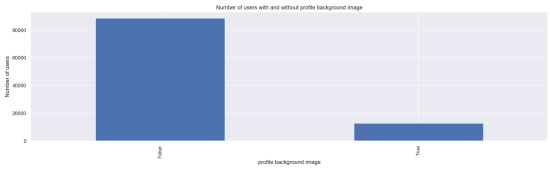

Fig. 10. Number of users with and without background image (Real News)

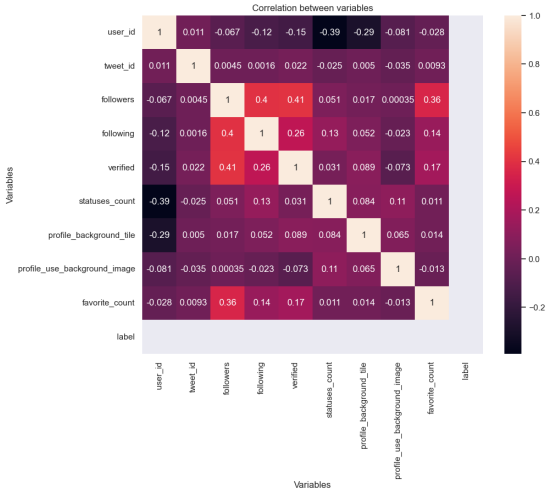

Fig. 11. Correlation between variables (Real News)

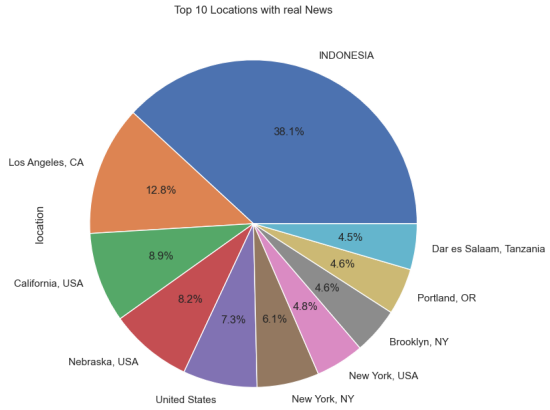

Fig. 12. Top 10 locations with Real News

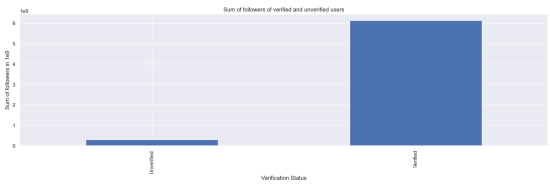

Fig. 13. Sum of followers of verified and unverified users (Real News)

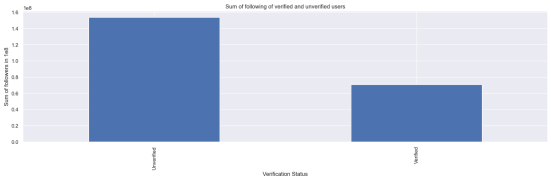

Fig. 14. Sum of following of verified and unverified users (Real News)

are also among the top contributors of Real news
Another insight gained from the graphs is that the following of unverified users is more than the following of verified users. However, the sum of followers of verified users is greater than the sum of followers of unverified users.