**Visualizing Relationships between Tweet Numbers and Real News Labels in FakeNewsNet**

**Table of**

[1. Introduction 3](#_Toc186727705)

[1.1 Problem statement and research motivation 3](#_Toc186727706)

[1.2 Data set 3](#_Toc186727707)

[1.3 Research question 3](#_Toc186727708)

[1.4 Null hypothesis and alternative hypothesis 3](#_Toc186727709)

[2. Background Research 4](#_Toc186727710)

[2.1 Previous Researche 4](#_Toc186727711)

[2.2 Research gap 4](#_Toc186727712)

[3. Visualization 4](#_Toc186727713)

[3.1 Plot for the RQ output of an R script 5](#_Toc186727714)

[4. Analysis 9](#_Toc186727715)

[4.1 Statistical test used to test the hypotheses and output 9](#_Toc186727716)

[4.2 The null hypothesis rejected /not rejected 9](#_Toc186727717)

[5. Evaluation – Group’s Experience at 7COM1079 9](#_Toc186727718)

[5.1 What went well 9](#_Toc186727719)

[5.2 Points for Improvement 9](#_Toc186727720)

[5.3 Group’s Time Management 10](#_Toc186727721)

[5.4 Project’s Overall Judgement 10](#_Toc186727722)

[5.5 Comment on GitHub Log Output 10](#_Toc186727723)

[6. Conclusion 10](#_Toc186727724)

[6.1 Results Explained 10](#_Toc186727725)

[6.2 Interpretation of the result 10](#_Toc186727726)

[6.3 Future Work and Limitations of Your Research 10](#_Toc186727727)

[Reference 11](#_Toc186727728)

[Appendix 12](#_Toc186727729)

# 1. Introduction

## Problem statement and research motivation

The analysis to be carried out will involve attempting to understand the correlation between the variable tweet\_num that relates to news articles and the variable real to determine whether the identified news belongs to fake news or real news using information from the FakeNewsNet.csv dataset (Hani et al., 2020). The motivation originates from an awareness of the pragmatic query of whether the involvement (by the number of tweets) varies between real and fake news. This should be of concern and significance when studying the manipulation and credibility of information online. Also, it indicates the type of the relationship between two variables and distribution of these types might also play a role in their properties.

## Data set

The feature set includes the following independent variables: “title” (character), “news\_url” (character), “source\_domain” (character), “tweet\_num” (numeric) and “real” (numeric, Real=1, Fake=0). Alphabetic and numerical data types are intermingled in this set.

## Research question

* To test the effect of the authenticity of news stories as a between-subjects variable (real news vs fake news) on the number of tweets linked to each article.

## Null hypothesis and alternative hypothesis

**Hypothesis for Correlation between “number of tweets” & “real and not real news stories”**

**( H0 ) (Null Hypothesis):** It is not possible to identify any marked distinctions between the two categories to denote the difference between real and not real news stories through the number of the tweeters.

**H1 (Alternative Hypothesis):** The results indicated that real news stories produced more tweets than not real stories did.

# 2. Background Research

## 2.1 Previous Researche

Gupta and Potika, (2021) researched how fake news spreads in a social network in comparison to true news. Using graphs that mimic the propagation of bogus and authentic news, we structure the topic as a binary classification problem and use social network analysis to find different characteristics. .

Soga et al., (2024) researched the crucial problem of identifying and elucidating the propagation of fake news in light of the growing dissemination of false information through social media platforms. While early detection techniques concentrated on text analysis, more modern algorithms have examined network graphs of news sharing by taking advantage of the unique ways that fake news spreads.

Choudhary et at., (2021) researched that on the Internet, fake news has grown to be a serious problem. It should be recognized since it has a direct impact on people. Several data-driven methods based on Machine Learning (ML) and Deep Learning (DL) have been proposed recently for the categorization of fake news. Most machine learning-based methods make use of manually created features that are taken from the input text.

## 2.2 Research gap

The analysis provides an initial exploration of the relationship between tweet counts and news authenticity, but several areas remain unexplored. It lacks causal analysis, deeper textual examination of news content, and time-based analysis. There’s no control for potential confounders, and the heatmap may not capture non-linear patterns. Additionally, the results haven't been externally validated, and no predictive modeling for real vs fake news based on tweet numbers was conducted.

# 3. Visualization

## 3.1 Plot for the RQ output of an R script

# Histogram for tweet\_num

ggplot(FakeNewsNet, aes(x = tweet\_num)) +

geom\_histogram(bins = 30, fill = "skyblue", color = "black") +

labs(title = "Histogram of Tweet Number", x = "Tweet Number", y = "Frequency")

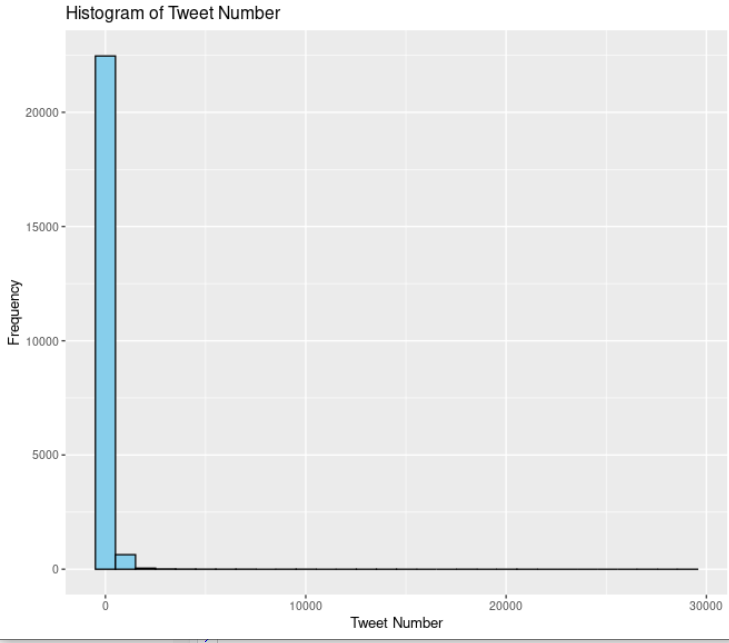


Figure 1 : Histogram (Tweet number)

# Histogram for real (binary distribution)

ggplot(FakeNewsNet, aes(x = real)) +

geom\_histogram(bins = 2, fill = "lightcoral", color = "black") +

labs(title = "Histogram of Real vs Fake", x = "Real (1) or Fake (0)", y = "Frequency")

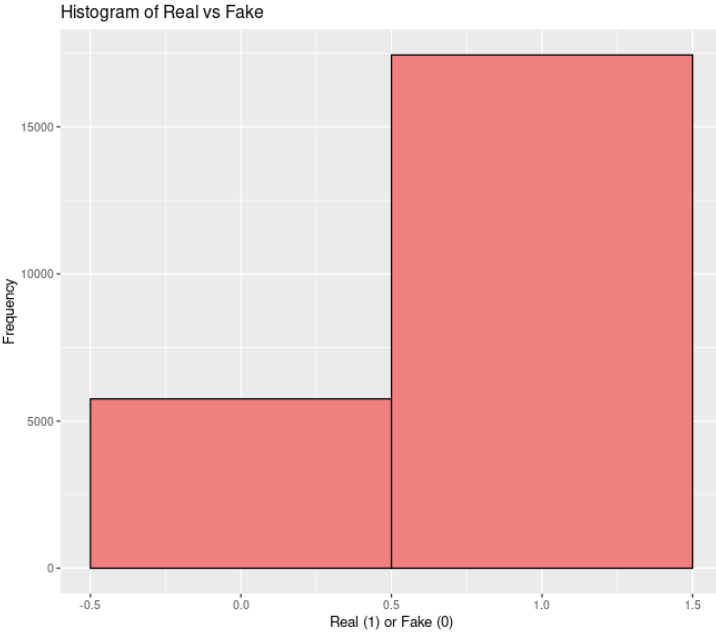


Figure 2 : Histogram ( Real / Fake)

# 3. Scatter plot

# Scatterplot: Real (Independent) vs Tweet Number (Dependent)

ggplot(FakeNewsNet, aes(x = real, y = tweet\_num)) +

geom\_point(color = 'blue') +

labs(title = "Scatterplot: Real vs Tweet Number", x = "Real (1) or Fake (0)", y = "Tweet Number")

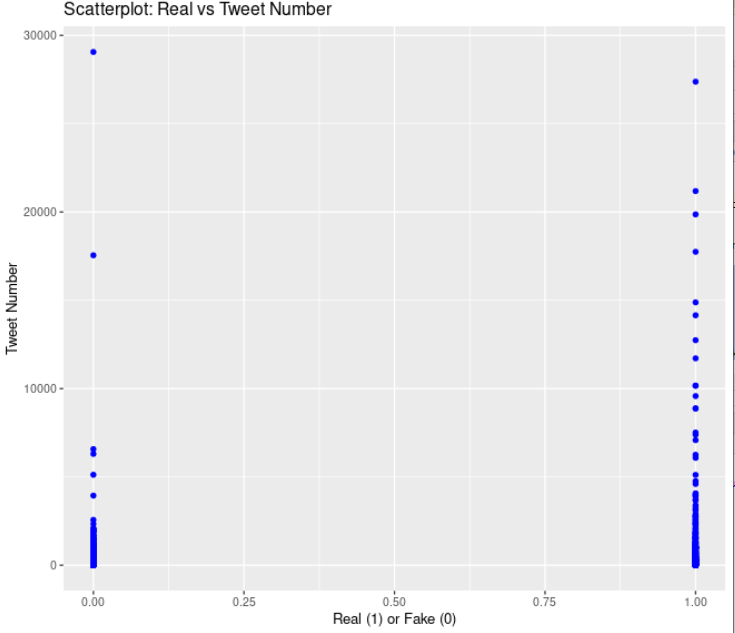


Figure 3 : Scatter plot

# 2. Correlation

# Calculate Spearman correlation between tweet\_num (independent) and real (dependent)

spearman\_correlation <- cor(FakeNewsNet$tweet\_num, FakeNewsNet$real, method = "spearman")

# Print the Spearman correlation

cat("The Spearman correlation value is:", spearman\_correlation, "\n")

# Create the Spearman correlation matrix

spearman\_matrix <- matrix(c(1, spearman\_correlation, spearman\_correlation, 1),

nrow = 2, ncol = 2,

dimnames = list(c("tweet\_num", "real"), c("tweet\_num", "real")))

# Melt the matrix to long format for ggplot

spearman\_melted <- melt(spearman\_matrix)

# Plot the heatmap with variable names

ggplot(spearman\_melted, aes(Var1, Var2, fill = value)) +

geom\_tile() +

scale\_fill\_gradient2(low = "red", high = "blue", mid = "lightgray", midpoint = 0, limit = c(-1, 1)) +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

labs(title = "Spearman Correlation Heatmap", x = "Variables", y = "Variables") +

geom\_text(aes(label = round(value, 2)), color = "black", size = 6)

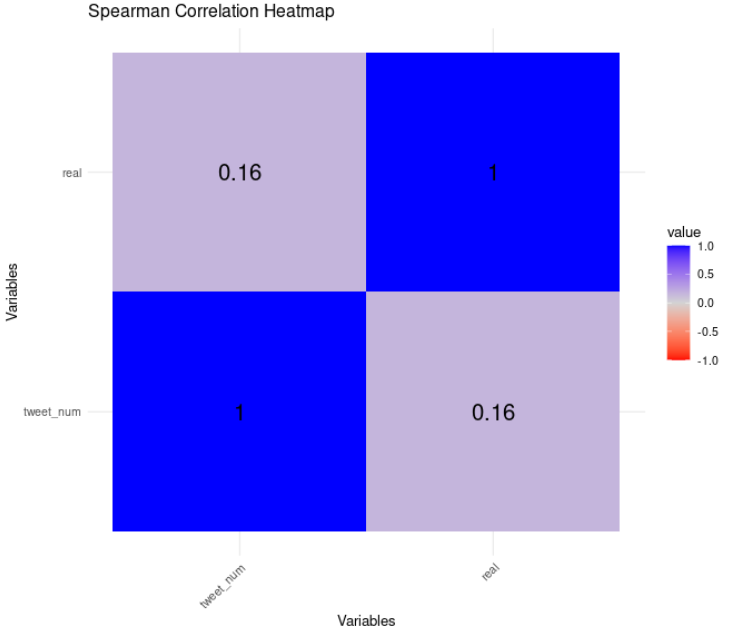


Figure 4 : Correlation Of heatmap

based on the histogram result in the dataset, the non-parametic method is preferred to analyze the data, by using Spearman's correlation to determine the relationship between tweet\_num and real.

The visualizations give a measure of the data within the data set. The histogram showing the numbers of tweets denotes right-skewed distribution where most of the articles are found to have less number of tweets (less than or around 1000 and more than it). The histogram of real as well as fake news looks like a bimodal distribution and makes it clear that there are many more real news articles as compared to the fake ones. First of all, plotting scatter real and fake news articles they depicted at a low tweet number Hence, no contrast between the two groups. The Spearman correlation heatmap reveals that tweet numbers have a very low positive correlation of 0.16 with news authenticity proving the existence of a very low correlation between the two.

# 4. Analysis

## 4.1 Statistical test used to test the hypotheses and output

* Correlation

The R code also calculates and plots the non-parametric statistical test, Spearman rank correlation test where the two variables are tweet\_num and real, ranging between zero and one where one is real news and zero is fake news. Spearman correlation which is non parametric and scale free measure and suitable for skewed distribution measures monotonicity. A 2 x 2 matrix of correlations is used, diagonal elements equal to 1 (indicating perfect correlation) and off diagonal elements equal to the Spearman coefficient, and cor() function estimates the coefficient of the correlation. The reshape2 package’s melt() function reconverting the entire matrix back into long format suitable for visualization. The heatmap is made using ggplot2 with geom\_tile() for tiles and scale\_fill\_gradient2() for color representation where red is negative, blue is positive and gray close to zero correlation. The clean theme and geom\_text() improve sentence visibility and add the correlation values.

## 4.2 The null hypothesis rejected /not rejected

* While, the Spearman correlation tends toward positive, but as it was stated this small number testifies that the connection is rather week (Nguyen et al., 2020). Normally when you get a low correlation coefficient, this will not enable one to reject a null hypothesis.

**Dependent Variable (DV):**

* **Tweet\_num:** This variable means the extent of its conversation, or in this context, the extent of the people’s attention to that particular news story as indicated by the number of tweets relating to news story.

**Independent Variable (IV):**

* **Real:** This variable records the story as real in case it is (1) or fake in case it is not (0). It can be used to understand how truthfulness of the news affects the public attention as measured by the tweet volumes.

# 5. Evaluation – Group’s Experience at 7COM1079

## 5.1 What went well

The group was useful in loading and exploring the data where things such as skewed distributions in tweet\_num are clearly observable (Sharma et al., 2022). They used the right kind of analysis for non-parametric correlation, Spearman, and made proper useful graphs forming histograms, both IV & DV scatter plot, and also heatmap.

## 5.2 Points for Improvement

There is some evidence, which should have been subjected to more imperative hypothesis testing such as p-value analysis. One could have conducted an additional statistical analysis of Spearman ‘s correlation coefficient to know the level of statistical significance of the correlations obtained the dataset could also have been cleaned beyond the basic cleaning procedures to get more meaningful results where more text analysis was required.

## 5.3 Group’s Time Management

Time was well managed within the group when it came to the exploitation and visualization of the data (Alghamdi et al., 2024). Nevertheless, more time was needed not only for statistical analysis but also to perform hypothesis testing, text analysis and be more thorough in the results analysis.,

## 5.4 Project’s Overall Judgement

Over time the group was able to effectively share time in exploring data and visualizing it. However, more time should have been spent in analysis of data such as statiscal analysis, hypothesis testing, text analysis etc so as to get deeper understanding of data.

## 5.5 Comment on GitHub Log Output

This approach of data ignorance was done well for the project in terms of tweet counts and the evaluation of news authenticity (Nawaz et al., 2024). However, the analysis appears sound, but to generate more significant and compelling results, much more testing is needed alongside exploring the findings in more detail and with greater consideration paid to confounding factors.

# 6. Conclusion

## 6.1 Results Explained

Spearman correlation between the obtained tweet\_num and the real variable indicating the news article authenticity is low and barely significant revealing positive relation (approximately 0.16). Moreover, the distribution of tweet\_num is positively skewed, but the real variable is bimodal.

## 6.2 Interpretation of the result

The obtained weak link proves that the number of tweets is insufficient for determining whether particular news is real or fake. This means that other steps probably contribute more and other interpretations, etc. could be investigated in subsequent studies.

## 6.3 Future Work and Limitations of Your Research

**Future work**

The future work should be to improve hypothesis testing, minimize confounding effects, and engage in more extensive detailed textual analysis. Moreover there should be further research on other methods that are similar to Logistic Regression or a Machine Learning classifier to determine news authenticity.

**Limitations**

The major limitation of the analysis is that it is based on a single large dataset, correlational rather than causal, and has been not validated externally. Thus, it is still more exploratory and requires other types of formal testing, as well as wider validation.

# Reference

Gupta, K. and Potika, K., 2021, August. Fake News Analysis and Graph Classification on a COVID-19 Twitter Dataset. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 60-68). IEEE.

Soga, K., Yoshida, S. and Muneyasu, M., 2024. Graph-based interpretability for fake news detection through topic-and propagation-aware visualization. *Computation*, *12*(4), p.82.

Hani, A.B., Adedugbe, O., Al-Obeidat, F., Benkhelifa, E. and Majdalawieh, M., 2020, December. Fane-kg: A semantic knowledge graph for context-based fake news detection on social media. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-6). IEEE.

Nguyen, V.H., Sugiyama, K., Nakov, P. and Kan, M.Y., 2020, October. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1165-1174).

Sharma, S., Saraswat, M. and Dubey, A.K., 2022. Fake news detection on Twitter. *International Journal of Web Information Systems*, *18*(5/6), pp.388-412.

Alghamdi, J., Lin, Y. and Luo, S., 2024. Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. *Engineering Applications of Artificial Intelligence*, *137*, p.109240.

Nawaz, M.Z., Nawaz, M.S., Fournier-Viger, P. and He, Y., 2024. Analysis and classification of fake news using sequential pattern mining. *Big Data Mining and Analytics*, *7*(3), pp.942-963.

Choudhary, M., Chouhan, S.S., Pilli, E.S. and Vipparthi, S.K., 2021. BerConvoNet: A deep learning framework for fake news classification. *Applied Soft Computing*, *110*, p.107614.

# Appendix

# Histogram for tweet\_num

ggplot(FakeNewsNet, aes(x = tweet\_num)) +

geom\_histogram(bins = 30, fill = "skyblue", color = "black") +

labs(title = "Histogram of Tweet Number", x = "Tweet Number", y = "Frequency")

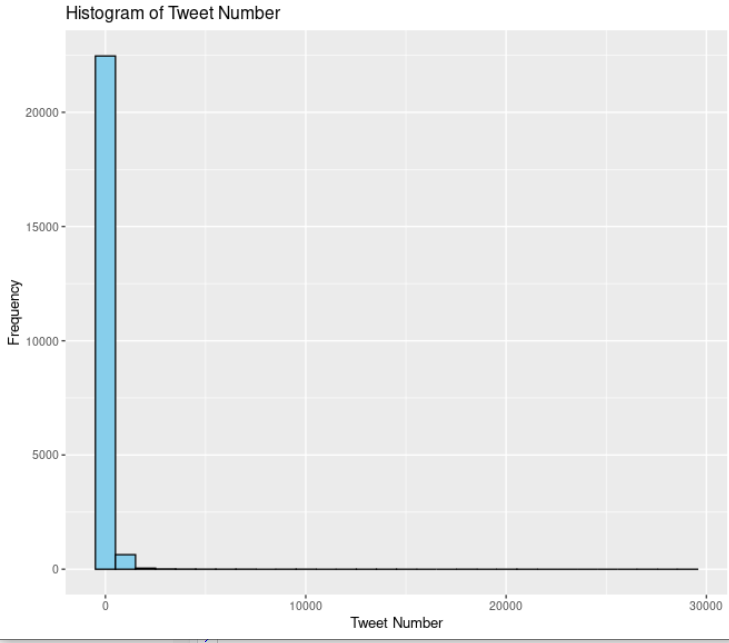


Figure 1 : Histogram (Tweet number)

# Histogram for real (binary distribution)

ggplot(FakeNewsNet, aes(x = real)) +

geom\_histogram(bins = 2, fill = "lightcoral", color = "black") +

labs(title = "Histogram of Real vs Fake", x = "Real (1) or Fake (0)", y = "Frequency")

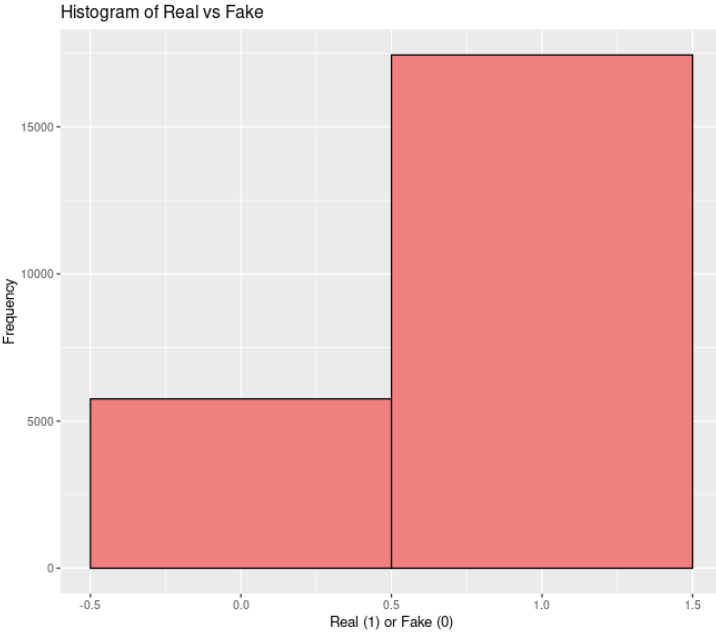


Figure 2 : Histogram ( Real / Fake)