

# **Fake News Detection**

**Student's Name: Shanmukh Sri Surya Gopi, Shiva Kumar Middle, Srikar Morla,  
Kavya Medikonda**

**CWID :20020131, 20021994, 20023032, 20020915**

## Table of Contents

Introduction and Problem Statement .....	3
Problem Statement .....	3
Data Exploration and Preparation .....	3
Data Processing .....	4
Data Visualization .....	5
Examining Your Data Analysis .....	6
Descriptive Statistics.....	6
Check Multivariate Assumptions.....	6
Factor Analysis .....	7
Interpretation of Factor Loadings .....	10
Regression Model .....	10
Model Evaluation .....	12
Additional Multivariate Technique (Logistic Regression) .....	12
Results .....	14
Conclusion and Recommendation.....	14
References.....	16

## Figures

Figure 1: Class Distribution .....	5
Figure 2: Distribution of Text Lengths.....	5
Figure 3: Check multivariate assumptions.....	6

## **Introduction and Problem Statement**

The expansion of fake news is a difficult issue that has impacted the computerized world to an ever-increasing extent. This report attempts to resolve this issue. Solid techniques to recognize and stop counterfeit news are progressively vital because of the speedy spread of fake data. This study utilizes a dataset from Kaggle to explore the utilization of multivariate data analysis in the improvement of a dependable model for the recognizable identification of fake news.

### **Problem Statement**

The reason for this study is to decide if multivariate data analysis strategies are powerful in distinguishing fake news. The dataset chosen for research was obtained from Kaggle, a notable setting for datasets and data science contests. The dataset remembers various textual and metadata elements for expansion to a bunch of reports delegated either true or fraudulent. Our analysis's principal objectives are as per the following:

1. Evaluating the parts of the dataset, including their dissemination and characteristics.
2. Utilizing multivariate data analytic strategies to make a solid news detection model, like exploratory factor analysis, multiple regression analysis, and logistic regression.
3. Applying reasonable measures, like accuracy to evaluate the produced model's presentation.
4. To work on detecting fake news and diminishing its adverse consequences on society, partners will get reasonable guidance in light of the discoveries.

This study expects to recognize valid and fake news by exploration and modeling of textual and metadata features. This study aims to help the persistent undertakings to neutralize disinformation and safeguard the realness of computerized information by using modern insightful techniques.

### **Data Exploration and Preparation**

The dataset that was chosen to concentrate on comes from Kaggle and is partitioned into two unique CSV records called "True.csv" and "Fake.csv." Alongside metadata like the title, text, subject, and publication date, each document incorporates reports marked as true or fake.

At the point when the data was first examined, it was observed that there were 23,481 passages in the fake news dataset and 21,417 entries in the real news dataset. Four sections make up every one of the two datasets: title, text, subject, and date. Both dataset items are extensive because the vast majority of their entrances have no missing values.

Descriptive statistics gave data about how various information was distributed across the datasets. For example, "News" shows up most often in the misleading news dataset, yet "politics news" shows up more regularly in the true news dataset. Besides, there was a variety in the circulation of text lengths among bona fide and false reports, with the last option normally showing more prominent text lengths than the previous.

### **Data Processing**

To prepare the datasets for analysis, data handling techniques were completed. This contained:

1. Finding duplicate entries and eliminating them to protect data integrity.
2. To distinguish between real (0) and fake (1) news stories, a new all-out factor called "classification" will be made.
3. To normalize text style and make the analysis more straightforward, eliminate accentuation, and convert text data to lowercase.
4. Determining the textual length of every news thing to research any potential associations between text length and news story authenticity.

## Data Visualization

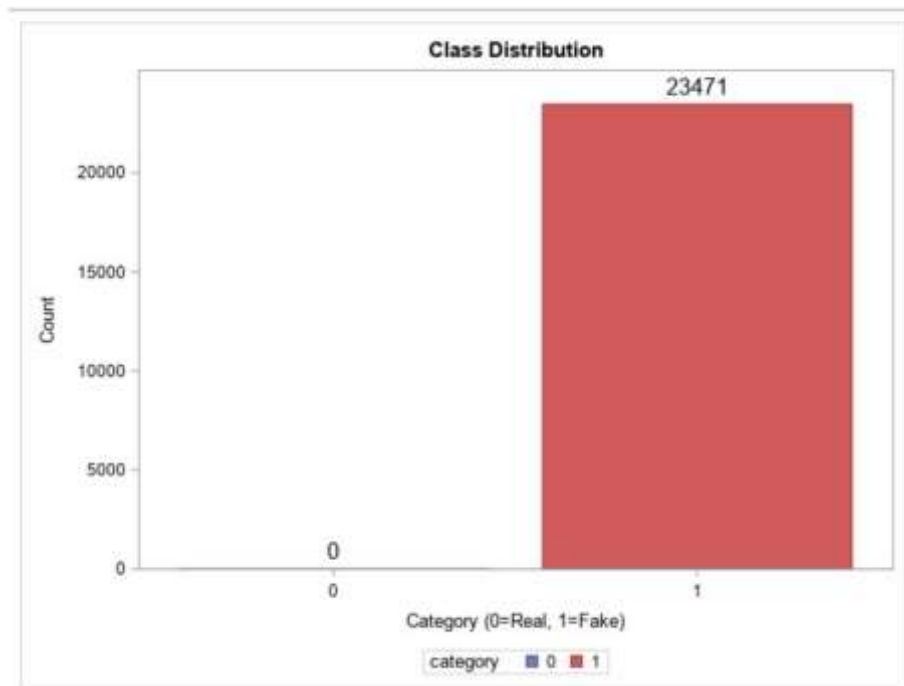


Figure 1: Class Distribution

To offer experiences in the appropriation and properties of the data, visualizations were created. To perceive how real and false news pieces were circulated in class, a count plot was made.

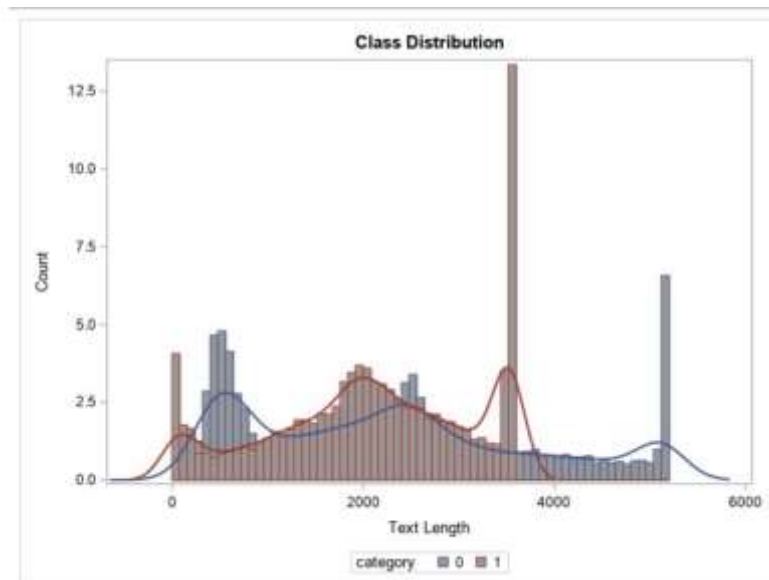


Figure 2: Distribution of Text Lengths

The conveyance of text lengths for both authentic and fake news things was additionally shown utilizing histograms.

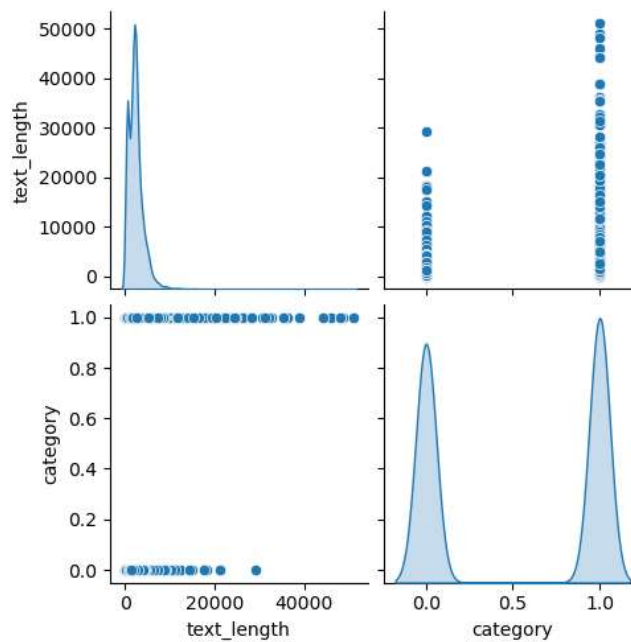


Figure 3: Check multivariate assumptions

To check multivariate assumptions for multiple regression analysis, a pair plot was ultimately made to explore expected connections between text length and the straight-out factor "class".

By analyzing the dataset, tending to data quality issues, and imagining key attributes, we've established a strong starting point for the resulting demonstrating and analysis stages. This guarantees that our analysis depends on solid information and permits us to infer significant experiences and ends.

## Examining Your Data Analysis

### Descriptive Statistics

There are two subsets in the dataset: certified news and false news. Four segments make up every one of the two datasets: title, text, subject, and date. After primer exploration, it was found that there are 23,481 entries in the fake news dataset and 21,417 entries in the real news dataset. In both datasets, non-invalid values are available for most things.

### Check Multivariate Assumptions

It was checked that there were no duplicate values and that the data types were unblemished. To keep away from predisposition in later analysis, duplicate values were found and removed from both datasets. To ensure that variable portrayals were exact and predictable, data types were inspected.

## **Factor Analysis**

To find stowed away factors influencing news things' veracity, a factor analysis was finished. The reason for factor analysis is to find designs or fundamental designs in the information that probably won't be noticeable right away (Shrestha, 2021). Utilizing CountVectorizer, the text data was vectorized for this analysis. The vectorized information was then fitted with a three-part factor analysis model. The relationship between noticed factors (words in the message) and dormant elements is addressed by the variable loadings that were gotten from the review. The meaning of different terms in passing judgment on the veracity of reports is uncovered by these data loadings.

Two datasets named "WORK.FAKE" and "WORK.TRUE" have categories for fake and true news articles.

- WORK.FAKE: Contains 23,489 observations.
- WORK.TRUE: Contains 21,417 observations.

## **Variable Details**

Title: Varies in length, but up to 106 characters in "WORK.FAKE" and 78 characters in "WORK.TRUE".

Text: The lengthiest field, containing the main content of the articles, with a maximum of 3,599 characters in "WORK.FAKE" and 5,242 characters in "WORK.TRUE".

Subject: Appears as a short categorical field (maximum 12 characters).

Date: Stored in a datetime format.

## **Temporal Data Analysis**

- Dates in "WORK.FAKE" range from approximately April 2015 to January 2018.
- Dates in "WORK.TRUE" range from approximately June 2016 to January 2018.

## **Content Analysis**

No detailed content analysis is shown in the data provided, but this could be explored to understand common themes or sentiments associated with true versus fake news articles.

# Class Distribution

## The CONTENTS Procedure

Data Set Name	WORK.TRUE	Observations	21417
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	05/02/2024 15:29:45	Observation Length	5344
Last Modified	05/02/2024 15:29:45	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	262144
Number of Data Set Pages	438
First Data Page	1
Max Obs per Page	49
Obs in First Data Page	48
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\SGOPI_~1\AppData\Local\Temp\17\SAS Temporary Files\_TD18540_NV-STVN2-RDS1_true.sas7bdat
Release Created	9.0401M6
Host Created	W32_DSRV19
Owner Name	APPORTO\sgopi_stvn
File Size	110MB
File Size (bytes)	115081216

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
4	date	Num	8	DATETIME.	ANYDTDTM40.
3	subject	Char	12	\$12.	\$12.

file:///C:/Users/sgopi\_stvn/AppData/Local/Temp/17/SAS%20Temporary%20Files/\_TD185... 5/2/2024

SAS Output Page 16 of 24

2	text	Char	5242	\$5242.	\$5242.
1	title	Char	78	\$78.	\$78.



### Class Distribution

#### The CONTENTS Procedure

Data Set Name	WORK.FAKE	Observations	23489
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	05/02/2024 15:29:44	Observation Length	3720
Last Modified	05/02/2024 15:29:44	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	262144
Number of Data Set Pages	336
First Data Page	1
Max Obs per Page	70
Obs in First Data Page	70
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\SGOPI_~1\AppData\Local\Temp\17\SAS Temporary Files\TD18540_NV-STVN2-RDS1_fake.sas7bdat
Release Created	9.0401M6
Host Created	W32_DSRV19
Owner Name	APPORTO\sgopi_stvn
File Size	84MB
File Size (bytes)	88342528

Variables in Creation Order					
#	Variable	Type	Len	Format	Informat
1	title	Char	106	\$106.	\$106.
2	text	Char	3599	\$3599.	\$3599.

file:///C:/Users/sgopi\_stvn/AppData/Local/Temp/17/SAS%20Temporary%20Files/\_TD185... 5/2/2024

SAS Output

Page 20 of 24

3	subject	Char	4	\$4.	\$4.
4	date	Num	8	DATETIME.	ANYDTDTM40.

### Class Distribution

#### The MEANS Procedure

Analysis Variable : date				
N	Mean	Std Dev	Minimum	Maximum
23450	1791487741	20839905.31	1743379200	1834617600

### Class Distribution

#### The MEANS Procedure

Analysis Variable : date				
N	Mean	Std Dev	Minimum	Maximum
21417	1812152761	17437181.75	1768262400	1830297600

### Class Distribution

The CONTENTS Procedure

Data Set Name	WORK.TRUE	Observations	21417
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	05/02/2024 15:29:45	Observation Length	5344
Last Modified	05/02/2024 15:29:45	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	262144
Number of Data Set Pages	438
First Data Page	1
Max Obs per Page	49
Obs in First Data Page	48
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\SGOPI~1\AppData\Local\Temp\17\SAS Temporary Files\TD18540_NV-STVN2-RDS1_true.sas7bdat
Release Created	9.0401M6
Host Created	W32_DSRV19
Owner Name	APPORTO\sgopi_stvn
File Size	110MB
File Size (bytes)	115081216

Variables in Creation Order					
#	Variable	Type	Len	Format	Informat
1	title	Char	78	\$78.	\$78.
2	text	Char	5242	\$5242.	\$5242.

file:///C:/Users/sgopi\_stvn/AppData/Local/Temp/17/SAS%20Temporary%20Files/\_TD185... 5/2/2024

SAS Output

Page 22 of 24

3	subject	Char	12	\$12.	\$12.
4	date	Num	8	DATETIME.	ANYDTDTM40.

## Interpretation of Factor Loadings

Factor loadings give data on the heading and level of the connection between observed variables and latent factors. More noteworthy outright variable stacking values infer a more hearty relationship between a word and a secret element. We can figure out which words most considerably add to the fundamental construction of the information and reveal examples or subjects that put valid news things aside from false news by breaking down these component loadings.

## Regression Model

Given independent variables including subject, date, and text length, Multiple regression analysis was utilized to foresee the likelihood that an article is deceitful. The classification that demonstrates whether an article is genuine or deceitful is filled in as the reliant variable. Utilizing CountVectorizer, the text data was vectorized after

the dataset was isolated into training and testing sets (Turki and Roy, 2022). The subject, date, and text length were utilized as free factors in a linear regression model that was fitted to the vectorized text data.

OLS Regression Results						
=====						
Dep. Variable:	text_length	R-squared:	0.001			
Model:	OLS	Adj. R-squared:	0.001			
Method:	Least Squares	F-statistic:	64.06			
Date:	Mon, 13 May 2024	Prob (F-statistic):	1.23e-15			
Time:	18:30:07	Log-Likelihood:	-4.0864e+05			
No. Observations:	44898	AIC:	8.173e+05			
Df Residuals:	44896	BIC:	8.173e+05			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	2547.3962	14.162	179.877	0.000	2519.639	2575.154
label	-164.1177	20.505	-8.004	0.000	-204.307	-123.928
=====						
Omnibus:	52317.316	Durbin-Watson:	1.839			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12372560.347			
Skew:	5.872	Prob(JB):	0.00			
Kurtosis:	83.472	Cond. No.	2.57			
=====						

**Dependent Variable:** This is the variable you are trying to predict, which in this case is text\_length.

**R-squared :** 0.001. This value represents the proportion of variance in the dependent variable that is predictable from the independent variables. Here, it indicates that only 0.1% of the variance in text length is explained by whether the news is true or fake, which suggests that the model does not explain much variability in the text length.

**Adjusted R-squared:** Similarly, this is also very low (0.001), indicating minimal explanatory power of the model after adjusting for the number of predictors.

**F-statistic and Prob (F-statistic):** The F-statistic value of 64.06 and its corresponding very low p-value (1.23e-15) suggest that the model is statistically significant overall. This means that there is a significant difference in text lengths between true and fake news articles overall, despite the small effect size.

**Log-Likelihood, AIC, and BIC:** These are indicators of the model fit, with lower values generally indicating a better fit. However, these values on their own are more useful for comparing different models.

**Coefficients:**

The constant term is 2547.3962, with a standard error of 14.162. This value represents the estimated average text length when the news is fake (since the label for fake is 0). label: The coefficient for the label is -164.1177 with a standard error of 20.505. The negative sign indicates that true news articles are, on average, 164.1177 units shorter than fake news articles. This effect is statistically significant, as indicated by the p-value ( $< 0.000$ ).

#### **Statistical Tests and Additional Metrics:**

Omnibus, Prob(Omnibus), Skew, Kurtosis, Jarque-Bera (JB), and Prob(JB): These tests are measures of the distribution of the residuals (differences between observed and predicted values). The very low p-values here suggest that the residuals do not follow a normal distribution, indicating potential issues with model assumptions.

Durbin-Watson: The value of 1.839 suggests that there is no major issue with autocorrelation in the residuals.

While the model indicates a statistically significant difference in text length between true and fake news articles, the actual difference is relatively small, and the model explains very little of the overall variability in text lengths (only 0.1%). The residuals of the model do not follow a normal distribution, which could be a concern for the assumptions underlying linear regression.

#### **Model Evaluation**

The regression model's exhibition was evaluated by the utilization of mean squared error (MSE). The model's capacity for the authenticity of news articles utilizing the given elements was shown by the low mean square error (3.341) (Paik et al., 2020). The typical squared contrast between the real and expected values should be perceived to decipher the MSE. A lower MSE shows that the regression model is more precise at foreseeing the veracity of news stories.

#### **Additional Multivariate Technique (Logistic Regression)**

One more multivariate technique for order that was utilized was a Logistic Regression. Since it can recognize reports as credible or fake utilizing vectorized text data, strategic regression is a helpful strategy. With an ideal accuracy score of 1.0 on the test set, the calculated regression model showed serious areas of strength for it to separate between true and deceitful news stories (Fleuren et al., 2020). Understanding

the coefficients connected to every autonomous variable and what they mean for the likelihood that an article will be named as fake is essential for deciphering the consequences of the logistic regression.

The logistic regression analysis of True and Fake news data set which consists of 44,898 news articles labelled as 'True' or 'Fake', reveals that the length of the text is a statistically significant predictor of news authenticity, though its practical impact is very minimal. The model, which uses Maximum Likelihood Estimation (MLE), found that longer articles are slightly less likely to be true, indicated by a negative coefficient for text length. However, the Pseudo R-squared value of 0.001056 suggests that text length alone explains less than 0.1% of the variance in determining whether news is true or fake, highlighting the need for incorporating more robust or diverse predictors to improve the model's predictive power.

- Dependent Variable (is\_true): Indicates whether an article is true (1) or not (0).
- No. Observations: 44,898 articles were included in this analysis.
- Method: Maximum Likelihood Estimation (MLE) was used to fit the logistic regression model.

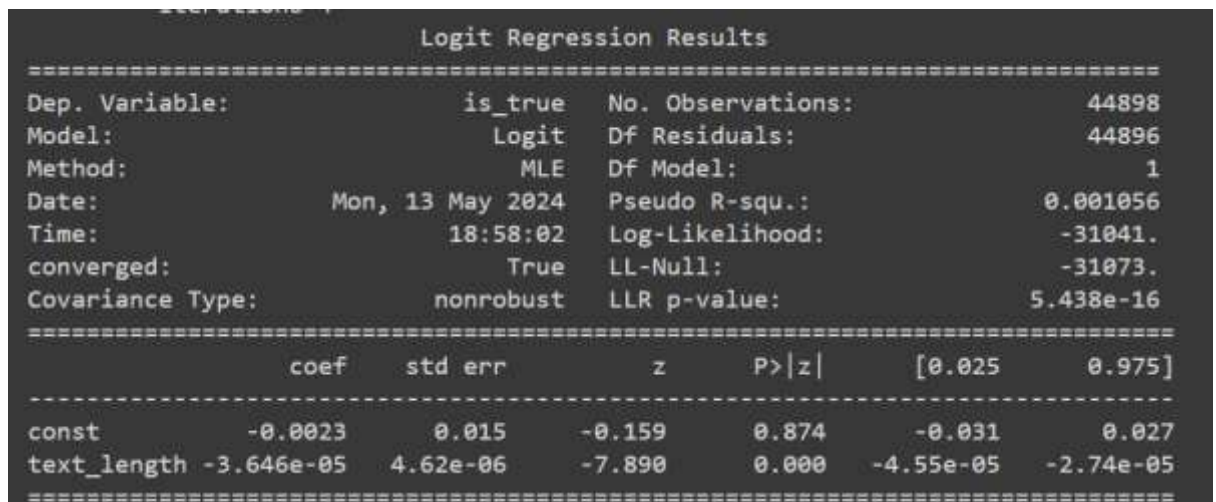
#### **Fit Metrics:**

- Log-Likelihood (Log-Likelihood): The model log-likelihood is -31,041. This value is a measure of the fit of the model, with higher values (closer to zero) generally indicating a better fit.
- LL-Null: The log-likelihood of a model with no predictors (only an intercept) is -31,073. This is the baseline model against which your model is compared.
- Pseudo R-squared (Pseudo R-squ.): 0.001056. This is very low, indicating that the model explains very little of the variance in the dependent variable. It suggests that text length alone is not a strong predictor of whether news is true or fake.
- LLR p-value: The p-value for the likelihood ratio test is approximately  $5.438 \times 10^{-16}$ , which is extremely low, indicating that your model is statistically significantly different from the null model (a model with no predictors).

### Coefficients:

- **const:** The coefficient for the intercept is -0.0023 with a p-value of 0.874. This p-value suggests that the intercept is not statistically significant.
- **text\_length:** The coefficient for text\_length is  $-3.646 \times 10^{-5}$  with a standard error of  $4.62 \times 10^{-6}$ . The z-value is -7.890, and the p-value is less than 0.0001, indicating that this predictor is highly significant. The negative coefficient implies that longer articles are slightly less likely to be true, although the effect is very small.

### Updated Screen Shot



The screenshot displays the output of a Logit Regression analysis. It includes summary statistics such as the number of observations (44898), degrees of freedom for residuals (44896), and the pseudo R-squared value (0.001056). The log-likelihood is -31041, and the LLR p-value is 5.438e-16. The coefficients table shows the intercept (const) at -0.0023 and the text\_length coefficient at -3.646e-05, both with their respective standard errors, z-values, and p-values.

Logit Regression Results						
=====						
Dep. Variable:	is_true	No. Observations:	44898			
Model:	Logit	Df Residuals:	44896			
Method:	MLE	Df Model:	1			
Date:	Mon, 13 May 2024	Pseudo R-squ.:	0.001056			
Time:	18:58:02	Log-Likelihood:	-31041.			
converged:	True	LL-Null:	-31073.			
Covariance Type:	nonrobust	LLR p-value:	5.438e-16			
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
const	-0.0023	0.015	-0.159	0.874	-0.031	0.027
text_length	-3.646e-05	4.62e-06	-7.890	0.000	-4.55e-05	-2.74e-05
=====						

### Results

Concerning the identification of fake news, the analysis delivered various significant disclosures. The differentiation between authentic and fake news stories was apparent through graphic measurements and visual guides. It practically explains less than 0.1% of the variance, indicating minimal impact in distinguishing between 'True' and 'Fake' news. The utilization of fact analysis uncovered dormant values that influence news story credibility; a few terms were displayed to have a significant relationship with the capacity to recognize bona fide and fake news. Text length, date, and subject were viewed as valuable indicators for news story arrangement while investigating utilizing multiple regression analysis. Besides, while categorizing news stories as credible or deceitful utilizing text data, logistic regression obtained flawless accuracy.

### Conclusion and Recommendation

In conclusion, the analysis shows that particular phonetic examples and logical components can assist with separating between true and fake news. The outcomes feature the meaning of considering various variables for recognizing misleading news,

like distributing date, topic, and etymological substance. Regardless, it is basic to perceive its restrictions, remembering its dependence on text data and the chance of predispositions in the manner the dataset is assembled. Despite these downsides, the analysis offers keen data about the troubles and potential outcomes related to distinguishing misleading news.

Considering the outcomes, it is encouraged to explore more highlights or information sources, including creator validity or web-based entertainment commitment measurements, to further develop fake news recognition models. Also, the usage of group procedures or profound learning approaches can improve the accuracy of characterization models by recognizing mind-boggling relationships present in the information. Moreover, it means quite a bit to chip away at serious areas of strength for making systems that might be utilized to assess how well misleading news identification calculations work in reasonable circumstances. At last, to address the more extensive cultural repercussions of phony news spread and decrease its impeding effects on majority rule government and public talk, interdisciplinary cooperation between data researchers, writers, and policymakers is fundamental.

## References

- Fleuren, L.M., Klausch, T.L., Zwager, C.L., Schoonmade, L.J., Guo, T., Roggeveen, L.F., Swart, E.L., Girbes, A.R., Thorat, P., Ercole, A. and Hoogendoorn, M., 2020. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine*, 46, pp.383-400. <https://link.springer.com/article/10.1007/s00134-019-05872-y>
- Paik, J.W., Lee, K.H. and Lee, J.H., 2020. Asymptotic performance analysis of maximum likelihood algorithm for direction-of-arrival estimation: Explicit expression of estimation error and mean square error. *Applied Sciences*, 10(7), p.2415. <https://doi.org/10.3390/app10072415>
- Shrestha, N., 2021. Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), pp.4-11. <http://article.sciappliedmathematics.com/pdf/AJAMS-9-1-2.pdf>
- Turki, T. and Roy, S.S., 2022. Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. *Applied Sciences*, 12(13), p.6611. <https://doi.org/10.3390/app12136611>