STAT 391 Final Project

Fake News Real News Classification

By Tao Zhang, Zhikai Li, ChenYang Yuan , Ziang Li (Leo)

Introduction:

The widespread use of the internet has facilitated the easy sharing of information but also the spread of fake news, posing a significant challenge. The low barrier to entry accelerates the spread of misinformation, undermining social discourse and institutional credibility. Hence, the development of tools capable of automatically detecting fake news is crucial. Such tools can help fight misinformation and enhance media literacy, leading to a more trustworthy information landscape.

Summary work of Our Project:

Our analysis involves combining two datasets—one for fake news and one for real news—comprising four variables: date, title, text, and subject. We add a 'truthfulness' attribute to each article and discover that 'subject' and 'date' aren't reliable indicators of veracity. Thus, we convert 'date' into the day of the week and merge 'title' and 'text' into a single variable. This combined text is then categorized by word count intervals of 100 words. We then focus on the mean-word-length, the top 25 single words, two-word phrases, and three-word phrases based on tf-idf values, ignoring others. Then we use those variables for logistic regression analysis. Using the R language, we assess variable relevance, multicollinearity, outliers, and high leverage points. Finally, we test this model with some current news from Reddit. We find that as time passes, the accuracy of predicting current news with past news decreases.

Introduction of Our Dataset:

For our study, we leverage the "Fake and Real News" dataset from Kaggle, which is openly available for download in the "Fake News Detection Datasets" section on the Kaggle website below.
Fake News Detection Datasets (kaggle.com)

This dataset is divided into two parts: "True.csv" and "Fake.csv". The "True.csv" file contains legitimate news articles collected from Reuters.com, while the "Fake.csv" file consists of articles classified as fake, gathered from sites flagged by Politifact and Wikipedia for their unreliability. Each record within the dataset encompasses information such as the article's title, text, genre, and date of publication from 2016 to 2017.

Research Questions:

- **RQ1:** What features classify fake news and real news? (By Ziang Li)
- **RQ2:** Can we use logistic regression to build a model using those features and what is the performance? (By Zhikai Li)
- **RQ3:** Are there any relationship between those features and how do they influence the model under the 5% significant level? (By Tao Zhang)
- **RQ4:** How does the model perform on the news nowadays? (By ChenYang Yuan)

Results:

RQ1: What features classify fake news and real news?

To explore RQ1, We adopted a structured approach to analyze the metadata (title, content, and date) from our dataset, aiming to identify distinguishing characteristics of fake versus real news. Our methodology is detailed as follows:

1. **Day of the week**:

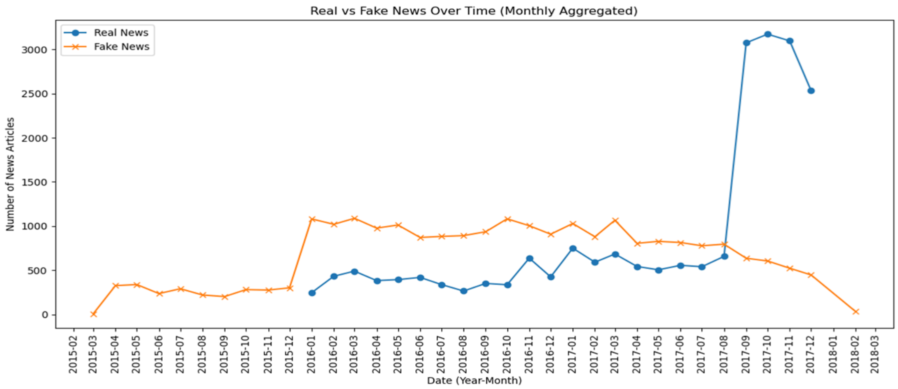


Figure 1: plot of number of news according to date

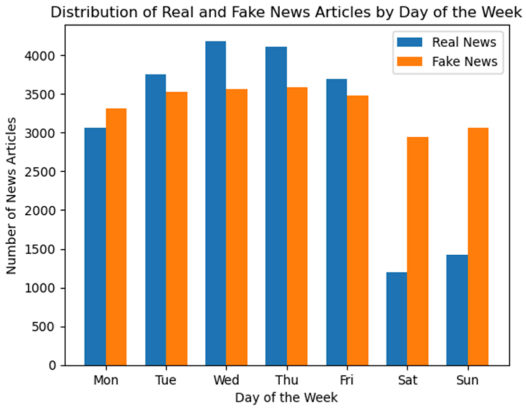


Figure 2: plot of number of news according to day of week

Initially, we considered the publication date as a potential distinguishing feature. However, we noticed a discrepancy in the date ranges of the fake and real news datasets, which could potentially bias our analysis. To address this, we converted the publication dates into the days of the week. This step was taken to investigate possible weekly patterns in news publishing that may differ between fake and real news sources, sidestepping the issue of mismatched date ranges and focusing on a broader temporal pattern that might be indicative of authenticity. Our analysis revealed a notable trend: the prevalence of fake news increases during weekends, compared to real news. This observation aligns with the rationale that real news production involves human journalists, who typically follow a standard workweek schedule, thereby reducing output on weekends. This pattern supports the hypothesis that fake news distributors may exploit weekends, when there is a dip in the production of legitimate news, to disseminate misinformation more freely.

2. **Categories:**

We also evaluated the potential of using news subjects as features for distinguishing fake from real news. However, from the plots on the right we found that the subjects present in our dataset for fake and real news do not match, making it impractical to use subjects as a reliable feature for classification.

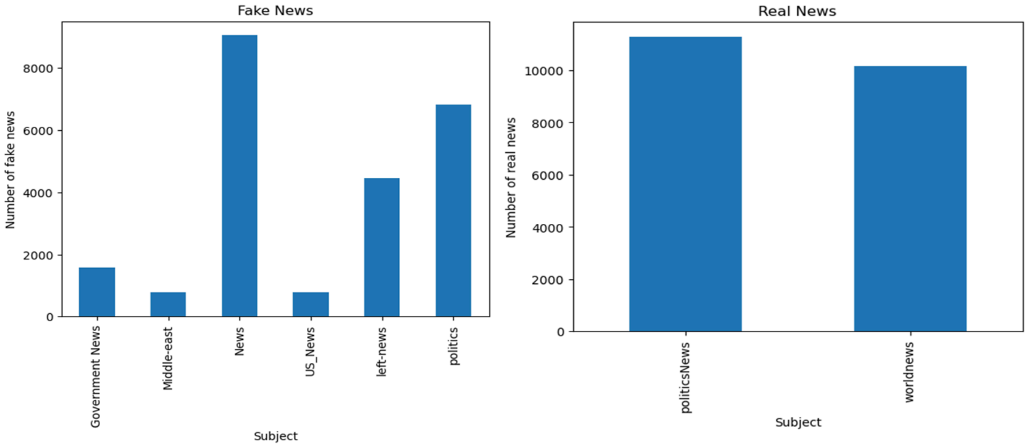


Figure 3: plot of number of news according to subject

3. **Word Count & Mean Word Length:**

Our hypothesis that fake news might differ in word count from real news was confirmed by analyzing our dataset. Real news articles generally have lower word counts, while fake news is more common in middle to higher ranges, especially beyond 700 words. This distinct pattern enables us to use article length as a significant feature in our classification model, effectively distinguishing between fake and real news.

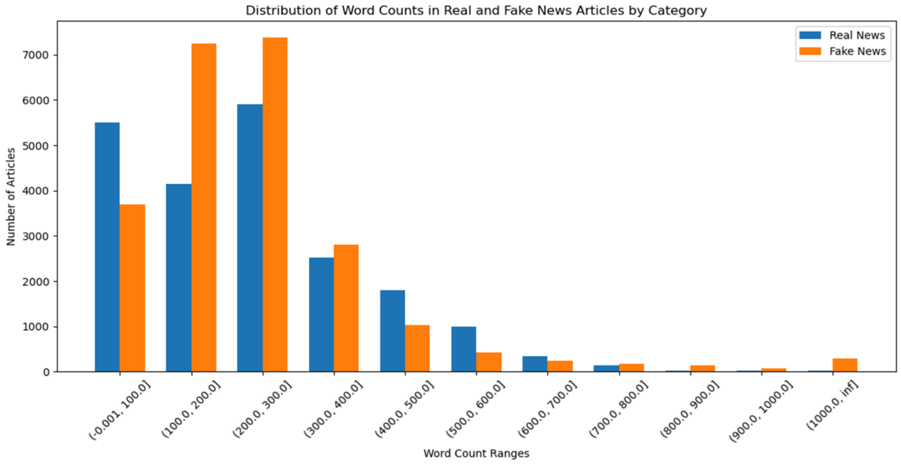
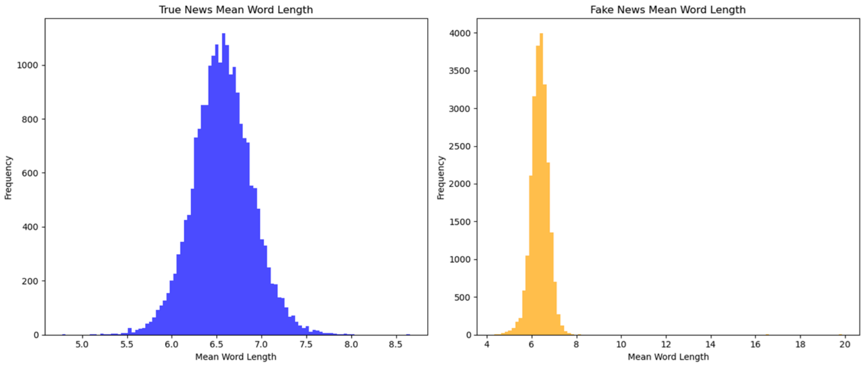


Figure 4: word count plot for news

Figure 5: mean word length plot for news



We further analyzed mean word length for each article, plotting the results. The analysis revealed that real news articles have a mean word length predominantly ranging from 5 to 8, centering around 6.6, suggesting a normal distribution. Conversely, fake news shows a broader spread, from 4 to 8, with outliers extending above 16, indicating a distribution that's slightly skewed to the left and a mean close to 6.5. This distinction in word length distribution between real and fake news offers another layer of differentiation for our classification.

4. **Top 30 words and TF-IDF Analysis:**

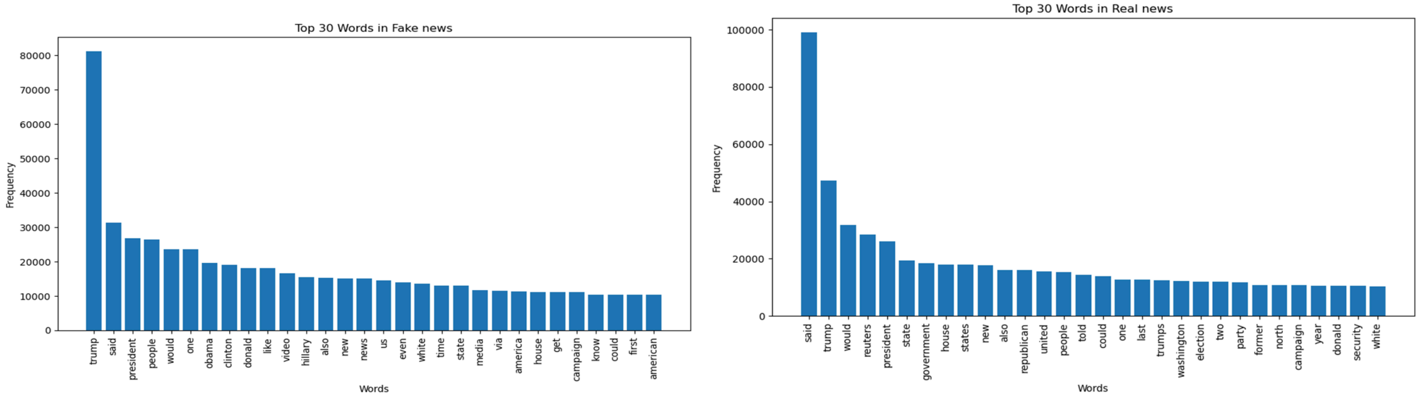
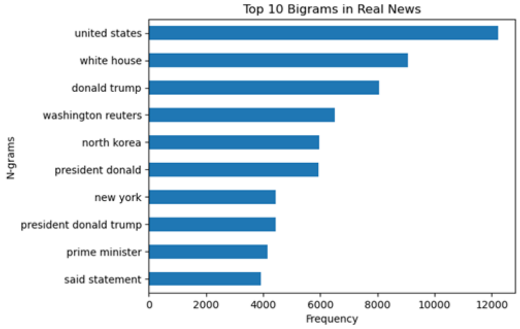
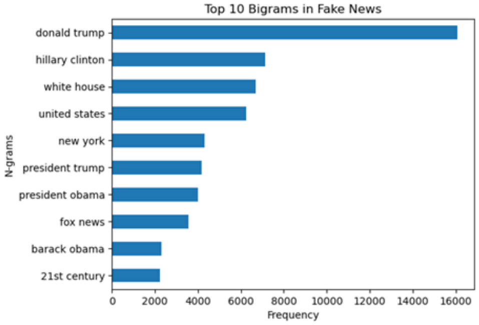


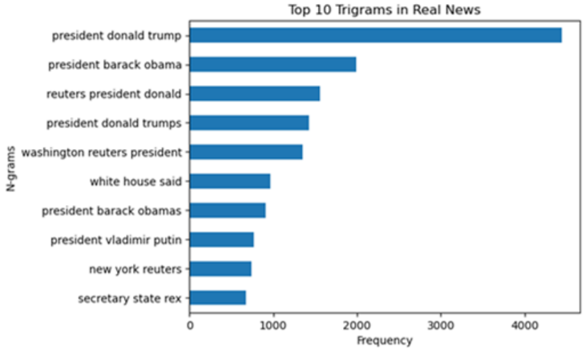
Figure 6: Top 30 words in both real and fake news

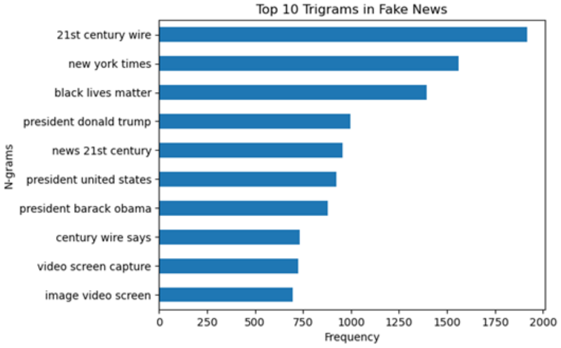
To delve deeper into the textual content, we analyzed the top-30 words in both real and fake news articles, followed by Term Frequency-Inverse Document Frequency (TF-IDF) analysis. This method assesses the significance of words within an article across a corpus, based on the premise that certain terms could be more common in either fake or real news. The TF-IDF scores for these top words were then utilized as features for classification, providing a nuanced approach to distinguish between fake and real news based on word importance.

5. N-gram Modeling:





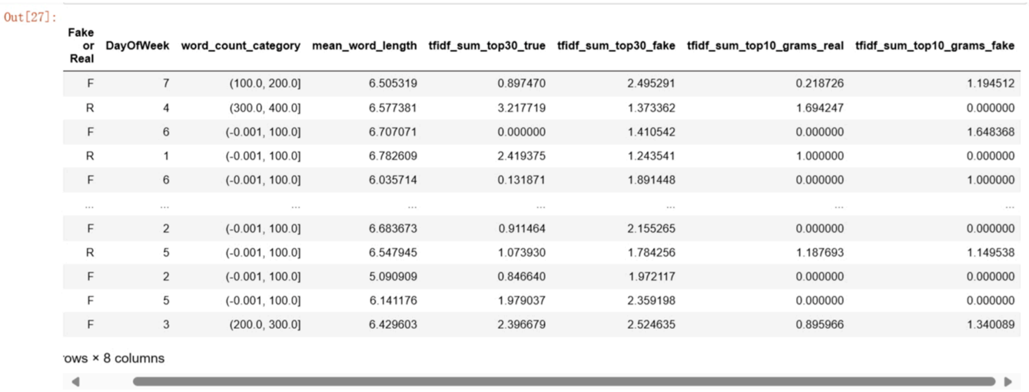




We incorporated N-grams—sequences of N contiguous words from the text—into our analysis, focusing on 2-gram (bigram) and 3-gram (trigram) models. These N-grams help identify phrases and expressions characteristic of either fake or real news, shedding light on stylistic and structural language differences. The TF-IDF of these N-grams was calculated and used as features in our classification model, enhancing our ability to differentiate between the two types of news based on linguistic patterns.

6. Final Features:

Following our comprehensive analysis, we've established a complete set of features for our model: the first column serves as the label, representing the target of our prediction. The second and third columns contain categorical features, while the subsequent columns are dedicated to numerical features.



Out[27]:

Fake or Real	DayOfWeek	word_count_category	mean_word_length	tfidf_sum_top30_true	tfidf_sum_top30_fake	tfidf_sum_top10_grams_real	tfidf_sum_top10_grams_fake
F	7	(100.0, 200.0]	6.505319	0.897470	2.495291	0.218726	1.194512
R	4	(300.0, 400.0]	6.577381	3.217719	1.373362	1.694247	0.000000
F	6	(-0.001, 100.0]	6.707071	0.000000	1.410542	0.000000	1.648368
R	1	(-0.001, 100.0]	6.782609	2.419375	1.243541	1.000000	0.000000
F	6	(-0.001, 100.0]	6.035714	0.131871	1.891448	0.000000	1.000000
...	...	...	...	...	...	...	...
F	2	(-0.001, 100.0]	6.683673	0.911464	2.155265	0.000000	0.000000
R	5	(-0.001, 100.0]	6.547945	1.073930	1.784256	1.187693	1.149538
F	2	(-0.001, 100.0]	5.090909	0.846640	1.972117	0.000000	0.000000
F	5	(-0.001, 100.0]	6.141176	1.979037	2.359198	0.000000	0.000000
F	3	(200.0, 300.0]	6.429603	2.396679	2.524635	0.895966	1.340089

ows × 8 columns

Figure 7: Final features

RQ2: Can we use logistic regression to build a model using those features and what is the performance?

Utilizing the key features identified from RQ1, we developed a logistic regression model. We divided our dataset into a training set and a test set—the latter is used to evaluate the model's accuracy. The training set achieved an accuracy of 0.90254, while the test set achieved 0.90305. The similar accuracies suggest a low risk of overfitting, indicating our model generalizes effectively. Consequently, we opted not to use a validation set or cross-validation methods for overfitting assessment.

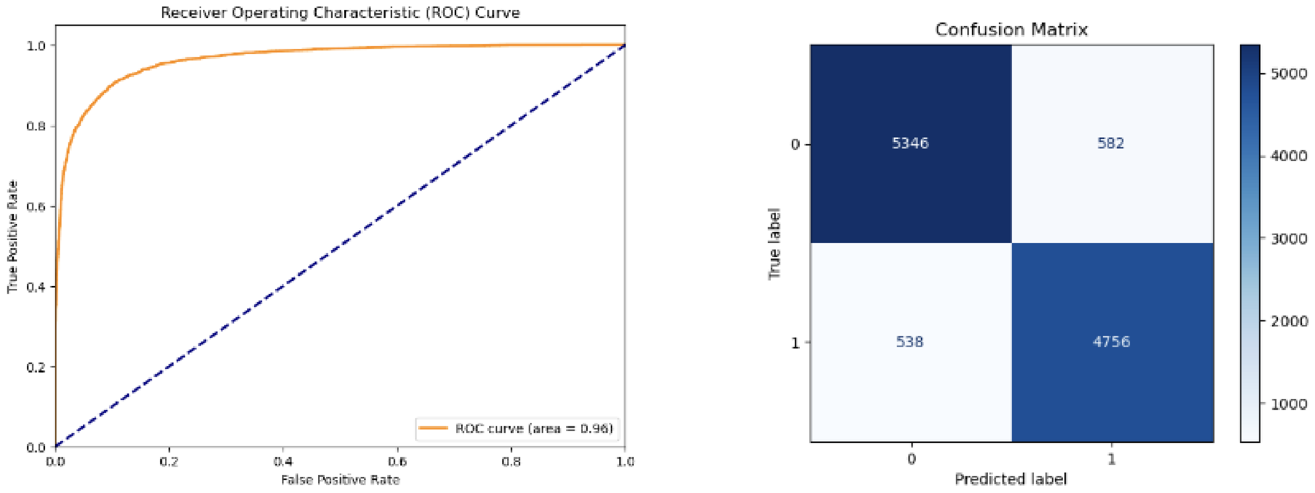


Figure 8: ROC and confusion matrix

We then plot the ROC curve and the confusion matrix. We can see that the ROC curve of the logistic regression model is consistently moving towards the upper left corner. The area under the ROC curve is large, which means that there is a significant portion we can explain with this model, far exceeding 50%. Regarding the confusion matrix: The accuracy is as high as 90.02%, meaning that 90% of all predictions made by the model are correct, indicating that the overall performance of the model is very good.

- Accuracy: 90.02%, indicating that the model correctly predicted about 90% of the samples, showing that the model's overall performance is very good.
- Error Rate: 9.98%, which is the proportion of predictions that the model got wrong, complementing the accuracy and further confirming the model's high accuracy.
- False Positive Rate (FPR): 10.90%, indicating that approximately 10.90% of all actual negative samples were incorrectly predicted as positive. Although this proportion is relatively low, it may still need to be reduced further in certain application scenarios, such as medical diagnosis.
- False Negative Rate (FNR): 9.14%, indicating that about 9.14% of all actual positive samples were incorrectly predicted as negative. The low FNR indicates that the model performs well in capturing positive samples.

Thus we believe that this logistic regression model has excellent accuracy and generalization ability.

RQ3：Are there any relationship between those features and how do they influence the model under the 5% significant level?

After training the logistic regression model, we began to explore whether these variables were associated with the final prediction outcome—the truthfulness of the news. We plan to exclude those variables with low correlationship and use hypothesis tests to determine whether the variables are related to the truthfulness of the news.

Initially, we explored their correlation. And we can see that the correlation between each variable and the truthfulness of the news does not seem to be very strong; none have reached a level of strong correlation. We skipped this step in determining the relationship between variables, because according to our hypothesis, the three consecutive words, two consecutive words, and the length of the text from the dataset are important features of the news. And we believe that the relationship between these variables and the truthfulness of the news may not be just a simple linear one.

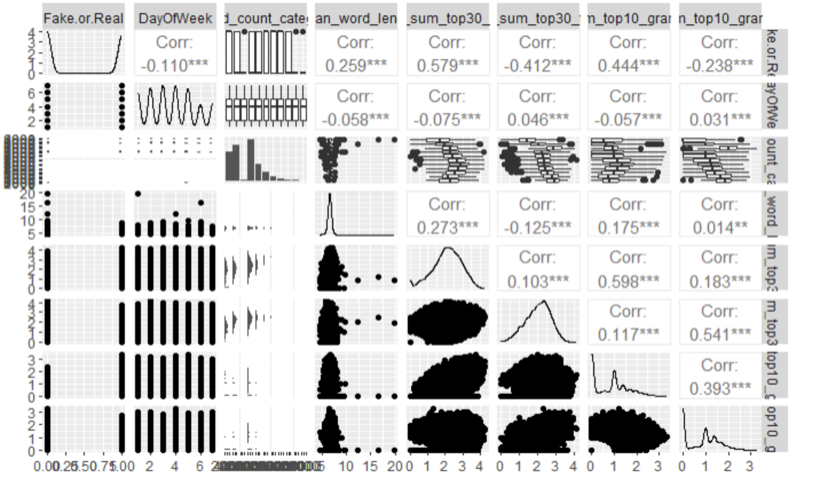
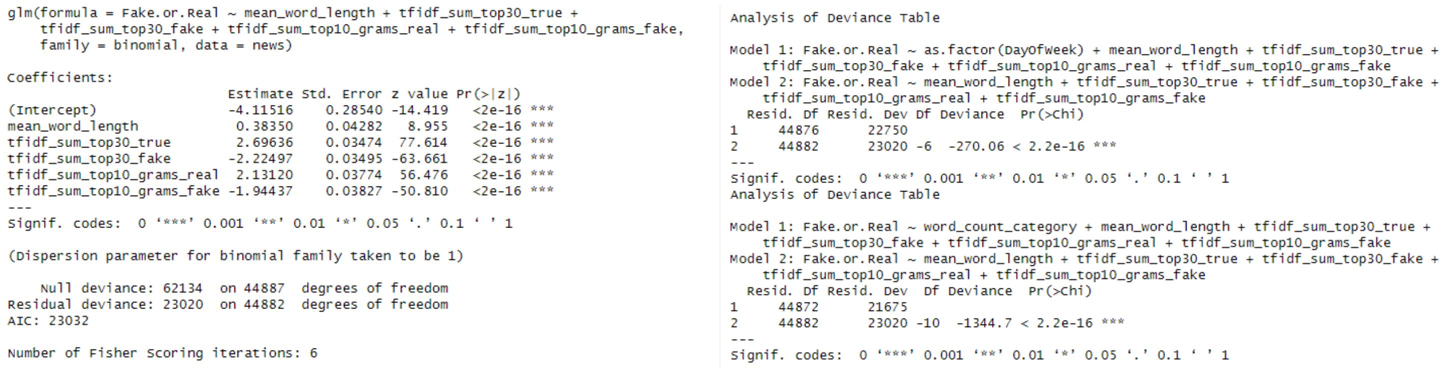


Figure 9: correlation between all variables

We then use hypothesis testing to determine whether the variables' coefficient should be zero or not.



```
glm(formula = Fake.or.Real ~ mean_word_length + tfidf_sum_top30_true +
    tfidf_sum_top30_fake + tfidf_sum_top10_grams_real + tfidf_sum_top10_grams_fake,
    family = binomial, data = news)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.11516    0.28540 -14.419   <2e-16 ***
mean_word_length            0.38350    0.04282   8.955   <2e-16 ***
tfidf_sum_top30_true        2.69636    0.03474  77.614   <2e-16 ***
tfidf_sum_top30_fake       -2.22497    0.03495 -63.661   <2e-16 ***
tfidf_sum_top10_grams_real  2.13120    0.03774  56.476   <2e-16 ***
tfidf_sum_top10_grams_fake -1.94437    0.03827 -50.810   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62134  on 44887  degrees of freedom
Residual deviance: 23020  on 44882  degrees of freedom
AIC: 23032

Number of Fisher Scoring iterations: 6
```

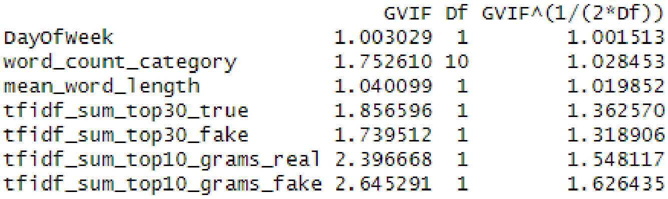
```
Analysis of Deviance Table

Model 1: Fake.or.Real ~ as.factor(DayOfWeek) + mean_word_length + tfidf_sum_top30_true +
    tfidf_sum_top30_fake + tfidf_sum_top10_grams_real + tfidf_sum_top10_grams_fake
Model 2: Fake.or.Real ~ mean_word_length + tfidf_sum_top30_true + tfidf_sum_top30_fake +
    tfidf_sum_top10_grams_real + tfidf_sum_top10_grams_fake
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     44876      22750
2     44882      23020 -6  -270.06 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Deviance Table

Model 1: Fake.or.Real ~ word_count_category + mean_word_length + tfidf_sum_top30_true +
    tfidf_sum_top30_fake + tfidf_sum_top10_grams_real + tfidf_sum_top10_grams_fake
Model 2: Fake.or.Real ~ mean_word_length + tfidf_sum_top30_true + tfidf_sum_top30_fake +
    tfidf_sum_top10_grams_real + tfidf_sum_top10_grams_fake
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1     44872      21675
2     44882      23020 -10  -1344.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: summary for the model

We use the summary() method in R on our logistic regression model, and as shown in the figure above, all numerical variables are significant. This means that the coefficients of all numerical variables should not be zero, and these numerical variables are related to the truthfulness of the news. We can use them to predict the truthfulness of the news. Then, for categorical variables, we used the ANOVA test to evaluate two different categorical variables: word_count_category and DayOfWeek, and determined that both are related to the truthfulness of the news.

Since our predictive variables are feature variables extracted from text, date, and title, we suspect that these variables might interact with each other, ultimately leading to multicollinearity. Thus, we tested the Variance Inflation Factor (VIF) among them.

The test results, as shown on the left, confirm that the GVIFs for all variables do not exceed 5, indicating that there is no significant correlationship between all predictive variables. We can safely use all variables to predict the truthfulness of the news.



```
                              GVIF Df GVIF^(1/(2*Df))
DayOfWeek                 1.003029  1        1.001513
word_count_category       1.752610 10        1.028453
mean_word_length          1.040099  1        1.019852
tfidf_sum_top30_true      1.856596  1        1.362570
tfidf_sum_top30_fake      1.739512  1        1.318906
tfidf_sum_top10_grams_real 2.396668  1        1.548117
tfidf_sum_top10_grams_fake 2.645291  1        1.626435
```

Figure 11: VIF of all variables.

To ensure that the final results are not influenced by high leverage points and outliers, we removed these and checked whether the logistic regression model met its assumptions. Regarding the residuals vs. leverage plot, it appears there are still many high leverage points, but this may actually be due to our very large sample size of the dataset, causing most to hover on the left side.

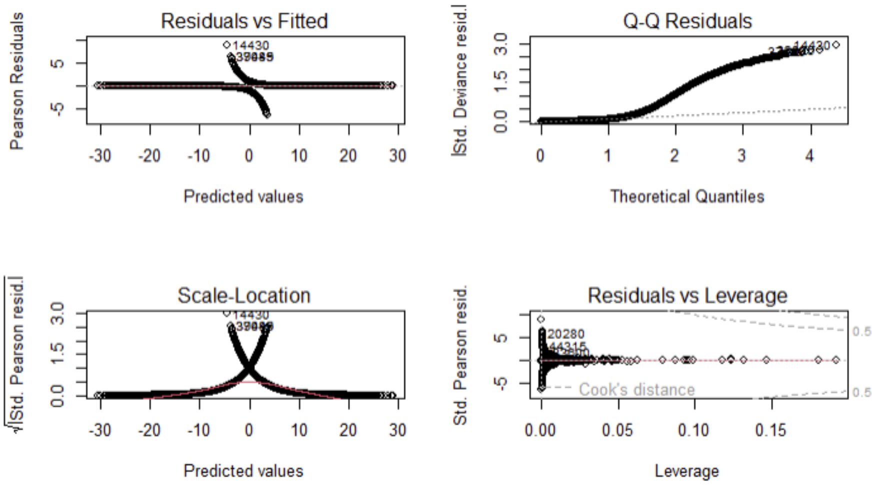
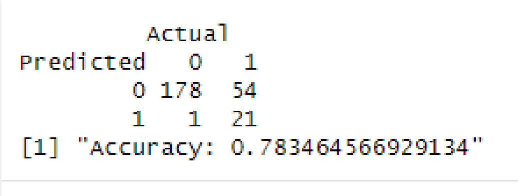


Figure 12: Residual and Leverage Plots of Data

RQ4 : How do those models perform nowadays?

Finally, since the dataset we used primarily comes from 2016 to 2017, we attempted to use our trained model to predict the truthfulness of today's news.

We found some publicly available news information from the Reddit website. This news primarily comes from several clubs: 'fakenews', 'newyorktimes', 'News_Politics', 'news', 'UpliftingNews', 'PupliftingNews', 'politics'. We found a total of 255 samples and conducted some organization.



```
         Actual
Predicted   0   1
        0 178  54
        1   1  21
[1] "Accuracy: 0.783464566929134"
```

Figure 13: Confusion matrix for news nowadays

This is the final confusion matrix. We've noticed a decrease in accuracy compared to before, but we believe this is within an acceptable range, as after all, we are using a logistic regression model trained on past news datasets. News is time-sensitive; the top 20 words in past news might include President Trump, while the current president is Biden. Using a model based on past news to predict current news might miss some key words and key figures, ultimately affecting the prediction results.

Here, according to the confusion matrix, most of our predictions are that the news is false. I believe the main reason is that most of the news from Reddit is fake. Therefore, we rarely predict the news to be true. However, once we consider a piece of news to be true, it is highly likely that the news is indeed true.

Conclusion:

Our project demonstrates the effectiveness of utilizing a model to distinguish fake news from real news with considerable accuracy. By analyzing article characteristics such as word count, word length, and the frequency of specific words and phrases, we developed a model that accurately identifies fake versus real news approximately 90% of the time.

However, its performance dipped when applied to newer articles, highlighting the dynamic nature of news and the evolving language it employs. This variability suggests the importance of continuously updating our model to adapt to current trends and vocabularies. Moving forward, enhancing the model with new indicators and ensuring it remains adaptive to contemporary news stories will be crucial. In essence, our research marks a promising step towards accurately identifying fake news, yet underscores the ongoing need for refinement and updates to maintain its effectiveness in a constantly changing news landscape.

References:

1. BozkuŞ, Emine. “Fake News Detection Datasets.” *Kaggle*, Kaggle, 7 Dec. 2022, www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets.
2. Emelyanov, G.M., Mikhailov, D.V. & Kozlov, A.P. The TF-IDF measure and analysis of links between words within N-grams in the formation of knowledge units for open tests. Pattern Recognit. Image Anal. 27, 825–831 (2017). https://doi.org/10.1134/S1054661817040058
3. Team, The Investopedia. “Variance Inflation Factor (VIF).” *Investopedia*, Investopedia, Sept. 2023, www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=A%20variance%20inflation%20factor%20(VIF)%20is%20a%20measure%20of%20the,adversely%20affect%20the%20regression%20results.
4. Stecanella, B. (2019, May 10). Understanding TF-ID: A simple introduction. MonkeyLearn Blog. https://monkeylearn.com/blog/what-is-tf-idf/