Part2:

Introduction:

In this part, we are using sklearn’s Logistic Regression package to classify the fake news and real news based on the fake news dataset.

Dataset:

The fake news dataset contains the texts and labels. Texts are the news that written by humans or computers. The label is either 0 or 1 that represents the text is written by computer (fake news) and human (real news) respectively.

First, we do the data cleaning process. To count the frequency of words and make it easy to train the model, we change all the words to lower case. Then we remove all the punctuation because we don’t think it can benefit the model. After the data cleaning, we visualize the data in the training dataset. Appendix.

Model training and the result:

We mainly use the LogisticRegression() function to train our model. Before we fit the data into this function, we call CountVectorizer() and TfidfTransformer() functions to converts the collection of text documents to a matrix of token counts[1] and convert the count matrix to a normalized tf or tf-idf representation[2]. Then we use a for loop to do the validation using the fake\_news\_val dataset. We mainly focus on the parameters in LogisticRegression() function. We set the solver = “liblinear” to see which of l1 and l2 penalty is better and we change C from 1 to 20 to find the best number for the model.

After the validation, as shown in figurexx we find that the l2 penalty with C = 17 can give us the highest validation score, which is 0.76. Then we use these best parameters to test our model using fake\_news\_test dataset and get 74.87% accuracy result.

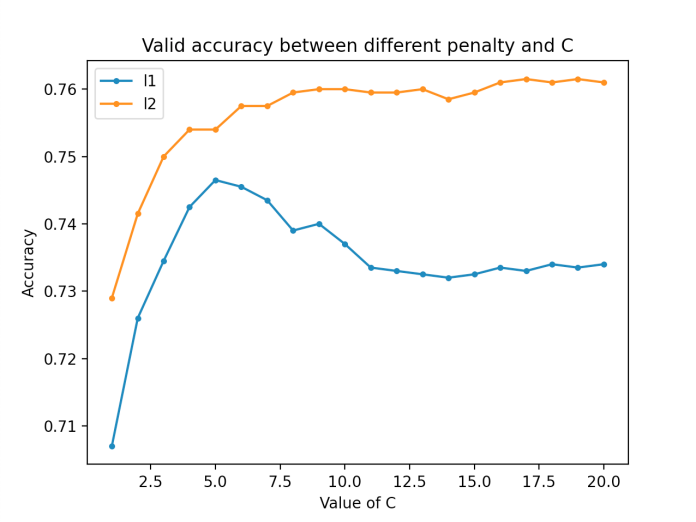


Figure XX. The validation accuracy between different penalty and C.

Reference

[1]https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html

[2]https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.TfidfTransformer.html

Appendix

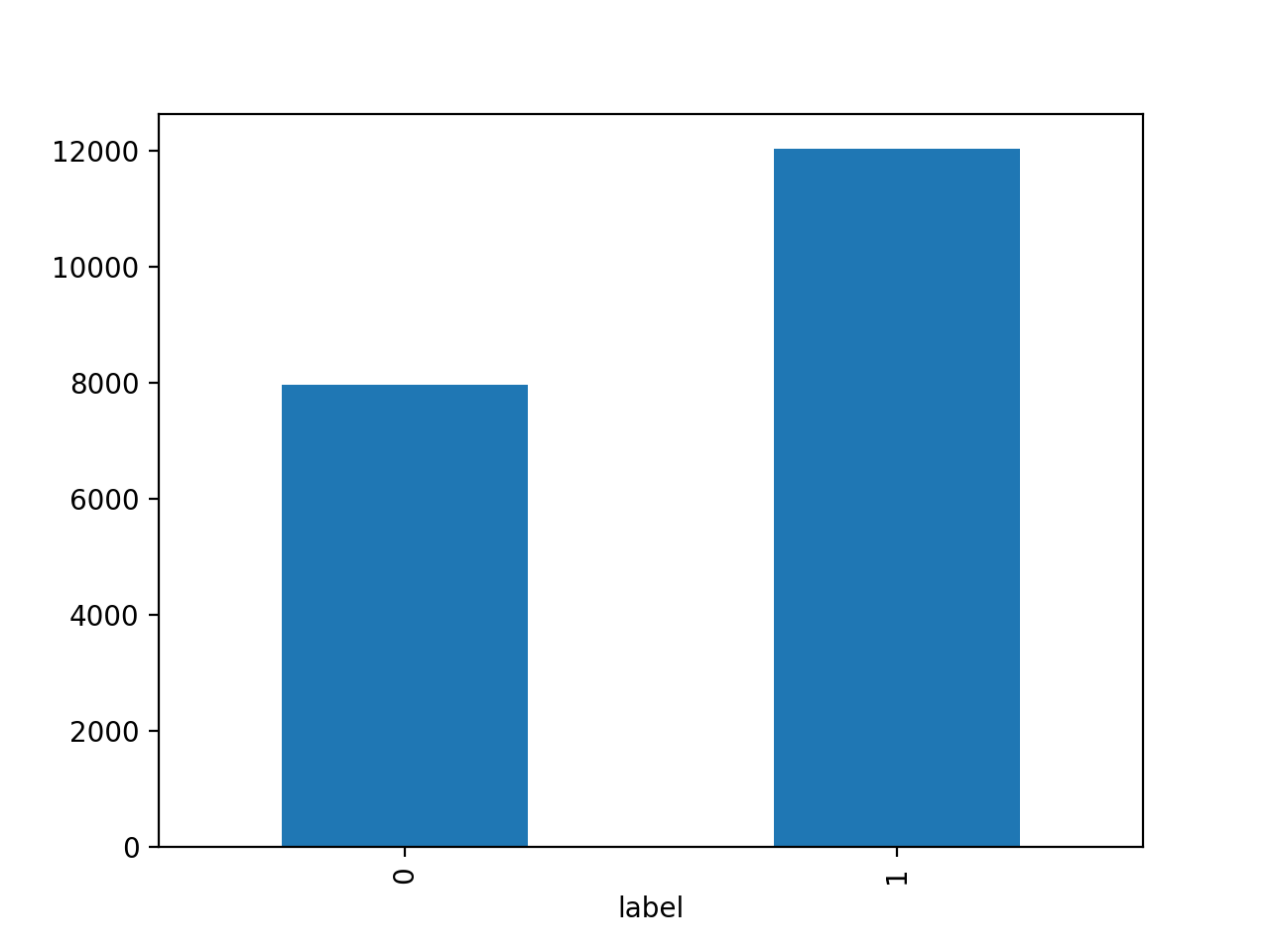


Figure1. visualized result of real news and fake news in fake\_news\_train dataset

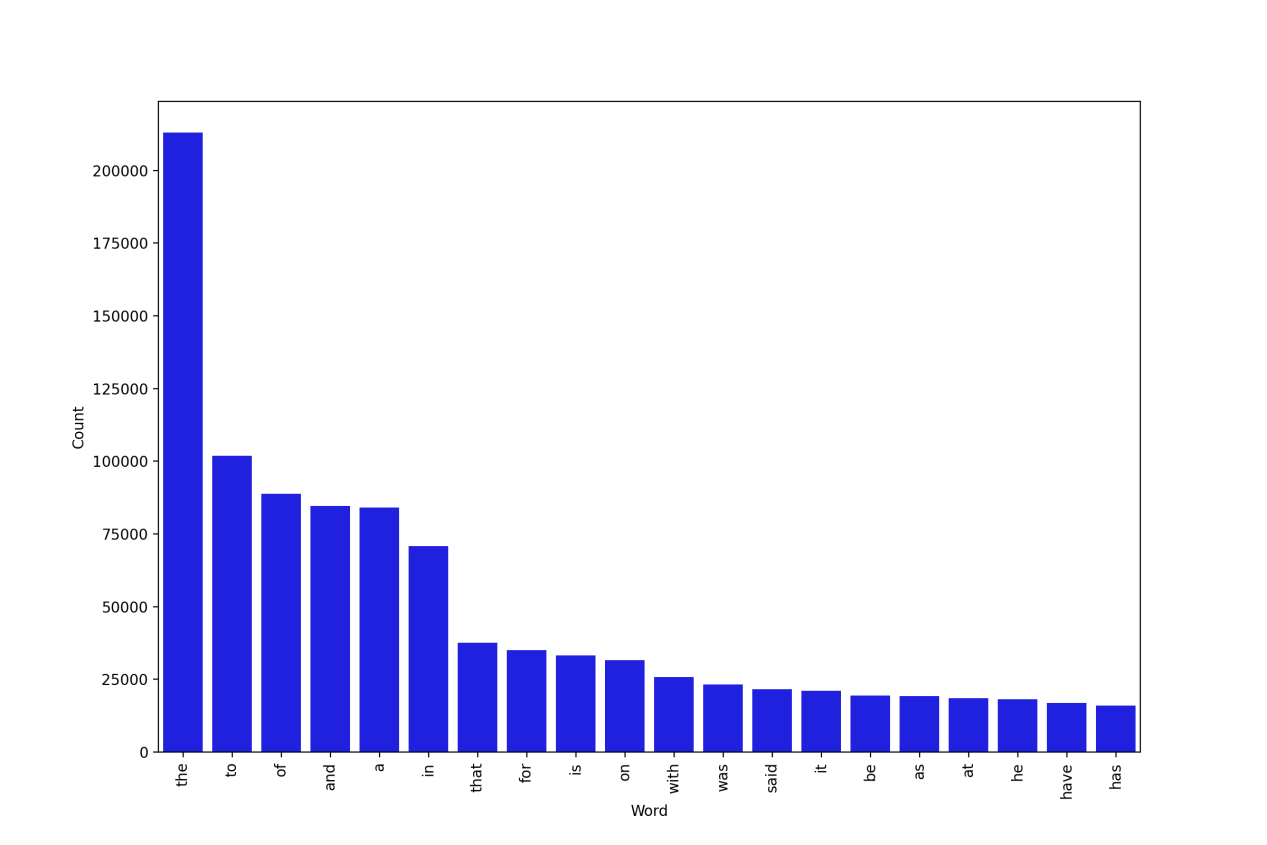


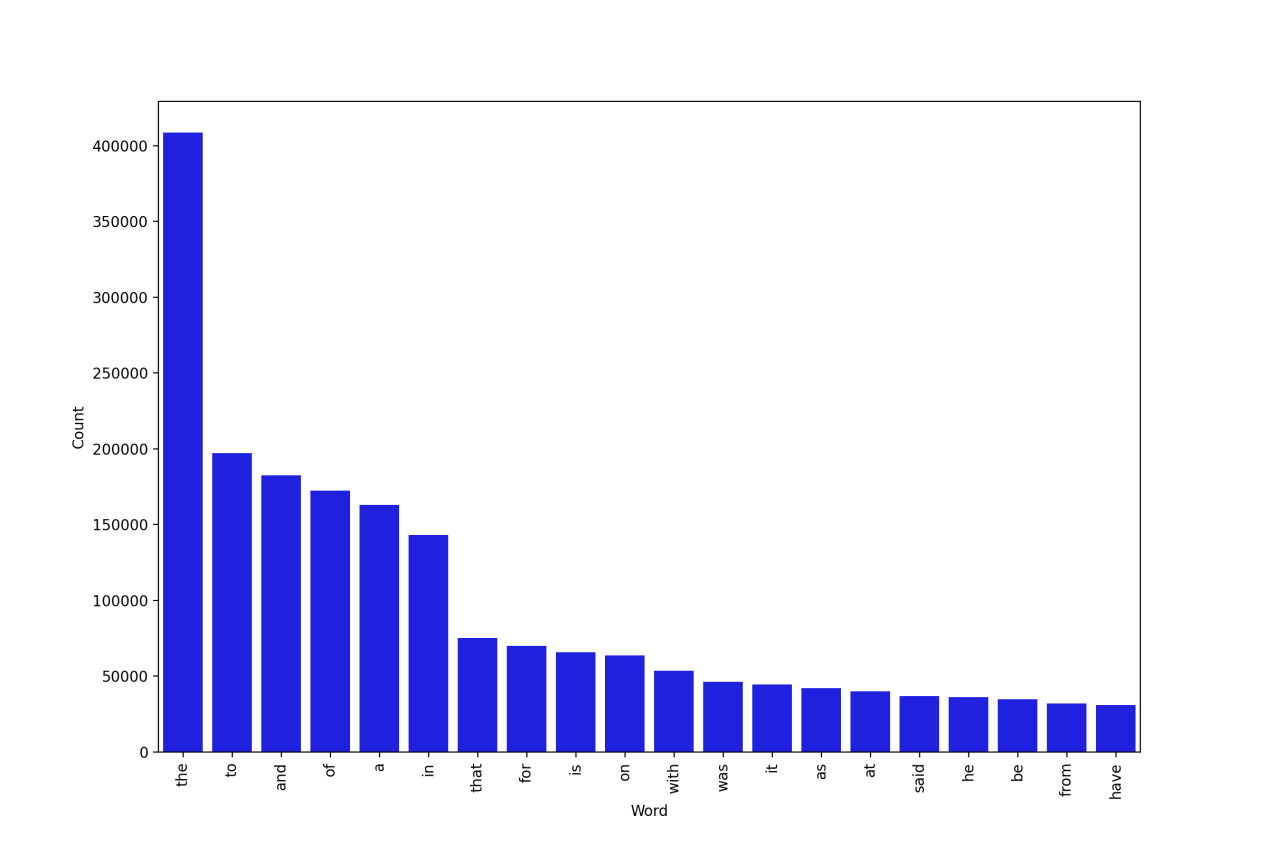
Figure 2. Top 20 words appeared in the fake news 

Figure 3. Top 20 words appeared in the real news