True and Fake News

By: Wenrui Jiang

Abstract:

The purpose of this research is to construct a model to determine if a given news text is true news or fake news. In order to achieve this goal, a preliminary task is necessary: transfer data text into variables that can be used in the model. The algorithm Term Frequency - Inverse Document Frequency (TF-IDF) is used to transfer a text paragraph into a dictionary that contains the frequency of different words appearing in the document. After eliminating the low-frequency words and frequent use of words in modern-day English, the words left were used to examine the reliability of the text. This is the list of models used in the research: logistic regression, decision tree, Naive Bayes classification, and support vector machine. In conclusion, the models can successfully distinguish the difference between true news and fake news, moreover, the logistic regression model had the best performance among all other models that were used in this research.

Introduction

In today's interconnected and information-driven world, the distinction between true and fake news has become increasingly vital. Technology has boosted information dissemination. It creates an efficient communication and diversity environment. However, the spread of fake information has also been boosted by technology. Like true information extends our vision, fake information not only narrows our vision but also causes serious consequences. This has happened many times in our history. One of them is the effect of “Yellow Journalism” on the

Spanish-American War. “Yellow journalism was a style of newspaper reporting that emphasized sensationalism over facts. During its heyday in the late 19th century, it was one of many factors that helped push the United States and Spain into war in Cuba and the Philippines, leading to the acquisition of overseas territory by the United States.” (*Milestones: 1866–1898 - Office of the Historian*). From history, we can understand how much damage fake news can bring to society. Moreover, in the current day, the amount of news is a huge number, therefore, we need to disguise the true information and fake information in an effective way.

Background

I. Text mining and text analytics

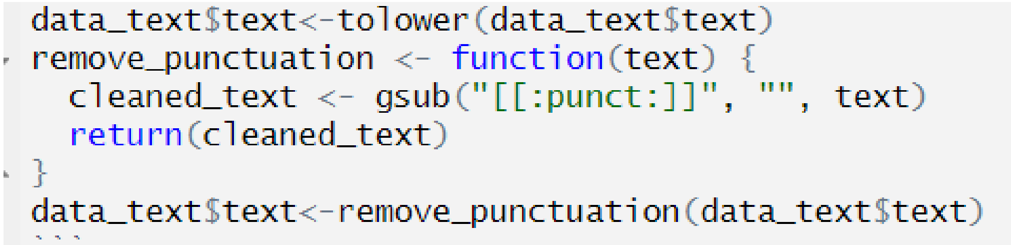
The text data is very different from the numeric and factor / categorical data that we usually study in data science. Even though the model is a classification model, we do not have a clear prediction variable as the ordinary regression and classification study. In data science there exists a field study focused on text data: text mining and text analytics. “Text mining and text analysis identifies textual patterns and trends within unstructured data …. By transforming the data into a more structured format through text mining and text analysis, more quantitative insights can be found through text analytics.” (*What Is Text Mining? | IBM*)) This provides a method for analyzing the text and transferring the text into another format that can suit the classical method in data science.

II. Data

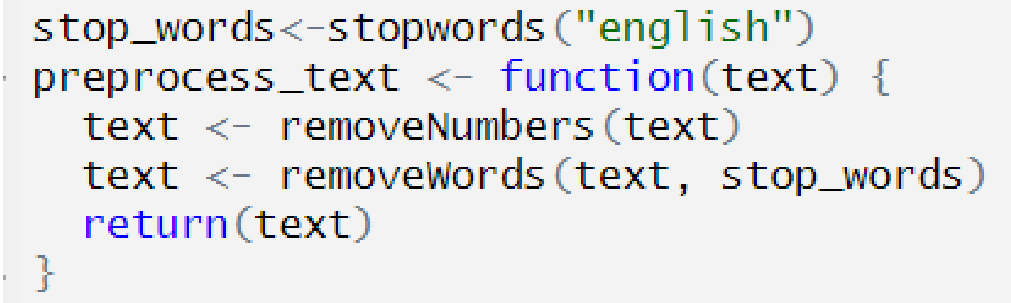
The Data is collected from **Kaggle.com** (AMMAR THABET,2023). The data set has two subsets: True News and Fake News. There are 23481 observations of fake news and 21417 observations of true news. Each data set contains 4 columns, “Title”, “Text”, “Subject”, and

"Date". "Title" columns record the news title. "Text" columns record the content / original news. "Subject" columns describe the field of news, such as political news, world news, and government news. "Date" columns record the post date of the news. The only variable needed is "text".

Before constructing the model, the text needs some preliminary work before analysis. First, the upper and lower case of letters can affect the result, so change all the upper case to lower case. Second, all the punctuation is unnecessary for the model, therefore, all the punctuation in the text should be removed. Then, remove all the numbers in the text. Last, there are many "stop words" that are unnecessary for analysis and need to be removed. (Stop words are commonly used in English) (Figure below shows the R code)



```
data_text$text<-tolower(data_text$text)
remove_punctuation <- function(text) {
  cleaned_text <- gsub("[[:punct:]]", "", text)
  return(cleaned_text)
}
data_text$text<-remove_punctuation(data_text$text)
```

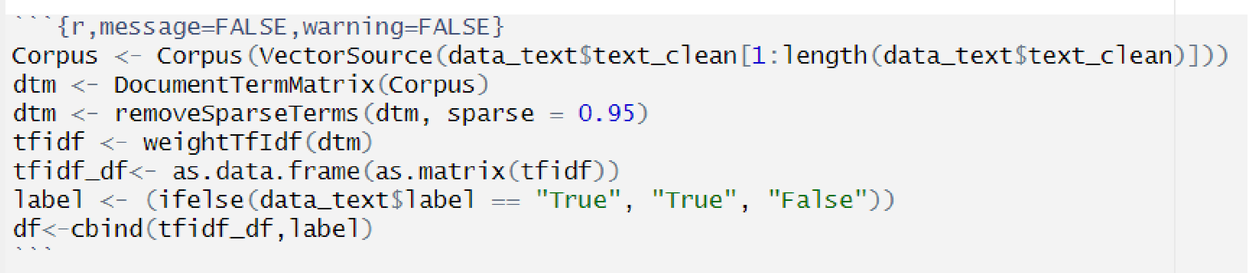


```
stop_words<-stopwords("english")
preprocess_text <- function(text) {
  text <- removeNumbers(text)
  text <- removeWords(text, stop_words)
  return(text)
}
```

Methodology

I. Term Frequency - Inverse Document Frequency (TF-IDF)

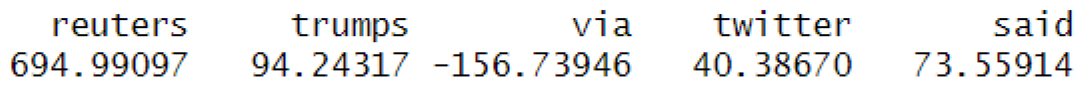
The text data is still not suitable for analysis, transferring an entire text into another format. First, create a dictionary that sorts the frequency of each word appearing in each individual text. Then, remove those words that do not appear in at least 5 percent of the texts in documents because they are insignificant, and removing them can reduce the time cost of the program. Next, use the formula of TF-IDF: (frequency of word x appears in text y) * log * (total number of text/number of text in a dataset containing word x) to create a matrix that shows the TF-IDF value of each word for each text. Converting the matrix into data frame format and combining the corresponding labels as the response variable. As a result, original text data is transferred into a data set that contains the TF-IDF value of 590 predictor variables (words appear in the data set) and 1 response variable (the news text is true or false).



~~~
```{r,message=FALSE,warning=FALSE}
Corpus <- Corpus(VectorSource(data_text$text_clean[1:length(data_text$text_clean)]))
dtm <- DocumentTermMatrix(Corpus)
dtm <- removeSparseTerms(dtm, sparse = 0.95)
tfidf <- weightTfIdf(dtm)
tfidf_df<- as.data.frame(as.matrix(tfidf))
label <- (ifelse(data_text$label == "True", "True", "False"))
df<-cbind(tfidf_df,label)
```
~~~

II. Logistic Regression

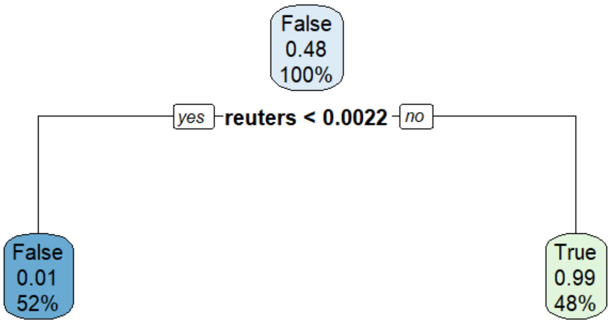
Logistic regression is a classification algorithm that uses regression analysis to classify the observation. It uses the predicted variables to generate a possibility of an event. The coefficient of the predicted variable tells us the increase/decrease probability when the value of that variable increases by 1. The figure below shows the top 5 lowest p-values predicted variables in the logistic model. The model suggests that when the TF-IDF value of the word "Reuters" increases by 1, it increases the probability that the text is true by the factor of exp(695). It suggests the word "Reuters" has a very strong connection with the true news in this data set. The variable "via" has a coefficient of -156.74, which suggests that frequently appearing "via" in the text will increase the chance of the news text being fake news in the data set. The accuracy rate of the model is 0.988.



```
   reuters      trumps          via     twitter        said
 694.99097    94.24317 -156.73946    40.38670    73.55914
```

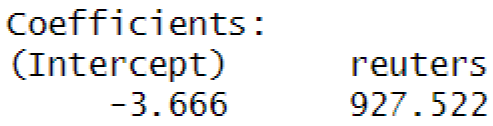
III. Decision Tree

A decision tree is a classification model that splits the dataset into subsets based on the most significant attribute. It contains the root node, tree node, and leaf. Where the root node is the input data, the tree node is the attribute criterion used to split the data set into the next step. The leaf is the final outcome of which class the data belongs. A decision tree is commonly used in classification analysis because its process is clear. The figure on the right is the decision tree plot for the research. The accuracy of the tree model in this research is 0.99. It shows that the only feature the model uses is the word "Reuters". If the TF-IDF value of "Reuters" in the news text is less than 0.0022, the news text will classify as fake news, otherwise, it will classify as true news. It matches some of the conclusions of the logistic model that "Reuters" is an important factor in identifying the news text in this data set. However, the tree model ignoring the importance of the rest variable is not convenient enough. Therefore deeper research is required.



IV. Revision logistic regression and decision tree

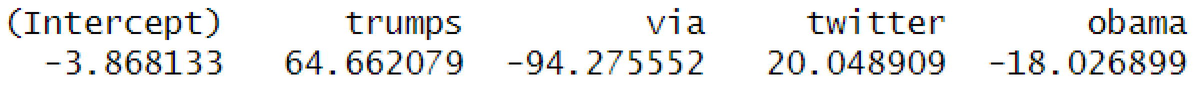
The decision tree above suggests the only important variable is "Reuters". In order to confirm this idea, some revision task was done for the research. First, the "Reuters" is the only important attribute in the decision tree model. Can it also work in the logistic regression model?



```
Coefficients:
(Intercept)      reuters
     -3.666      927.522
```

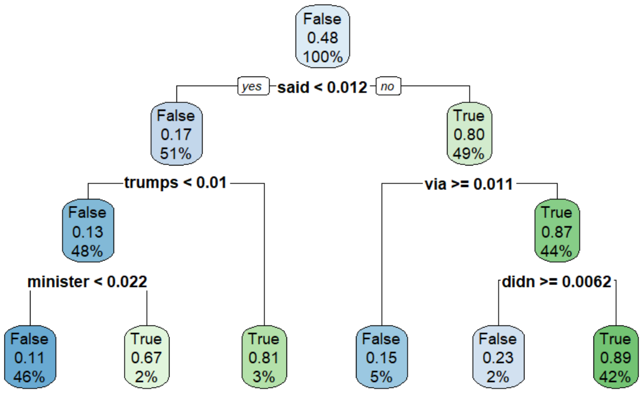
Thus, a new logistic regression model that only has “Reuters” as a predicted variable is used to test this. (Figure shown) The result has a similar coefficient as the first logistic regression, also the accuracy of the model is 0.985 the accuracy is not heavily decreased after dropping other variables. This supports the result of the decision tree model.

The “Reuters” clearly is an important factor for this research. Are other variables insignificant for this research? In order to answer this question, a new model without “Reuters” was needed in research. The figure below shows the top 5 least p-value variables of the new logistic model. Most of the words are also the words in the first logistic model. It demonstrates that text with a higher frequency of the word “Trumps” can highly increase the chance of news text being classified as true and a higher frequency of the word “via” will highly increase the chance of news text being classified as fake news. Moreover, the word “said” was replaced by “Obama” and "Intercept" replaced “Reuters”. The new model has an accuracy of 0.97, which is still accurate enough. This not only stays the result in the original logistic regression model but also shows that other variables are also important factors in determining whether the new text is true or fake in this data set.



```
(Intercept)      trumps         via     twitter       obama 
  -3.868133   64.662079  -94.275552   20.048909  -18.026899 
```

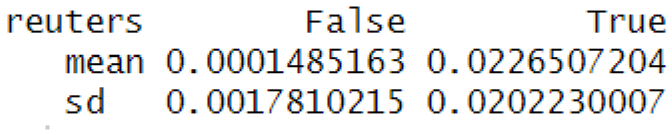
Will the decision tree function well after we drop the variable “Reuters”? The new decision tree model is shown at right. As the model suggests, there are multiple variables that can be used to predict the text. For example, if the TF-IDF value of “said” in the text is less than 0.012, the TF-IDF value



of “Trumps” is less than 0.01 and the TF-IDF of “minister” less than 0.022 will be classified as fake news. The model has an accuracy of 0.88. This supports the result that apart from “Reuters”, other words are also important for the classification model in the research.

V. Naive Bayes Classifier

Another popular text analysis model is the Naive Bayes Classifier. Naive Bayes Classifier based on the Bayes theorem, assumes all the features are conditionally independent given the class label. The classifier classifies the data by choosing the highest probability of the label of the text. Each variable contributes to the model by class-conditional probabilities: P(X1,X2,...Xn | Y) = P(X1|Y) P(X2|Y) …P(Xn|P). An example of the variable of the Naive Bayes Classifier in the research is shown in the figure. The figure shows the mean and standard deviation of the distribution of “Reuters” in both classes. The formula P(X = x | Y = y) = (sqrt * (2 * pi * (standard deviation) * y))^(-1) * exp(-((x-mean)^2) / (2*(standard deviation)*y)). However, the accuracy of the Naive Bayes Classifier is 0.878, which is relatively low compared to the logistic regression model.



```
reuters         False         True
   mean 0.0001485163 0.0226507204
   sd   0.0017810215 0.0202230007
```

VI. Support Vector Machine (SVM)

The last model that was used in this research is the support vector machine. The Support Vector Machine model has the basic idea of finding the best boundary that can separate the two classes. The SVM maps each observation as a vector in a hyper-dimensional space. Then use a hyperplane to separate these vectors. The hyperplane is located at the farthest distance for the closest class points for each class near it. The accuracy of the support vector machine model for the data set is 0.988. It suggests the support vector machine is an accurate model.

Conclusion

As a result, the logistic regression model is the best model in the research. It not only has high accuracy, but its model includes all the variables in the text and clearly explains how each variable will affect the outcome of the prediction. In the model, the TF-IDF value of “Reuters” in the text has a relatively high influence on classifying the news text as true news. The TF-IDF value of “via” in the text has a relatively high influence on classifying the news text as fake news. These conclusions are also supported by other models (decision tree & naive Bayes). This demonstrates that “Reuters” news is reliable.

The reasons that may lead to this conclusion are: Reuters firm may be a high firm and the data set is made by purpose. Reuters is a famous news agency that provides news and finance information. The text includes “Reuters” maybe: 1. The news text is written by Reuters or the text quotes the information from Reuters. If Reuters is a reliable news agency, it can explain the situation in the model. Another possibility is the data set is designed by Reuters since the data set is organized on a website, and it just includes some news in world history. Therefore, there is a chance that this data set is collected by Reuters to make the impression that Reuters is a reliable news agency.

There are some points that can improve the research. First, group the synonyms word into a variable or record the combination of phases in the text. This may help the research because they focus on a higher dimension of how humans understand the text. Humans understand the context by the sentence and phase rather than keywords. Second, the response variable can be replaced by the probability of true or fake news, since the true and fake news is too simple. Last, collecting more news from different news agencies, diverse and more data can always improve.

Citation:

1. *Milestones: 1866–1898 - Office of the Historian*. history.state.gov/milestones/ 1866-1898/ yellow-journalism.
2. *What Is Text Mining? | IBM*. www.ibm.com/topics/text-mining.
3. “Fake and True News.” *Kaggle*, 27 Feb. 2023, www.kaggle.com/datasets/ammarthabet/fake-and-true-news.