Machine Learning Coursework Report

Yogi Bhoot - 001189309

Abstract

Nowadays, there are many fake rumours posted on social platforms, newspapers, etc. This topic is the most important because there is an increasing number of misleading and misinformation among the people. People easily believe in that news which has lucrative and greedy types of information. Moreover, there are a number of crimes increasing in the modern world like money fraud, sharing private data with unknown people and many more. For these reasons, people and the government need to develop a fact checking model or software which is necessary to understand suspicious news to verify the fraud from genuine news. According to this category dataset (true and false), the logistic model is the best suitable for clarification.

1. Introduction

There are many types of fake news as a result of intentional attempts by user groups to spread rumours or misinformation, while propaganda is based on conspiracy theories. To evaluate the data that allows us to extract the fraud and original information. we have two datasets True news and the fake news dataset. Both datasets contain a title, subject, text and date. there is no number and no available value in the dataset. After that, we observed the relationship between the data. Then, analyse data and recognise and detect the false and true news and construct the machine learning model to validate the dataset.

2. Observation of Dataset

Data is defined as the core element expressed as facts, statistics in the form of numerical, or collected as primary data, stored as secondary data for observation (Carpineto and Romano,2004). There are many different types of data forms like scientific data, geographical data, financial data, statistical data, operations data etc. For the research work, the field data is the raw data that is transformed with the use of statistical tools help to reveal the relation of variables and their meaning (Hair, 2009). Here, There are two types of data: one is a true news dataset and the other one is a fake news dataset. Both have the same collection of parameters like subject, title, date and text.

2.1. Data collection method:

Any investigation ought to collect information in any shape in arrange to demonstrate the investigative points and targets. Data is central to proving research goals, and you also need to choose the right method to prove the relationship between variables. Data collection can be done primarily from respondents or secondarily from printed or digitally published documents.

2.2. Fake news:

The definition of fake news defines it as a deliberate attempt to spread misinformation over the original news (Rubin et al. 2015). Fake news reduces the value of original news. It has no base and clarification to prove context and purpose. It is just biased information that has been made to fool people. According to Wikipedia, These are seven types of fake news

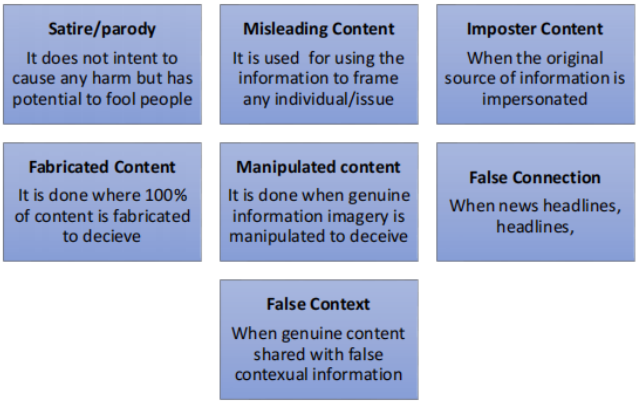

Fig 1: Type of fake news

2.3. True news:

The definition of true news is whole content and title that has proven and identical proof evidence. Real news provides genuine text and not any misleading information. It 100% represents original issues and problems with the right heading with relative description.

3. Data analysis

We have a data set containing 44,919 fake and real news date tags related to political news, world news, US news, government news, and Middle East news. We have 23,502 fraudulent (fake) news and 21,417 real news. We are assigning class 0 stands for fake messages and class 1 stands for true news. All of these analysis methods perform after merging true and fake dataset.

3.1. True and Fake news count

Based on our analysis of the data set, we have classified news on various topics. Here 0 represents fake data, and 1 represents real data. We clearly see some political news, world news, government news US news, etc. In the following graph, shows us the number of fake and true information.

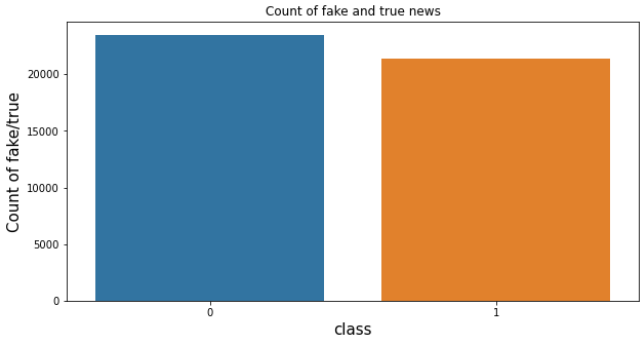

Fig 2: Type of fake news

3.2. Different news according to class

Here, As we see in the over chart, it gets to be apparent that the genuine news is for the most part clustered around the legislative issues news and world news where as fake news is conveyed over the distinctive subjects.

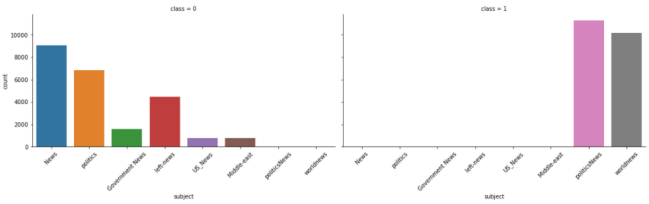

Fig 3: Type of fake news

3.3. Type of news

The below picture tells us that among all the news most of the news is from political news, at that point world newsthan news than politics than left news than government news than middle east and US news. Political news are more than 10000 and news and worldnews are nearly same around 10000 and 9000

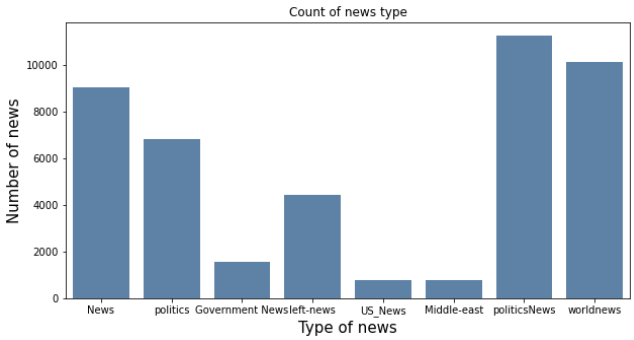

Fig 4: Type of fake news

3.4. Fake: Common name observation

Here, In this method identify the most common name used in fake datasets. This method uses a stopword, wordcloud as well as background color parameter .

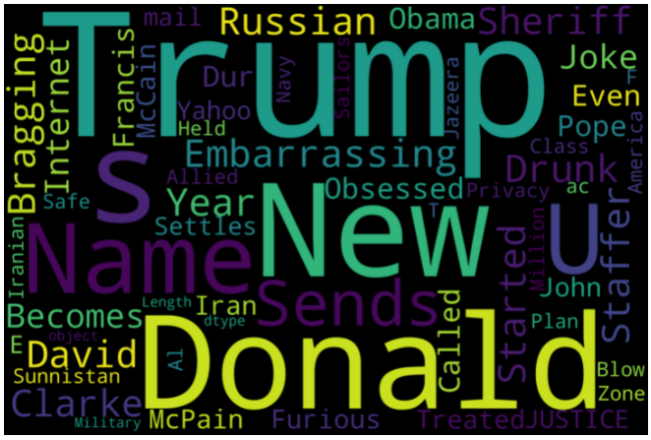

Fig 5: Common name in fake

3.5. True: Common name observation

Here, we perform the same operation for the real common name in the true dataset as we perform above in the common name in the fake dataset.

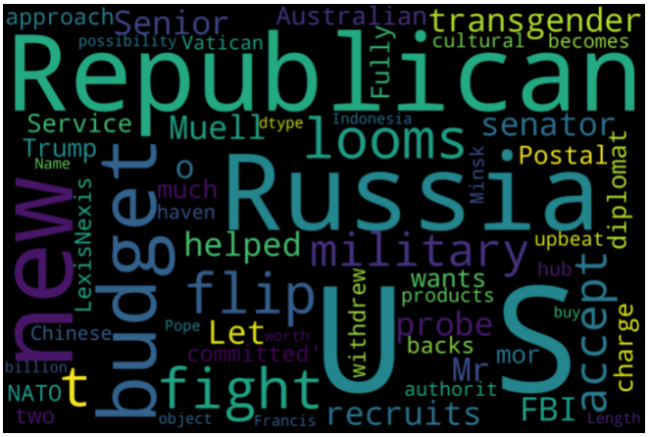

Fig 6: Common name in true

3.6. Total common name

Here, we perform to find the common name in the merged dataset. The highest common name in both datasets is president donald trump

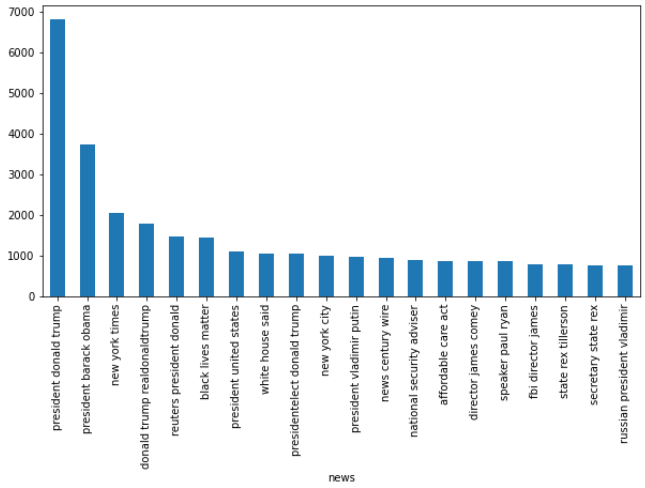

Fig 7: Common name in true

4. Mode Fitting Section

We have used two classification methods for each case. i.e. we have to use

- a) Logistic regression,
- b) KNN.

4.1. Logistic Regression

Logistic regression is a traditional method that divides observations into two or more categories. In this particular problem, our target variable is a binary variable (because it only needs the values ”true” and ”false”). This is a special case of logistic regression, we usually call it binary logistic regression, because the dependent variable can only take binary values. Logistic regression has been observed to produce stable and accurate results in many classification problems. The goal is to create the best model as always to predict the probability of the dependent variable. In order to obtain the predicted probability, the binary logic classifier uses the sigmoid function. Logistic regression methods use logistic functions and log odds to perform binary classification tasks. In fact, in our research, we will also notice that in most cases, logistic regression uses The whole model is stable.

4.2. KNN

KNN stands for K-Nearest Neighborhood. It is also a supervised machine learning approach that builds a model based primarily on the ”k” closest matches in the training dataset. The principle of this classification method is that it looks for similarities between the new case and the available datasets and, accordingly, places the new data in the category that is most similar to the available categories. This algorithm uses a distance measure to determine similarity/proximity. Most often we use the Euclidean distance measure. When using KNN to classify text, the distance between feature vectors of unlabeled text data and vectors extracted from the dataset will be calculated. The KNN classification method is also sometimes referred to as the lazy learner algorithm because, instead of immediately learning, it first saves a dataset and then performs an action on the data during classification. This method is also a nonparametric method. Like the two above, KNN is also one of the popular classification algorithms methods. In our study, since we want to split the text into two cases, we prefer to use KNN along with other algorithms and compare the result.

We divide the data into training and testing batches, where 70% of the data is used to build the model, and the remaining 30% is used for the testing part. The following is a confusion matrix for each of the 16 models.

5. Accuracy Score

In this report, we use two models logistics and KNN model. Both models fit with the dataset which model is the best for the dataset is defined by the highest accuracy score.

5.1. Logistic Accuracy score

The logistic accuracy score is 0.9359174277864409

5.2. KNN Accuracy score

KNN accuracy score is Logistic accuracy score is 0.726665181554912

After finding the accuracy score. We say that the logistic model is the best suitable model for fake and true datasets.

5.3. Confusion Matrix

5.3.1. Logistic confusion perform

For logistic regression N = 3, in this one confusion matrix, 6745 predicted false labels are real false labels, 288 predicted true labels are real false labels, and 575 predicted false labels. You can see that the fake label is the real true label and the predicted true label of 5859 is really true. Label. i e Logistic Regression Algorithm: [[6745 288] [5755859]].

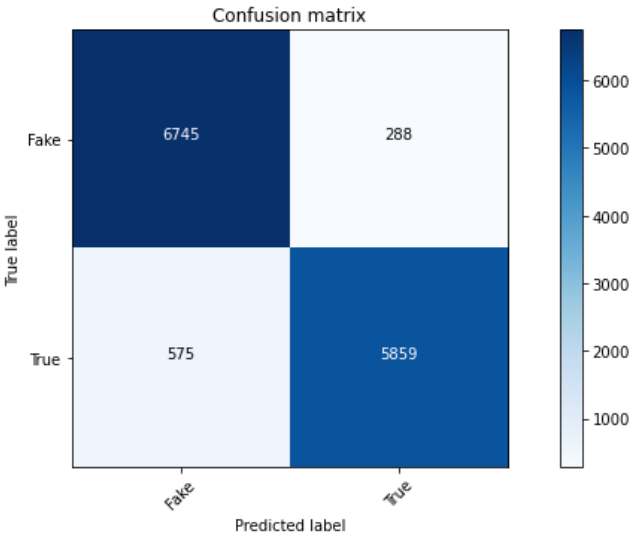

Fig 7: Logistic confusion matrix

5.3.2. KNN CONFUSION PERFORM

In this one confusion matrix, the 6775 predicted false label is a real false label, the 258 predicted true label is a real false label, the 3423 predicted false label is the real true label, and the 3011 predicted true label is the real label. I understand this. That is, the KNN algorithm: [[6775 258] [3423 3011]]

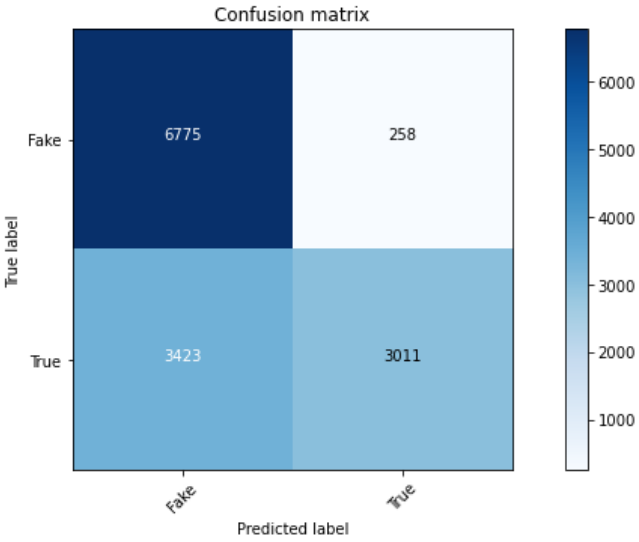

Fig 7: KNN confusion matrix

6. Conclusion

The actual significance of the analysis and data of this research means that false data is widespread in any form, so the database needs to be determined. False data detection needs to detect that the first individual can perceive the data, and corrective measures need to be taken. As organizations use data in their daily work processes, the question of how and when to capture fake data is urgently needed. The ubiquitous negative trend of data dilution affects macro and micro decision-making, leading to erroneous decision-making processes. Generally speaking, it will affect the perception of the majority of the people, which witnessed the changes in the government composition of Russian hackers in the 2016 U.S. President Trump election.