CISC839 Project-5: Fake News Analysis

Aliaa Faisal kashwa[1], Ragia Hisham Aboutaleb[2], and Ahmed Ibrahim Salem [3]

[1]**21afae@queensu.ca**
[2]**21rhma@queensu.ca**
[3]**21aisa@queensu.ca**

1 BACKGROUND AND OBJECTIVE

False information on the Internet has caused many social problems due to the raise of social network and its role in different domains such as politics. In this Project, we are going to predict if a specific reddit post is fake news or not, by looking at its title. Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers. Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This project analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites and links containing false and misleading information. We use simple and carefully selected features such as the title and post to accurately identify fake posts.

*** The hypothesis question we provide is:**
- Is the ratio of fake news in Emergent fact-checking site significantly higher than the ratio of fake news by all other fact-checking sites?
- Observation: From calculating the ratio for the 3 fact-checking sites, we found that the ratio for snopes is 0.25 , the ratio for politifact is 0.166 and the ratio for emergent is 0.441, so according to this observation, emergent site has provided the highest ratio of fake news. So Emergent is pessimistic (comparing to the other two sites) when evaluating each news, because we expect to see almost similar ratio of fake news when using any of these sites.
-The benefit of answering this question is: Helping the public using social media to determine which sites that are used to detect false news are more credible than others.

*** The regression question is:**
- Predicting the number of user comments for each news.
-The benefit of answering this question is: Helping knowing that which type of news users can interact with it and may be indicative of its importance or impact.

* **The question that can be answered via predictive analysis is:**
- Predicting fake news with/without urls.
-The benefit of the this question that, knowing the type of url whether it contain fake news or not. So the user can distinguish between them. And that if the user know that the url was fake for one news we may predict that the other news that has the same link is also fake.

2 DATASET

In this project, A large dataset for fake news detection using social media news and its related comments from Reddit. This data is provided from kaggle website which is available on (https://www.kaggle.com/datasets/deepnews/fakenews-reddit-comments).
Data is collected from Reddit which is a social news website.

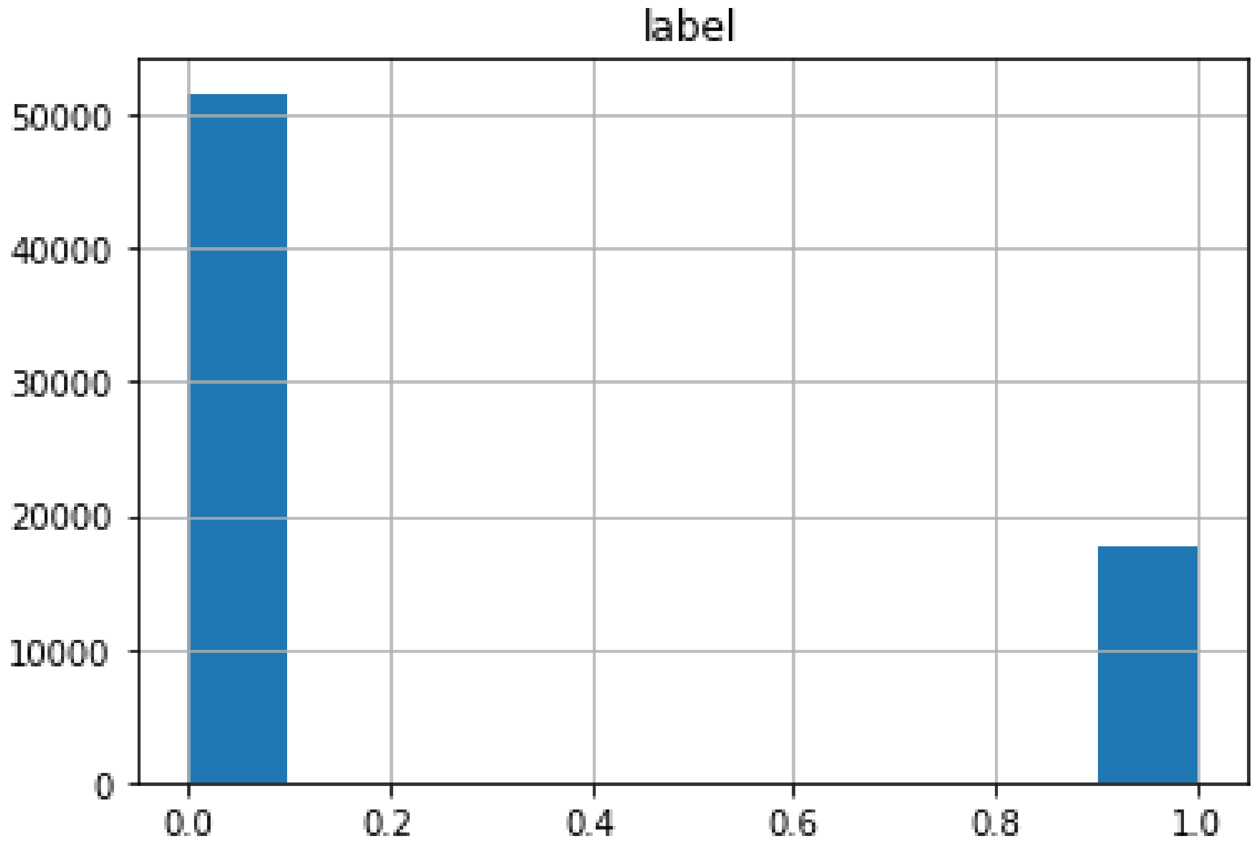

Figure 1. label histogram

The dataset consists of 69396 records (of which 74 percentage is labelled as true news, the rest is labelled as fake ones) from three different sources (Snopes, Poltifact and Emergent) which they are fact checked news.
Given a record in textual format, our goal is to automatically detect whether it is fake or not.
The dataset is in json format.

3 BASIC DATA EXPLORATION

Number of data rows= 69396
Number of data columns= 6
Names of columns: ['label', 'reddit comments', 'researched by', 'text', 'title', 'url']
'label' column has int datatype and the other columns have object datatype.
The only column that has missing values is: reddit comments that equal to 64161 the other column have 0 missing values. and when we check duplicate value were equal zero.

We have 5 categorical columns out of 6 columns that may be converted their values to dummy values.

4 DATA ANALYSIS PIPELINE

-According our dataset having enough information to address our main goal in this project, we saw thet after an initial look at the data, we concluded that there is an important column that contains most of the missing data and is likely to be deleted because it contains 92 percentage of the missing data, which reduces the chance of better prediction than whether the news is fake or not. But there is a column that contains the same news, and there is also a column that contains the title of the news, so it is possible through them to predict what is required. - We judge data quality based on checking missing values, duplicated data and outliers. At first, handling the missing values will be by seeing the percentage of missing values in each column that can contain missing values. Accordingly, it will be decided whether to delete that column or deal with it in an appropriate way to fill it. This is important because entries with missing values will lead models to misunderstand features. At Second, we deal with duplicated values by dropping them because the data sets that contain duplicates that may contaminate training data with the test data or vice versa. At third, we deal with duplicated values by dropping them because the outliers will general the training process – leading our model to "learn" patterns that do not exist in reality.

-Feature Engineering: We can extract new column from reddit comments column by taking the most important part from it that can help us more in process of detecting whether the text is fake or not. This part is the bodies which indicate the related comments to our text we want to check. The text that we want to check its reliability and its title attributes may be the most suitable features to accurately detect whether or not the text is fake news.

- The techniques that are used for text classification in machine learning include:

- **Support Vector Machines**
- **Naïve Bayes**
- **K-nearest neighbors**
- **decision tree**
- **Random forest (1)**

-The techniques that are used for text classification in deep learning : Deep learning techniques have great prospect in fake news detection task. There are very few studies suggest the importance of neural networks in this area:

- **hybrid neural network model which is a combination of convolutional neural networks and recurrent neural networks**
- **ANN or RNN or its variations with this problem.(2)**

- we will use in our project some important libraries such as
scikit-learn or sklearn (is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling),
nltk(Natural Language Toolkit),
Pandas (is a Python library for data analysis).
Matplotlib (is a comprehensive library for creating static, animated, and interactive visualizations in Python.),
and we will use Flask App to convert our project to webpage
in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

- our project model is classification model so we can use one of the following to evaluate the performance:
Accuracy(The overall accuracy of a model is simply the number of correct predictions divided by the total number of predictions)
Confusion Matrix(is an extremely useful tool to observe in which way the model is wrong (or right!). It is a matrix that compares the number of predictions for each class that are correct and those that are incorrect)

1.1) AUC/ROC(A classifier such as logistic regression will return the probability of an observation belonging to a particular class as the prediction output. For the model to be useful this is usually converted to a binary value e.g. either the sample belongs to the class or it doesn't.)

2.1) Precision(Precision measures how good the model is at correctly identifying the positive class.)

3.1)Recall(Recall tell us how good the model is at correctly predicting all the positive observations in the dataset.)
4.1) F1 score(The F1 score is the harmonic mean of precision and recall. The F1 score will give a number between 0 and 1. If the F1 score is 1.0 this indicates perfect precision and recall. If the F1 score is 0 this means that either the precision or the recall is 0.)

5.1) Kappa(The kappa statistic compares the observed accuracy to an expected accuracy or the accuracy expected from random chance.)

- For instance, the dataset used to train the model contains news articles from specific period to another specific period that newer.

. So, as an example, if the news about COVID-19 was to be presented to the model for inference, what would its prediction be? News, by its nature, is ever-changing. It is straightforward to assume that model staleness is an important consideration here.

Another issue is that the method by which the text is extracted in the original dataset is not shown. It is, most probably, not the same method I used to extract content in production. So, there is some training/serving skew here that cannot be avoided.

But we must have ways to assess it and to mitigate it in the long run, when retraining the model with data gathered in production. Finally, we should also consider the differences between text preprocessing during training and serving. If I remove stopwords during training, I have to make sure I'm also removing them in production. If additional preprocessing steps are introduced in production, then I should retrain the model. Another example of training/serving skew.(4)

5 DISTRIBUTION OF WORKLOAD

person number 1- Make Visualization and data exploration
person number 2- Do data prepossessing
person number 3-Build the Model
Note: we will do the tasks together

REFERENCES

1-journal=1-https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf(1)
2-https://www.sciencedirect.com/science/article/pii/S266682702100013X(2)