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 text. 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[1].

*** 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 is used from Reddit.
This data is provided from kaggle website which is available on (https://www.kaggle.com/datasets/deepnews/fakenews-reddit-comments)and is collected from Reddit which is a social news website.
The dataset consists of 69396 records (of which 74 percentage is labelled as real 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.

2.1 Data Preprocessing

For the text column we will make:

- **Expand Contractions**
- **Lower Case**
- **Remove Punctuations Remove words and digits containing digits**
- **Remove Stopwords**
- **Stemming**
- **Remove White spaces Remove tags.**

Regarding data prepossessing we will focus on the text column on this data which contains the news part so so we will modify this text column to extract more information to make the model more predictable.

2.2 Basic Statistics of the Dataset

After displaying data info, we observed that:
• 'label' column has int datatype
• and the other columns have object datatype.

After checking missing values, we observed that:
• reddit comments that equal to 64161
• the other column have 0 missing values.

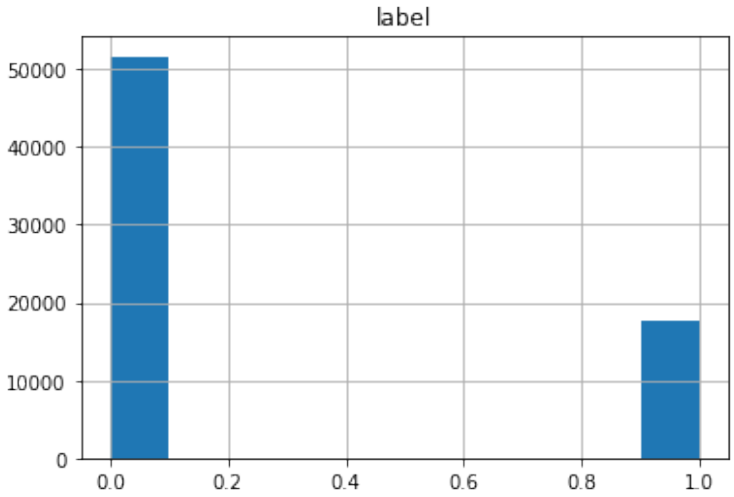

Figure 1. Description of events dataset

3 ANSWERS TO THE RESEARCH QUESTIONS

1.First Question:

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?
* *Motivation* :The benefit of answering this question is helping the public using social media to determine which sites that are used to detect fake news are more credible than others.
* *Approach*: Calculating ratio of news that checked by each fact-checking site are fake.
* *Findings*:NFrom 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

2. Second Question:
• Predicting the number of user comments for each news.
* *Motivation*: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.
* *Approach:* We extracted the length of comments of each news as a new feature, then used it as label to predict later the expected number of comments on each news and applied this through 2 scenarios:

1- Using the whole data (69396 rows) that has missing values in reddit comments column assuming that the remained news have no comments (0)

* Findings: We have tested these different models in the 2 scenarios. according to our obtained results that they were unsatisfactory, we conclude that it is impossible to predict the number of comments according to the news content.

Before converting the predicted values that less than zero to zero		
model name	**MSE**	**MAE**
Baseline model	14107.77857	16.244761
Linear Regression	24121.66292	54.1132
K-Nearest Neighbor Regression	15081.60	10.16561
Decision Tree Regression	14835.7741526	15.7921898

After converting the predicted values that less than zero to zero		
model name	**MSE**	**MAE**
Baseline model	14107.77857	16.24476
Linear Regression	22041.08544	34.524439
K-Nearest Neighbor Regression	15081.6095792	10.165619
Decision Tree Regression	14835.77415	15.792189

2- Using part of the data that has comments on each news without the ones that has missing values assuming that their comments are missed, so delete them, and use that 5235 rows to train/test our model.

Before converting the predicted values that less than zero to zero		
model name	**MSE**	**MAE**
Baseline model	247282.69434	163.42013
Linear Regression	438954.62748	327.92770
Lasso Regression	247282.694340	163.42013
Ridge Regression	248400.2516	165.4099
Elastic Net Regression	245078.3679	162.03333
K-Nearest Neighbor Regression	268638.8339	152.24450
Decision Tree Regression	251767.5684	162.65568
Random Forest Regression	252835.10101	165.77209
Gradient Boosting Regression	293442.002988	162.91259

After converting the predicted values that less than zero to zero		
model name	**MSE**	**MAE**
Baseline model	247282.69434	163.420133
Linear Regression	438954.62748	327.92770
Lasso Regression	247282.694340	163.42013
Ridge Regression	248370.067723	164.998652
Elastic Net Regression	245078.3679	162.03333
K-Nearest Neighbor Regression	268638.8339	152.24450
Decision Tree Regression	251767.5684	162.65568
Random Forest Regression	252835.10101	165.77209
Gradient Boosting Regression	293442.002988	162.91259

3.Last Question:

• Predicting fake news with/without URLs.

* *Motivation:* The benefit of 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.

* *Approach:* We tried to predict fake news in two different scenarios:

1: predicting it by some features but url.

* *Findings:* Without using url feature , we obtained better performance for predicting the fake news than using it. And we got the best performance from logistic regression model with test accuracy 78 percentage.

the results of our models			
model name	**Train Score is**	**Test Score is**	**F1 Score is**
LogisticRegression	0.8691906	0.7807486	0.2506527
MultinomialNBModel	0.9294944	0.7669977	0.3813387
PassiveAggressiveClassifier	0.997959	0.770817	0.4140625
Random Forest	0.9984130	0.77005347	0.06230529

2: predicting it by the same some features including the url.

* *Findings:* From comparing the models performance, we got the best performance from logistic regression model with test accuracy 76 percentage.

the results of our models			
model name	**Train Score is**	**Test Score is**	**F1 Score is**
LogisticRegression	0.83359782	0.7677616	0.3090909
MultinomialNBModel	0.791883	0.734148	0.330769
PassiveAggressiveClassifier	0.99773	0.721161	0.388609
Random Forest	0.748583	0.7669977	0.012944

We used hold-out Method to split our data (training set= 80 percentage, testing set= 20 percentage).

We encoded our input(text data) with TF-IDF Vectorizer that transforms the text into a usable vector and label encoder technique for categorical columns.
we use default hyper-parameter in all models but some models as ”Ridge Regression we used cv=3” , ”Decision Tree Regression we used min-samples-split=45, min-samples-leaf=45, random-state = 10 ” , ”Random Forest Regression we used n-jobs=-1, n-estimators=1000, min-samples-leaf=10, random-state = 10”

4 LIMITATIONS

In our answer to the second question ’Predicting the number of user comments for each news’, we met difficulty in extracting new feature from ’reddit comments’ column because each record in it is represented as a list, each list

contains number of dictionaries, each dictionary represents a user comment information that contain the comment body that we interest to extract to calculate the length of comments on each news to use it as a label to answer the question. Another obstacle in this question answer is that, there were comments bodies that has empty strings or 'deleted' word that indicate there is no comment or it is deleted so we cleaned that and recalculate the number of comments on each news again, also the results on this question were not satisfactory.

5 TAKE-AWAY MESSAGES

5.1 CONCLUSION :

•We imported our data from kaggle to our colab notebook then we made different preprocessing for the used features for each different scenario. After that we provided our different models for each scenario of the two questions followed by the results for each of them to compare their performance.
As a summary to our questions:
In the first question we have tested different models in 2 scenarios using some features and the results were not satisfactory, so we colclude that it is impossible to predict the number of comments according to the news content. In the second question, we could classify the news whether they are fake or not throught 2 scenarios (with/without urls) and colclude that the accuracies without using the url feature are higher than using it.

5.2 OUR IMPACT IN THE FUTURE :

Our work came up with a solution that can be utilized by users to detect and filter out sites and links containing false and misleading information. This can reduce the tendency of rumors that cause many unsatisfactory results for their owners or for topics related to them. Including the protection of privacy by preventing the circulation of such news on social media.

6 REPLICATION PACKAGE

colab notebook here

You should get your API key (kaggle.json) from kaggel account throw going to your 'Account' and create New API Token then you can run easily our attached co-lab notebook.

7 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 did the tasks together

REFERENCES

[1]• URL: %22https:1-journal=1-https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf(1).