Fake news project

Submitted by:

Pallavi Bajpai

ACKNOWLEDGMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and references. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

INTRODUCTION

• Business Problem Framing

Fake news and hoaxes have always existed, even before the Internet. Fake news on the internet is defined as “fictional content created with the intent of deceiving readers.” Fake news is published on social media and news sites to promote readership or as psychological warfare. In general, the objective is to benefit on clickbait's. Clickbait's utilise flashy headlines or graphics to encourage visitors to click links to increase ad income. This study shows that the prevalence of fake news stories considering the advancements in communication made possible by the rise of social networking sites. The aim of the project is to provide a system that consumers may use to detect and filter out sites that contain incorrect and misleading information. To accurately identify fraudulent postings, we employ basic and carefully chosen elements of the title and content.

• Conceptual Background of the Domain Problem

A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

• Review of Literature

Many automatic detection approaches for fake news and deception posts have been reported in the literature. Because there are many different components to fake news detection, from employing chatbots to disseminate disinformation to utilising clickbait's to propagate rumours. There are several clickbait's accessible on social media networks such as Facebook, which increase the sharing and like of content, therefore spreading false information. A great deal of effort has gone into detecting false information.

The Author [1] discussed Linguistic Cue Methods with Machine Learning, Bag of Words Approach, Rhetorical Structure and Discourse Analysis, Network Analysis Approaches, and SVM Classifiers. These algorithms are only text-based and offer little or no improvement over previous approaches.

The Author [2] examined the concepts, methodologies, and algorithms used for classifying false and manufactured news articles, authors, and subjects from virtual communities, as well as assessing their scope and efficiency. The report also proposed research challenges based on previously unknown aspects of false news and varied relationships between news pieces, writers, and subjects. The authors of the study discuss Fake Detector, an automated fake news inference algorithm. It is based on textual classification and employs a deep diffusive network model to concurrently learn these same portrayals of news items, authors, and subjects. Fake Detector focuses on two key elements: The wide diffusional proposed approach will be composed of representations features extraction and reliability labelling inference. Fake Detector.

- ## Motivation for the Problem Undertaken

The project's goal is to find a solution that can be used to recognise and filter out sites that contain whether it is real news or fake news, therefore assisting consumers in avoiding being enticed by clickbait. It is critical that such solutions be discovered since they will be beneficial to both readers and the IT firms engaged in the issue.

In this project, a model based on the count vectorizer or a tfidf matrix (i.e., word tallies related to how frequently they are used in other articles in our dataset) might be effective. Because this is a text classification problem, it is recommended to use a Naive Bayes classifier, which is typical for text-based processing. The primary plan is to create a model for text transformation (count vectorizer versus tfidf vectorizer) and decide which type of text to utilise (headlines vs full text). The next step is to extract the most optimal features for count vectorizer or tfidf-vectorizer, which is done by using an n-number of the most used words and/or phrases, lower casing or not, removing stop words such as “the,” “when,” and “there,” and only using words that appear at least a given number of times in each text dataset.

- ## Data Sources and their formats

The dataset consists of both numerical and categorical variables. There is a total of 4 explanatory variables with 6334 instances describing every aspect of the fake news detection. Overview of the dataset:

```
#printing the first several rows of news data

news_dataset.head()
```

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

```
#printing the last several rows of news dataset

news_dataset.tail()
```

	Unnamed: 0	title	text	label
6330	4490	State Department says it can't find emails fro...	The State Department told the Republican Natio...	REAL
6331	8062	The 'P' in PBS Should Stand for 'Plutocratic' ...	The 'P' in PBS Should Stand for 'Plutocratic' ...	FAKE
6332	8622	Anti-Trump Protesters Are Tools of the Oligarc...	Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	4021	In Ethiopia, Obama seeks progress on peace, se...	ADDIS ABABA, Ethiopia —President Obama convene...	REAL
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's W...	Jeb Bush Is Suddenly Attacking Trump. Here's W...	REAL

- Data Pre-processing Done

Inspecting the vectors

```
count_df =  pd.DataFrame(count_train.A, columns=count_vectorizer.get_feature_names())
```

```
tfidf_df = pd.DataFrame(tfidf_train.A, columns=tfidf_vectorizer.get_feature_names())
```

```
print(count_df.head())
```

```
   00  000  0000  00000031  000035  00006  0001  0001pt  000billion  000ft  \
0   0    0     0         0       0      0     0       0           0      0
1   0    3     0         0       0      0     0       0           0      0
2   0    1     0         0       0      0     0       0           0      0
3   0    1     0         0       0      0     0       0           0      0
4   0    0     0         0       0      0     0       0           0      0

   ...  والمرضى  هذا  من  محاولات  ما  لم  عن  عربي  حلب  ŭade
0  ...        0    0   0        0   0   0   0     0    0     0
1  ...        0    0   0        0   0   0   0     0    0     0
2  ...        0    0   0        0   0   0   0     0    0     0
3  ...        0    0   0        0   0   0   0     0    0     0
4  ...        0    0   0        0   0   0   0     0    0     0

[5 rows x 57870 columns]
```

```
print(tfidf_df.head())
```

```
    00       000  0000  00000031  000035  00006  0001  0001pt  000billion  \
0  0.0  0.000000   0.0       0.0     0.0    0.0   0.0     0.0         0.0
1  0.0  0.041696   0.0       0.0     0.0    0.0   0.0     0.0         0.0
2  0.0  0.031448   0.0       0.0     0.0    0.0   0.0     0.0         0.0
3  0.0  0.014377   0.0       0.0     0.0    0.0   0.0     0.0         0.0
4  0.0  0.000000   0.0       0.0     0.0    0.0   0.0     0.0         0.0

   000ft  ...  والمرضى  هذا  من  محاولات  ما  لم  عن  عربي  حلب  ﻻade
0    0.0  ...  0.0  0.0  0.0  0.0  0.0    0.0  0.0  0.0    0.0    0.0
1    0.0  ...  0.0  0.0  0.0  0.0  0.0    0.0  0.0  0.0    0.0    0.0
2    0.0  ...  0.0  0.0  0.0  0.0  0.0    0.0  0.0  0.0    0.0    0.0
3    0.0  ...  0.0  0.0  0.0  0.0  0.0    0.0  0.0  0.0    0.0    0.0
4    0.0  ...  0.0  0.0  0.0  0.0  0.0    0.0  0.0  0.0    0.0    0.0

[5 rows x 57870 columns]
```

```
difference = set(count_df.columns) - set(tfidf_df.columns)
```

```
print(difference)
```

```
set()
```

```
print(count_df.equals(tfidf_df))
```

```
False
```

- ## Hardware and Software Requirements and Tools Used

- Python code was written in Jupiter notebook. Below are the libraries needed in the process.
- Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model_selection.cross_val_score, sklearn.linear_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion_matrix,accuracy_score, sklearn.metrics.f1_score, scipy.stats, seaborn.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

- Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
- Outliers were detected with the help of scatterplot and boxplot and were removed.
- Skewness was checked using histogram.

- Testing of Identified Approaches (Algorithms)

MultinomialNB algorithms were used for training and testing of data

- Run and Evaluate selected models

simple NLP,complex problems

```
alphas = np.arange(0,1,0.1)
```

```
def train_and_predict(alpha):
    nb_classifier = MultinomialNB(alpha=alpha)
    nb_classifier.fit(tfidf_train,y_train)
    pred  = nb_classifier.predict(tfidf_test)
    score = accuracy_score(y_test,pred)
    return score
```

```
for alpha in alphas:
    print('Alpha: ',alpha)
    print('Score: ',train_and_predict(alpha))
    print()
```

```
Alpha:  0.0
Score:  0.8858495528669121

Alpha:  0.1
Score:  0.9042609153077328

Alpha:  0.2
Score:  0.9011046817464492

Alpha:  0.30000000000000004
Score:  0.8953182535507628

Alpha:  0.4
```

```
Alpha:  0.4
Score:  0.8921620199894792

Alpha:  0.5
Score:  0.8884797475013151

Alpha:  0.6000000000000001
Score:  0.8826933193056287

Alpha:  0.7000000000000001
C:\Users\PRANAY\AppData\Roaming\Python\Python38\site-packages\sklearn\naive_bayes.py:508: Use
ric errors, setting alpha = 1.0e-10
  warnings.warn('alpha too small will result in numeric errors, '
Score:  0.875854813256181

Alpha:  0.8
Score:  0.8695423461336139

Alpha:  0.9
Score:  0.8679642293529721
```

- Key Metrics for success in solving problem under consideration

Accuracy score was the key metric used to finalize the model.

- Interpretation of the Results

Saving The Model

```
import pickle
```

```
with open('model.pkl', 'wb') as handle:
    pickle.dump(pipeline, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

```
```

CONCLUSION

- Key Findings and Conclusions of the Study

Accuracy score was highest in Gradient Boosting Classifier hence we trained the model with the same.

- ## Learning Outcomes of the Study in respect of Data Science

  - Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
  - Outliers were detected with the help of scatterplot and boxplot and were removed.
  - Skewness was checked using histogram.

- ## Limitations of this work and Scope for Future Work

From this project we can also extend this project in future by employment of a technology that can identify and eliminate fraudulent sites from the results presented to a user by a search engine or a social media news feed is advocated as a solution to the issue of false news. The user can download the tool and then add it to the browser or programme they use to get news feeds. Once operational, the tool will use various techniques including those related to the syntactic features of a link to determine whether the same should be included as part of the search results.