MACHINE LEARNING APPLIED: FAKE NEWS CLASSIFICATION

Submitted by:

Vijaysingh Arjunsingh Pardeshi

ACKNOWLEDGMENT

Foremost, I would like to express my sincere gratitude to Data Trained team for the continuous support of my Data Science study and research, for the patience, motivation, enthusiasm, and immense knowledge. The guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Data science study.

Besides Data Trained, I would like to thank Flip Robo Team, for their encouragement, insightful internship, and help to understand the study.

My sincere thanks also go to **Vaishali Singh and Khushboo Garg** for offering me the internship opportunities in their ally and leading me working on diverse exciting projects. I am over helmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

Last but not the least, I would like to thank my family, my parents, for being a morale support to me at the first place and backing me up in any down time, for being there in all hard situations making me to fight against any difficulty I face.

INTRODUCTION

Business Problem Framing

Fake news is false or misleading information presented as news. It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. However, the term does not have a fixed definition, and has been applied more broadly to include any type of false information, including unintentional and unconscious mechanisms, and also by high-profile individuals to apply to any news unfavourable to his/her personal perspectives.

Once common in print, the prevalence of fake news has increased with the rise of social media, especially the Facebook News Feed. Political polarization, post-truth politics, confirmation bias, and social media algorithms have been implicated in the spread of fake news. It is sometimes generated and propagated by hostile foreign actors, particularly during elections. The use of anonymously-hosted fake news websites has made it difficult to prosecute sources of fake news for libel. In some definitions, fake news includes satirical articles misinterpreted as genuine, and articles that employ sensationalist or clickbait headlines that are not supported in the text.

Fake news can reduce the impact of real news by competing with it; a Buzzfeed analysis found that the top fake news stories about the 2016 U.S. presidential election received more engagement on Facebook than top stories from major media outlets. It also has the potential to undermine trust in serious media coverage. The term has at times been used to cast doubt upon legitimate news, and former U.S. president Donald Trump has been credited with popularizing the term by using it to describe any negative press coverage of himself. It has been increasingly criticized, due in part to Trump's misuse, with the British government deciding to avoid the term, as it is "poorly-

defined" and "conflates a variety of false information, from genuine error through to foreign interference".

Multiple strategies for fighting fake news are currently being actively researched, and need to be tailored to individual types of fake news. Effective self-regulation and legally-enforced regulation of social media and web search engines are needed. The information space needs to be flooded with accurate news to displace fake news. Individuals need to actively confront false narratives when spotted, as well as take care when sharing information via social media. However, reason, the scientific method and critical thinking skills alone are insufficient to counter the broad scope of bad ideas. Overlooked is the power of confirmation bias, motivated reasoning and other cognitive biases that can seriously distort the many facets of immune mental health. Inoculation theory shows promise in designing techniques to make individuals resistant to the lure of fake news, in the same way that a vaccine protects against infectious diseases.

Conceptual Background of the Domain Problem

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

In the below blog we are going to see about how we are classifying the fake news with the genuine news; we are going to use several machine learning techniques and we will plot and analyse how to identify a news as fake. I have tried using several NLP techniques and arrived at a model that will classify news is fake or genuine.

Review of Literature

The purpose of the literature review is to:

1. Identify the News basis on the content and author to tell weather it is fake or not
2. Stop the spread of fake news which will potentially spread incorrect information amount the people.

To solve this problem, we are now building a model using our machine learning technique that identifies all the Fake news, using the same the news companies can avoid the spread of fake news in all the mediums.

I have used 8 different Classification algorithms and shortlisted the best on basis on the metrics of performance and I have chosen one algorithm and build a Machine Learning model in that algorithm.

Motivation for the Problem Undertaken

Fake news is a topic that has gained a lot of attention in the past few years, and for good reasons. As social media becomes widely accessible, it becomes easier to influence millions of people by spreading misinformation. As humans, we often fail to recognize if the news we read is real or fake. A study from the University of Michigan found that human participants were able to detect fake news stories only 70 percent of the time. But can a neural network do any better? Keep reading to find out.

The goal of this article is to answer the following questions:

- What kinds of topics or keywords appear frequently in real news versus fake news?
- How can we use a deep neural network to identify fake news stories?

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

I start analysis on this project in importing the data set and simple play around with the data and identifying the characteristics of each column.

I noticed that there are five columns “title”, “text”, “subject”,date and “target”.

Importing Dataset

```
df_fake = pd.read_csv("Fake.csv")
df_true = pd.read_csv("true.csv")
```

```
df_fake.head(10)
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

I dropped date and title columns as these columns won’t help us in predicting.

```
# Removing the title
data.drop(["title"],axis=1,inplace=True)
data.head()
```

	text	subject	target
0	Senator John McCain was overheard saying a sic...	politics	fake
1	Donald Trump s two month old presidency has be...	News	fake
2	Melania Trump just accepted damages of approxi...	left-news	fake
3	(In this Jan. 31 story, in 11th paragraph cor...	politicsNews	true
4	Yeah you know the Obama regime is serious abou...	Government News	fake

Post this I have checked about the null values in the data

I changed the null values in writtern_by as “not_avaliable” and dropped the null values in the remaining data.

```
df_fake.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
dtype: int64
```

```
df_true.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
dtype: int64
```

Then post this I analysed the label column which is our target variable and I understood that label column has two variables ‘0’ and ‘1’. ‘0’denotes true news and 1 denotes fake news.

```
# How many fake and real articles?
print(data.groupby(['target'])['text'].count())
data.groupby(['target'])['text'].count().plot(kind="bar")
plt.show()
```

```
target
fake    23481
true    21417
Name: text, dtype: int64
```

On further analysis of the label data, I understood that we have a balanced data almost 50% of data with fake news and not fake news. Balance data will help us in building a perfect machine learning model and we also avoid the model to overfit and underfit with the data.

Data Pre-processing Done

I started the pre-processing with cleansing the data, filtering out all the Bash data and I like to keep only the required data for our analysis.

I started with importing the required libraries. And I have declared stop words and lemmatize to a variable.

```
# Removing stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')

data['text'] = data['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)
```

```
data.head()
```

	text	subject	target
0	senator john mccain overheard saying sick anti...	politics	fake
1	donald trump two month old presidency disastro...	News	fake
2	melania trump accepted damages approximately 3...	left-news	fake
3	jan 31 story 11th paragraph corrects show two ...	politicsNews	true
4	yeah know obama regime serious taking isis sen...	Government News	fake

Then I have created a function to clean the data as like below.

Then I passed my data in this function to get the data cleaned.

```
# Remove punctuation

import string

def punctuation_removal(text):
    all_list = [char for char in text if char not in string.punctuation]
    clean_str = ''.join(all_list)
    return clean_str

data['text'] = data['text'].apply(punctuation_removal)
```

```
# Check
data.head()
```

	text	subject	target
0	senator john mccain was overheard saying a sic...	politics	fake
1	donald trump s two month old presidency has be...	News	fake
2	melania trump just accepted damages of approxi...	left-news	fake
3	in this jan 31 story in 11th paragraph correc...	politicsNews	true

To understand how much data, I have removed I calculated with the length of the column before cleansing and I calculated length of the columns after cleansing

Nearly 35% of the data I have cleaned further I have split the data into X and y before training and converted the data into vectors by TFIDF vectorizer.

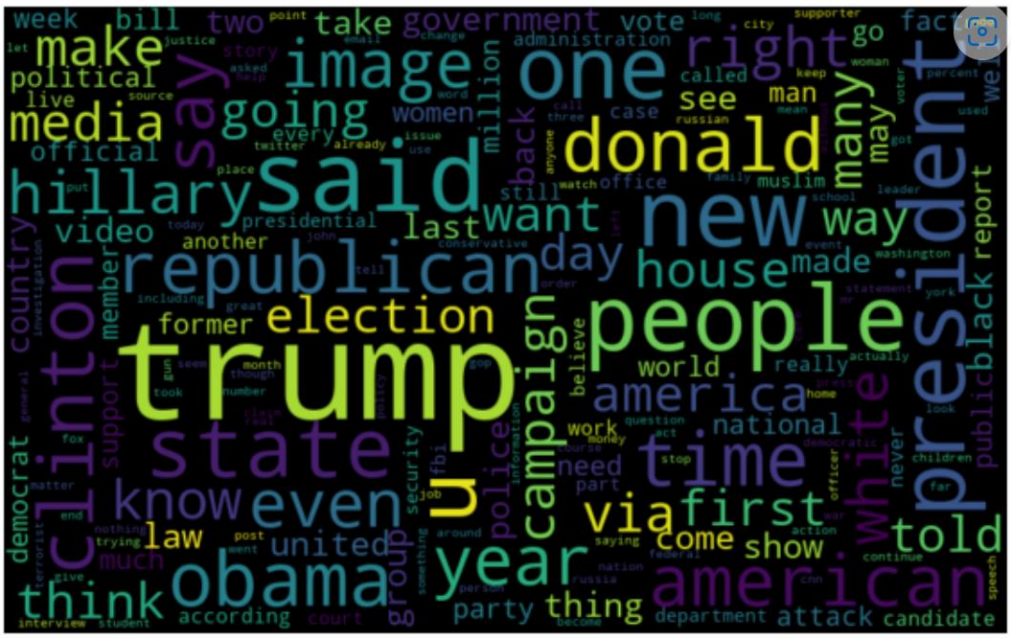
```
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
```

```
subject
Government News     1570
Middle-east          778
News                9050
US_News              783
left-news           4459
politics            6841
politicsNews       11272
worldnews          10145
Name: text, dtype: int64
```

Data Inputs- Logic- Output Relationships

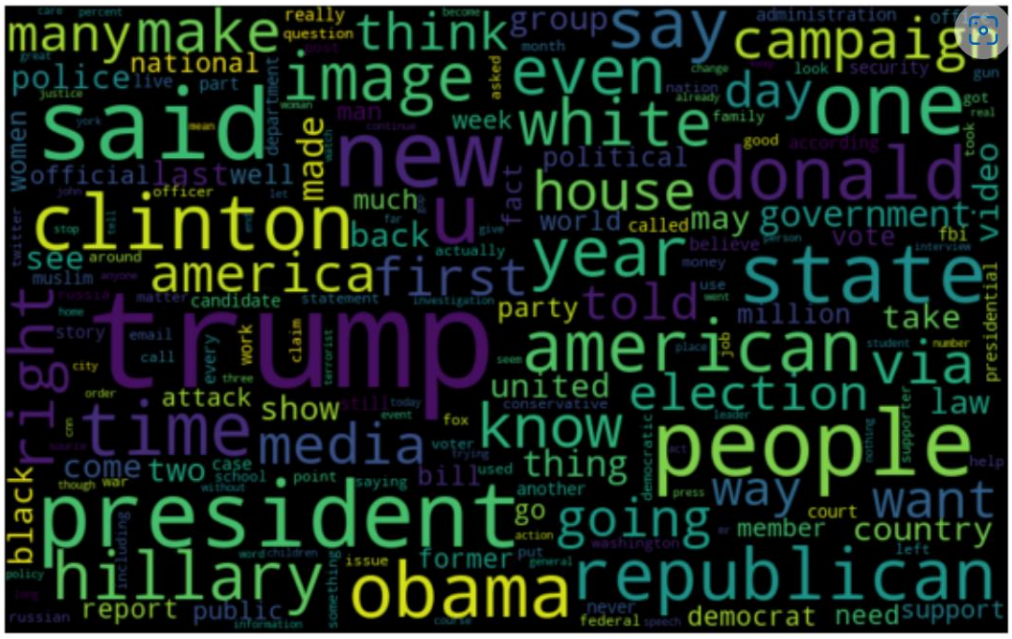
To understand the input and output technique I used Word Cloud Plot to understand what are the repeated words in a category.

loud words in real News - Headline



Key Observations:

1. We can evidently see that most of the fake news are from "not_avaliable", "Dtype', "Length", "Editor" which means these fake news sources are not available.
2. News without a proper author name is being a fake news.



Key Observations:

1. Fake news headlines have most repeated keywords Trump, Prison, win outcome November. So, any news on these keywords might be fake.
2. This data should be taken at the time of election or before the time of election because we link the key words, we can understand these news's are fake propaganda.

Key Observations:

1. Fake news's has most repeated keywords Washington, republican, plan, Society so any news on these keywords might be fake
2. As live we guessed before there are more chances that this data might have taken before the time of election or might be some funded individuals might have speeded this news's.
   We will now see about the not Fake News highlighting words.

Key Observations:

1. Not Fake news have most repeated keywords Washington, republican, plan, Society which are as same as repeated words in fake new. So, news won't be much helpfully in our prediction.

2. And we also see week, Augusta, sonny, telling, Georgian are being the most repeated words in not Fake news.

Hardware and Software Requirements and Tools Used

1. Python 3.8.
2. NumPy.
3. Pandas.
4. Matplotlib.
5. Seaborn. 6. Data science.
6. SciPy
7. Sklearn.
8. Anaconda Environment, Jupyter Notebook.Model/s Development and Evaluation

Testing of Identified Approaches (Algorithms)

I have started the training in selecting the best random state parameter for the model as follows.

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
```

```python
def train_classifier(clf, X_train, y_train):
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)

    return accuracy, precision
```

```python
train_classifier(svc, X_train, y_train)
```

```
(0.9930957683741648, 0.9906648451730419)
```

```python
accuracy_scores = []
precision_scores = []

for name, clf in clfs.items():
    current_accuracy, current_precision = train_classifier(clf, X_train, y_train)
    print("For", name)
    print("Accuracy", current_accuracy)
    print("Precision", current_precision)

    accuracy_scores.append(current_accuracy)
    precision_scores.append(current_precision)
```

```
For SVC
Accuracy 0.9930957683741648
Precision 0.9906648451730419
For KN
Accuracy 0.707238307349666
Precision 0.9144079885877318
For NB
Accuracy 0.9477728285077951
Precision 0.9400225479143179
For DT
Accuracy 0.9962138084632517
Precision 0.9934030937215651
For LR
Accuracy 0.9951002227171493
Precision 0.9927140255009107
For RF
Accuracy 0.9978841870824053
Precision 0.9977132403384404
For AdaBoost
Accuracy 0.9962138084632517
Precision 0.9952054794520548
```

So, like above I have run all the algorithms with the data.

Key-Metrics for success in solving problem under consideration.

I have taken key metrics as Accuracy and precision.

```
performance_df
```

	Algorithm	Accuracy	Precision
5	RF	0.997884	0.997713
7	BgC	0.996993	0.997026
10	xgb	0.997884	0.996804
6	AdaBoost	0.996214	0.995205
9	GBDT	0.996102	0.993851
3	DT	0.996214	0.993403
4	LR	0.995100	0.992714
8	ETC	0.995323	0.992269
0	SVC	0.993096	0.990665
2	NB	0.947773	0.940023
1	KN	0.707238	0.914408

As we can see above Random Forest tops the chart, I have selected Random Forest model as my final model and I have Hyper parameter tuned the same to increase the performance of the model and have achieved the accuracy of 99.8 % and I have saved the model.

CONCLUSION

Key Findings and Conclusions of the Study

The finding of the study is that when the news's are being published on a bogus name, the author names not available that news are end up being Fake, and also, we can understand this fake news's are desperately being spread among the public to create a fake image of

an individual, or to get profit out of it or to destroy the good deeds of the target person.

Learning Outcomes of the Study in respect of Data Science

The universe of “fake news” is much larger than simply false news stories. Some stories may have a nugget of truth, but lack any contextualizing details. They may not include any verifiable facts or sources. Some stories may include basic verifiable facts, but are written using language that is deliberately inflammatory, leaves out pertinent details or only presents one viewpoint. "Fake news" exists within a larger ecosystem of mis- and disinformation.

Misinformation is false or inaccurate information that is mistakenly or inadvertently created or spread; the intent is not to deceive. Disinformation is false information that is deliberately created and spread "in order to influence public opinion or obscure the truth"

-(https://www.merriam-webster.com/dictionary/disinformation).

As per our evaluation, we found that lesser number of Authors or bogus names or authors unknown have released fake news. We trained 20800 observations for five context categories using a Random Forest algorithm for context detection. Then, the system classifies the fake news in one of the trained contexts in the text conversation. In our testbed, we observed 48.41% of records have fake news but if we search for the authors names in fake news only 10% of the authors spread almost all the fake news. Hence, our proposed approach can identify the Fake news and the authors who spread fake news, as discussed usually on a no source news or on a bogus name these fake news’s are spread.

Limitations of this work and Scope for Future Work

The limitation of the study is that this data was taken in a shorter time frame on a current trend which might help us in a prediction for a shorted period of time. So, if the prediction of fake news was done with very old data with our model there are chances that the prediction won't be accurate. Same applies for not immediate future data. So, in such case if we have analysis the trend of the news, and if we split the news category as politics, sports arts, general, local, international then we might get some accurate prediction.