Fake News Detection

Manthan Patel
Department of Computer Engineering
University of Guelph
Guelph, Canada
manthana@uoguelph.ca

Abstract—Computer scientists and social scientists both find fake news detection to be fascinating topics. Recent increases in fake news on social media have had a significant impact on society. Numerous users all around the world have access to a vast amount of information from numerous sources. A lot of people use social media sites like Facebook, Twitter, and WhatsApp because they can quickly and effectively convey interesting data. It is getting more and more important and difficult to develop a method that can identify false information on these platforms. To implement this model various algorithms can be used like support vector machine, logistic regression algorithm, decision tree etc. From above mentioned algorithm I have used the logistic regression algorithm to implement this model.

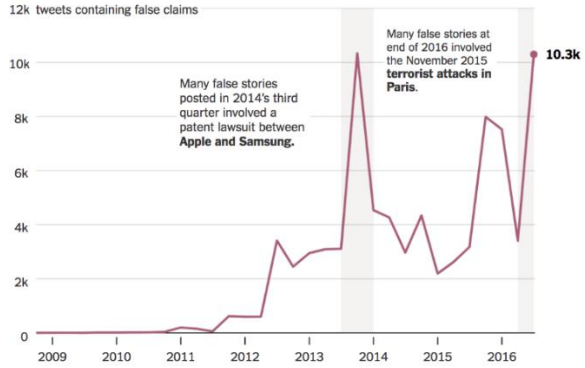
Keywords—Fake news, Logistic regression algorithm, social media

I. Introduction

Fake news exists way before from social media but it multifold when social media was introduced. Fake news is a news designed to deliberately spread hoaxes, propaganda, and disinformation. Fake News stories are usually spread through social media sites like Facebook, Twitter, etc. People used to read news from hard copies of a newspapers. These hard copies were distributed by the source which we can trust. But nowadays most people read the news on the internet. Some of this news on the internet is posted by credible sources while others are posted by deceptive sources. Therefore, it has become increasingly difficult to trust news sources.

II. About Fake news

Due to high levels of digital illiteracy and low levels of internet adoption, false news has grown to be a severe issue. Fake news has advantages and disadvantages just like any other social phenomenon. Figure 1 shows an increase in a number of fake news between the years 2009 and 2016. It is observed that between 2010 and 2011 number of fake news is negligible as compare to 2014. Many false stories posted in 2014's third quarter involved a patent lawsuit between Apple and Samsung. This event increased the number of fake news over 10k. Many false stories at end of 2016 involved the November 2015 terrorist attacks in Paris. Over 10k news were also reported during that time. It is concluded that whenever any major incident happens, the number of fake news rises. Some sources, including official government websites, are reliable, but others require a licence (libertarian). The true source's identity can be quickly determined in both of these situations. The issue arises when the authorities are unable to determine the source of a news story and social media enters the picture. Social media is a decentralised information source with low reliability.



(Fig 1. Comparison Graph)

III. Problem Statement

Fake news is frequently written and published with the purpose to deceive in order to harm a company, organization, or person and/or to profit financially or politically. To attract readers, fake news frequently uses sensationalist, dishonest, or outright fraudulent headlines. The American people were shocked by the abundance of "fake news" pieces during the most recent presidential election, which changed the narrative (and possibly the results) of the contest. The headlines on the papers and social media posts were absurd, and they made ludicrous accusations about the contestants. peoples are disseminating information from alternative sources without confirming it because they no longer trust the conventional media. They genuinely believe they are spreading the truth while doing this. Social media platforms are used to distribute false information like kid abduction and cow slaughter, which incites mob agitation. People decide to become vigilantes as a result of communications from an unknown source. This situation has led to disobedience of the law and authorities as well as the murder of numerous innocent people without any just legal intervention.

IV. Scope and Importance of the project

This initiative will help launch a new revolution against the propagation of fake news, one of the most pervasive hazards. It will eradicate the same from the ground up. The project will contribute to raising public knowledge to a new degree and

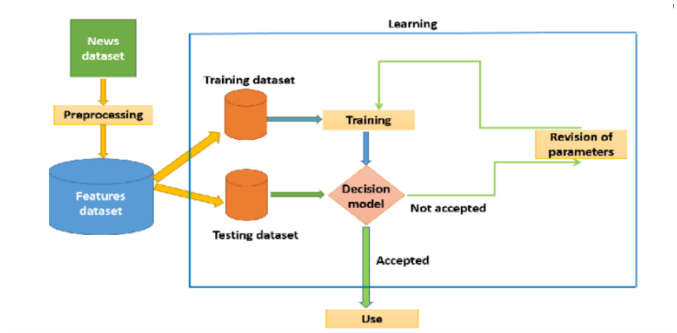
fostering greater civic responsibility. The nation's citizens will benefit from this effort by being better equipped to make important choices.

V. LITERATURE SURVEY

Large IT businesses and computer science students have recently started trying to stop the spread of false information online. Both Google and Facebook made pronouncements about how they would stop the spread of fake news on their own platforms. [1] Google looks prepared to implement policy changes that will restrict advertising to purposefully deceptive websites, eliminating the financial incentives to spread false information. Facebook is taking a similar stance on policy. [2] Daniel Sieradski also garnered news attention for creating a Chrome browser extension that contrasts the current online page with a list of dubious websites. [3] Last but not least, a Stanford student created and distributed a tool that makes the claim that it uses neural network machine learning techniques to distinguish between credible and dubious news sources. [4] The developer withheld the implementation information of how it uses neural networks in the detection after it was launched in 2016. College students made an attempt to create a browser extension to identify bogus news at Hack Princeton 2016. The project is still in the early stages of development, therefore additional contributions are required for it to function. [5] The strategies used by Google and Facebook to combat false news are based on policy decisions rather than a technical fix. We can utilize Sieradski's BS Detector as a starting point, but it depends on people to compile a list of unreliable news sources. We are interested in the Stanford student's answer since it claims to utilize machine learning to determine how trustworthy a news site is, but it withholds any information about its methodology or algorithms. To discover base less news stories at network speed, the community requires an open solution that uses statistical analysis and maybe machine learning. We work to put up a research of bogus news and employ various methods to locate it. The various accuracy levels of the various machine learning models utilized in the project have also been taken into consideration.

VI. SYSTEM ARCHITECTURE

The System Architecture of the model is shown in figure 2. First step of this process is News dataset that means collection of data. Samples of various news will be collected and stored in the dataset. This dataset needs to be clean as it contains many null values and missing values. This process will be done in Preprocessing stage. After preprocessing the data, we will get Features dataset. In the next step this feature dataset will be divided into Training dataset and Testing dataset. Prediction model will be trained using training dataset. Once the model has been trained it will be evaluated using testing dataset or the data that is has never seen before. During evaluation when an input is given to model and it will be accepted by the model then the given news will be real news otherwise if model will not accept the news, then it will pass through Revision of parameters phase and go to Training phase again.

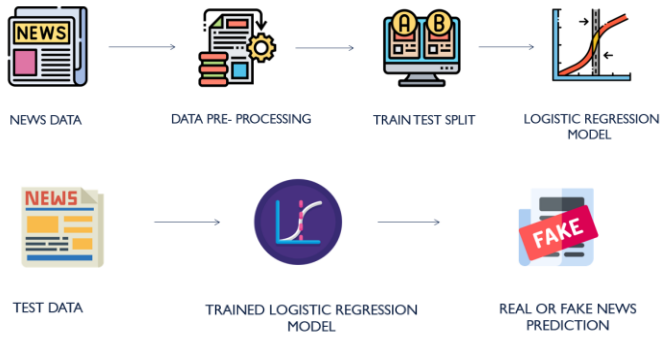


(Fig 2. System Architecture)

VII. WORK FLOW

There are seven steps in the work flow of this model.

A. News Data
B. Data Pre-processing
C. Train Test Split
D. Logistic Regression Model
E. Test Data
F. Trained Logistic Regression Model
G. Real or Fake News Prediction



(Fig 3 Work Flow)

A. News Data

This is a collection of data. In this step the data is collected from various sources and stored in a dataset.

```
# Loading the dataset to a pandas DataFrame
dataset = pd.read_csv("D:\Guelph_study\ML\Project\data.csv", low_memory = False)

data=dataset.iloc[:,0:5]

data
```

	title	text	subject	date	label
0	Bye Bye Cowboys! Crowd Boos As Owner Jerry Jon...	The Dallas Cowboys tried to have it both ways ...	Government News	25-Sep-17	1
1	U.S. lawmakers seek missing information in rev...	LONDON, (Reuters) - The chairman of a congress...	politicsNews	8-Aug-17	0
2	SELF-ADMITTED SEXUAL PREDATOR Who Supported Wi...	The hypocrisy of these liberal entertainers ...	politics	20-Nov-16	1
3	Trump travel curbs pose revenue challenges for...	NEW YORK/SAN FRANCISCO (Reuters) - President D...	politicsNews	1-Feb-17	0
4	MARINE ARRESTED FOR Complaining About Governme...	This is some pretty surreal stuff In the four ...	Government News	27-Nov-15	1

(Fig 4)

B. Data Pre-processing

Collected data may contains redundant values or null values. So, these data need to be clean. There are multiple functions available to make our dataset clean.

1. Counting missing values from dataset and replacing all values with empty string.

```
# counting the number of missing values in the dataset
data.isnull().sum()
```

```
title      44
text       49
subject    70
date       70
label      70
dtype: int64
```

```
# replacing the null values with empty string
data = data.fillna('')
```

(Fig 5)

2. Applying Stemming Function to the dataset. Stemming is the process of reducing a word to its Root word. By applying stemming function only the root words will be considered and prefix and suffix from the words will be eliminated.

Example: actor, actress, acting →act

```
port_stem = PorterStemmer()
```

```
def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content
```

```
data['content'] = data['content'].apply(stemming)
```

(Fig 6)

3. The last step of data pre-processing is transformation of data. The dataset contains textual information which computer cannot understand. Vectorizer function will be used to transform data from textual form into numerical data. This Function will convert textual data into computer understandable language.

```
# converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)
```

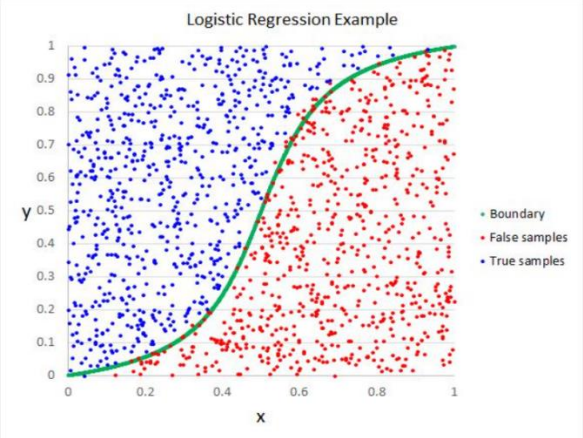
(Fig 7)

C. Train Test Split

Once the dataset has been preprocessed it will generate feature dataset. This feature dataset will be divided into Training dataset and Testing dataset. Training dataset will be used to train prediction model. Accuracy of the model will be measured by using Testing dataset.

D. Logistic Regression Model

In this step, Training dataset and Machine learning algorithm will be used to train the prediction model. In this project Logistic regression algorithm will be used to train the model. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. In figure 6, green line indicates the boundary and it will separate the dataset into two parts. One part contains true samples while the other one contains false samples.



(Fig 8)

E. Test Data

This Test Data contains the news or information that the model has never seen before. This data will be fed to the model in order to measure its performance.

F. Trained Logistic Regression Model

Trained Logistic model will be tested in this phase. Test data will be given to this model as input and it will generate the output.

G. Real or Fake news prediction

This is the last step of work flow. It will tell whether the news is real or fake. If it shows 0 as an output then the news is real and if it shows 1 as an output then the news is fake.

```
X_pred = X_test[782]

prediction = model.predict(X_pred)
print(prediction)

if(prediction[0]=='0'):
    print('The news is Real')
else:
    print('The news is Fake')
```

```
['1']
The news is Fake
```

(Fig 9)

VIII. Python Libraries Used

The entire system uses various packages and machine learning libraries for the training and predicting. Major components involved in the machine learning process includes NumPy, pandas, re, nltk, scikit-learn. The following provides more details about each single component:

1. Numpy: A scientific computing package generating N-dimensional array objects. It also has functions for working in domain of linear algebra, matrices and Fourier transform.

2. Pandas: Pandas library is used for data analysis and manipulation. It offers data structures and operations for manipulating numerical tables and time series.

3. Re(regular expression): This python library is used find particular text from the data. It is also used to check whether a particular string matches a given regular expression or not.

4. Nltk(Natural language toolkit): This python library used to build python programs that works with human language data for applying in statistical natural language processing.

5. Sci-kit learn: A Python library built on Numpy. This project uses it mainly for data classification.

IX.Algorithms that can be used

1. Support Vector Machine (SVM)

   Support Vector Machine is a powerful linear classification algorithm. By solving the maximum margin problem, it aims at finding hyperplane decision boundary in the middle the two classes. It is a part of supervised machine learning.

2. Decision Tree

   In this algorithm, the whole dataset is divided into sub dataset. Each sub dataset will be evaluated and output will be generated. It follows the path from root node to a leaf node. When it reaches a leaf node, the result stored in will be the final answer.

3. Logistic Regression Algorithm

   This algorithm is used for categorical data and most of the time for the data that needs binary classification. In this project output must be either True or False. I have used this logistic regression algorithm to train and test the model.

X. Future Scope

This project has several possibilities that could be taken into account for future development.
The possible scopes of improvements are as follows:

1. The proposed project currently using the dataset that contains title, text, subject, date and label column. In order to improve the model's accuracy two more columns "source" and "author" should be added to the dataset. By adding these two columns, it will provide the name of source by which the news was published and if the source will be genuine then there are higher chances that the published news would be real. Also, by adding author name to the dataset, it will provide the name of author by whom the news was collected or written and if author's name found to be popular then chances of fake news detection will be decreased.

2. It also makes sense to use deep learning instead of machine learning, because in deep learning there are two modules named CNN(Convolutional Neural Network) and RNN(Recurrent Neural Network) that will help to improve the performance of the model.

XI. Conclusion

It is crucial that we have a system in place for identifying false news, or at the very least, that we are aware that not all of what we read on social media and other websites is accurate. This project will increase public knowledge. It will help to launch a new revolution in opposition to one of the most pervasive hazards, namely fake news. It will help to eradicate the same from the ground up.

XII.References

Y. A. S. H. A. R. T. H. SHEKHAR, V. I. K. R. A. N. T. SAXENA, and D. E. V. A. S. Y. A. SRIVASTAVA, "A project report on fake news detection,"

pdfcoffee.com, May-2019. [Online]. Available: https://pdfcoffee.com/a-project-report-on-fake-news-detection-pdf-free.html. [Accessed: 30-Nov-2022].

Sharma, Uma & Saran, Sidarth & Patil, Shankar. (2020). Fake News Detection Using Machine Learning Algorithms. 2320-2882.

A. Jathan, "Fake news detection using machine learning," *prezi.com*, 11-Nov-2018. [Online]. Available: https://prezi.com/p/fh62zznlt7ay/fake-news-detection-using-machine-learning/. [Accessed: 30-Nov-2022].

L. Szeto, "An algorithm to detect fake news," *College of Engineering*, 04-Nov-2022. [Online]. Available: https://engineering.ucdavis.edu/news/algorithm-detect-fake-news. [Accessed: 30-Nov-2022].

Ahmed, Alim Al Ayub & Aljarbouh, Ayman & Donepudi, Praveen & Choi, Myung. (2021). Detecting Fake News using Machine Learning: A Systematic Literature Review. Psychology (Savannah, Ga.). 58. 1932-1939. 10.17762/pae.v58i1.1046.