Predicting Fake Information Using Supervised Learning and Deep Learning Algorithms

Jasmika Vemulapalli
Dept of Computer Science.
Georgia State University
jvemulapalli1@student.gsu.edu

Yaswanth Reddy Kurre
Dept of Computer Science.
Georgia State University.
nkurre1@student.gsu.edu

Kanaka Subhash Gumpena
Dept of Computer Science.
Georgia State University.
kgumpena1@student.gsu.edu

Chandana Movva
Dept of Computer Science
Georgia State University
cmovva1@student.gsu.edu

ABSTRACT

The web and internet-based life have led the entrance to news data, a lot less demanding and agreeable. Mass media affects the life of the public as it frequently occurs. There are few individuals that exploit these privileges. This prompts the creation of news articles that are not totally evident or indeed. People intentionally spread these counterfeit articles with the help of web-based social networking sites. For example, it is mostly observed that fake news had been spread during the pandemic situation where the number of deaths accountable shown much lesser than the actual deaths. So, to curb such counterfeit news from being spread we need some tools to automate the process and find efficient ways to classify it. The fundamental objective of fake news sites is to influence popular belief on specific issues. The main goal of fake news websites is to affect public opinion on certain matters. Our aim is to find a reliable and accurate model that classifies a given news article as either fake or true using algorithms like Random Forest Classifier, Pipeline Algorithm, LSTM, and Bi-directional LSTM.

KEYWORDS

News, Fake, Genuine, Natural Language Processing (NLP), Machine Learning (ML), and Supervised Learning (SL).

I. INTRODUCTION

Online social networks have transformed into a primary source for people around the world to access various types of news articles and popular content. The way news information is delivered has drastically evolved from traditional newspapers to modern websites. The spread of fake news, which refers to low-quality content containing deliberately inaccurate facts that are disseminated by individuals or bots for gossip or political purposes, has made it crucial to differentiate between genuine and fake news. Instead of the laborious manual labeling of news, a tool for the automatic classification of news articles is utilized. In this paper, we utilized the Fake or Real News 2020 dataset consisting of 72,134 rows and 4. The data set is from Kaggle, which is an open source.
https://drive.google.com/file/d/1tlcUxml7LdNbOUebrINRcgnDd2pTo-Pe/view?usp=share_link

title	text	label
You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fellow…..	FAKE
Watch The Exact Moment Paul Ryan…	Google Pinterest Digg,Linkedin Reddit…..	FAKE
Kerry to go to Paris in gesture of …	U.S. Secretary of State John F. Kerry……	REAL

Table 1: Few rows of the data set

II. LITERATURE REVIEW

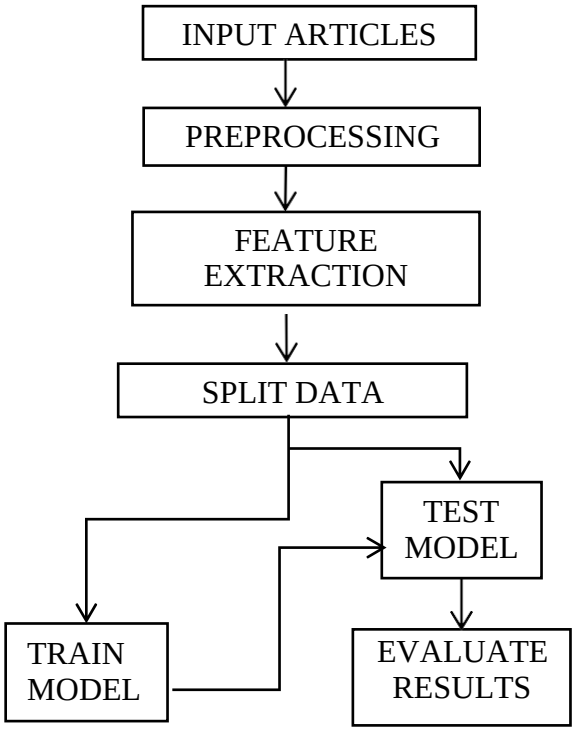
Algorithms used in different papers	**Conclusion**
Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects	The algorithms used LSTM, Passive Aggressive, Random Forest, and Naïve Bayes
Fake News Detection using Bi-directional LSTM-Recurrent Neural Network	The algorithms used in the paper's deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) can detect complex patterns in textual data.

The literate survey includes some techniques from various papers that we have researched. It states some merits and demerits of the algorithms used by other authors in their papers. Classification algorithms are mostly used for small-scale data and for huge datasets ensemble methods and CNN, and RNN algorithms are preferred. It also depends upon the datasets; different algorithms work efficiently for a particular set of data.

III. NEWS CLASSIFICATION PROCESS

The methodology in this paper starts from datasets and they are preprocessed using natural language processing tool kit libraries and techniques.

Methodology



Flowchart 1. The methodology used in our paper

After the data has been preprocessed, 75% of it is divided into training and the rest into test sets, then the model is trained upon the training set and the test set is used to evaluate the accuracy of the output [8]. Then the data is converted to a 2D matrix for making the process of analyzing easier because the matrix is in the form of numbers that are easily understood by machines. Finally, the after the data is trained using different supervised machine learning algorithms it is then evaluated using metrics.

A. Importing required libraries and data sets

We have different packages and libraries for NLP Libraries in Python, the NLTK is a useful library for tasks like tokenization, stemming, parsing, and so on. It's essentially your primary tool for NLP and ML [10]. SCI-KIT LEARN is a useful NLP library that gives developers access to a variety of ML techniques and deals with text categorization problems utilizing the bag-of-words method of building features. To read the datasets pandas library is a prerequisite and the NumPy library is for simplifying mathematical operations.

A. Text Pre-processing

The preprocessing of the data involves tokenization, stemming, stop word Removal.

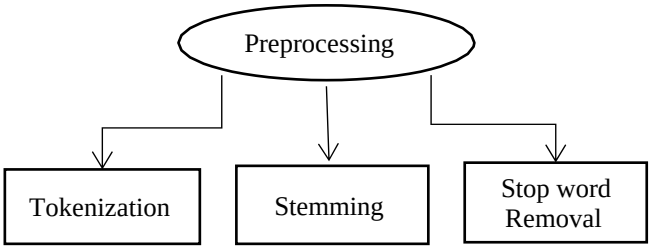
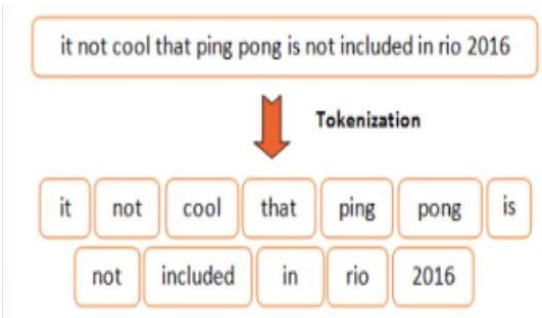


Figure 1. PREPROCESSING TECHNIQUES

TOKENIZATION

This process divides a large piece of continuous text into distinct units or tokens basically this process is often known as tokenization [6]. For breaking strings into tokens, NLTK has a method named word tokenize () (nominally words). It divides tokens based on punctuation and white space.

Figure 2. Tokenization Example



	text	target
0	[Donald, Trump, just, couldn, t, wish, all, Am...	0
1	[House, Intelligence, Committee, Chairman, Dev...	0
2	[On, Friday, ,, it, was, revealed, that, forme...	0
3	[On, Christmas, day, ,, Donald, Trump, announc...	0
4	[Pope, Francis, used, his, annual, Christmas, ...	0

STEMMING

This is the idea of removing the suffix of a word and reducing different forms of a word to a core root. The requirements for stemming are Stemmer from nltk.stem package. There are different stemmers in this package Snowball, Porter, and Lancaster. For example, several forms of wait like waits, waiting, and waited can be reduced to the word wait by removing their suffixes.

Figure 3. STEMMING

STOPWORD REMOVAL

Stop words are removed, then the entire The document is converted to lowercase for consistency. This method removes any special characters that could cause an anomaly in the document. Stop words are words that aren't relevant and have minimal lexical significance. The words that are typically avoided include: who, of, a, what, and so on.

```
0    [donald, trump, just, couldn, wish, all, ameri...
1    [hous, intellig, committe, chairman, devin, nu...
2    [friday, was, reveal, that, former, milwauke, ...
3    [christma, day, donald, trump, announc, that, ...
4    [pope, franci, use, his, annual, christma, day...
```

Figure 4. Output of STOPWORD REMOVAL on our dataset

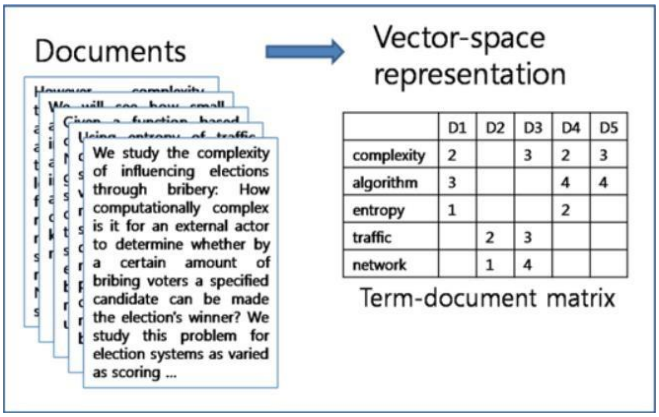
B. Vectorization

The scikit-learn library provides simple methods for extracting features from text data, a process known as vectorization. It is a technique used to convert textual data to a numerical format. Term Frequency (TF) is defined as the frequency of a word in the document when compared to all other words [2].

$$TF = \frac{number\ of\ time\ word\ appear\ in\ document}{total\ number\ of\ words\ in\ the\ document}$$

TD-IDF basically informs you how important a word is in a corpus or dataset [3]. It is a different concept for determining the significance of a word. W(I, J)=weight which signifies how important a word is for individual text messages.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 5. Converting to a 2D matrix

C. Supervised ML Algorithms

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Pipeline Algorithm

Pipelines have been growing in popularity, and now they are everywhere you turn in data science, ranging from simple data pipelines to complex machine learning pipelines. The overarching purpose of a pipeline is to streamline processes in data analytics and machine learning. ML pipeline makes the process of inputting data into the ML model fully automated.

D.Deep Learning Algorithm

Long Short-Term Memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture usedin the field of deep learning. Unlike standard feed-forward neural networks, LSTM has feedback connections. It can process notonly single data points (such as images) but also entire sequences of data (such as speech or video). A general LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and three gates regulate the flow of information into and out of the cell. LSTM is well-suited to classify, process, and predict the time series given of unknown duration.

E. Metrics

Some of the evaluation criteria are critical. in establishing the correctness of the models are applied. The categorization algorithms' results are shown in the confusion matrix (CM). It is a performance metric for ML classification problems for categorical variables as output.

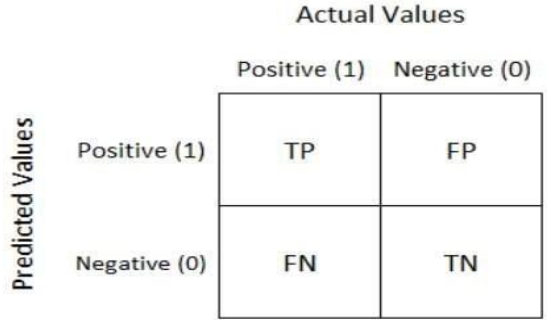


Figure 6. Confusion Matrix

True Positive: Interpretation: You predicted. that it would be positive, and you were correct.

True Negative: Interpretation You correctly anticipated negatively.

False Positive:Interpretation: You expected something positive, but it turned out to be untrue.

False Negative: Interpretation: You projected a negative outcome, but it was incorrect.

Recall: This metric indicates how well the "1" values were correctly classified as "1" values.

$$Recall = \frac{TP}{TP + FN}$$

Precision: This metric's goal is to figure out what percentage of the anticipated "1" values are true "1" values.

$$Precision = \frac{TP}{TP + FP}$$

F1 Score: The HM of Precision and Recall is the F1 Score where HM is harmonic mean. It's a new metric for determining original accuracy.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy: Accuracy is the proximity of measurement results to the true value.

IV. RESULTS

Confusion Matrices for the above algorithms used are as follows:

Predicted/Actual values	Positive(1)	Negative(0)
Positive(1)	10543	620
Negative(0)	852	9614

Table 2: CM for RF Algorithm

Predicted/Actual values	Positive(1)	Negative(0)
Positive(1)	10896	365
Negative(0)	267	10101

Table 3: CM for Pipe Algorithm

Predicted/Actual values	Positive(1)	Negative(0)
Positive(1)	9790	1223
Negative(0)	1077	9372

Table 4: CM for LSTM Algorithm

Predicted/Actual values	Positive(1)	Negative(0)
Positive(1)	9728	1156
Negative(0)	1139	9439

Table 5: CM for BI- LSTM Algorithm

From each of the above matrices we can observe the values along the diagonal are True positive and True Negative as described above and the accuracy of each algorithm depends on these values i.e.; the values closer to the original values mentioned are said to be more accurately predicted.

Algorithms/Metrics	Accuracy	Precision	Recall	F1 Score
Random Forest	93.2	92.5	94.4	93.5
Pipeline	**97.08**	**97**	**98**	**97**
LSTM	89.2	88	92	90
Bi-Direc LSTM	89.5	90	89	90

Table 6: Comparison table using metrics

From the above observations, it can be observed that Pipeline seems to produce more accurate results when compared to the remaining algorithms.

V. CONCLUSION

Detecting fake news is a difficult task as news stories are very dynamic. This research proposes a new robust method to tackle fake news or misinformation. In our methodology, first, we use Preprocessing techniques to clean the data for simple analysis through machine learning algorithms. Next, the natural language processing analyzes the retrieved news, which results in feature data that are well distinguished. Lastly, machine learning receives the feature data and classifies the news articles into two classes: real and fake. The number of data samples in each group is balanced. We separate the data into three sets: training set, and test set, each for 70% and 30% respectively. The machine learning models used in the study were Random Forest, Pipeline, LSTM, and Bi-Directional LSTM. We found that Pipeline Algorithm was the best model that achieved 97.08% on test data measured by accuracy, precision, recall, and f-measure.

VI. REFERENCES

[1]. A Taxonomy of Fake News Classification Techniques: Survey and Implementation

[2]. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network by Pritika Bahada, Preeti Saxenaa ,Raj Kamalb

[3]. H. A. Ahmed, N. Z. Bawany and J. A.Shamsi,"CaPBug-A Framework for Automatic Bug Categorization and Prioritization Using NLP and Machine Learning Algorithms," in IEEE Access, vol. 9, pp. 50496-50512, 2021 doi: 10.1109/ACCESS.2021.3069248.

[4]. Huynh, Hiep & Nguyen, Vu & Duong Trung, Nghia & Pham, Van-Huy & Phan, Cang. (2019). Distributed Framework for Automating Opinion Discretization From Text Corpora on Facebook. IEEE Access. PP. 1-1.10.1109/ACCESS.2019.2922427.

[5]. Z. Tan, J. Chen, Q. Kang, M. Zhou, A. Abusorrah and K.Sedraoui, "Dynamic Embedding Projection-Gated Convolutional Neural Networks for Text Classification," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.3036192.

[6]. A. M. Alghoson, "Medical Document Classification Based on MeSH," 2014 47th Hawaii International Conference on System Sciences, 2014, pp. 2571-2575, doi: 10.1109/HICSS.2014.324.

[7]. S. Li and Y. Liu, "News Video Title Extraction Algorithm Based on Deep Learning," in IEEE Access, vol.9, pp.12143-12157,2021,doi:10.1109/ACCESS.2021.3051613.

[8]. Saleh, Hager & Alharbi, Abdullah & Alsamhi, Saeed. (2021). OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection. IEEE Access. PP. 1-1.10.1109/ACCESS.2021.3112806.

[9]. D. Rogers, A. Preece, M. Innes and I.Spasić, "Real-Time Text Classification of User-Generated Content on Social Media: Systematic Review," in IEEE Transactions on Computational Social Systems, doi:10.1109/TCSS.2021.312013.

[10]. "Rapid Detection of fake news based on machine learning methods", Volume 192,2021, Pages 2893-2902, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.09.060.