JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS DETECTION

Ch. Karthik, P.S.S Pranav, Ch. Teja, B. Praveen

Guide Name: K. Subhadra (Associate Professor)

Department of CSE, Gandhi Institute of Science and Technology, Visakhapatnam, Andhra Pradesh, India

Abstract: In today's age where internet usage is increasing rapidly, we rely on various online news sources. Due to the huge volume and speed of data transmission over the Internet, it is impossible to have every article analysed by an expert. Even an expert in a certain field has to examine several aspects before passing a verdict on the veracity of an article , such as creating opinions in favour or against certain candidates, spammers using interesting headlines to make money with click-bait ads. Our goal is to perform a binary classification of various news articles and provide the user with the ability to classify news as fake or real using the concepts of natural language processing and machine learning.

INTRODUCTION:

The advancement of digital technology has outpaced all other innovations in human history, and while the digital age has numerous benefits, it also has drawbacks. Several issues exist in our digital environment. The fake news is one of them. False news is easily propagated by others. To harm the reputation of a person or business, fake news is disseminated. Propaganda against other groups, such as political parties or organizations, is a possibility. It also doesn't help that there are more and more online outlets for disseminating false information. A subfield of artificial intelligence called machine learning aids in the development of autonomous learning and action-capable systems. supervised, unsupervised, and reinforcement learning are all types of machine learning algorithms. A data set known as the train data set must first be used to train the algorithm. These algorithms can be utilized to carry out numerous tasks once they have been taught. Many industries utilize machine learning to carry out various tasks. The majority of machine learning algorithms are employed to make predictions or find items that are concealed.

Throughout time, fake news has increased. So, it's important to spot bogus news. Algorithms for machine learning are created with this objective in mind. After being educated, machine learning algorithms can recognize bogus news automatically. This makes its use very beneficial. In this procedure, we've processed the data, created features using a vectorizer, then trained the model using them after gathering a sizable enough collection of records to obtain an accurate accuracy score. Test out several classification algorithms. We can use an algorithm to identify bogus news after we know which algorithm has a higher accuracy score.

LITERATURE SURVEY:

There are several algorithms for detecting fake news. For this we can analyse using different classifiers. In general, after considering the dataset to work on, we choose a single algorithm and use it to solve our purpose of detecting fake news.

Instead of using a single algorithm to detect fake news, we decided to consider three algorithms and then study them to see which one provides high accuracy.

Brief information about the algorithms is as follows

Logistic Regression:
It is a statistical method, used for binary classification.

It varies in categorical in 2 methods , either true or false.

We use the sigmoid function , like s- shaped curve.

Random Forest:

It is a classifier , it is used to solve the both classification and regression problems.

It takes less time to training the dataset , it check the every decision tree to predict the accurate value.

Naive Bayes:

It is used to solve only the classification problems. it helps in making the model fast prediction and also used in text classification in high dimensional dataset It predicts based on probability of an object.

PROBLEM IDENTIFICATION AND OBJECTIVES:

The Problem:

When there is a lot of data on the Internet, the amount of fake news also increases.

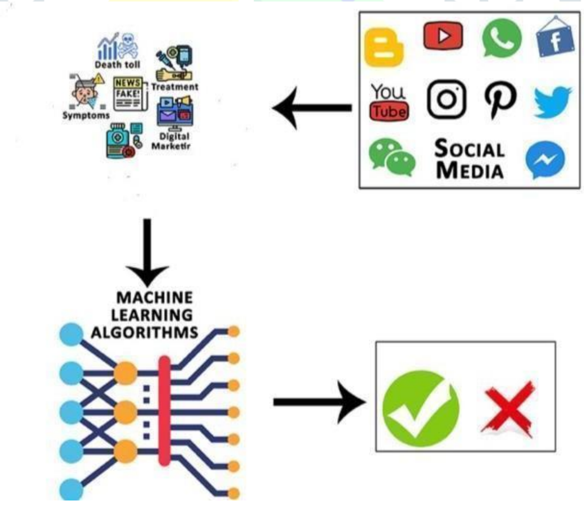
The consequences of the spread of fake news have a very huge impact on our daily lives. It is very difficult and also impossible to analyse every message whether it is real or fake. So there is a need to classify the messages as fake or real to avoid many adverse effects.

The Proposed Solution:

We can very well use machine learning to classify messages as real or fake. There are machine learning algorithms that can classify messages as genuine or fake. With a large set of records, models can be trained using algorithms to produce greater accuracy in detecting fake messages. The solution is to implement different classification of news articles and give users the ability to classify news.

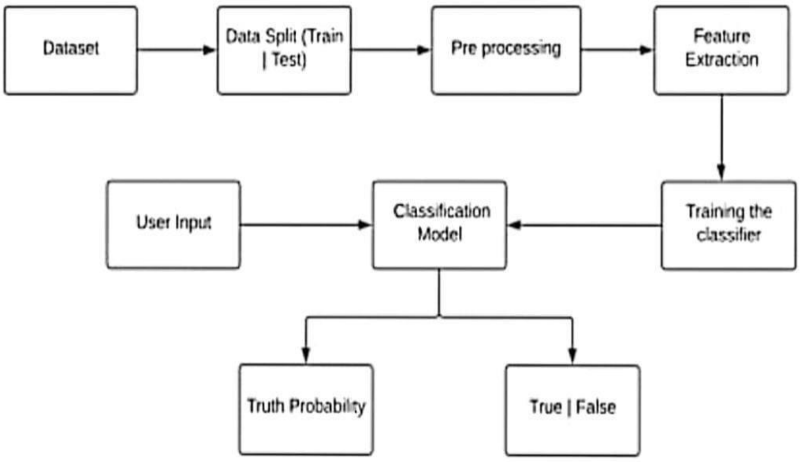
SYSTEM DESIGN:

Proposed System Architecture:



It shows the fake news detection system and how it will be implemented. As we have seen, news can be obtained from various social media platforms. There are guides that can be linked to news. After collecting a large amount of data, we help a good machine learning algorithm in the classification and use it to train our model to identify fake or real news.

Proposed System Flowchart:

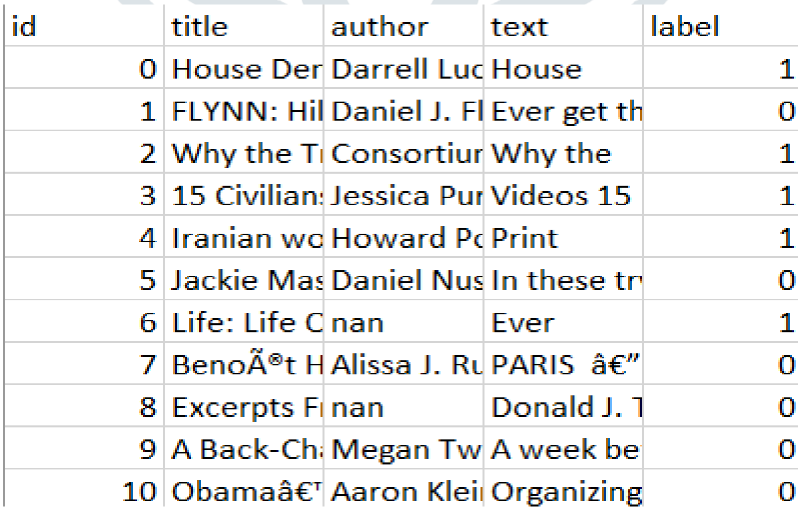


We pre-processed the data where we used natural language processing concepts to remove stop words and punctuation marks by reducing words to their root form through a stemming process. Then, we use the TF-IDF vectorizer, which converts text into digital data in the element creation phase. The learning model is now tested against the test data set and its accuracy is checked.

METHODOLOGY:

- About the dataset(sample)
- Data Pre-processing
- Feature Generation
- Algorithms
- Evaluation of the model

About the Dataset:



id	title	author	text	label
0	House Der	Darrell Luc	House	1
1	FLYNN: Hil	Daniel J. Fl	Ever get th	0
2	Why the Tı	Consortiur	Why the	1
3	15 Civilian:	Jessica Pur	Videos 15	1
4	Iranian wo	Howard Pc	Print	1
5	Jackie Mas	Daniel Nus	In these tr	0
6	Life: Life C	nan	Ever	1
7	BenoÃ®t H	Alissa J. Ru	PARIS â€”	0
8	Excerpts Fı	nan	Donald J. T	0
9	A Back-Ch:	Megan Tw	A week be	0
10	Obamaâ€	Aaron Klei	Organizing	0

Here is an illustration of a collection of five qualities:

Id: Special identifier for the news item

Title: The news article's heading

Author: The writer of the piece

Text: The article's text might not be full.

Label: The news article's label, either 1 for true or 0 for false.

Data Preprocessing:

• **Remove Punctuation:** Punctuation marks can give sentences grammatical context that aids in our comprehension. Nevertheless, because the value for the vector we use to count the number of words is missing from the context, we ignore all special characters.

Example: How are you today? -> How are you today?

• **Eliminate stop words:** These are frequent terms that can be found wherever in writing. We eliminate them because they don't provide much information about our data.

I enjoy watching both cricket and football, for instance**.**

Feature Generation:

• Number of words, frequency of big words, frequency of unusual words, etc. We can use text data to create certain features, such as where we use the vectorization process.

• Vectorization is the conversion of text into digital form. In NLP

• Vector data can be used to train our model

TF-IDF Vectorizer:

•The TF-IDF vectorizer calculates the "relative frequency" where a word appears in the entire document compared to its frequency in the document.

• The higher the TF value, the higher the frequency

• IDF stands for inverse document frequency: A word is not very useful if it is in all documents.

• "A", "an", "the", etc. Some words, like IDF measure the importance of this term and increase the importance of rare words.

• IDF(t,d) = Total number of documents

• TFIDF(t,id) = TF(t,d)*IDF(t), for example in the form of 10 [20]

ALGORITHMS:

Logistic Regression:

It predict the probabilistic value either True or False .

In this we use sigmoid function , if the value is 1 then it is real or if the value 0 then it is fake , if value lies in between 0 and 1 then it is threshold value.

How does Logistic regression work:

Logistic regression model based on sigmoid curve and sigmoid curve sigmoid function. In the sigmoid curve we have Ŷ (approximate value).

y-axis and the independent variable (X) on the x-axis. This is the formula

Calculate the value of the sigmoid function.

$$\boxed{\hat{Y} = \frac{1}{1+e^{-Z}}}$$

where Z = w*X + b

Y_hat = Predicted value

X = Independent variable

w = weight

b = bias

To determine the value of z, we need the help of Gradient descent, where we get the weight and the bias value.

Gradient Descent:

This is used to update the teaching model parameters.

each iteration determines step by step while moving towards the minimum loss function.

$$dw = \frac{(\hat{Y} - Y)*X}{m}$$

$$db = \frac{(\hat{Y} - Y)}{m}$$

where dw and db are partial derivatives of the cost function and partial formation of costs.

Work flow of the Logistic Regression model:

Step 1: Set the learning speed and number of repetitions; initialize random weight and random weight value.

Step 2: Construct the logistic regression function (sigmoid function)

Step 3: Update parameters using gradient descent

Step 4: Use the predicted value to plot the sigmoid graph.

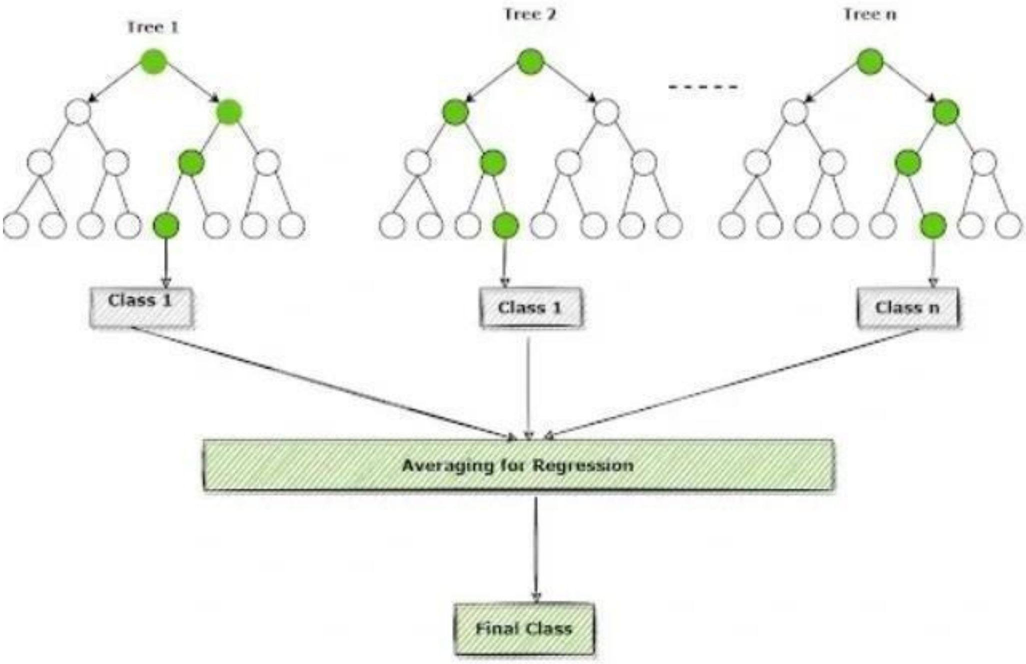
Random Forest:

The random forest is classifier , it has many decision trees .

It checks the every tree possibility to be sure which one gives the correct accuracy.

How does Random Forest algorithm work:

The random forest technique works in two stages: the first involves combining N decision trees to create a random forest.



The Random Forest algorithm's process:

Step-1:First, choose K data points at random from the training set.

Step 2:Choose N for the decision tree you wish to construct.

Step-3:Repeat steps 1 and 2.

Step 4: Locate each decision tree's forecast for the new data point, then group the new data point into the category with the most support.

Naïve Bayes:

Mainly used in textual measurement, including high-dimensional training databases.

How does Bayes Theorem Work:

Working with conditional probability, Bayes' theorem. The conditional probability is the probability of anything occurring given what has already occurred.

Process of Naive Bayes:

Step-1:Determine the standard probability for a given class label in step one.

Step-2:Determine the probability distribution for each class for each attribute.

Step-3:Calculate the posterior probability using these values and the Bayes algorithm.

Step-4:Establish that the input belongs to the class with a high probability and determine which class it does.

Evaluation of the Trained Model:

After the model is trained with different algorithms, we check the accuracy score of each algorithm.

The model is now evaluated on a test database.

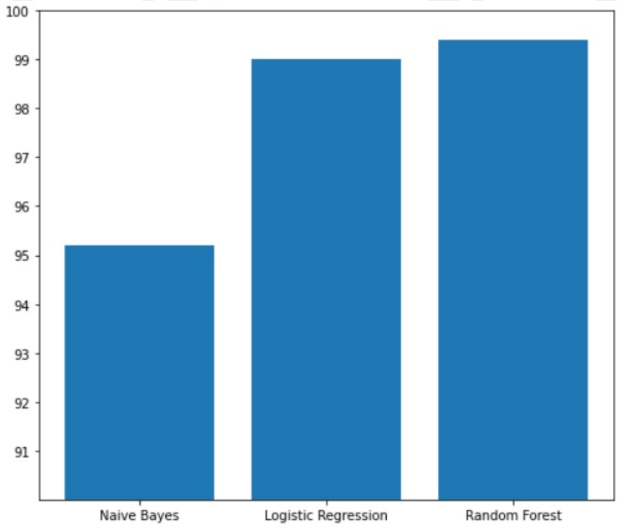
The accuracy score method is used to calculate the accuracy of fractions or correct guesses in learning Python Scikit.

Accuracy = Number of correct guesses * 100

The total number of data points

Comparison of Algorithms:

After testing model using our test database, we obtained the accuracy scores of various algorithms. Now we can compare the accuracy of the algorithms and understand which algorithm is better at determining whether a given message is real or fake. Therefore, we can use better algorithms to identify incoming messages as genuine or fake.



CONCLUSION

As the flow of information on the internet is increasing day by day, it is important to be aware of the authenticity of the information. Even today, there are many people who believe fake news and make up their own minds, which affect their decisions. The consequences of fake news can sometimes be very serious. Also, it is impossible to determine whether each news article is genuine or not. In this scenario, machine learning can be a great tool to prevent people from believing fake news. We aim to create a model that detects fake news with high accuracy using different algorithms.