

**Fake News Detection Project**



**Submitted By:**

**Ashish Kumar Rathor**

**Acknowledgment**

I would like to express gratitude for presenting this report on the “ **Fake News Detection**” Project. Working on this project was a good experience to give me a basic knowledge of the **Machine Learning Model On NLP.**  This project helps me research whether I know many new things related to this.

At the commencement of this project report, I would like to thank my SME “**Mr. Mohd. Khasif**”. Without her guidance, decision, and valuable comments and correction on my all submitted project, I try to do better compared to last time. It would possible to reach that point without his direction.

I would like to thank Flip Robbo Technologies and Data Trained for providing a suitable environment and guidance to complete my project work.

**REFERENCE**

I have also used a few external resources that helped me to complete this project successfully. Below some resources are available that I am used to completing my project

1. https://www.javatpoint.com/nlp

2-https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp

3-https://www.youtube.com/watch?v=5ctbvkAMQO4

4-https://www.youtube.com/watch?v=X2vAabgKiuM

**TABLE OF CONTAINS**

## 1. Introduction

1. Business Problem Framing
2. Conceptual Background of the Domain Problem
3. Review of Literature
4. Motivation for the Problem Undertaken

## 2. Analytical Problem Framing

1. Mathematical/ Analytical Modelling of the Problem
2. Data Sources and their formats
3. Data Pre-processing Done
4. Data Inputs- Logic- Output Relationships
5. Hardware & Software Requirements & Tools Used

## 3. Model/s Development and Evaluation

* + - 1. Identification of possible Problem-solving approaches

1. (Methods
   * 1. Visualizations
     2. Testing of Identified Approaches (Algorithms)
     3. Run and Evaluate Selected Models
2. Key Metrics for success in solving the problem under consideration
   * 1. Interpretation of the Results

## 4. Conclusion

* + 1. Key Findings and Conclusions of the Study
       1. Learning Outcomes of the Study in respect of Data Science
    2. Limitations of this work and Scope for Future Work

**INTRODUCTION**

Fake News contains misleading information that could be checked. This maintains lies about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest in some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues by confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, the development of a system to provide a credible automated index scoring, or rating for the credibility of different publishers, and news contexts. This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources, and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that is manually classified and guaranteed. Then, according to confusion matrix results, feature selection methods are applied to experiment and choose the best-fit features to obtain the highest precision. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

**BUSINESS PROBLEM FARMING**

1. **Context**

Fake news has become one of the biggest problems of our age. It has a serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

1. **Content**

What's inside is more than just rows and columns. Make it easy for others to get started by describing how you acquired the data and what period it represents, too.

**What is Fake News?**

Fake news simply means incorporating information that leads people to the wrong path. Nowadays fake news spreads like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is necessary to detect fake news.

1. **Workflow**

In this project, we are using some machine learning and Natural language processing libraries like NLTK, re (Regular Expression and ), and Scikit Learn.

1. **Natural Language Processing**

Machine learning data only works with numerical features so we have to convert text data into numerical columns. So, we have to pre-process the text and that is called natural language processing.

In-text pre-process we are cleaning our text by steaming, lemmatization, removing stop words, removing special symbols and numbers, etc. After cleaning the data, we have to feed this text data into a vectorizer which will convert this text data into numerical features.

1. **Dataset**

You can find many datasets for fake news detection on Kaggle or many other sites. I download these datasets from Kaggle. There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. You have to insert one label column zero for fake news and one for true news. We combined both datasets using pandas’ built-in function.

**MOTIVATION FOR THE PROBLEM UNDERTAKEN**

The main objective of this study is to investigate which method from a chosen set of machine learning techniques performs the best, So, far, we have a range of publicly available models served through the Perspective API, including fake news. But the current models still make errors and don’t allow users to select the type of fake words they are interested in finding.

The project which is given by Flip Robo as a part of the internship program gives the insight to identify major factors that lead to fake text. The exposure to real-world data and the opportunity to deploy my skillset in solving a real-time problem has been the primary objective. However, the motivation for taking on this project was that it is relatively a new field of research. Here we have many options but less concrete solutions. The main motivation was to classify the news to bring awareness and reduce unwanted chaos and make a good model to help us know such kinds of miscreants. Our goal is to build a prototype of an online text classifier that can be used to classify fake and true comments.

**ANALYTIC PROBLEM FARMING**

**MATHEMATICAL/ANALYTICAL MODELLING OF THE PROBLEM**

Machine learning data only works with numerical features so we have to convert text data into numerical columns. So, we have to pre-process the text and that is called natural language processing.

In-text pre-process we are cleaning our text by steaming, lemmatization, removing stop words, removing special symbols and numbers, etc. After cleaning the data, we have to feed this text data into a vectorizer which will convert this text data into numerical features.

**DATA SOURCES AND THEIR FORMATS**

You can find many datasets for fake news detection on Kaggle or many other sites. I download these datasets from Kaggle. There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. You have to insert one label column zero for fake news and one for true news. We combined both datasets using pandas’ built-in function.

**DATA PRE-PROCESSING DONE**

Data pre-processing is the processing of converting raw data into a well-readable format to be used by a machine learning model. Data pre-processing is an integral step in machine learning as the quality of the data and the useful information that can be derived from it directly affects the ability of our model to learn, therefore we must pre-process our data before feeding it into our model. I have used the following pre-processing steps:-

1. Importing necessary libraries and loading datasets into in data frame
2. After loading the data checked some statistical information like the shape of the dataset, the number of unique values, info about the dataset, null values, blank place values, values of the count of the target, duplicated values, etc
3. There are no null values present in the dataset, but white spaces were present in our dataset near 627 rows. So, we drop it directly else during modeling lots of errors are present and our accuracy and precision score get down.
4. Visualized each feature using seaborn and matplotlib libraries by plotting pie charts, heatmaps, counterplots, etc.
5. Done text pre-processing techniques like Removing Punctuations and other special characters, splitting sentences into word tokenizers, and Removing stops words by Stemming and Lemmatization techniques.
6. In last we apply Regmax for removing a special character.
7. After getting cleaned data we save it into “fake news pre.csv” format to save time to run all cells.
8. After this we apply the TF-IDF vectorizer. It will help to transform the text data to a feature vector which can be used as input in our modeling process

**DATA INPUT-LOGIC OUTPUT RELATIONSHIP**

The dataset contains two variable types one news is true and the other news is false and the target column is dependent as the values of our independent variables changes as our target variable.

To know about the scene of a loud and the most common word in each target. We use a word cloud that gives the words frequented in the target.

I know about the scene of a loud and the most common word in each target we use word cloud which gives the words frequented in the target.

**HARDWARE & SOFTWARE REQUIREMENTS & TOOLS USED**

To build the machine learning project it is important to have the followings software requirements and tools

|  |  |
| --- | --- |
| Hardware | 1. Processor: intel core i5  2. RAM: 8GB  3. SSD:512GB |
| Software | 1. Distribution: Anaconda Navigator  2. Programming Language: Python  3. Browser-based language shell: Jupyter Notebook |

**MODELS DEVELOPMENT AND EVALUATION**

**IDENTIFICATION OF POSSIBLE SOLVING**

**APPROACHES(METHOD)**

In this project, some features define the type of text like fraud news.

We extract the dependent feature y as the target and the independent feature as x. In this NLP-based project, we need to predict the binary-type labels. I have converted text into feature vectors using TF-IDF vectorizer and separated our feature and target columns. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into machine-learning models.

**TESTING OF IDENTIFIED APPROACHES(ALGORITHMS)**

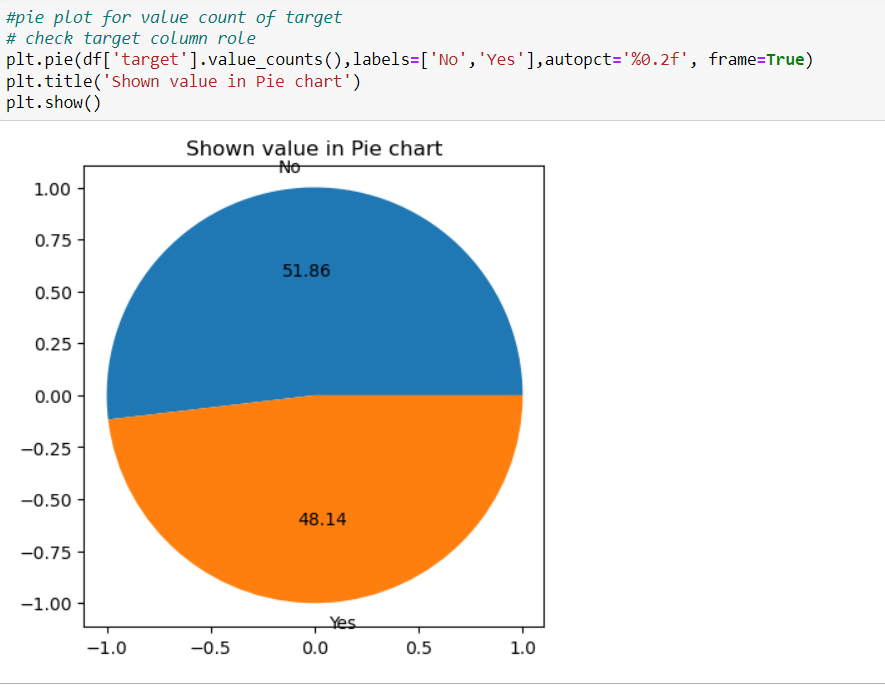
Since the target variable is categorical in nature, from this I can conclude that it is a classification-type problem hence I have used the following classification algorithms, After the pre-processing and data cleaning. I have two columns one is the target and 2nd is text nlp after applying to pre-process of NLP.

The algorithms used for training the data are as follows:

1. Logistic Regression
2. MultinomialNB
3. GaussianNB
4. BernoulliNB
5. Decision Tree Classifier
6. Kneighbors Classifier
7. Ada Boost Classifier
8. Gradient Boosting Classifier
9. Random Forest Classifier
10. Voting Classifier

**VISUALIZATIONS**

I used pandas profiling to get the overviewed visualization of the pre-processed data. Pandas is an open-source python module that can do an Exploratory Data Analysis to get a detailed description of the features and it helps in visualizing and understanding the description of the features and it helps in visualizing and understanding the distribution of each variable. I have used word cloud to get the sense of loud words in the target.



The percentage of true data is 48.14% and the percentage of fraud news is 51.86%.

**OBSERVATION**

From the pie chart, we can notice approximately 51.86% of the news is a fraud, and 48.14% of the news is true. The count of fraud news is very high compared to true news. Here no major unbalancing problem so we solve this problem by SMOTE method.

**MODEL BUILDING**

After cleaning and processing data, I performed a training test split to build the model. I have built multiple models. I have multiple classification models to get the accurate accuracy score and evaluation metrics like precision score, confusion matrix, classification matrix, etc. I got a Random Forest classifier as the best model. When I performed the Hyperparameter tuning on the Random Forest classifier but I did not get the highest accuracy score and the precision score or an f1 score.

So, I mixed the three highest accuracies and precision scores as estimators Decision Tree Classifier, Gradient Classifier, and Ada boost Classifier and put them into Voting Classifier for accuracy, then I got an accuracy score of 99.52 and a precision score is 99.71 and recall accuracy score is 99.36.

I saved my final model and got good prediction results for the test dataset.

CONCLUSION

**Learning Outcomes of the Study in respect of Data Science**

**Limitations:** This project was amazing to work on, it creates new ideas to think about but there were some limitations in this project like an unbalanced dataset. Every effort has been put into it for perfection but nothing is perfect and this project is no exception. Certain areas can be enhanced.

**Future work:** In future work, we can focus on performance and error analysis of the model as lots of text is misclassified into the fraud category. Previous work has achieved success using various algorithms on data in the English language but in the future, we can consider having data in regional languages. We can also work on after work of the detection of fraud news like automatic blocking of the user, and auto-deletion of harmful text on news, mail, and message platforms. Fraud news detection is an emerging research area with few public datasets. So, a lot of work needs to be done in this field