

# FAKE NEWS DETECTION




Presentation by Shashank

# **FAKE NEWS ?**

- Fake news is news designed to deliberately spread hoaxes, propaganda, and disinformation.
- It denotes a type of yellow journalism which intentionally presents misinformation or hoaxes spreading through both traditional print newsmedia and recent online social media.
- This is often done to further or impose certain deals and is often achieved with political agendas.
- Often, fake news will mimic real headlines and twist the story.



# WHY IS FAKE NEWS DETECTION SO IMPORTANT?

- Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing.
  - This is an evolution in human history, but at the same time it unfocusses the line between true media and maliciously forged media. Today anyone can publish content, credible or not, that can be consumed by the world wide web. Sadly, fake news accumulates a great deal of attention over the internet, especially
- 

# HOW TO RECOGNIZE A FAKE NEWS STORY?



- Watch for headline and content typos.
- Watch for excessive punctuation.
- Watch for biased vocabulary.
- Example: "Immigrants" vs. "Illegals".

# DETECTING FAKE NEWS WITH NATURAL LANGUAGE PROCESSING (NLP)




As human beings, when we read a sentence or a paragraph, we can interpret the words with the whole document and understand the context.

Given today's volume of news, it is possible to teach to a computer how to read and understand the differences between real news and the fake news using Natural Language Processing (NLP). The building blocks are Data Set and Machine Learning Algorithms.



# OBJECTIVE

- This project helps to find a way to utilize Natural Language Processing (NLP) to identify and classify fake news articles  
The main objective is to detect the fake news, which is a classic text classification problem.
  - We gathered our data, preprocessed the text, and converted our articles into features for use in supervised models.  
It is needed to build a model that can differentiate between "Real" news and "Fake" news.
- 

# METHODOLOGY

**1.**

**Data collection and preparation,** which includes getting the training data cleaning it and prepare it to the process of features extracting. This step important to get good results.

**2.**

**Feature selection,** which consists of identifying the features that are most useful for the problem under examination.

**3.**

**Algorithm choice,** given the dataset, after selecting the features, we should choose the suitable algorithm to extract these features from the dataset (classification).

**4.**

**Parameter and model selection,** which means choosing the machine learning model and setting its parameters to guarantee the best performance with the extracted features.

**5.**

**Build a model,** as given the dataset, algorithm, and parameters, training uses computational resources in order to build a model of the data in order to predict the outputs on new data.

**6.**

**Testing the model,** as before the model is used, it needs to be tested and evaluated for accuracy on data that it was not trained on.



# WORKFLOW





# 1. DATA COLLECTION

In this project, the dataset is being taken from kaggle.com. The size of the dataset is  $23471 \times 5$ . It means that there are 23471 rows along with 5 columns. The name of the columns are 'Title', 'Text', 'Subject', 'Date' and 'Class'.

## 2. PREPROCESSING THE TEXT

The performance of a text classification model is highly dependent on the words in a corpus and the features created from those words. Common words (otherwise known as stopwords) and other "noisy" elements increase feature dimensionality but do not usually help to differentiate between documents.

# 3. FEATURE EXTRACTION

To analyze and model text after it has been preprocessed, it must first be converted into features. Techniques include Bag of Words and TfidfVectorizers.

- **Bag of Words:** This model analyzes the text from all input documents and converts it in a Bag-of-Words form. For example, for more than one text (set of text documents), we can have one bag of words which will contain all distinct words from all texts in one bag.
- **Term Frequency - Inverse Document Frequency (TF-IDF):** It increases proportionally with the number of times a word appears in a document, but it is offset by its frequency in the overall corpus. While TF-IDF is a good basic metric for extracting descriptive terms, it does not take into consideration word's position or context.

# 4. CLASSIFICATION

## DECISION TREE CLASSIFIER

it is a machine learning algorithm used for classification problems where the goal is to predict the class label of an input based on its features. It works by creating a tree-like model of decisions and their possible consequences. Each node in the tree represents a decision based on a feature, and the branches represent the possible outcomes of that decision.

## GRADIENT BOOSTING CLASSIFIER

it is a machine learning algorithm used for classification problems where the goal is to predict the class label of an input based on its features. It works by creating an ensemble of decision trees, where each tree is trained to correct the errors of the previous tree. The algorithm learns the optimal decision trees by minimizing the loss function using gradient descent.

## RANDOM FOREST CLASSIFIER

it is a machine learning algorithm used for classification problems where the goal is to predict the class label of an input based on its features. It works by creating an ensemble of decision trees, where each tree is trained on a random subset of the data and features. The algorithm learns the optimal decision trees by using a random selection of data and features to reduce overfitting and improve generalization.

# LIBRARIES USED

## PANDAS

Working with "relational" or "labeled" data can be simple and intuitive thanks to the Python module pandas, which offers quick, adaptable, and expressive data structures.

## NUMPY

The Python package NumPy is used to manipulate arrays. Additionally, it has matrices, Fourier transform, and functions for working in the area of linear algebra.

## SEABORN

A package called Seaborn uses Matplotlib as its foundation to plot graphs. In order to see random distributions, it will be used.

## SKLEARN

It includes a variety of classification, regression, and clustering methods, such as support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is built to work with Python's NumPy and SciPy scientific and numerical libraries.

# LIBRARIES USED

## **TRAIN\_TEST\_SPLIT()**

Machine learning algorithms applicable to prediction-based algorithms and applications are evaluated using the train-test split. We can compare the output of our own machine-learning model to that of other machines using this quick and simple process.

## **ACCURACY\_SCORE**

This function computes subset accuracy in multilabel classification: the set of labels predicted for a sample must exactly match the corresponding set of labels in `y true`.

## **RE**

The functions in this module allow you to determine whether a given text fits a given regular expression, known as a regular expression.

## **CLASSIFICATION\_REPORT**

A classification report is used to assess the accuracy of a classification algorithm's predictions. How many predictions are correct and how many are incorrect? True Positives, False Positives, True Negatives, and False Negatives are specifically utilized to predict the metrics of a classification report.

# CONCLUSION

fake news detection is a critical task in today's world where the spread of misinformation and propaganda can have severe consequences on society. With the abundance of information available on the internet and social media, it is becoming increasingly difficult to distinguish between what is true and what is not. Machine learning algorithms such as logistic regression, decision tree classifier, gradient boosting classifier, and random forest classifier have shown promising results in detecting fake news based on features such as the language used, source credibility, and social engagement.







# REFERENCES

## 🔍 REFERENCES 1

GeeksforGeeks

url: <https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/>

## 🔍 REFERENCES 2

Simplilearn

url: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/how-to-create-a-fake-news-detection-system>

## 🔍 REFERENCES 3

Ray Oshikawa, Jing Qian, William Yang Wang, “A Survey on Natural Language Processing for Fake News Detection” published in March 2020.

# THANK YOU

Presentation by Shashank....

