Project Name: Fake News Detection Using NLP

Phase1:Problem Definition and Design Thinking

In this project,We understand the fake news has been rapidly increasing in numbers. It is not a new problem but recently it has been on a great rise. According to Wikipedia Fake news is false or misleading information presented as news. Detecting the fake news has been a challenging and a complex task. It is observed that humans have a tendency to believe the misleading information which makes the spreading of fake news even easier. According to reports it is found that human abimlity to detect deception without special assistance is only 54%. Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted. Recent reports showed that due to the rise of fake news that was being created online it had impacted the US Presidential Elections. Fake news might be created by people or groups who are acting in their own interests or those of third parties. The creation of misinformation is usually motivated by personal, political, or economic agendas.

Problem Definition:

The problem of fake news is,

Since a lot of time is spent by users on social media and people prefer online means of information it has become difficult to know about the authenticity of the news. People acquire most of the information by these means as it is free and can be accessed from anywhere irrespective of place and time. Since this data can be put out by anyone there is lack of accountability in it which makes it less trustable unlike the traditional methods of gaining information like newspapers or some trusted source. Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted.

Design Thinking:

For solving the problem,the steps that we are proceeding is given below.

1. Description of Dataset:

The dataset used in this paper is ISOT dataset. In this dataset, there are two types of articles: fake news and real news. The dataset was gathered from real-world sources, and true articles were retrieved via crawling articles from Reuters.com. The fake news articles came from a variety of sources. Politifact and Wikipedia were used to gather the fake news items. Although the majority of the articles in the collection are about politics and foreign events, they cover a wide range of topics. The dataset consists of two CSV files. True.csv is the first file, and it contains almost 12,600 reuter.com stories. Fake.csv, the second file, comprises about 12,600 items obtained from various fake news sites.

1. Web Scraping:

Large volumes of data can be automatically gathered from websites via web scraping. The majority of this data is unstructured in HTML format and is transformed into structured data in a database or spreadsheet so that it can be used in multiple applications.Here, web scraping Is done on 4 websites to get the present news. It is further added into the dataset to detect the present news as fake or not and also to increase the efficiency of detecting the fake news.

1. Text Cleaning and Pre-processing:

1.Tokenization: Tokenization is the process of breaking down a stream of text into tokens, which can be words, phrases, symbols, or any other significant items. This step’s major purpose is to extract individual words in a sentence. The tokenization is done on each text in the dataset.

2.Stop Words: Stop words are the commonly used words and are removed from the text as they do not add any value to the analysis. These phrases have little or no meaning. A list of terms that are regarded as stop words in the English language is included in the NLTK library. All the stop words from the texts are removed.

3.Capitalization: Sentences can have a combination of capital and lowercase letters. A written document is made up of multiple sentences. One of the method for reducing the issue space is to convert everything to lower case. This aligns all of the words in a document in the same location. Using the python function, all the words are converted to lower case.

4.Stemming: Stemming is the process of reducing the words to its root form by eliminating extraneous characters. PorterStemmer is one of the stemming model which is used here to convert the words into its root form.

5.Lemmatization: Text lemmatization is the process of removing a word’s superfluous prefix or suffix and extracting the basic word. All the suffixes and prefixes from the words are removed to reduce space.

D. Feature Extraction:

TF-IDF stands for Term Frequency-Inverse Document Frequency and it is a measure, used in the fields of information retrieval and machine learning that can quantify the importance or relevance of string representations in a document amongst a collection of documents. The Bag of Words technique, which is useful for text classification or for assisting a machine read words in numbers, is outperformed by the TF-IDF technique when it comes to understanding the meaning of sentences made out of words. Each feature’s TF-IDF weights are computed and recorded in a matrix with columns denoting features and rows denoting sentences.

E.Dimensionality Reduction

Dimensionality refers to how many input features, variables, or columns are present in a given dataset, while dimensionality reduction refers to the process of reducing these features. In many circumstances, a dataset has a significant number of input features, which complicates the process of predictive modelling.

F. Classification Techniques:

The classification algorithms used in this paper is,

1.Rocchio Classification

2.Bagging

3.Gradient Boosting

4.Passive Aggressive Classifier