FAKE NEWS DETECTION USING NLP

Team Leader: Lokesh.T

Team Member-1: Harish.C

Team Member-2: Manikandan.S

Team Member-3: Richard Aaron.S

Phase 1: Problem Definition and Design Thinking

Problem Definition:

The problem is to develop a fake news detection model using a Kaggle dataset. The goal is to distinguish between genuine and fake news articles based on their titles and text. This project involves using natural language processing (NLP) techniques to preprocess the text data, building a machine learning model for classification, and evaluating the model's performance.

Detecting fake news using NLP involves developing a system that can analyze news articles to distinguish between reliable and unreliable information. The problem is to use linguistic patterns, sentiment analysis, and contextual clues to create algorithms that assess the authenticity of news sources and claims. By employing Natural Language Processing, the goal is to provide a tool that helps users, journalists, and fact-checkers identify misinformation, promoting accurate news dissemination in an era where the spread of fake news is a significant concern.

By addressing the fake news problem, this technology aims to enhance information integrity, empower users to make informed decisions, and preserve the trustworthiness of news sources in the digital landscape. It's a dynamic field that continually evolves to counter emerging tactics used by purveyors of misinformation.

We are going to analyse the datatset present in kaggle “Fake and real news dataset”.

We planned to implement algorithms to clean and preprocess the raw data. This may include handling missing data, outlier detection, and normalization.

Using statistical analysis we will find unknown trends, patterns, and anomalies in the data.

Design Thinking:

- **Data Source**: Identify an available dataset containing fake news and real news.
- **Data Preprocessing**: Clean, transform, and prepare the dataset for analysis.
- **Feature Extraction**: Extract relevant features and metrics from the fake and real news data.
- **Model Selection**: Choose a suitable NLP model for classification.

  Common choices include:

  Logistic Regression
  Naive Bayes
  Support Vector Machines
  Deep Learning models (e.g., LSTM, BERT)
- **Model training:** Train the selected model using the preprocessed and vectorized data. The model learns the patterns that distinguish fake from real news.

- **Evaluation**: Assess the model's performance on a separate test dataset to ensure it generalizes well to new, unseen data.

1.Data Source: We are going to consider the dataset available in kaggle website “Fake and real news dataset”.

Dataset link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

2. Data Preprocessing: We need to identify and handle missing data, outliers, and duplicates and ensure correct data types and formats.

Generate summary statistics and visualizations.

3.Feature Extraction: We need to Identify relevant features that can provide insights to detect fake and real news.

It involves Linguistic Patterns,Sentiment Analysis,Named Entities,Source Credibility Analysis and Contextual Analysis.

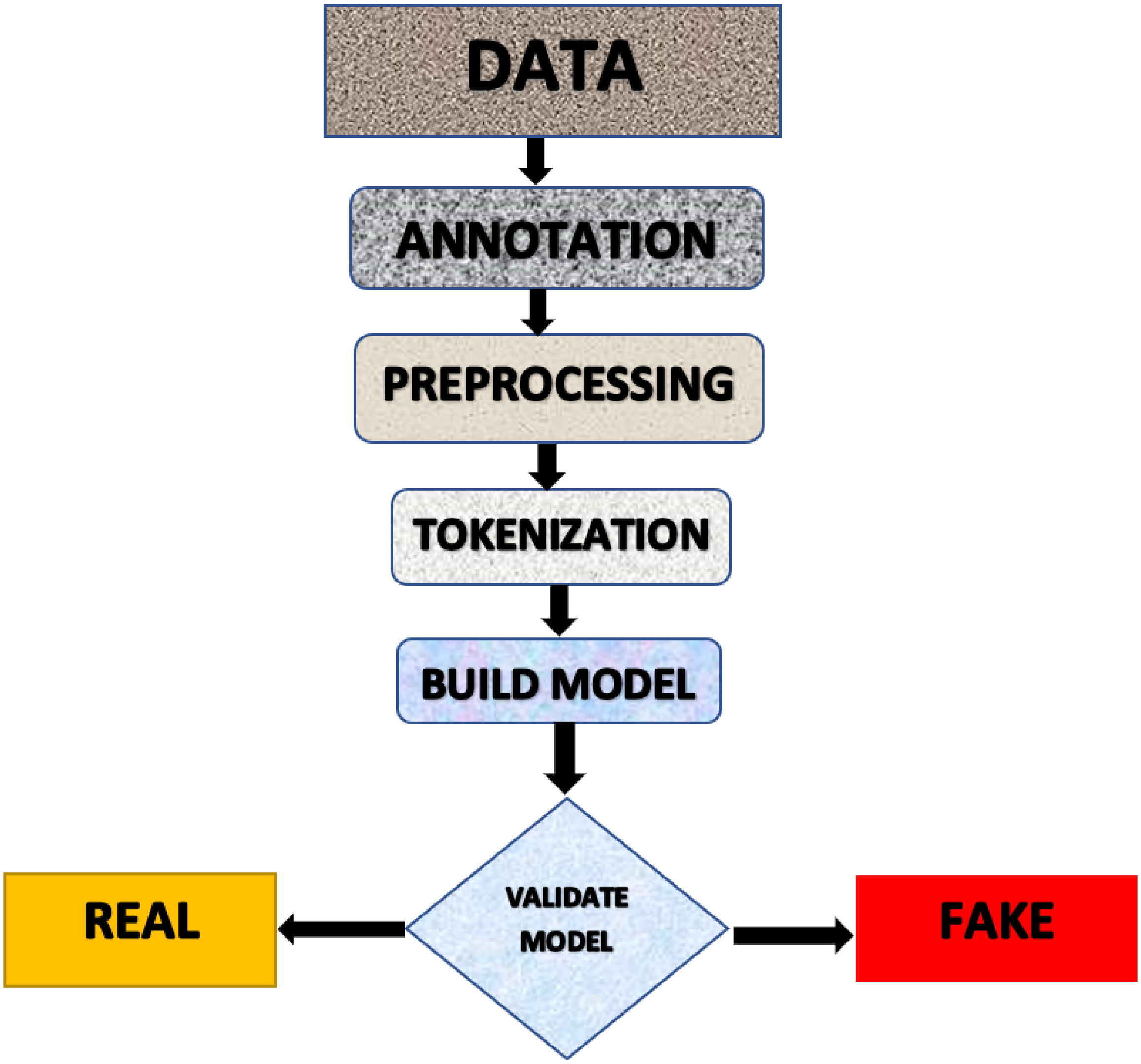
4.Model Selection: Select an NLP model (e.g., Naive Bayes, LSTM, BERT). Teach the model to recognize patterns in the data.

Test the model on new data to see how well it performs.

5. Model Training: Based on the available datasets the model is trained which is then used to predict the fake and real news for the user data.

6.Evaluation: The evaluation of data includes Testing of data,Predictions,Metrics,Confusion matrix and Real world testing.

Graphical representation:


DATA
ANNOTATION
PREPROCESSING
TOKENIZATION
BUILD MODEL
VALIDATE MODEL
REAL
FAKE