PROJECT: 8

PROJECT ID: Proj_225023_Team_1

NAME: Dhineshkumar M

FAKE NEWS DETECTION USING NLP

PHASE 1 – PROBLEM DEFINITION AND DESIGN THINKING

PROBLEM DEFINITION

The given problem requires development of a fake news detection model using Kaggle dataset. The primary objective of the project is to create a model that distinguishes fake and real news articles with high accuracy. The news articles are classified based on textual content within their titles and the body text. This project will involve the usage of Natural Language Processing (NLP) techniques for steps such as data preprocessing, feature extraction, machine learning model selection, and performance evaluation through various metrics.

This project will be immensely helpful for detecting fake news that we come across in our daily life. It helps us mitigate the spread of misinformation.

DESIGN THINKING

In creating a robust model for Fake News Detection, a methodical approach steeped in Design Thinking principles will be followed for maximum efficiency. To execute this project the following steps are followed.

Step 1: Data Source

- The dataset required for the project is given in the Kaggle website. Two separate files Fake.csv and Real.csv are provided for us to work with. In both the files data for title, text, subject and date are given.
- These files can be imported into our Google Colab Notebook and used for testing and training of the model.
  Dataset Link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Step 2: Data Preprocessing

After collecting required datasets for our project, the next step is to preprocess the datasets. Preprocessing tasks such as text cleaning and removing HTML tags, special characters, punctuation, and stopwords are carried out. This is done so that the text data is in the best format for analysing.

Step 3: Feature Extraction

- In feature extraction, techniques like TF-IDF (Term Frequency-Inverse Document Frequency) are used to convert the textual data into numerical features.
- Alternatively, another technique Word Embeddings can also be used instead of TF-IDF to represent words as dense vectors and it captures **semantic relationships** between words.

Step 4: Model Selection

- Once the features are extracted, a machine learning model is chosen from the scikit-learn module.
- At the basic level this project is a classification algorithm, meaning we are classifying fake and true news articles. So, a classification algorithm is used for the project.
- More than one algorithm is chosen and the algorithm with best accuracy metrics will be chosen for the project.

Step 5: Model Training

- First of all, the dataset is split into training and testing sets to ensure model generalization.
- The selected classification model is trained using the pre-processed datasets. It may take some time for training the model.

Step 6: Evaluation

- Since we are using more than one classification algorithms, the best one among them can be found out by comparing their performance metrics.
- For this project, metrics such as accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic – Area Under the Curve).
- With these values, we can determine how well the fake news detection model is working.

Additional Features

With a little extra code the project can be used by others for distinguishing whether the news they read on the internet is real or fake with a good level of accuracy and precision.