Fake News Detection using NLP

Phase 1

Phase 1: Problem Definition and Design Thinking

Problem Definition

- Fake News Detection is a critical challenge in today's information age, where the spread of misinformation can have significant consequences. The goal of this project is to develop an automated fake news detection system using Natural Language Processing (NLP) techniques. This system aims to identify deceptive news articles and promote the consumption of reliable information.

Importance

1) Enhancing the credibility of news sources.
2) Empowering individuals to make informed decisions.
3) Countering the harmful effects of misinformation on society.

Design Thinking

- Design thinking is an iterative approach that focuses on understanding user needs, brainstorming solutions and continuous refinement. It ensures user-centricity and adaptability.

Data Source Selection

To build an effective fake news detection system, we need a diverse and credible dataset.

✓*The given dataset link* : https://www.kaggle.com/datasets/clmentbisailion/fake-and-real-news-dataset

- This dataset contains a list of articles considered as “fake” news.

Data Preprocessing

- Data preprocessing is a crucial step to prepare the text data for analysis. It involves:

1. **Cleaning**: Removing special characters, punctuation, and HTML tags from the text.
2. **Tokenization**: Splitting text into individual words or tokens.
3. **Stop word Removal**: Eliminating common words that carry little meaning.
4. **Stemming/Lemmatization**: Reducing words to their base form.
5. **Handling Missing Data**: Dealing with articles lacking necessary information.

Feature Extraction

- Feature extraction is essential to convert text data into a numerical format suitable for machine learning models.

Techniques

✓**TF-IDF** (Term Frequency-Inverse Document Frequency):

To represent the importance of words in documents.

✓**Word Embeddings** (e.g., Word2Vec):

To capture semantic relationships between words.

Model Selection

- Choosing the right model is critical for accurate fake news detection.

Machine Learning Models

✓Logistic Regression
✓Random Forest
✓Neural Networks

Model Training and Evaluation

- The model will be trained on the preprocessed data with a division into training, validation, and test sets. We will evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score and ROC-AUC. Cross-validation will be employed to ensure robustness.