KINGSTON
ENGINEERING COLLEGE

COLLEGE CODE: 5113

Batch Members:

1. AKSHAYA PRATHIKSHA C K - (au511321106004) - akshayaprathikshaya@gmail.com
2. K PAVITHRA - (au511321106015) – kp262990@gmail.com
3. SONIA S – (au511321106022) - soniashekar2027@gmail.com
4. DIVYA A – (au51132110612)- arulkavi2003@gmail.com

ARTIFICIAL INTELLIGENCE

Project No.8 – FAKE NEWS DETECTION USING NLP

PREDICTIVE MODEL FOR FAKE NEWS DETECTION USING NLP : PROBLEM UNDERSTANDING AND DESIGN APPROACH

Problem Definition:

In an age where information dissemination is swift and ubiquitous, the proliferation of fake news has become a significant societal concern. Fake news, characterized by deliberately false or misleading information presented as genuine news, can have detrimental effects on individuals, communities, and even entire nations. To address this issue, the primary objective is to develop an accurate and reliable fake news detection model using natural language processing (NLP) techniques.

Design Thinking:

1. Data Source:

The foundation of any machine learning project lies in the quality and relevance of the dataset. Kaggle, a popular platform for data science competitions, offers a diverse range of datasets. For the purpose of developing a fake news detection model, selecting a Kaggle dataset containing articles with titles, text, and corresponding labels (genuine or fake) is essential. These datasets are curated and labeled, providing a reliable starting point for the project.

2. Data Preprocessing:

The raw textual data extracted from the dataset requires preprocessing to prepare it for analysis. Text cleaning involves removing unnecessary elements like special characters, numbers, and symbols, ensuring uniformity by converting all text to lowercase, and eliminating stop words to focus on meaningful content. Tokenization, the process of breaking down text into individual words or tokens, and lemmatization or stemming to reduce words to their base forms further refine the text for analysis.

3. Feature Extraction:

To effectively utilize the text data for machine learning, it needs to be converted into numerical features. TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings are popular techniques for this purpose. TF-IDF represents the importance of words in a document based on their frequency, providing a numerical representation of the text. Word embeddings, on the other hand, capture semantic relationships between words by mapping them into a continuous vector space.

4. Model Selection:

Choosing an appropriate classification algorithm is a crucial step in the development of a fake news detection model. Commonly used algorithms include Logistic Regression, Random Forest, Support Vector Machines (SVM), Naive Bayes, and various deep learning models such as Neural Networks, LSTM (Long Short-Term Memory), or transformer-based models like BERT.

5. Model Training:

Once the data is preprocessed and the features are extracted, it's time to split the dataset into training and testing sets. The model is then trained using the training set, allowing it to learn the patterns and relationships between the features and labels. The chosen algorithm, along with the preprocessed data, is fed into the model for training.

6. Evaluation:

The model's performance is evaluated using a variety of metrics to assess its effectiveness. Accuracy, representing the overall correctness of the model's predictions, is a fundamental metric. Precision, recall, and F1-score provide insights into the model's ability to correctly classify genuine and fake news. Additionally, the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score assesses the model's ability to distinguish between the classes.

Conclusion:

In addition to these fundamental steps, further enhancements can be made to the model. Hyperparameter tuning involves optimizing the model's parameters to improve performance. Cross-validation ensures the model's robustness and generalizability by validating its performance across different subsets of the dataset. Ensemble methods, which combine multiple models, can be employed for improved accuracy and stability. Deployment of the trained model allows it to be used for real-time predictions on new data. Continuous improvement involves periodic updates and retraining with new data to ensure the model's relevance and accuracy over time.