FAKE NEWS DETECTION USING NLP

PHASE 5

PROBLEM STATEMENT

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.

PHASES OF DEVELOPMENT

- Understanding the problem statement
- Data preprocessing
- Feature extraction
- Model selection
- Model training
- Model evaluation

DATASET & PREPROCESSING

- I obtained the dataset from Kaggle consisting of two sets; one containing fake news data and the other true news data. Both datasets have four columns; subject, date, text and title.
- PREPROCESSING TECHNIQUE USED:
- Removing duplicates
- Removal of special characters
- Removal of HTML tags
- Stopword removal
- Tokenization
- Labelling subjects
- Changing column name from text to news

FEATURE EXTRACTION

- Feature extraction is a process in machine learning and data analysis where you transform raw data, which may be in the form of text, images, or other types of data, into a format that is suitable for modeling. The goal of feature extraction is to select, transform, or create meaningful features (variables or attributes) from the raw data that can capture important patterns, information, or characteristics, and are useful for machine learning algorithms.
- TF-IDF (Term Frequency-Inverse Document Frequency) is a popular technique used in natural language processing (NLP) and text mining for feature extraction from text data. It's particularly useful for representing and quantifying the importance of words or terms in a collection of documents.

MODEL SELECTION

- For this project I have selected logistic regression and random forest.
- Logistic Regression: Logistic regression is a suitable choice for this project due to its simplicity, interpretability, and effectiveness in binary classification tasks. It is a linear model that works well when there is a clear separation between classes, making it particularly useful in distinguishing between genuine and fake news.
- Random Forest: Random Forest, on the other hand, is a versatile ensemble learning algorithm known for its robustness and ability to handle complex datasets. In the context of fake news detection, it excels at capturing non-linear relationships and interactions among features. By aggregating the predictions of multiple decision trees, it often leads to improved accuracy and robustness against overfitting.

MODEL TRAINING AND EVALUATION

- Then by training the model Random Forest, an ensemble learning method, proved to be a compelling choice. It demonstrated superior performance during the training phase, exhibiting a remarkable ability to capture complex patterns and relationships within our dataset. The ensemble approach, which aggregates predictions from multiple decision trees, contributed to its robustness and effectiveness. In particular, it showcased a notable increase in precision, which is a crucial metric in fake news detection, as we aim to minimize false positives and ensure that authentic news is not misclassified.
- While Logistic Regression is a reliable and interpretable model, it was outperformed by the Random Forest model in our specific context.
- The evaluation of our models was primarily based on key metrics, including accuracy, precision, recall, and the F1-score. These metrics allowed us to comprehensively assess the performance of our models, understanding their ability to correctly identify fake news while minimizing false positives. In particular, the higher precision achieved with the Random Forest model ensures that our model effectively identifies fake news with a reduced risk of misclassifying authentic news.