Big data technologies

**Final Assignment**

Fake News Detection with Random Forest Classification

Yertayev Ansar

Izmir Bakircay University

2024

Table of contents

**Description of the dataset and variables3**

Material and methods3

Results of the method and interpretation of success5

**Description of the dataset and variables.**

The dataset consists of two files, true.csv (21417 true news) and fake.csv (23481 fake news), which contain news articles categorized as true and fake, respectively. These datasets are combined to form a single dataset for analysis. This data is used to train a machine learning model to determine the authenticity of news articles.

Variables of both csv

* Title: The title of the news article.
* Text: The text of the news article.
* Subject: The category or topic of the news article.
* Date: The date the article was published.

The variable ‘date’ was deleted because it does not contain important information for classifying news as real or fake. Additionally, removing it reduces the time required to build the model and speeds up the training process.

Then the variable ‘label’ has been added to data to identify whether the news is real of fake (1 for real news and 0 for fake)

***Data preprocessing***

Preprocessing is required to use this data in a machine learning model. What kind of preprocessing used:

* Duplicate Removal: Removal of duplicate records.
* Missing Value Processing: Checking and removing/replacing missing values.
* Data Concatenation: Concatenating the two datasets.
* Text conversion: Converting text to lower case, removing stop words and lemmatization.

**Material/Method**

***Theoretical Explanation of the Applied Method***

***Random Forest Classifier***

Overview:

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. It was introduced by Leo Breiman in 2001 and has since become one of the most powerful and popular machine learning algorithms due to its simplicity and versatility.

***Key Concepts:***

Ensemble Learning: Ensemble learning involves combining predictions from multiple models to improve the overall performance. Random Forest is a type of ensemble learning where the base learners are decision trees.

Decision Trees: A decision tree is a flowchart-like structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

Bagging (Bootstrap Aggregating): Random Forest uses a technique called bagging to train each decision tree on a different subset of the training data. This is achieved by randomly sampling the data with replacement (bootstrap sampling). Bagging helps in reducing variance and prevents overfitting.

Feature Randomness: In addition to bagging, Random Forest introduces another layer of randomness by selecting a random subset of features to split on at each node of the tree. This further decorrelates the trees and enhances the model's robustness.

***Algorithm Steps:***

Bootstrap Sampling: Randomly sample the training data with replacement to create multiple subsets. This step is automatically handled by the RandomForestClassifier in the sklearn library. The library automatically creates multiple subsets from the training dataset with replacement.

Train Decision Trees: Train a decision tree on each bootstrap sample. During the training of each tree, a random subset of features is selected for splitting at each node. When you create an instance of RandomForestClassifier and call the fit method, the sklearn library automatically trains multiple decision trees on different subsets of the data. For each node, a random subset of features is selected for splitting.

Aggregate Predictions: For classification, aggregate the predictions of all individual trees using majority voting. For regression, average the predictions of all trees. For classification, RandomForestClassifier aggregates the predictions of all the individual trees using majority voting. This is also done automatically when you call the predict method.

***Advantages:***

Accuracy: Random Forest generally achieves high accuracy due to the ensemble of multiple decision trees.

Robustness: It is robust to overfitting, especially when the number of trees is large.

Versatility: It can handle both classification and regression task and works well with large datasets.

Feature Importance: Random Forest can provide estimates of feature importance, which can be useful for understanding the data.

***Disadvantages:***

Complexity: The model can become complex and less interpretable compared to a single decision tree.

Computational Cost: Training multiple decision trees can be computationally expensive and require more memory.

***Application in This Study:***

This study uses the Random Forest classifier to predict whether a news article is real or fake. The dataset is made up of two classes: real news and fake news. For classification, the features are created from the texts of articles including title and body text. The textual information is preprocessed through lowercase conversion, removal of stop word and stemming. The transformed data is then subjected to vectorization using term frequency-inverse document frequency (TF-IDF) as it will be converted into numerical features that can be used to train the Random Forest model. To assess the performance of the model, it is trained on a subset of this data and evaluated on another separate test set. Accuracy, precision, recall and F1- score are major metrics used during evaluation which give insights on how well our model classifies true news articles from wrong ones. Therefore, by exploiting Random Forest as described above, this study seeks to develop a robust and accurate detection model for identifying fake news which plays an important role in combating misrepresentation of facts and ensuring trustworthiness in information provision.

**Results of the method and interpretation of success**

***Model Performance***

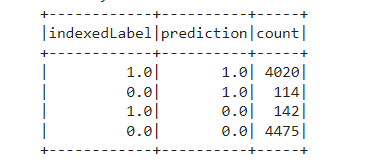
* Accuracy: The model achieved an accuracy of 0.9707462004342361.
* Classification Report:

Precision: 97.07%

Recall: 97.07%

F1-Score: 97.07%

* Confusion Matrix:



The model’s performance metrics show that the Random Forest classifier is very good at differentiating between real and fake news stories. The accuracy of 97.07% represents the overall effectiveness of the model. A precision of 97.07% and recall value of 97.07% implies that a model is precise in its predictions and can accurately identify most of the false news articles. The F1-score, which is 97.07%, gives a fair estimate of how accurate and strong this model is to make predictions on new set of examples.

