NAME: Dinesh Kumar K

NM ID: au513521106005

TITLE: Fake news detection using NLP

1. FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING

Most text and documents contain many terms that are redundant for text classification, such as stop words, misspellings, slangs, and so on. Hence, data pre-processing has to be done before the data is sent to the classification models. After that, the dataset's dimensionality is decreased in order to save time and storage space. When the dimensions are reduced, it becomes easier to visualise. The data is then used to train classification models, which can be used to predict whether or not the presented data is fraudulent.

2. Tokenization: Tokenization is the process of breaking down a stream of text into tokens, which can be words, phrases, symbols, or any other significant items. This step's major purpose is to extract individual words in a sentence. The tokenization is done on each text in the dataset.

3. Stop Words: Stop words are the commonly used words and are removed from the text as they do not add any value to the analysis. These phrases have little or no meaning. A list of terms that are regarded as stop words in the English language is included in the NLTK library. All the stop words from the texts are removed.

4. Capitalization: Sentences can have a combination of capital and lowercase letters. A written document is made up of multiple sentences. One of the method for reducing the issue space is to convert everything to lower case. This aligns all of the words in a document in the same location. Using the python function, all the words are converted to lower case.

Rocchio Classification: A type of Rocchio relevant feedback is 5. Rocchio classification. The centroid of the class of relevant documents is the average of the relevant documents, which corresponds to the most important component of the Rocchio vector in relevance feedback. Rocchio classification, which uses centroids to define the boundaries, is used to compute good class boundaries. Rocchio classification calculates the centroid for each class. When a new text data is given, it calculates the distance from each of the centroid and assigns the data point to the nearest centroid.

6. Bagging: When the goal is to reduce the variance of a decision tree classifier, bagging is utilised. The goal is to construct different subsets of data from a training sample that was picked at random and replaced. Their decision trees are trained with each group of data. As a result, we have a collection of various models. The average of all the forecasts from various trees is used which is more robust than a single decision tree classifier.

7. Gradient Boosting: A method for creating a collection of forecasts is called boosting. In order to reduce training errors, boosting is an ensemble learning technique that combines a number of weak learners into a strong learner. A random sample of data is chosen, fitted with a model, and then trained successively in boosting; each model attempts to make up for the shortcomings of the one before it. The weak rules from each classifier are joined during each iteration to create a single, powerful prediction

rule. Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

8. Passive Aggressive Classifier: For large-scale learning, passive-aggressive algorithms are commonly used. It is one of the few 'online-learning algorithms'. In contrast to batch learning, where the full training dataset is used at once, online machine learning algorithms take the input data in a sequential order and update the machine learning model step by step. This is quite helpful when there is a lot of data and training the entire dataset is computationally difficult because of the size of the data. Since, the web scraping is used in this method, it adds the data to the dataset, and the size of the dataset becomes large which makes the Passive Aggressive Classifier model to work efficiently.

9. RESULT

To assess the effectiveness of the suggested technique on diverse datasets, we ran a number of simulations and experiments using different classifiers. The dataset was divided into training and test set. 80 percent of the dataset is regarded as the training data, and the remaining 20 percent is taken as the test data. The performance of several approaches was compared using the classification's accuracy as the criterion.

10. ACKNOWLEDGMENT

We, the authors would like to express our profound appreciation to our guide Prof. Ashritha R Murthy for her invaluable mentoring, as well as for her helpful suggestions and encouraging words, which motivated us to work even harder. Due to her forethought, appreciation of work involved and continuous imparting of useful tips, this research has been successfully completed.

11. .Conclusion

The manual classification of false political news requires for a deeper understanding of the field. The problem of predicting and categorizing data in the fake news detection issue needs to be confirmed using training data. Reducing the amount of these features could increase the accuracy of the fake news detection algorithm because the majority of fake news datasets have many attributes, many of which are redundant and useless. As a result, this research suggests a technique for dimensionality reduction-based fake news detection. The dimension-reduced dataset is constructed using the final set of features. After specifying the final set of features, the next step involves utilizing classification models like Rocchio Classification, Bagging, Gradient Boosting Classifier, and Passive Aggressive Classifier to forecast the fake data. We assessed the performance of the suggested method on the dataset after it had been implemented. With the classification methods, we achieved the highest accuracy with the 94.67 percent accuracy of the TF-IDF feature extraction and the bagging classifier technique.