Task 1: NLP Project

Mark Chindudzi 19/05/2024 InternCareer

Project Overview

Objective:

AI interns will work on a Natural Language Processing (NLP) project with the goal of developing an NLP model capable of performing a specific task such as sentiment analysis, text classification, or named entity recognition.

Detailed Explanation of the Code and Project Workflow

1. Project Selection:

- The project focuses on text classification to differentiate between genuine and fake news articles.

2. Data Collection:

- Various datasets containing true and fake news articles are collected. These datasets are typically available from sources like Kaggle.

3. Data Preprocessing:

- Combining Text: The title and text of news articles are combined into a single text column for more comprehensive feature representation.

- Assigning Categories: Labels are assigned to the news articles (1 for genuine news and 0 for fake news).

- Concatenating Datasets: True and fake news datasets are concatenated into a single dataset.

- Handling Missing Values: Any missing values in the datasets are removed to ensure data quality.

- Removing Blank Texts: Blank or whitespace-only texts are removed to prevent issues during model training.

4. Text Cleaning and Preprocessing:

- Preprocessing: Text is converted to lowercase, URLs, HTML tags, punctuation, and digits are removed. This step also removes newline characters.

- Cleaning: The text is further cleaned by replacing contractions (e.g., "I'm" to "I am") and removing special characters. Lemmatization is performed to reduce words to their base form, and stopwords (commonly used words that do not contribute much meaning) are removed.

5. Feature Extraction and Model Building:

- Feature Extraction: TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is used to transform the text data into numerical features suitable for machine learning algorithms.

- Model Building: Several machine learning models are defined and trained, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Linear Support Vector Classifier (SVC).

- Evaluation: The models are evaluated using metrics such as confusion matrix, accuracy, precision, recall, and F1-score. This helps determine the performance of each model in classifying news articles correctly.

6. Manual Testing:

- A function is provided to manually input a news article and receive predictions from each trained model. This allows for testing the models on new, unseen data to check their real-world applicability.

Results Presentation

- Performance Metrics: The results include detailed performance metrics for each model, showcasing their accuracy and classification capabilities.

  Confusion Matrix: A confusion matrix is provided to visualize the model's performance in terms of true positives, true negatives, false positives, and false negatives.
  Classification Report: A comprehensive classification report that includes precision, recall, and F1-score for each class (genuine and fake news).

[Output]

...

SVM Model

Confusion matrix:

[[6956 39]

[27 6445]]

Overall accuracy: 0.9950991312096236

	precision	*recall*	*f1-score*	*support*
0.0	*1.00*	*0.99*	*1.00*	*6995*
1.0	*0.99*	*1.00*	*0.99*	*6472*
accuracy			*1.00*	*13467*
macro avg	*1.00*	*1.00*	*1.00*	*13467*
weighted avg	*1.00*	*1.00*	*1.00*	*13467*

Logistic Regression

Accuracy: 0.9883418727259227

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	6995
1.0	0.99	0.99	0.99	6472
accuracy			0.99	13467
macro avg	0.99	0.99	0.99	13467
weighted avg	0.99	0.99	0.99	13467

Decision Tree Classifier

Accuracy: 0.9964357317888171

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	6995
1.0	1.00	1.00	1.00	6472
accuracy			1.00	13467
macro avg	1.00	1.00	1.00	13467
weighted avg	1.00	1.00	1.00	13467

Random Forest Classifier

Accuracy: 0.990346773594713

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	6995
1.0	0.99	0.99	0.99	6472
accuracy			0.99	13467
macro avg	0.99	0.99	0.99	13467
weighted avg	0.99	0.99	0.99	13467

Gradient Boosting Classifier

Accuracy: 0.9945793420954927

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	6995
1.0	0.99	1.00	0.99	6472
accuracy			0.99	13467
macro avg	0.99	0.99	0.99	13467
weighted avg	0.99	0.99	0.99	13467

Enter your Article: (Reuters) - A lottery drawing to settle a tied Virginia legislative race that could shift the statehouse balance of power has been indefinitely postponed, ...

Logistic Regression Prediction: Genuine News

Decision Tree Classifier Prediction: Genuine News

Random Forest Classifier Prediction: Genuine News

Gradient Boosting Classifier Prediction: Genuine News

Process finished with exit code 0

...

Conclusion

This project demonstrates a comprehensive approach to solving an NLP classification problem, covering essential steps from data collection and preprocessing to model development, training, evaluation, and results presentation. By leveraging various machine learning models and thorough text preprocessing techniques, the project aims to effectively distinguish between fake and genuine news articles, providing valuable insights into the effectiveness of different models and preprocessing strategies in the context of text classification tasks.