TASK 01

Fake News Detection: An Internship Report

Intern: Kshetrimayum Galax Singh

Date: 19-02-2024

Company: Devtern (https://devtern.tech/)

Domain: Data Science

Batch: February-March

GitHub repo: https://github.com/galax19ksh/Detecting-fake-news

Contact: 7005788363

Mail: galaxkshetrimayum16@gmail.com

Introduction:

The spread of misinformation and fake news is a growing concern in today's digital age. It poses significant societal and democratic challenges. Misinformation can distort public perception, sway election outcomes, incite violence, and undermine trust in institutions. As a Data Science Intern at Devtern Company, I was given the first task of developing a machine learning model to detect fake news. The objective was to analyze the textual content of news articles and classify them as either real or fake. To accomplish this task, I utilized the TfidfVectorizer for feature extraction and PassiveAggressiveClassifier for classification. The entire project was executed in Google Colab, an online platform for running Python code, particularly well-suited for machine learning tasks. As a data science intern at Devtern Company, I was responsible for the entire data preprocessing, model building, and evaluation process. This report summarizes the key findings and outcomes of the project.

Data Source: https://365datascience.com/resources/downloadables/Python-Projects-Detecting-Fake-News.zip

Platform: Google Colab

Libraries used: pandas, numpy, matplotlib, seaborn, scikitlearn

Tools used: TfidfVectorizer, PassiveAggressiveClassifier

Data Preprocessing:

- Handle missing values
- Inspect sample column entries
- Balance target labels if any class imbalances

Data Splitting and Model Training:

- Split training data and test data in the ratio 8:2.
- Implement TfidfVectorizer to convert raw text into a matrix of TF-IDF features.
- Fit and transform the vectorizer on the train set and transform the vectorizer on the test set.
- initialize a PassiveAggressiveClassifier for classification. I fit this on vectorized training set and target training set.

Evaluation:

- Predict on the test set from the TfidfVectorizer and calculate the accuracy with accuracy_score() from sklearn.metrics. I got an accuracy of 94% with this model.
- Also build confusion matrix and plot heatmap for it.
- By classification report, we also got an aggregate score of 94%.

Conclusion:

In this Task 01, I learned to detect fake news with Python. We took a large dataset, implemented a TfidfVectorizer, initialized a PassiveAggressiveClassifier, and fit our model. I ended up obtaining an accuracy score of 94% in magnitude. I am looking forward to next tasks.

Future Work:

- Task 2: Uber Trips Analysis
- Task 3: The Discovery of Handwashing