Natural Language Processing Challenge: Fake vs Real News Classification

Ironhack Project

Marc Jahnert

Problem Definition

Data

- **Training data: Headlines labeled 0 (fake) and 1 (real).**
- **Testing data: Headlines with unknown labels (label 2), where the model will predict 0 or 1.**

Task

- **Develop a text classification model.**

Data Overview

Training Data: Contains news headlines with corresponding labels.

- **Label Distribution**: Approximately equal distribution:
    - Fake news (0): 17,572
    - Real news (1): 16,580

Test Data: News headlines without labels (need to be predicted).

Preprocessing:

- Text cleaning (lowercasing, removing non-alphabetic characters).

Data Preprocessing

Text Cleaning Steps:

- Convert all text to lowercase.
- Remove non-alphabetical characters.
- Remove extra spaces.

Vectorization:

- **TF-IDF Vectorization**: Uses unigrams + bigrams (to capture important word relationships).

Model Choice

Model Used: Logistic Regression

- **Why Logistic Regression**:
  - Simpler and fast compared to other models like Random Forest.
  - Suitable for binary classification tasks like fake/real news.

Cross-Validation: 5-fold cross-validation to estimate model performance.

- **Average Accuracy**: 89.49%

Model Evaluation

Cross-validation results:

- Accuracy scores:
  - 86.68%, 91.08%, 85.36%, 93.07%, 91.27%
- **Mean Accuracy**: 89.49%

Conclusion: The model has a strong predictive performance and generalizes well.

Predictions

Test Predictions: The model predicts whether news headlines are fake (0) or real (1).

Example Output:

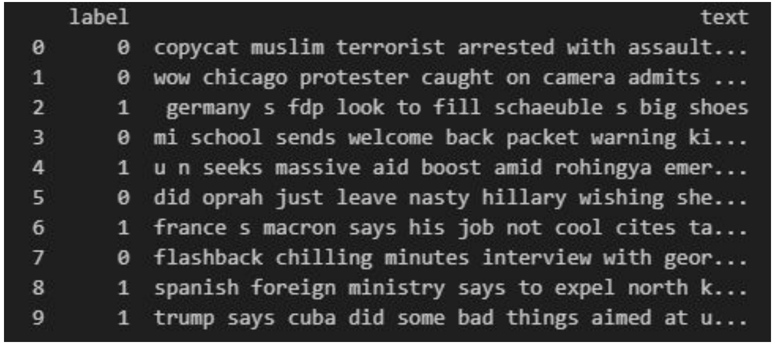
- Real News (1): "Germany’s FDP looks to fill Schaeuble’s big shoes"
- Fake News (0): "Copycat Muslim terrorist arrested with assault charges"

Final Results

Predictions: Model successfully replaced label 2 with 0 (fake) or 1 (real).

Output File: `predicted_results.csv` with the correct predictions.

Sample Preview:



```
   label                                               text
0      0  copycat muslim terrorist arrested with assault...
1      0  wow chicago protester caught on camera admits ...
2      1   germany s fdp look to fill schaeuble s big shoes
3      0  mi school sends welcome back packet warning ki...
4      1  u n seeks massive aid boost amid rohingya emer...
5      0  did oprah just leave nasty hillary wishing she...
6      1  france s macron says his job not cool cites ta...
7      0  flashback chilling minutes interview with geor...
8      1  spanish foreign ministry says to expel north k...
9      1  trump says cuba did some bad things aimed at u...
```

Summary and Conclusion

Approach:

- Text cleaning and TF-IDF vectorization with bigrams.
- Logistic Regression model for binary classification.
- High accuracy (89.49%) after cross-validation.

Final Output: Model predicts fake/real news, stored in CSV format.

Next Steps: Could explore additional techniques (e.g., word embeddings) for improved accuracy.

THANK YOU!

Marc Jahnert