NLP Exam Prep Report

Natural Language Processing (E25.147222U005.A)
Carl Emil, Liv, Julie & Hannah Mai

The Fake and Real News Dataset by Clement Bisaillon on Kaggle is a data collection of news articles made for binary text classification[1]. Specifically, it is for training machine learning models to tell real news apart from fake news. It includes around 45,000 articles split into two files (fake.csv and true.csv) each with the article's title, full text, topic, and date. A possible ethical issue is that it is not well described how the dataset distinguishes between "real" and "fake" news. We made the assumption that the labelling of the data is correct.
Furthermore, upon visual inspection of the two files, it became apparent that the "true.csv" included the source of the news article, whereas the "fake.csv" lacked this information. Despite this limitation, the dataset provides a foundation for exploring whether a language model can accurately differentiate between real and fabricated news articles. This leads to our research question;
To what extent machine learning models (specifically a distilBERT classifier) accurately distinguish between fake and true news articles?

To investigate our research question, we first combined the two datasets and added a 'label' column where 1 is real news and 0 is fake news. Then some datacleaning was completed in python using pandas, scikit-learn, and transformers libraries. For efficiency purposes the data was downsampled to 1 % of the original data, and then split into 80 % training and 20 % test data while making sure to keep class balances.

We use the DistillBERT (uncased) model, which does not distinguish between uppercase and lowercase letters. However, this prevents the classifier from interpreting capitalization patterns (such as the potential more frequent use of capital letters in fake news) as meaningful features. We specified batch size to 8 and number of epochs to 1. Full model evaluation is seen in Table 1.

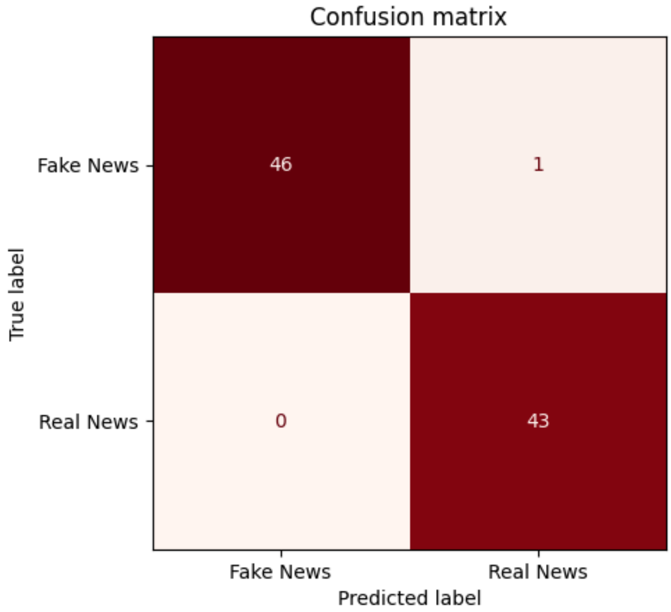
Table 1. *Classification results.*

Evaluation loss	0.066
Evaluation f1	0.989
Evaluation runtime	3.283
Evaluation samples per second	27.415
Evaluation steps per second	3.655

[1] https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?select=True.csv

Lastly, a confusion matrix was created to show the performance of our model, see Figure 1.

Figure 1. *Confusion matrix of classification results*



For further study we could consider excluding the source information from the text column of the real data, as this might be a feature that tells real and fake apart without needing to understand the content of the articles. Indeed it seems that our predictive accuracy is unnaturally high. Another possible issue could be data contamination, because the dataset was published online before the model: Data is from 2015-2018 and distilBERT/roBERTa is from 2019.

We could investigate whether there is a bias in the dataset - from visual inspection, it could seem that many of the “fake” articles are negative about Trump or the republican government in the US. If this bias is in fact in the training data, then our classifier could use negative sentiment towards these politicians as a feature to predict truthfulness of news, biasing our classifier towards calling negative news about them fake. One way to investigate this is to analyze sentiment towards Trump in the fake and real datasets, and compare the two.

Link to our repository: https://github.com/HMHojgaard/CogSci_F25

Data and code can be found in CogSci_F25/NLP/Report