Classifying Fake News in the WELFake Dataset

Lorran Caetano Machado Lopes
Insper
São Paulo, Brazil
lorrancml@al.insper.edu.br

I. Dataset

We used the *WELFake Dataset* [1] from *Kaggle*, containing around 72,000 articles for fake news detection, sourced from multiple datasets.

II. Classification Pipeline

The pipeline includes:

- **Preprocessing**: Tokenization, stopword removal, and lemmatization using NLTK.
- **Vectorization**: Text is converted to numerical form via TF-IDF Vectorizer.
- **Classification**: Logistic Regression [2] is used, with coefficients indicating word importance.

III. Evaluation

We evaluated the model's performance using both a single train-test split and 5-fold cross-validation. The accuracy on the initial test set was **95.17%**, and across 5-fold cross-validation, the model achieved a mean accuracy of **95.23%**. To further analyze the model's performance, we generated a confusion matrix (shown in **Figure 1**) based on the initial train-test split, highlighting its ability to differentiate between fake and real news.

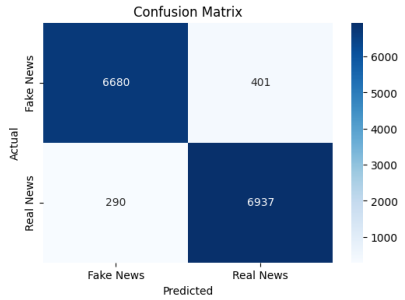

Fig. 1. Confusion Matrix for Fake News Detection: The model correctly classified 6680 out of 7081 fake news articles (94.34%) and 6937 out of 7227 real news articles (95.99%). This highlights the model's high accuracy in detecting both fake and real news.

A. Important Words

- **Fake News**: Words like "via," "us," and "hillary" were key indicators, but over-reliance on structural elements (e.g., "via," "com") could lead to misclassifications.
- **Real News**: Words like "reuters," "said," and "but" indicated real news, reflecting journalistic language and attribution to reputable sources.

IV. Dataset Size

Following the learning curves from [3], we found that expanding the dataset beyond 20,000–25,000 samples leads to diminishing returns in accuracy, as testing error stabilizes (see **Figure 2**).

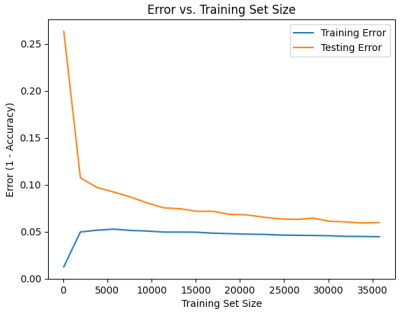

Fig. 2. Error vs. Training Set Size: The figure shows how training and testing errors evolve as the size of the training set increases. Initially, the testing error drops rapidly, but beyond approximately 20,000 samples, the improvement in accuracy diminishes, indicating a saturation point where increasing dataset size no longer leads to significant gains.

V. Topic Analysis

To analyze the topics in the dataset, we applied Non-Negative Matrix Factorization (NMF) on the document-term matrix. Five distinct topics were identified, and the results show that classification accuracy varies across topics. Topics related to *US politics* and *official investigations* achieved the highest accuracy (above 96%), while more generic topics, such as those using common terms like "people" and "know," had lower accuracy (around 93%). This indicates that classification is more effective when the language is topic-specific. The two-layer classifier, which first classifies by topic using NMF and then uses Logistic Regression for topic-specific classification, performed similarly to the single-layer model but provided more insights into the topic-based variations in classification performance.

References

[1] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, 'WELFake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021. doi: 10.1109/TCSS.2021.3068519.

[2] J. S. Cramer, 'The Origins of Logistic Regression," Tinbergen Institute Working Paper No. 2002-119/4, Dec. 2002. Available at SSRN: https://ssrn.com/abstract=360300.

[3] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, "Learning curves: Asymptotic values and rate of convergence," in *Advances in Neural Information Processing Systems (NIPS)*, 1993, pp. 327–334.