Truth or Lie? Fake News Classification using the LIAR dataset.

Academic Year 2024/2025

Chiara Martina

1. Introduction

The spread of fake news has emerged as a challenge, especially within political and social contexts.

The difficulty in automatically detecting it lies both in the variety of sources (political debates, social media, interviews, news articles) and in the lack of large, well-labeled datasets.

The LIAR dataset, developed by William Yang Wang in 2017[1], addresses this challenge by providing 12.8K short statements manually labeled into six truthfulness classes, enriched with metadata such as speaker, context, party, and historical accuracy.

This project is inspired by previous works on fake news detection based on the LIAR dataset, such as Marcelo Scatena's study[2], which compared multiple machine learning algorithms and NLP embeddings[3].

In this project, we implement a hybrid classifier that combines DistilBERT with a Multi-Layer Perceptron. The results demonstrate that this approach achieves more balanced performance across classes compared to traditional models.

2. The dataset

The LIAR dataset is a collection of short statements gathered from PolitiFact.com[4] between 2007 and 2016, designed for automatic fake news detection and fact-checking tasks. It contains 12,836 statements manually labeled across six truthfulness levels: pants-fire, false, barely-true, half-true, mostly-true, and true.

There are 14 features (such as ID, label, statement, speaker, party, state, venue, justification).

The dataset is divided into the following splits:

Training set size	10,269 (80%)
Validation set size	1,284 (10%)
Testing set size	1,283 (10%)

This division follows a standard supervised learning scheme, with the majority of data allocated for training and two smaller,

[1] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.

[2] Github Pages: https://github.com/moscatena/Fake-News-Classification?tab=readme-ov-file

[3] From classical models like Logistic Regression and Random Forest to Neural solutions using the Universal Sentence Encoder

[4] Available at the following link: https://www.politifact.com

similarly sized sets used for validation and testing.

Most texts are short (~18 tokens), typical of social media posts and political speeches.

The LIAR dataset exhibits a moderately balanced distribution of truthfulness classes. However, as noted by the dataset's creators[5], pants-fire class has significantly fewer examples compared to intermediate classes like half-true or false. This imbalance can lead models to more frequently predict the majority classes, reducing accuracy in recognizing minority classes.

3. Related Work

This work is inspired by a previous project (Scatena, 2022), which chose to focus exclusively on two main variables: the statement (the text of the claim) and the label (the truthfulness level).

In that project, the primary evaluation metric was accuracy, complemented by precision as a secondary measure.

The baseline model used was Logistic Regression, which showed clear signs of overfitting, as noted by the author, highlighting the need for regularization and tuning strategies.

To address these limitations, several approaches and models[6] were explored; however, many of them tended to overfit or did not yield significant improvements. This behavior can be attributed to several factors:

1. First, LIAR is intrinsically noisy: the linguistic distinction between a true and a false statement is often minimal and not always associated with specific keywords.
2. Second, there is a relatively small number of examples per class. Classes such as pants-fire are underrepresented, favoring memorization of training patterns at the expense of generalization ability.

The best results were obtained with models based on sentence embeddings, although after reproducing the preprocessing, performance decreased slightly. The final choice fell on a neural network with embeddings, achieving good results in recognizing true news but with nearly random performance on fake news.

4. Workflow and Architecture

Scatena (2022) shows through exploratory analysis that no clear lexical differences emerge between true and false statements. Instead, metadata such as speaker, topic, and context often play a decisive role, while the text itself requires models able to capture complex semantic nuances beyond simple word frequency.

To provide the model with additional context and reduce exclusive reliance on the statement

[5] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.

[6] Such as Random Forest, Extra Trees, SVC, Gradient Boosting, neural networks and other traditional classifiers

text, the main statement was concatenated with all relevant metadata. This approach allows the model to access additional information, which is particularly useful in cases where context, such as the speaker or political affiliation, plays a fundamental role. However, as noted by Wang et al. (2018), the inclusion of contextual data introduces some challenges: patterns associated with a specific speaker or context can vary over time, reducing the model's ability to generalize to future data[7].

For classification, following the approach of the reference project, a binary classification was adopted.

Regarding the binary class, we have 44% False and 56% True. This implies that the model "prefers" the True class.

For this project, DistilBERT was chosen as the base model, as it represents a good compromise between performance and computational cost. DistilBERT is a lighter and faster version of BERT, while still maintaining the ability to capture semantic and contextual relationships between words in a sentence. This choice was also guided by the limitations of the development environment, as the experiment was conducted on Google Colab Free, using an NVIDIA T4 GPU.

The model underwent fine-tuning, updating both the DistilBERT weights and the classification head during training.

The classification head is a multilayer perceptron (MLP) that progressively reduces the dimensionality of the sentence embedding. This gradual reduction has two main effects: on one hand, it compresses the information by selecting the most relevant features; on the other hand, it acts as implicit regularization, reducing the risk of overfitting. The MLP layer weights were initialized using Xavier Normal (Glorot & Bengio, 2010), while biases were set to zero, ensuring stability and neutrality at the start of training.

The activation function selected for the model is GELU (Gaussian Error Linear Unit) (Hendrycks & Gimpel, 2016; Devlin et al., 2018). GELU was chosen because it introduces a smooth non-linearity, weighting negative values rather than abruptly zeroing them, which improves the performance of transformer-based models such as DistilBERT.

Dropout blocks were inserted between layers to prevent reliance on specific pathways[8] and reduce overfitting, hereby enhancing the model's ability to generalize to unseen data.

[7] Buchholz, M. G. (2023). *Assessing the effectiveness of GPT-3 in detecting false political statements: A case study on the LIAR dataset.*

[8] This forces the network not to depend too much on specific paths and reduces overfitting, which is a high risk with small to medium sized LIAR datasets.

Mean pooling was adopted to aggregate the embeddings produced by DistilBERT for each token into a single vector representing the entire sequence. This method computes the average of valid token embeddings while excluding padding tokens, ensuring greater numerical stability and improving gradient flow during the early stages of training (Maini, 2020).

To optimize the fine-tuning of DistilBERT, a gradual unfreezing strategy with discriminative learning rates was applied. In this approach (Howard & Ruder, 2018), the weights of the pre-trained layers are progressively unfrozen, which helps preserve the model's prior knowledge and reduces the risk of catastrophic forgetting, while allowing the MLP head to gradually adapt to the pre-trained embeddings.

An important aspect of the training process was handling class imbalance. To mitigate this issue, the loss function was weighted based on the inverse class frequencies, ensuring that errors on the minority class (Fake) carried greater significance during optimization. This prevents the model from being biased toward the majority class (True) and improves its ability to correctly identify the minority class.

Training was carried out with a batch size of 16, chosen as a trade-off between computational efficiency and the limitations of the available GPU resources.

5. Results and Conclusion

For this project, evaluation metrics including Precision, F1-score, and ROC-AUC were selected. Accuracy was also considered, as in the reference study; however, it may overestimate performance in the presence of class imbalance, as observed here.

The best performance results are reported in *Table 1*, compared with those obtained from other models:

Model	**Accur**	**Precision**	**Test Roc-Auc**	**F1 score**
Neural Network	0.61	Class 0: 0.48 Class 1: 0.73		
Logistic Regression (TF+IDF)	0.61	Class 0: 0.52 Class 1: 0.69		
Distil BERTMLP Classifier	0.67	Class 0: 0.62 Class 1: 0.70	0.71 *(fig.1)*	0.70

Table 1. The performance of the model compared to the others.

Considering the class imbalance, it is evident that the DistilBERTMLPClassifier presents more balanced values. Although the Neural Network and Logistic Regression with TF-IDF show similar accuracy, they exhibit asymmetry in class precision, giving greater importance to the "True" class at the expense of the "False" class. improved precision for the "False" class, demonstrating enhanced capability in correctly identifying negative instances. In contrast, the DistilBERTMLPClassifier improved precision for the "False" class, demonstrating enhanced capability in correctly identifying negative instances and overall stronger discriminative ability.

This performance is supported by methodological choices: leveraging DistilBERT, integrating metadata, applying class-frequency-weighted loss, gradual unfreezing, and layer-wise discriminative learning rates. These strategies promote stable, balanced training and improve generalization despite the LIAR dataset's slight skew toward the "True" class.

The training and validation phases showed smooth and stable convergence, without evident overfitting, with continuous improvements in both accuracy and metrics such as F1 and ROC-AUC (*fig. 1*).

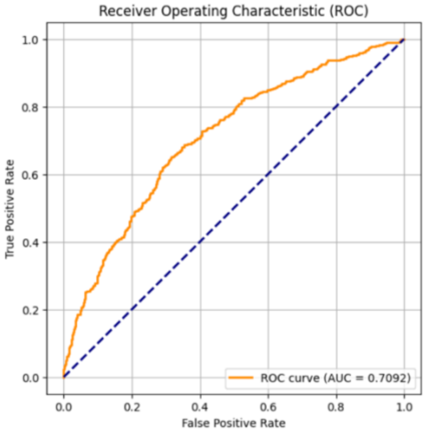

Figure 1.

The model achieved a ROC-AUC of 0.71, indicating good discriminative capability between true and false statements.

Finally, the project includes an example application of the model, where a simple function extracts predictions and decides whether to mark a statement as fake news. Given the characteristics of the LIAR dataset, the system is implemented to trigger an alert when the predicted probability of "false" exceeds a defined confidence threshold. This approach allows for the automated identification of potentially false statements while maintaining flexibility in handling uncertain or borderline cases.

Bibliography

- Buss, J. (2020). Activation function GELU in BERT. iq.opengenus.org
- Chen, Q., Ling, Z.-H., & Zhu, X. (2018). Enhancing Sentence Embedding with Generalized Pooling. arXiv:1806.09828
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics.
- Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). arXiv:1606.08415
- Idrees, H. (2024). Fine-Tuning Transformers: Techniques for Improving Model Performance. Medium
- Jeremy Howard and Sebastian Ruder. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339.

- Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts.
- Maini, P., Kolluru, K., Pruthi, D., & Mausam. (2020, May 1). Why and when should you pool? Analyzing pooling in recurrent architectures (preprint). arXiv.
- Pinto, G. V. S., Silva, R. dos S., Dazzi, R. L. S., Teive, R. C. G., Fernandes, A. M. da R., & Parreira, W. D. (2024). Classificação de Fake News utilizando o dataset LIAR. *Anais do XV Computer on the Beach*, 15, 230–235. https://doi.org/10.14210/cotb.v15.p230-235
- Ismailvanak. (2025). Understanding Embeddings with ModernBERT: Comparing Mean Pooling via Transformers and SentenceTransformers. Medium
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., … Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP architecture for vision. arXiv:2105.01601.
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2022). Improving stability of fine-tuning pretrained language models via component-wise gradient norm clipping. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.