# Fake News Classification on Liar Dataset with DistilBERT

## **Academic Year 2024/2025**

Chiara Martina

## Introduction

The spread of fake news has become a critical problem with significant political and social impacts. 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]](#footnote-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]](#footnote-2), which compared multiple machine learning algorithms and NLP embeddings (from classical models like Logistic Regression and Random Forest to neural solutions using the Universal Sentence Encoder).

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.

## The dataset

The LIAR dataset is a collection of short statements gathered from PolitiFact.com[[3]](#footnote-3) 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). Most texts are short (~18 tokens), typical of social media posts and political speeches, and the dataset is split into training, validation, and test sets.

The LIAR dataset exhibits a moderately balanced distribution of truthfulness classes. However, as noted by the dataset[[4]](#footnote-4), creators, the 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.

## 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, however, 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[[5]](#footnote-5) 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 obvious 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.

## Workflow and Architecture

In this project, in order to provide the model with more context and reduce exclusive reliance on the statement 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[[6]](#footnote-6).

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

Immagine che contiene schermata, testo, Diagramma, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

*Figure 1.*

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 chosen activation function is GELU (Devlin et al., 2018; Hendrycks & Gimpel, 2016), which introduces smooth non-linearity and promotes better gradient propagation even for negative values (Buss, 2020). Dropout blocks were inserted between layers to prevent reliance on specific pathways[[7]](#footnote-7) e ridurre l’overfitting.

Since DistilBERT produces an embedding for each token in the sentence, it was necessary to reduce these representations to a single vector describing the entire sequence. Mean pooling was adopted for this purpose, calculating the average of valid tokens while excluding padding tokens, ensuring greater stability and improving gradient flow in the early training stages (Maini, 2020).

To optimize the fine-tuning of DistilBERT, a gradual unfreezing strategy with discriminative learning rates was applied. In the first epochs, only the MLP head weights are updated while DistilBERT remains frozen. Subsequently, the upper layers of DistilBERT are progressively unfrozen until all parameters are trainable. This approach (Howard & Ruder, 2018) preserves the model’s pre-trained knowledge and reduces the risk of catastrophic forgetting, allowing the MLP head to gradually adapt to the pre-trained embeddings.

An important aspect concerns handling class imbalance. To address this, the loss function was weighted according to the inverse frequency of the classes, so that errors on the minority class (Fake) have a greater impact.

Training was conducted with a batch size of 16, compatible with the available GPU resources. Optimization was performed using AdamW with discriminative learning rates: layers closer to the input receive slower updates, preserving pre-trained information, while deeper layers and the MLP head use higher rates to allow rapid task adaptation. Finally, a CosineAnnealingWarmRestarts scheduler (Loshchilov & Hutter, 2016) was employed, which oscillates the learning rate with periodic restarts, improving training stability and helping the model escape local minima during optimization.

## Results and Conclusion

In this project, metrics such as Precision, F1-score, and ROC-AUC were chosen. Accuracy was also considered since it was used in the reference project, but it can overestimate performance when classes are unbalanced, as is evident here.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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.2) | 0.70 |

Table 1. The performance of the model compared to the others used in the campaign

Considering the class imbalance, it is evident that the DistilBERT MLP Classifier 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. In contrast, the DistilBERT MLP Classifier demonstrates stronger discriminative ability and more effective balance between precision and recall.

The use of DistilBERT, the inclusion of all metadata, and the application of techniques such as class-frequency-weighted loss, gradual unfreezing during fine-tuning, and a layer-wise discriminative learning rate optimizer contributed to more stable and balanced training. This improved the generalization for both classes, even in the presence of a dataset slightly skewed toward the “True” class, as in the LIAR dataset.

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. 2).

Immagine che contiene testo, linea, Diagramma, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

*Figure 2.*

In particular, the ROC-AUC indicated that the model has a good ability to distinguish 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.

1. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.  [↑](#footnote-ref-1)
2. Github Pages: <https://github.com/moscatena/Fake-News-Classification?tab=readme-ov-file> [↑](#footnote-ref-2)
3. Available at the following link:<https://www.politifact.com> [↑](#footnote-ref-3)
4. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.  [↑](#footnote-ref-4)
5. Quali Random Forest, Extra Trees, SVC, Gradient Boosting, reti neurali e altri classificatori tradizionali [↑](#footnote-ref-5)
6. Buchholz, M. G. (2023). *Assessing the effectiveness of GPT-3 in detecting false political statements: A case study on the LIAR dataset*.  [↑](#footnote-ref-6)
7. 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. [↑](#footnote-ref-7)