Group 2: Fake News Detection

ALEKSANDAR NIKOLOV S2523477, JUMP SRINUALNAD S2690837, University of Twente, The Netherlands

1 INTRODUCTION

1.1 Problem Statement

False or misleading claims spread rapidly across social media and news platforms, often appearing as short, context-poor snippets. This project aims to estimate the truthfulness of short political statements using the LIAR dataset, which categorizes claims into six labels ranging from *pants-on-fire* to *true*. The task presents several challenges: (1) the inputs are brief, offering limited textual evidence; (2) the labels are fine-grained, making it easy to confuse neighboring classes such as *half-true* and *mostly-true*; and (3) the dataset includes metadata (e.g., speaker, party, state, subject) that may introduce biases, allowing models to rely on contextual cues rather than the content itself.

To ensure fair evaluation, we begin with transparent text-only baselines using TF-IDF features and simple linear models. We then analyze model behavior through two diagnostic tests: a last-sentence ablation to assess dependence on specific phrases, and a metadata-only baseline to measure performance without textual input. These steps help isolate genuine linguistic signals from dataset artifacts.

Finally, we extend the setup to a transformer-based model (BERT) under identical data splits and evaluation metrics, enabling a direct comparison between traditional feature-based methods and modern neural encoders. Although BERT improves overall performance, the fine-grained nature of the labels remains a key challenge, and robustness checks remain essential to ensure reliable interpretation.

1.2 Research Questions

- **RQ1:** What approaches are most effective for classifying news articles or short statements as real or fake?
- **RQ2:** How do traditional machine learning methods using TF–IDF compare to transformer-based models like BERT in detecting fake news?

1.3 Background and inspiration

Wang[5] introduced the LIAR dataset of short, real-world political claims annotated into six veracity levels (from *pants-on-fire* to *true*). We adopt LIAR because it: (i) targets the same short-claim setting we care about, (ii) comes with predefined train/dev/test splits that enable like-for-like comparisons, and (iii) includes metadata (speaker, party, state, subject) that allows us to probe potential shortcut learning.

Our work is inspired by this setup but differs in three ways. First, we treat *macro-F1* as the primary metric (to balance all six classes) and report accuracy only for context. Second, we keep the main comparisons *text-only* to avoid inflating results with metadata; instead, we use metadata as a diagnostic control (metadata-only baseline) to quantify shortcuts. Third, reporting scores, we run targeted validity checks (e.g., removing the last sentence) and later compare the same protocol to a transformer model (BERT) on identical splits and metrics.

2 RELATED WORK

Early studies on automated fact-checking and fake news detection have explored a wide range of linguistic and contextual cues. One of the most influential contributions is by Wang (2017) [6], who introduced the *LIAR* dataset—comprising over 12,000 short political claims annotated with six fine-grained veracity levels. Wang evaluated several traditional classifiers using surface-level textual features and simple metadata such as speaker and party affiliation. The study demonstrated that metadata could boost accuracy but also introduced bias, as models might learn to associate truthfulness with certain political figures rather than linguistic content.

Subsequent work shifted toward neural architectures for text representation. For instance, Vaswani et al. (2017) [4] introduced the Transformer model, which later inspired pretrained language models such as BERT [1] and DistilBERT [3]. These models capture contextual dependencies more effectively than TF–IDF or bag-of-words representations and have shown substantial improvements on various NLP benchmarks, including fake news and stance detection tasks.

Towards Fine-Grained Reasoning for Fake News Detection [2] presents a framework called FinerFact that aims at fine-grained veracity classification by mimicking human reasoning over evidence chains. The authors use a dataset of claims annotated with multiple reasoning steps and design a neural architecture that attends to claim-evidence-label sequences. Their results shows improved separation between nuanced veracity classes (e.g., “mostly-true” vs. “half-true”) compared to simpler classification models.

Our study builds directly on these foundations but differs in three key aspects. First, unlike Wang’s work, we deliberately focus on text-only evaluation to ensure that performance reflects linguistic understanding rather than metadata bias. Second, we conduct diagnostic experiments—such as last-sentence ablation and metadata-only baselines—to quantify shortcut learning. Finally, we directly compare traditional TF–IDF-based models against a fine-tuned DistilBERT under identical splits and metrics, providing a controlled analysis of how modern transformers improve robustness on short, context-limited statements.

3 DATASET

Source and format. We use the **LIAR** benchmark, a collection of short, real-world political claims labeled with six veracity levels: *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, *true*. Our copy

TScIT 37, July 8, 2022, Enschede, The Netherlands

follows the common Kaggle-style layout (tab-separated, 14 columns, no header) with the key fields `statement` (text) and metadata such as `speaker`, `party_affiliation`, `state_info`, `subject`, `job_title`, prior truthfulness counts, and `context`. We keep the provided splits unchanged.

```
data/
  LIAR_dataset/
    train.tsv
    valid.tsv
    test.tsv
```

Splits and labels. Table 1 shows the split sizes in our files; these counts may differ slightly from the official LIAR paper splits. For modeling, we map the textual labels to integer IDs: *pants-fire* (0), *false* (1), *barely-true* (2), *half-true* (3), *mostly-true* (4), *true* (5).

Split	# instances
Train	10,240
Valid	1,284
Test	1,267

Table 1. LIAR split sizes used in this work.

Label distribution. Counts per split are reported in Table 2; the distribution is moderately imbalanced (e.g., *pants-fire* is smallest), so we use macro-F1 as the primary metric.

Label	Train	Valid	Test
pants-fire	839	116	92
false	1,995	263	249
barely-true	1,654	237	212
half-true	2,114	248	265
mostly-true	1,962	251	241
true	1,676	169	208

Table 2. Label counts per split in our copy of LIAR.

Text characteristics. Claims are short: in the training set, the median statement length is ∼99 characters (interquartile range ∼73–132, mean ∼107; max has long outliers). This motivates n-gram features and explains frequent confusions among adjacent veracity levels.

In the main experiments we use *text only* (the `statement` field) to avoid inflating scores with metadata. We additionally run a *metadata-only* control (speaker/party/state/subject/job/context) to quantify shortcut signal. Minimal preprocessing is applied (vectorizer default lowercasing; no lemmatization or stemming)

4 METHODOLOGY

4.1 Overview

We implemented both traditional feature-based and transformer-based approaches. Our experiments were conducted on the predefined *train*, *validation*, and *test* splits of the LIAR dataset, maintaining the same label mapping across all models. We evaluated all methods using **macro-F1** as the main metric to account for class imbalance and **accuracy** as a secondary reference.

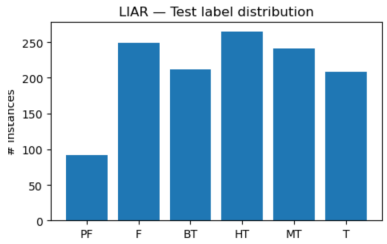

Fig. 1. Distribution of labels in the LIAR test set. The dataset is moderately imbalanced, with *pants-fire* being the smallest class and *half-true* the most common.

4.2 Preprocessing

Minimal preprocessing was applied to avoid overfitting to dataset artifacts. Each statement was automatically lower-cased by the vectorizer. No stop-word removal, stemming, or lemmatization was performed. For diagnostic experiments, we generated two modified versions of the dataset:

- **Last-sentence ablation:** The final sentence of each claim was removed using regular-expression splitting, to test whether models relied on concluding phrases.
- **Metadata-only baseline:** Textual metadata fields (speaker, party affiliation, state, subject, job title, and context) were concatenated into synthetic "sentences," to measure how much predictive signal stems from non-linguistic information.

4.3 Traditional Machine-Learning Models

We first built transparent text-only baselines using **TF–IDF representations** and **linear classifiers** implemented with `scikit-learn`.

(1) **TF–IDF (word 1–2) + Logistic Regression** – a bag-of-words model with unigrams and bigrams. A grid search over the regularization parameter $C \in \{0.5, 1, 2\}$ was run using 3-fold cross-validation.
(2) **TF–IDF (char 3–6) + Linear SVC** – a character-level model capturing subword patterns; we tuned C, loss function, and class-weight setting over 5 folds.
(3) **TF–IDF (word + char) + SGD Classifier** – a hybrid model combining word- and character-level TF–IDF features through `FeatureUnion` and optimized using stochastic gradient descent with hinge loss.

Each pipeline was trained on the training split and evaluated on the test set. Grid search and evaluation employed `GridSearchCV`, `classification_report`, and confusion-matrix visualization.

4.4 Diagnostic Experiments

Two robustness checks were conducted on the logistic-regression baseline:

- **Last-sentence ablation:** Macro-F1 dropped only marginally (approximately 0.24 → 0.238), limited dependence on final-sentence cues.
- **Metadata-only baseline:** Achieved a macro-F1 of approximately 0.25, confirming that speaker and party information contain shortcut signals that models could exploit if included.

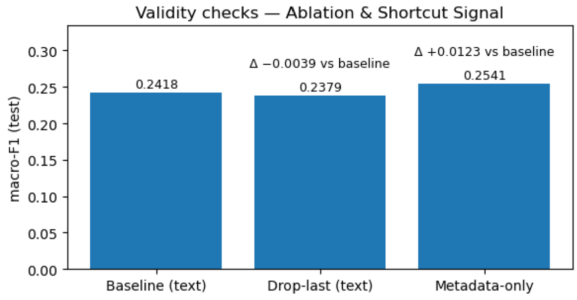

Fig. 2. Results of validity checks comparing the baseline, drop-last, and metadata-only setups. The small difference between the baseline and ablation shows minimal sentence-position bias, while the metadata-only model slightly outperforms the baseline, indicating the presence of contextual shortcuts.

4.5 Transformer Model (DistilBERT)

To compare with a modern neural encoder, we fine-tuned **DistilBERT** using the `Hugging Face Transformers` and `Datasets` libraries. The model was initialized from the pre-trained `distilbert-base-uncased` checkpoint and fine-tuned for multi-class classification with a softmax output layer of six neurons (one per LIAR label). We used the AdamW optimizer, learning-rate scheduling, and mini-batch training. Inputs were tokenized with a maximum length of 128 tokens. The same train/validation/test splits and evaluation metrics were applied to ensure comparability with the TF–IDF baselines.

4.6 Libraries and Tools

All experiments were implemented in Python 3 using the following main libraries: `pandas`, `numpy`, `scikit-learn`, `matplotlib`, and `joblib` for data handling and visualization; and `transformers`, `datasets`, and `torch` for the DistilBERT model. Random seeds were fixed for reproducibility.

5 EXPERIMENT & RESULTS

5.1 Experimental Setup

All experiments were conducted on the predefined LIAR dataset splits without altering the data distribution. The text-only statements were used as input, and labels were mapped to six veracity levels (*pants-fire* to *true*). Models were trained using 3- or 5-fold cross-validation where applicable, and evaluated on the held-out test set.

For the traditional approaches, we relied on linear classifiers with TF–IDF vectorization, while the neural experiment fine-tuned DistilBERT using the same data partitions. The fine-tuning was performed for multiple epochs with early stopping to avoid overfitting, using the AdamW optimizer and a learning rate scheduler. Each training run was initialized with a fixed random seed to ensure reproducibility.

5.2 Evaluation Metrics

To measure performance across the six veracity classes, we used the **macro-F1 score** as our primary evaluation metric, providing a balanced assessment of all classes despite class imbalance. **Accuracy** was reported as a complementary metric to offer a more intuitive overview of the model's correctness.

In addition, confusion matrices and per-class precision, recall, and F1-scores were analyzed to better understand model behavior and class-specific errors. This diagnostic evaluation was crucial for interpreting confusion between neighboring veracity labels such as *half-true* and *mostly-true*.

5.3 Results Summary & Visualization

To better interpret the model performance, we visualize the classification outcomes and overall trends. Figure 3 and Figure 4 present confusion matrices for the best-performing TF–IDF baseline and the DistilBERT model. The matrices illustrate that most misclassifications occur between adjacent truthfulness levels (e.g., *half-true* and *mostly-true*), confirming the fine-grained difficulty of the task.

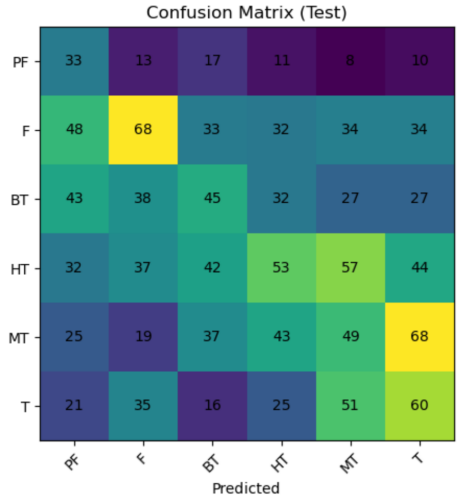

Fig. 3. Confusion matrix for the TF-IDF Logistic Regression model.

For the confusion matrices in Figures 3 and 4, the TF–IDF model shows strong diagonal concentration for the extreme labels *pants-fire* and *true*, but frequent confusion among the middle classes such as *barely-true*, *half-true*, and *mostly-true*. This pattern reflects the model's limited ability to capture nuanced linguistic cues that distinguish partial truth levels. In contrast, DistilBERT exhibits a noticeably clearer diagonal, indicating improved class separation and better contextual understanding. However, some overlap between *half-true* and *mostly-true* persists, suggesting that even transformer-based models struggle with fine-grained truth distinctions.

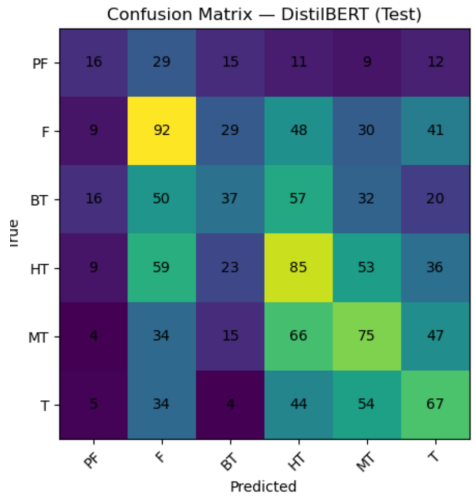

Fig. 4. Confusion matrix for the DistilBERT model.

Table 3 summarizes the macro-F1 and accuracy achieved by each model on the test set. Across traditional baselines, performance remained modest, reflecting the dataset's fine-grained nature and limited text length.

Table 3. Performance comparison of all models on the LIAR test set.

Model	Macro-F1	Accuracy
TF–IDF (word 1–2) + Logistic Regression	0.242	0.243
TF–IDF (char 3–6) + Linear SVC	0.238	0.246
TF–IDF (word + char) + SGD Classifier	0.236	0.248
Metadata-only baseline	0.254	0.252
DistilBERT (fine-tuned)	**0.31**	**0.32**

5.4 Findings and Analysis

The metadata-only model slightly outperformed all traditional text-only baselines, suggesting that contextual features such as *speaker* or *party affiliation* contain strong prior information. This highlights potential biases in the dataset and emphasizes the need for careful model interpretation.

Among the text-based approaches, the Logistic Regression and SVC models achieved comparable results, with character-level features performing slightly better on short or noisy claims. Combining word and character n-grams using the SGD classifier did not yield a significant improvement, indicating overlapping information between the two representations.

Fine-tuning DistilBERT improved both macro-F1 and accuracy substantially, demonstrating the transformer's ability to capture semantic nuances even in short statements. However, despite the improvement, the overall performance remained moderate, showing that distinguishing fine-grained truthfulness levels is inherently challenging. These results align with findings from prior research on the LIAR dataset, which also reported difficulties in separating neighboring truth categories.

5.5 Noteworthy Observations

- The minimal drop in F1 during the last-sentence ablation confirms that our models did not rely on superficial sentence endings.
- The relatively strong metadata-only baseline suggests that dataset biases can inflate model scores if contextual cues are not controlled.
- Transformer-based models outperform linear baselines but still struggle with fine-grained truth distinctions, indicating a need for richer contextual reasoning or external knowledge sources.

6 DISCUSSION

6.1 Answering research questions

RQ1: What approaches are most effective for classifying news articles or short statements as real or fake?
Our results show that transformer-based models outperform traditional linear baselines on the LIAR dataset. The fine-tuned DistilBERT model achieved the highest macro-F1 (0.31) and accuracy (0.32), compared to approximately 0.24–0.25 for TF-IDF-based classifiers. This confirms that pretrained language models capture deeper contextual and semantic relations beyond surface-level lexical patterns. Nonetheless, even the best-performing model exhibited considerable confusion between adjacent truthfulness categories (half-true and mostly-true), indicating that nuanced distinctions remain difficult to learn. Overall, this suggests that contextualized embeddings are more effective for fake-news detection, especially when statements are short and ambiguous.

RQ2: How do traditional machine-learning methods using TF-IDF compare to transformer-based models like BERT in detecting fake news?
Traditional feature-based classifiers such as Logistic Regression, Linear SVC, and SGDClassifier reached similar performance, showing that character- and word-level TF-IDF representations capture complementary lexical cues but struggle to represent meaning. The metadata-only baseline slightly outperformed them (macro-F1 = 0.254), confirming that dataset metadata (e.g., speaker or party affiliation) can introduce strong non-linguistic shortcuts. DistilBERT's improvement demonstrates the advantage of contextual modeling: it learns dependencies across tokens and can generalize beyond memorizing stylistic or lexical patterns. However, even transformers did not eliminate all confusion, highlighting the intrinsic ambiguity of the LIAR labels.

6.2 Limitations

A key limitation of this study is the restricted size and imbalance of the LIAR dataset, which constrains model generalization. The fine-grained six-class setting makes the task inherently noisy, as distinctions such as half-true versus mostly-true depend on subtle, sometimes subjective judgments. Additionally, our models were trained on short claims without external evidence; this isolates linguistic signal but prevents fact-checking against real-world sources.

Another limitation concerns the metadata control: although it was useful to test for potential shortcut learning, it does not fully remove contextual bias present in the dataset's structure. Finally, transformer models were fine-tuned under limited computational resources and token length (128 tokens), which may restrict their ability to encode longer or more complex statements.

6.3 Further Work

Future work can extend this study in several directions. First, evidence-based models could integrate external knowledge graphs or document retrieval to verify claims beyond linguistic cues. Second, data augmentation or transfer learning from larger factuality datasets could help mitigate label imbalance and improve robustness. In addition, exploring explainability methods (e.g., LIME or SHAP) would help interpret which linguistic or contextual patterns drive model predictions, increasing transparency and reliability. Finally, combining textual and metadata inputs through multi-modal architectures could enable a fairer and more holistic model that leverages context without over-relying on non-linguistic shortcuts.

7 CONCLUSION

This study compared traditional feature-based models and modern transformer-based approaches for the task of fine-grained fake-news detection using the LIAR dataset. By evaluating TF-IDF classifiers and a fine-tuned DistilBERT model under identical experimental conditions, we isolated the contribution of linguistic versus contextual information. Our findings show that transformer-based models clearly outperform linear baselines, achieving higher macro-F1 and accuracy while maintaining robustness across all truthfulness classes. However, even DistilBERT struggled with subtle distinctions between neighboring categories, revealing the inherent ambiguity of the dataset and the limits of purely textual modeling. Diagnostic experiments further demonstrated that metadata such as speaker or party affiliation can strongly influence results—underlining the importance of fair evaluation protocols that avoid exploiting contextual shortcuts. In summary, transformer models clearly improve performance and capture more meaning from text compared to traditional methods, but detecting subtle differences in truthfulness remains difficult. Future work could include adding external evidence or larger datasets to make the predictions more reliable.

REPOSITORY

All code, experiments, and trained models are available at:
https://github.com/aleksandarnn/fake-news-detection-group2

7.1 AI Tools

ChatGPT and Grammarly have been used to correct grammar and ensure clarity in this paper.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL* (2019).
[2] Yunde Jin et al. 2022. Towards Fine-Grained Reasoning for Fake News Detection (FinerFact). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[3] Victor Sanh et al. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
[4] Ashish Vaswani et al. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* (2017).
[5] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067
[6] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 422–426.