proof of concept for Fact-Checking Model Development

Leen Ajjan Alhadid

2025

Structured Roadmap for Fact-Checking Model Development

1 Introduction

This roadmap outlines the development of a fact-checking system trained on the LIAR dataset (a benchmark of 12.8K manually labeled short statements with six truthfulness categories) . The system evolved through multiple stages of experimentation, feature selection, and classification refinement to improve accuracy and reliability.

2 Dataset Overview

The dataset used was LIAR dataset, is a benchmark dataset for fact-checking research, containing 12,836 short statements from PolitiFact. Each statement is manually labeled with one of six truthfulness categories (true, mostly-true, half-true, barely-true, false, pants-fire). The dataset includes additional metadata such as:

- **Statement**: The claim being fact-checked.
- **Speaker**: The individual who made the claim.
- **Job Title**: The speaker's position (if applicable).
- **Party Affiliation**: Political alignment of the speaker.
- **Context**: Background information on the claim.
- **Historical Counts**: The speaker's past record, including the number of true, false, and misleading statements.

3 Initial Model Experimentation

Several pre-trained language models were fine-tuned and evaluated for this task:

- **Tasksource/DeBERTa-Base-Long-NL**: Selected for its ability to handle longer text sequences efficiently.
- **DistilBERT/DistilBERT-Base-Uncased**: Chosen for its computational efficiency and fast inference.
- **ProsusAI/FinBERT**: Evaluated due to its strong performance on financial text classification tasks.

Each model was fine-tuned with different learning rates (initially 3e-5, then 2e-5, and finally 1e-5). The best performance was achieved at the lowest learning rate 1e-5). However, none of these models performed as well as RoBERTa, which showed better generalization and accuracy on the dataset.

4 Data Preprocessing and Feature Engineering

To enhance classification, multiple input features were explored during preprocessing:

- **Speaker Party Affiliation**: Considered initially but found irrelevant to the truthfulness of statements.
- **Context:**: Tested as an additional feature but showed little correlation with a statement's veracity.
- **Statement Subject**: Included in early experiments but removed due to low impact on predictive power.
- **Speaker History**: Found to be the most relevant feature, as a speaker's historical credibility tends to influence the likelihood that their new statement is truthful.

The final preprocessing pipeline included the following steps:

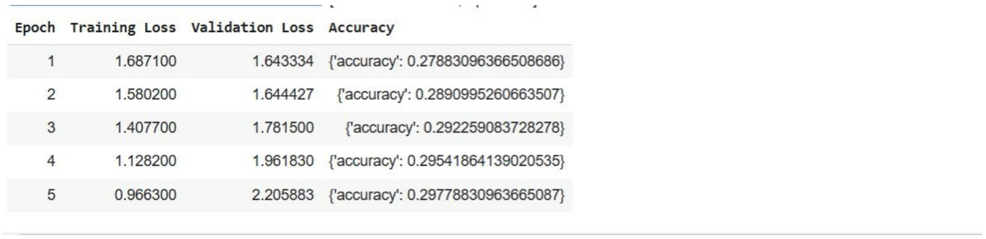
1. **Text Cleaning**: Removed unnecessary characters and standardized the text format for each statement.
2. **Tokenization**: Used RoBERTa's subword tokenizer to convert statements into token sequences suitable for the model.
3. **Label Mapping**: Originally used six distinct labels, but later converted them into binary labels (True/False) to improve classification accuracy.

5 Transition to Binary Classification

Initially, the system categorized statements into six classes (true, mostly-true, half-true, barely-true, false, pants-fire). However, model accuracy remained low due to high variance and ambiguity between similar categories.

5.1 Observed Issues in Multi-Class Classification

- **High Misclassification Rate**: The model struggled to differentiate between adjacent labels (e.g., mostly-true vs half-true), leading to frequent misclassifications.
- **Lower Accuracy (29%)**: Validation accuracy plateaued around 29–30%, comparable to published baselines (27%).
- **Confusion Between Neighboring Classes**: Many statements were placed into the wrong adjacent category, highlighting subjectivity in fine-grained truthfulness ratings.



Epoch	Training Loss	Validation Loss	Accuracy
1	1.687100	1.643334	{'accuracy': 0.27883096366508686}
2	1.580200	1.644427	{'accuracy': 0.2890995260663507}
3	1.407700	1.781500	{'accuracy': 0.292259083728278}
4	1.128200	1.961830	{'accuracy': 0.29541864139020535}
5	0.966300	2.205883	{'accuracy': 0.29778830963665087}

Figure 1: preformance Matrix of the Multi-Class Model

5.2 Implementation of Binary Classification

To improve classification accuracy and reduce label ambiguity, the original six-class label system was consolidated into two broader categories:

- **True (0)**: Includes statements originally labeled as true, mostly-true, and half-true.
- **False (1)**: Includes statements originally labeled as barely-true, false, and pants-fire.

This modification allowed the model to focus on a more distinct separation between truthful and false statements, leading to a significant improvement in accuracy.

5.3 Implementation of Binary Classification

To improve classification accuracy and reduce label ambiguity, the original six-class label system (true, mostly-true, half-true, barely-true, false, pants-fire) was consolidated into two broader categories:

- **True (0)**: This group includes statements that were originally labeled as true, mostly-true, and half-true.
- **False (1)**: This group includes statements that were originally labeled as barely-true, false, and pants-fire.

This modification was necessary because the model struggled to differentiate between similar categories, particularly within the middle ground (mostly-true vs. half-true or barely-true vs. false). By reducing the number of classes, the model was able to focus on a more distinct separation between truthful and false statements, leading to a significant improvement in accuracy. Additionally, speaker history counts were also restructured to provide better contextual features for classification. Instead of treating each historical category separately, past statements made by the speaker were aggregated into two numerical groups:

- **True history counts**: The combined count of mostly-true, half-true, and a portion of barely-true statements.
- **False history counts**: The combined count of false, pants-fire, and a portion of barely-true statements.

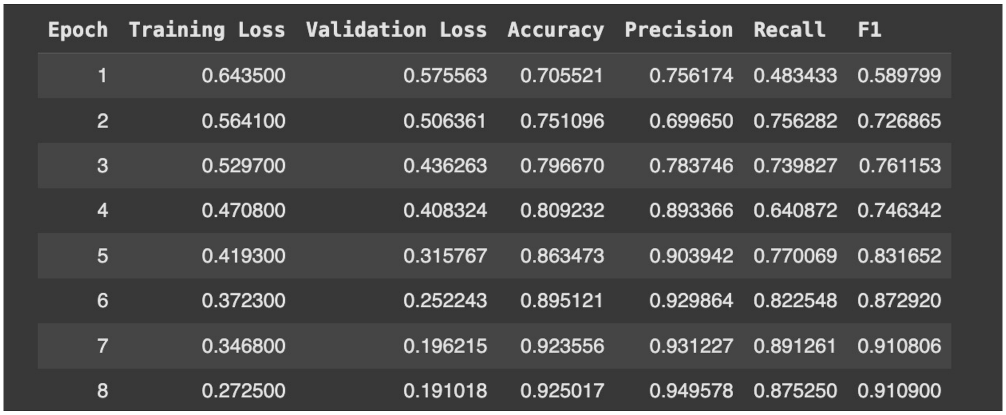
This transformation allowed the model to incorporate a speaker's credibility history as a continuous numerical feature rather than a categorical label. While the classification task itself became binary, the history counts remained a useful supporting feature, helping the model assess whether a speaker had a tendency toward truthful or deceptive statements.

By combining binary classification for statement labels with grouped speaker history counts, the system achieved a balance between simplified decision boundaries and the retention of valuable contextual information. This approach ultimately contributed to the observed improvements in accuracy and F1-score.

5.4 Binary Classification Model Performance

This binary relabeling resulted in several improvements:

- **Higher Accuracy**: Achieved 92.5% accuracy after eight training epochs, a dramatic improvement over the multi-class approach.
- **Better Generalization**: The binary model focused on a clear true/false distinction, avoiding the subjective middle-ground labels and generalizing better to new statements.
- **Lower Validation Loss & Higher F1 (0.91)**: The validation loss decreased steadily, and the F1 score reached 0.91, indicating balanced precision and recall for both classes.



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.643500	0.575563	0.705521	0.756174	0.483433	0.589799
2	0.564100	0.506361	0.751096	0.699650	0.756282	0.726865
3	0.529700	0.436263	0.796670	0.783746	0.739827	0.761153
4	0.470800	0.408324	0.809232	0.893366	0.640872	0.746342
5	0.419300	0.315767	0.863473	0.903942	0.770069	0.831652
6	0.372300	0.252243	0.895121	0.929864	0.822548	0.872920
7	0.346800	0.196215	0.923556	0.931227	0.891261	0.910806
8	0.272500	0.191018	0.925017	0.949578	0.875250	0.910900

Figure 2: preformance matrix and Performance Metrics of the Binary Classification Model

In contrast to the six-class model, the binary classification model's performance was much stronger. This confirms how simplifying the label space leads to a more robust model, as evidenced by the jump in accuracy (to 92%) and the high F1-score (0.91) after adopting binary labels. This approach helped the model learn what is truly true and what is truly false, regardless of how much a statement leans toward truth or falsehood. Instead of struggling with subtle differences between mostly-true and half-true, or barely-true and false, the model focused on correctly distinguishing truth from falsehood. As a result, the predicted labels are closer to the correct category compared to the six-class model, even if the exact multi-class accuracy is not significantly higher.

5.5 Test Set Performance

After training and validation, the model was evaluated on the test set, achieving the following results:

Metric	**Value**
Test Accuracy	70.3%
Test Loss	0.82
Precision	70.4%
Recall	55.5%
F1-score	62.1%

Table 1: Test set performance metrics of the binary classification model

These results indicate that while the model generalizes well beyond training data, further refinements such as confidence-based six-class mapping and class imbalance handling could enhance overall reliability and performance.

By leveraging this performance analysis, we can determine potential areas for further optimization, particularly in handling class imbalances and refining the model's ability to distinguish between ambiguous truthfulness categories.

6 Correlation Between Features, Hyperparameters, and Model Performance

6.1 Features vs. Accuracy

We observed the following about feature inclusion and model accuracy:

- Adding speaker meta-data like party affiliation or job title did not yield any significant improvement in predicting statement veracity.

- Incorporating speaker history (counts of past truthful/untruthful statements) proved helpful, slightly improving the model's precision and recall by providing context on the speaker's credibility.
- Eliminating unnecessary features reduced input noise and complexity, which helped the model focus on the most relevant information and improved overall performance.

6.2 Learning Rate vs. Performance

The choice of learning rate had a clear impact on training stability and outcome:

- **3e-5**: This higher learning rate caused slower convergence and a higher training loss, indicating the steps were too large for the model to learn effectively.
- **2e-5**: Training was more stable at this rate and converged better than 3e-5, but some instability in validation accuracy was still observed.
- **1e-5**: This lower rate provided the best results, striking a good balance between convergence speed and final accuracy. The model learned steadily without overshooting, leading to improved validation metrics.

6.3 Model Structure Impact on Performance

- **Multi-Class (6-way) Structure**: With six outcome classes, the validation accuracy stalled around 29–30%, and improvements were minimal beyond a certain point. This aligns with known difficulty in fine-grained truthfulness classification on LIAR.
- **Binary Classification Structure**: Simplifying the problem to true/false boosted accuracy to 92%, and the model's F1-score became much more stable. The reduced label complexity allowed the model to converge to a better solution and avoid confusion.

7 Future Improvements and Next Steps

While switching to binary classification improved performance, further refinements can be made to address label imbalance and enhance classification granularity.

7.1 Training a Confidence-Based Reclassification Model

While transitioning to binary classification improved performance, efforts were also made to map predictions back to six classes to recover the lost granularity. This was attempted by leveraging the confidence scores from the binary classification model and applying manual thresholding to reassign labels:

- If a statement was classified as true with high confidence ($\geq 80\%$), it was assigned to the true category.
- If classified as true but with moderate confidence (40%-80%), it was mapped to mostly-true or half-true, depending on confidence intervals.
- The same approach was applied for false predictions, splitting them into barely-true, false, and pants-fire based on probability scores.

This manual thresholding helped reintroduce a structured label hierarchy but had limitations—mainly that it relied on static confidence thresholds rather than learning optimal decision boundaries dynamically. A potential improvement would be to train a secondary model that takes binary classification outputs, confidence scores, and speaker history to fine-tune six-class predictions automatically rather than relying on manually defined cutoffs. This would allow the model to refine its certainty in borderline cases and adapt dynamically to variations in statement phrasing and contextual information.

7.2 Addressing Class Imbalance

The dataset has a significant imbalance, particularly in the pants-fire category, which has far fewer samples compared to other labels. This causes the model to favor frequent labels (e.g., half-true, false, mostly-true) and struggle with underrepresented ones. To mitigate this issue, we can:

- **Use Class Weights**: Penalize errors in rare categories more heavily to ensure balanced learning.
- **Oversample Minority Classes**: Increase instances of low-frequency labels (pants-fire) to give the model more exposure.
- **Try Focal Loss**: Reduce the dominance of easy-to-classify examples and focus on harder, underrepresented cases.

7.3 Multi-Stage Classification for Improved Accuracy

The binary model can act as a first-stage filter, determining if a statement is likely true or false. A second-stage model can then reclassify statements into six categories, learning decision boundaries dynamically rather than relying on manually set thresholds. This two-step pipeline would refine predictions and improve model interpretability.

By implementing these improvements, the model would achieve better accuracy, fairness across all classes, and a more structured fact-checking system.