Explainable Fake News Detection with DistilBERT: A Lightweight Framework for Cross-Domain Misinformation Classification

Gaddafi Tukur Usman* and Ankesh Gupta
Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, India
gaddafitusman@gmail.com*, gupta.awin66@gmail.com

Abstract. The majority of automated fake news detection systems are usually evaluated using one in-domain accuracy metric, which may be misleading when the data has artificially exploitable features. In addition to a TF-IDF Logistic Regression baseline, this work fine-tunes DistilBERT for binary fake news classification on a popular Kaggle news corpus with an in-domain accuracy of 98.62% and 100% respectively. These numbers are not considered as a measure of success, but rather as a diagnostic tool on correctly classified articles, using SHAP (SHapley Additive exPlanations) attribution. Only a small percentage of the attribution in a representative real article is due to the semantic content of the article itself (up to 83% can be attributed to the wire-service tag "(Reuters)"). The model is then tested on the structurally different LIAR dataset of short political statements, and it achieves a 43.65% accuracy (MCC = 0.0000), with not a single one of 714 real statements correctly predicted. The results indicate that near-perfect in-domain accuracy can be a consequence of the formatting of the data and not a true measure of the detection ability, and that explainability audits and cross domain evaluation should be considered as a mandatory requirement, and not an optional one, when reporting the performance of fake news detection.

Keywords: Fake News Detection, DistilBERT, Explainable AI, SHAP, Cross-Domain Generalization, Dataset Artifacts, Natural Language Processing

1 Introduction

Social media has not only revolutionized the way news is disseminated, it has revolutionized the way attention is sold. The false news spreads faster than the verified news: Vosoughi et al. [1] found that false news spreads about six times faster on Twitter than verified news, even in the absence of bot activity. The impact is substantial – false health information during COVID has been associated with vaccine hesitancy and riskier treatment decisions [1] and political misinformation has been associated with a reduction in institutional trust [17]. Misinformation is often written like the real news, which makes it hard to find manually at scale, and is a current research area in NLP.

Classical machine learning techniques like Logistic Regression, Naive Bayes, and SVMs with TF-IDF features are 80–90% accurate on benchmark datasets [2] but do not take into account context and rhetorical framing. Later deep learning architectures

© The Author(s) 2026
A. Shastri et al. (eds.), *Proceedings of the Fourth Workshop on Applications of Computational Intelligence and Interactive Systems (WACIIS 2026)*, Advances in Intelligent Systems Research 217,
https://doi.org/10.2991/978-94-6239-832-0_10

(CNNs, LSTMs) achieved 85-90% accuracy on FakeNewsNet [4] with significantly higher computational complexity and without any improvement in interpretability. Transformer models reduced the computational cost, and models like BERT [5] and DistilBERT [8] (a smaller, distilled version of BERT) gave good in-domain accuracy, but neither solved the interpretability issue. One of the lesser talked about issues: a close to perfect benchmark result is not necessarily a true detection capability. If the model is good enough, it will look for any systematic difference in its training set, regardless of whether it's related to truthfulness or not, a formatting detail, a source identifier. The accuracy exhibited is true; what it seems to exhibit is not necessarily true. This is the main result of this study and the reason for the methodology in Section 3.

This study aims at four objectives: (i) create a baseline TF-IDF Logistic Regression model that is interpretable; (ii) fine-tune DistilBERT for binary fake news classification on the same corpus; (iii) provide a diagnostic interpretation of the model, that is, learn what the model learned, not just visualize its behavior; and (iv) test the fine-tuned model without retraining it on the structurally different LIAR dataset to see if any apparent capability generalizes beyond the training distribution.

2 Related Work

2.1 Classical and Deep Learning Approaches

In datasets like FakeNewsNet and LIAR [2, 4] the classical machine learning techniques such as TF-IDF with Logistic Regression, Naive Bayes, SVMs and Random Forests had an accuracy of 80–90%. These methods are simple and easy to interpret, but as a bag-of-tokens approach, they cannot distinguish semantically similar statements that use the same words (e.g., negated claims). This is a structural limitation, not an optimization problem, according to Capuano et al. [3] who found that such features are not able to detect misinformation that mimics authentic writing. Around 2020, neural architectures achieved 85–90% accuracy on FakeNewsNet and BuzzFeed, but with a significant increase in computational costs, without any accompanying interpretability [3] [4].

2.2 Transformers and the Explainability Gap

BERT [5] further improved accuracy, with variants achieving nearly 90% accuracy on fake news test sets [6]; DistilBERT [8] subsequently provided similar accuracy with reduced computational requirements. However, interpretability did not catch up. However, as decisions need to be explained, high performance alone is not sufficient, and Szczepanski et al. [7] demonstrated that meaningful patterns can be uncovered in BERT-based classifiers using post-hoc attribution. Men and Mariano [9] demonstrated that SHAP could be applied to a BERT-based COVID-19 detector, and that this approach could enhance user trust without impacting accuracy, suggesting that trust and accuracy are distinct properties. However, generalization has also been found to be challenging: Ni et al. [10] report that BERT-based models achieve only 61–73%

accuracy on five corpora when trained adversarially, indicating that in-domain accuracy is not necessarily reliable.

2.3 The Dataset Artifact Problem

Most popular fake news benchmarks rely on the fact that real news comes from a few wire services while fake news comes from a few partisan or satirical sites, which means that the two classes are structurally as well as semantically distinct. Bozarth and Budak [13] formalized this criticism, pointing out that when such structure exists, in-corpus accuracy can be higher than in real-world detection, and advocated for more stringent evaluation, which has not yet been widely adopted, with most published studies reporting a single in-domain accuracy figure. This study directly addresses this criticism in Section 4.3.

2.4 Research Gap

In this literature, accuracy has been steadily improving, but with little cross-over: classical models are interpretable, but limited; deep learning and transformer models are more capable, but opaque; lightweight variants are more efficient, but less interpretable; and cross-dataset validation, though well known as important, is not widely performed. To the authors' knowledge, there is no previous research that has employed explainability as a diagnostic tool, tested the hypothesis with a structurally different dataset, and reported the results honestly when they showed that there is a fundamental limitation. This study fills that void.

3 Methodology

An experimental design is designed to address a single concern, which in this case is whether high accuracy on one dataset is a true reflection of the detection capability. There are four components, each addressing this in turn: a classical baseline sets a clear baseline, fine-tuning DistilBERT is the primary experimental model, SHAP attribution asks what the model has actually learned, and cross-dataset validation on LIAR checks whether the model has learned what is generalizable.

3.1 Datasets

The Bhavik Jikadara fake news dataset from Kaggle, which was extracted from the Bisaillon dataset [16] and has been used as a benchmark in previous studies [2, 6], is used as the main training and evaluation corpus. It has 44,898 articles, 23,481 of which are fake and 21,417 are real, from political and general-interest news. One question that has not been studied in detail using this corpus is the distribution of the sources of the authentic articles, as there are more articles from Reuters wire service than from other sources, whereas fabricated articles are from a different pool of sources. Perhaps this structural asymmetry is what models trained on this corpus learn, not semantic content, as is assumed here and tested.

The LIAR dataset [11] was selected for cross-dataset validation because it is structurally different from the main corpus: Wang [11] created it from 10 years of PolitiFact.com fact checks, which are short, conversational political statements

(collapsed here to a binary true/false label, and there is no wire-service formatting at all). The test set here consists of 1267 statements, 714 real and 553 fake. If the primary model has actually learned to detect misinformation, its LIAR performance should be significantly above chance.

3.2 Preprocessing

Both corpora were stripped of HTML tags, whitespace and punctuation were normalized and all text was lowercased. The DistilBERT pipeline was tokenized with the HuggingFace Transformers tokenizer [15] and sequences longer than 512 tokens were truncated, which is the limit of the pipeline in practice, as at least one of the training articles was 772 tokens long. This truncation is not expected to impact the artifact findings in Section 4.3 because the Reuters byline is at the beginning of the article, which is typically within the truncation window. It is expected to have less impact on the model's ability to reason about very long documents. The main data set was divided into 70/20/10 training/validation/testing sets with a fixed random seed. The Logistic Regression test set consisted of 8,980 articles, and DistilBERT was tested on a 1,000 article subset, which was balanced approximately to the class distribution of the corpus.

3.3 Logistic Regression Baseline

TF-IDF features with Logistic Regression were used to create a clear and transparent reference point: the weight of each token can be viewed directly in this non-neural model. This is also a diagnostic: If a shallow bag-of-words classifier is able to perform close to a well-tuned transformer, the closeness itself is a hint that the underlying task might not require as much semantic demand on a model as it should for real fake news detection.

3.4 DistilBERT Fine-Tuning

Google Colab was used to fine-tune DistilBERT for sequence classification using GPU acceleration. The output of the [CLS] token of the final layer of the encoder was passed to a linear classification head, following the approach taken by BERT-based classifiers [5]. The training configuration is summarized in table 1.

Table 1. DistilBERT fine-tuning configuration.

Hyperparameter	Value
Base model	distilbert-base-uncased
Classification head	Linear layer on [CLS] token
Optimizer	Adam
Learning rate	2e-5
Epochs	3 (early stopping on validation loss)
Max sequence length	512 tokens
Train / val / test split	70% / 20% / 10%
Compute environment	Google Colab (GPU)

At the end of the first epoch, the validation loss was already 0.000061, and after the third epoch, it had decreased to 0.000017. This is a very deep and fast convergence, and for a really hard classification problem, it is unusual that the model got a strong, consistent signal so early on. SHAP was used specifically to determine whether the model had learned something semantically meaningful about misinformation, or whether it had learned something incidental to the construction of the corpus (Section 4.3).

3.5 SHAP Explainability

SHAP was selected over other methods like LIME or raw attention visualization, as it is based on cooperative game theory: Shapley values allocate credit to contributing tokens based on their average marginal contribution over numerous feature subsets, not based on a heuristic. It was applied to a sample of correctly classified test articles, selected to represent model behaviors on high-confidence predictions (a corpus-wide attribution analysis was not conducted in this study, but is described in Section 6). In much published work, explainability is added on as a post hoc, demonstrating that a model pays attention to plausible features. It was used here as an audit, not to explain the general workings of the model, but to check whether the outputs are consistent with actual linguistic features of credibility — which is important for the interpretation of the results in Section 4.

3.6 Cross-Dataset Evaluation on LIAR.

The DistilBERT model was fine-tuned and then tested on the LIAR test set without any additional fine-tuning on the LIAR training set, a strict design choice. A model that has learned something truly generalizable about misinformation should be able to transfer at least partially to a new domain, even with some loss of accuracy. A model trained on a corpus of artifacts should, on the other hand, not be able to recognize that artifact on a corpus without it. There are no Reuters-style datelines in LIAR, so this is a direct empirical test of the artifact hypothesis proposed in Section 1.

3.7 Evaluation Metrics

All the classification metrics were used, such as accuracy, precision, recall, F1 score, AUC-ROC and precision-recall AUC. Two additional metrics are used to specifically prevent the artifact scenario being studied. The Matthews Correlation Coefficient (MCC) is a measure that punishes a model that correctly classifies the majority class, which is what a model that always classifies “fake” would do on LIAR. Cohen's Kappa also adjusts for the possibility of chance agreement. Both should be zero if the model has no true predictive signal; this was indeed the case as reported in Section 4.4.

4 Results

Results are presented in four parts: in-domain performance (4.1), confusion matrices and probability analysis (4.2), SHAP diagnostic findings (4.3), and LIAR cross-dataset evaluation (4.4).

4.1 In-Domain Performance

Table 2 summarizes accuracy on the held-out in-domain test sets. Logistic Regression achieved 98.62% accuracy, an F1-score of 0.9869, and AUC-ROC of 0.9982 on 8,980 test articles. DistilBERT achieved 100% across every measured metric on its 1,000-article test set.

Table 2. In-domain performance comparison.

Model	Accuracy	Precision	Recall	F1	AUC-ROC	MCC	κ
LR + TF-IDF	0.9862	0.9892	0.9846	0.9869	0.9982	0.9723	0.9723
DistilBERT	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

At first glance these are excellent results — a 98% baseline and a perfect transformer. Figure 1 depicts this comparison across six key metrics, showing both models at or near ceiling. Section 4.3 addresses the central question: whether these figures represent a model that has genuinely learned to identify misinformation, or something else entirely.

4.2 Confusion Matrices, Curves, and Probability Analysis

Table 3 breaks down the in-domain confusion matrices. Logistic Regression made 124 errors across 8,980 test articles — 51 false positives and 73 false negatives. DistilBERT made none.

Table 3. In-domain confusion matrix values.

Model	TP	TN	FP	FN
LR + TF-IDF	4660	4196	51	73
DistilBERT	524	476	0	0

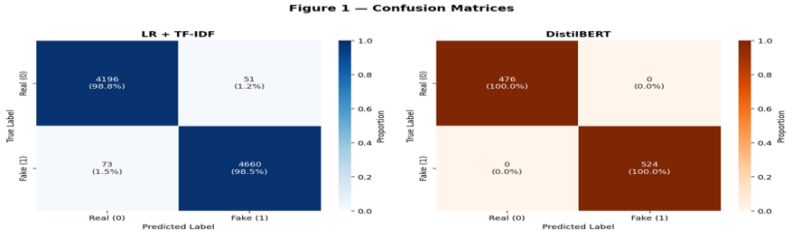

Fig. 1. In-domain confusion matrices for both models.

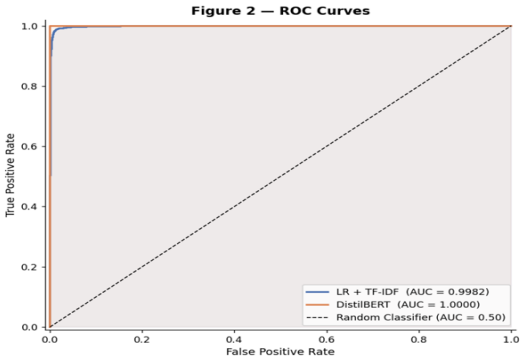

Fig. 2. In-domain ROC curves for both models.

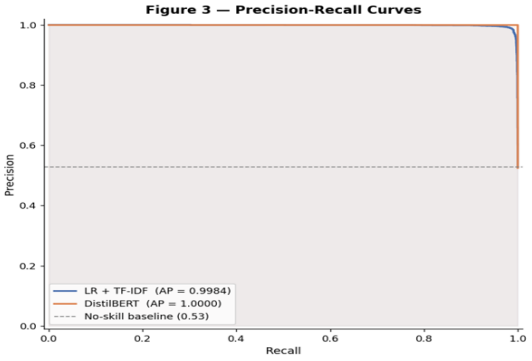

Fig. 3. In-domain precision-recall curves.

The predicted probability distributions (Fig. 5) reveal an asymmetry the aggregate metrics do not show: DistilBERT's real-class predictions cluster near 0.5 rather than 0.0, while its fake-class predictions are highly confident. This points to a strong positive signal tied to the presence of the Reuters tag — driving “real” predictions with high confidence — but only a weak, diffuse signal in its absence, foreshadowing the finding developed in Section 4.3.

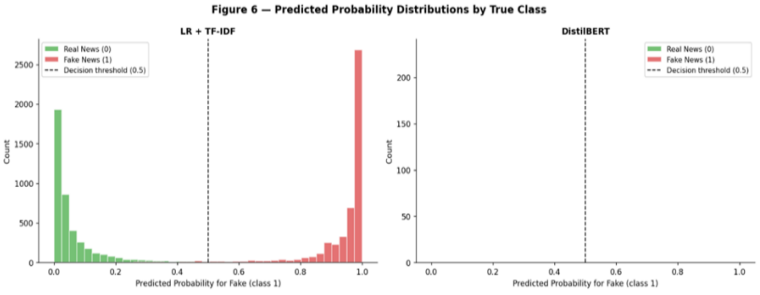

Fig. 4. Predicted probability distributions by true class.

4.3 SHAP Analysis: What the Model Actually Learned

SHAP attribution was applied to correctly classified test articles to see which tokens drove each prediction. Results for a representative real article — a Reuters report on South Korean aid to North Korea, classified with predicted probability 0.99998 — are shown in Fig. 6.

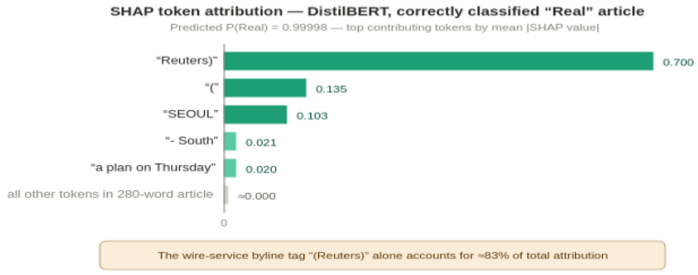

Fig. 5. SHAP token attribution for a correctly classified real article. The wire-service tag "(Reuters)" alone accounts for approximately 83% of total attribution.

The token "Reuters)" has a SHAP value of 0.700, the token "(" has a SHAP value of 0.135, and the token "SEOUL" has a SHAP value of 0.103. These three tokens (parts of the standard Reuters dateline format "CITY (Reuters) –") make up about 93% of the total positive contribution of the model to the real label. The overall attribution of all the substantive factual claims in the article is near zero, however.

There are three lines of evidence that lead to the same conclusion. First, the validation loss dropped to 0.000061 in just one epoch — that's too quick of a convergence for a truly nuanced semantic pattern, but exactly what you would expect if the model had stumbled onto a reliable, simple shortcut. Second, a zero-misclassification confusion matrix is the ideal result that should be expected when a test set has a label that is easily predictable by one formatting cue, and this is an artificial problem, not a difficult one. Third, the known composition of the primary corpus [16] (real articles are overwhelmingly from Reuters, fake articles are from elsewhere) is enough to explain why the "(Reuters)" tag is a near-deterministic proxy for the label, even if the claim is true.

This is also the case for Logistic Regression baseline. The difference between 98.6% and 100% is relatively small for a shallow bag-of-words classifier, and much smaller for a deep contextual model, on a task that is not truly difficult; only a deep contextual model should be able to capture rhetorical framing or source plausibility. That near parity implies that both models are addressing a much simpler problem than the one they were designed to address.

4.4 Cross-Dataset Validation: LIAR

The LIAR evaluation tested a specific prediction: if the primary model's performance is being driven by the Reuters artifact, performance should collapse on LIAR, where that artifact is absent. Results appear in Table 4 and Fig. 7.

Table 4. In-domain vs. LIAR cross-dataset evaluation.

Metric	In-domain (DistilBERT)	LIAR (DistilBERT)	LIAR (LR+TF-IDF)
Accuracy	1.0000	0.4365	0.4388
Precision	1.0000	0.4365	0.4361
Recall	1.0000	1.0000	0.9747
Specificity	1.0000	0.0000	0.0238
F1-Score	1.0000	0.6077	0.6026
AUC-ROC	1.0000	0.5322	0.5576
MCC	1.0000	0.0000	-0.0048
Cohen's κ	1.0000	0.0000	-0.0013

The prediction held completely. DistilBERT scored 43.65% on LIAR — almost exactly the proportion of fake statements in the test set (553/1,267 = 43.6%). The model classified every single real article as fake: zero true negatives, a specificity of 0.0000, an MCC of exactly 0.0000. Operationally, this is indistinguishable from a model that ignores its input entirely and outputs the same label every time.

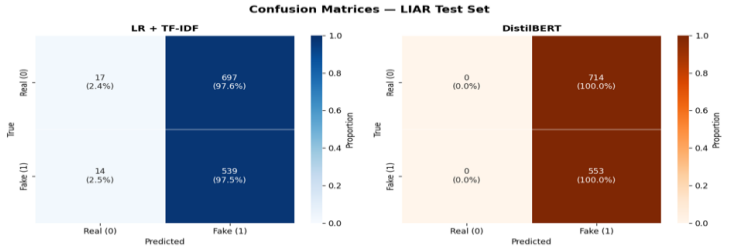

Fig. 6. Confusion matrices for both models on LIAR. DistilBERT's matrix (right) places every real article in the "fake" column.

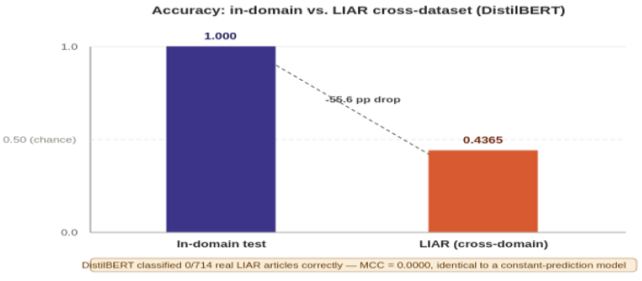

Fig. 7. In-domain vs. LIAR accuracy comparison.

The Logistic Regression baseline did marginally better — 44.0% accuracy, a specificity of 0.0238, correctly identifying 17 of 714 real articles — but was no more useful in practice. The 1.4-point in-domain gap between the two models had effectively disappeared; both were performing at or near chance level on the cross-domain test.

5 Discussion

If you look at Table 2 alone, you might think that both models perform well, with DistilBERT being the superior model. The SHAP analysis and LIAR evaluation demonstrate that this reading does not hold true: neither model has learned anything meaningful about fake news. Both models actually learned to identify which corpus an article was from, recognizing a wire-service dateline format, but not judging the truthfulness of a claim. All accuracy rates in Table 2 should therefore be interpreted as the ability of each model to distinguish between the Reuters-formatted text and the fabrications from other sources, rather than as measures of the ability to detect fabrications.

One obvious question that arises from such a study is whether it is an actual contribution. This type of artifact exploitation has been reported for other NLP sub-fields, such as reading comprehension [13] and machine translation, in addition to fake news detection. There are not as many studies that report the accuracy figure, but the actual artifact. Here, as a contribution to the practice of evaluation, a high accuracy figure is reported and the reasons for its likely informativeness as evidence of the detection capability are explained, followed by direct experimental evidence.

The results indicate that caution should be taken when using a primary dataset from the Kaggle fake/real news family as used here, because in-domain accuracy does not imply that a system is ready for deployment. The specific artifact that is identified, the Reuters dateline, is just one example of a larger class of problem, in which the two classes in a dataset are not distinguishable on the basis of the truth of the content of the data. The practical implication: Use an attribution method on a subsample of correctly classified instances prior to reporting results and test on at least one dataset from a different source distribution.

6 Threats to Validity and Future Work

This section separates the findings of this study from what it suggests as the next steps.

6.1. Demonstrated in this study

The following are experimental results, not predictions: (i) both DistilBERT and Logistic Regression achieve near-ceiling in-domain accuracy on the primary dataset (Section 4.1); (ii) SHAP attribution on a sample of correctly classified articles shows the wire-service tag "(Reuters)" accounts for approximately 83–93% of attribution toward the real-class prediction in the illustrative example examined (Section 4.3); and (iii) the trained DistilBERT model, evaluated with no additional fine-tuning on the structurally distinct LIAR dataset, performs at chance level, with an MCC of exactly 0.0000 (Section 4.4).

6.2 Threats to Validity

The artifact claim is based on a small, illustrative sample of correctly classified articles, and not on the entire test set, which reduces the statistical strength of the claim. The 512-token truncation limit was reached during training, which is not likely to impact the central findings, but is a limitation on reasoning across long documents. This pattern has been evaluated with only two model architectures, which restricts the extent to which this pattern can be generalized to other architectures. Lastly, the structural difference between the primary corpus (short statements) and LIAR (long articles) may also account for some of the observed collapse, as it may be domain mismatch as well as artifact dependence; this study cannot fully distinguish the two.

6.3 Proposed Future Work

Three extensions follow directly from these findings and are proposed as necessary next steps, not completed here. First, and most directly diagnostic: strip wire-service datelines from all training articles, retrain both models, and re-evaluate on both datasets. In-domain accuracy is expected to fall toward the 80–90% range reported elsewhere for less artifact-prone settings [2], [3], while LIAR performance may improve. Second, a corpus-wide SHAP audit — aggregating attribution across the full test set by predicted class and source — would clarify whether the Reuters artifact accounts for most decisions, or whether genuine linguistic signal is also present. Third, evaluating additional baseline architectures (e.g., Naive Bayes, Random Forest, TinyBERT, ALBERT) and additional cross-domain datasets (e.g., WELFake, BuzzFeed, or domain-specific misinformation corpora) would clarify whether this pattern is specific to these two models and this dataset pairing, or reflects a broader generalization failure.

7 Conclusion

The aim of this study was to develop an accurate, lightweight, and explainable fake news detection system. What it shows instead is more important: These three properties, alone, are not enough — and a fourth property, honesty about what the model has actually learned, is just as important.DistilBERT achieved 100% accuracy on the in-domain test set. This was not because this model understood misinformation, but because it had captured a formatting tag that occurred in a wire service that happened to be correlated with the label: the result was obtained using SHAP analysis. The LIAR cross-domain evaluation showed that the model had learned nothing that generalized: it returned an MCC of 0.0000, which is the same result as a model that just blocks out the input.This result is the key finding of this study, and not a warning about an otherwise successful project: if a model is not able to explain its predictions, then it should not be trusted; if a model has not been evaluated in a distribution outside of its training distribution, it has not been tested at all.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[2] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," Multimedia Tools and Applications, vol. 83, no. 17, pp. 51009-51067, 2024.

[3] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, "Content-based fake news detection with machine and deep learning: a systematic review," Neurocomputing, vol. 530, pp. 91-103, 2023.

[4] J. Li et al., "A brief survey for fake news detection via deep learning," Procedia Computer Science, vol. 214, pp. 1339-1344, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[6] N. Raza et al., "Enhancing fake news detection with transformer-based deep learning: a multidisciplinary approach," PLOS ONE, 2025.

[7] M. Szczepanski, M. Pawlicki, R. Kozik, and M. Choras, "New explainability method for BERT-based model in fake news detection," Scientific Reports, vol. 11, p. 23705, 2021.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.

[9] X. Men and V. Y. Mariano, "Explainable fake news detection based on BERT and SHAP applied to COVID-19," Int. J. Modern Education and Computer Science, vol. 16, no. 1, pp. 11-22, 2024.

[10] S. Ni et al., "Boosting generalization of fine-tuning BERT for fake news detection," Information Processing & Management, 2024.

[11] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in Proc. ACL, 2017, pp. 422-426.

[12] K. Shu et al., "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big Data, vol. 8, no. 3, pp. 171-188, 2020.

[13] L. Bozarth and C. Budak, "Toward a better performance evaluation framework for fake news classification," in Proc. Int. AAAI Conf. Web and Social Media, vol. 14, 2020, pp. 60-71.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[15] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP: System Demonstrations, 2020, pp. 38-45.

[16] C. Bisaillon, "Fake and real news dataset," Kaggle, 2020.

[17] Reuters Institute for the Study of Journalism, Digital News Report 2023, Univ. of Oxford, 2023.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

