International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Benchmarking machine learning and deep learning models for cross-domain fake news detection: performance, generalisation, and computational Trade-offs

Rajesh Chauhan[1], Akshay Bhardwaj[2], Rohit Kumar Verma[3]
[1]Department of Computer Science, University Institute of Technology, Himachal Pradesh University, Shimla, India
[2]Department of Computer Science and Engineering, University Institute of Technology, Himachal Pradesh University, Shimla, India

Abstract- The spread of false information on digital platforms has surged and there is a growing demand for the adoption of accurate and deployable automated false information detection systems. But models learned in one news domain can easily suffer significant performance drop when transferred to other domains out of the scope of its training. This study compares four classical machine learning models and five deep learning architectures for within and cross domain fake-news detection. Five publicly available benchmark datasets, which include over 150,000 labelled instances, are used for experiments: LIAR, ISOT Fake News, FakeNewsNet GossipCop, WELFake, and Fake and Real News Dataset. Their models are evaluated based on classification accuracy, F1 score, cross domain performance retention, computational cost, data requirements and interpretability. The best fine-tuned RoBERTa model obtained the highest accuracy score of 97.8% on ISOT and 84.9% on the transfer task from ISOT to GossipCop, outperforming the linear SVM model by 13.7 percentage points. However, classical models are still suitable in resource-limited and interpretability sensitive scenarios, and BiLSTM with additive attention is a balanced model. The results show that model selection should not only evaluate the predictive performance of the model but also take into account the operational constraints.

Keywords- Fake news detection; Cross-domain generalisation; Machine learning; Deep learning; Deployment trade-offs.

I. INTRODUCTION

The rapid expansion of digital news platforms and social-media networks has increased both the accessibility of information and the circulation of misleading or fabricated content. Fake news can influence public opinion, weaken trust in institutions, disrupt public-health communication, and affect political and economic decision-making. Because manual verification cannot process the volume of content generated online, automated fake-news

detection has become an important natural-language-processing and machine-learning task. The CSI framework demonstrated that detection performance can be improved by integrating article content, temporal user-response patterns, and source behaviour [1]. Research from a data-mining perspective has also emphasised the combined importance of news content, social context, and propagation information [2].

Automated fake-news detection is commonly formulated as a supervised classification problem in which a model learns patterns that distinguish false content from legitimate news. Recent advances in language representation have substantially improved text-classification performance. BERT introduced bidirectional contextual representations by considering both the preceding and following context of each token [3]. RoBERTa subsequently refined the BERT pretraining process through modifications to training data, masking strategy, batch size, and optimisation settings [4]. These developments established pretrained transformers as strong models for language-processing tasks. However, fake-news detection remains difficult because misinformation varies across linguistic form, topic, source, and social context [5].

Prior to the rise of deep learning, there was little research on fake-news detection that depended primarily on statistical text representations and classical machine-learning algorithms.

Lexical patterns as shown in [6] were used to distinguish fake news and real content by using N-gram features in combination with conventional classifiers. These methods are efficient in terms of resources and can be trained with relatively small hardware. However, the domain-specific vocabulary that they rely on might be mapped to domain-specific expressions rather than general ones that signal misinformation. Past experimental results reveal that models learned from one dataset can suffer significant performance losses when applied to a different one, meaning the results for an individual data set do not necessarily hold up for the other data set [7].

To enhance representation learning, beyond hand-engineered features, neural architectures were introduced. Self-attention allowed models to capture relationships between tokens, irrespective of their separation in a sequence and laid the groundwork for the modern transformer architecture [8]. Fake-news detection has also been extended with attention mechanisms, which are used to detect influential textual evidence. A claim representation, external evidence, and attention-based credibility assessment were integrated into the DeClarE framework [9]. Attention scores can give some indications of model decisions but cannot be used as a full explanation of model behaviour.

However, classical algorithms are still baselines that are important. A Naive Bayes based classifier showed that probabilistic text classification has the potential to give a useful fake-news detection performance at a low computation cost [10]. However, transfer-learning methods are able to learn more complex contextual information. In the detection of fake news, BERT-based representations have been found to be superior to multiple other baselines, including recurrent and conventional neural models [11]. The results show that it is crucial to consider an important trade-off: classical models are efficient and have relative interpretability, while the transformer models have better semantic representations but take longer to train, use more memory, and cost more in terms of inference time.

One important problem yet to be addressed is cross domain generalisation. Fake news is found throughout the political, entertainment, health, science and finance sectors, and each sector has its own vocabulary, style of writing, sources, and methods of annotation. The work in [12] has demonstrated that for cross domain fake-news detection, the ability to retain domain-specific and domain-independent information can enhance the transfer performance. The same, but with the introduction of emotional features, has been used to minimize the difference between the source and target distributions in the case of adversarial domain adaptation [13]. These methods are based on domain-adversarial learning in which the models are trained to aid classification but are worsened in their ability to distinguish between domains [14].

Recent reviews have validated that the field has shifted from feature-based classifiers to convolutional, recurrent, attention-based and transformer architectures. Content-based detection has been successfully accomplished with deep learning and transformer-based methods, with their performance being comparable in some studies, but difficult to compare among different studies due to the differences in data sets, preprocessing methods, and the evaluation methods [15]. Other challenges that have been noted as prevalent limitations in fake-news and rumour detection studies are those related to dataset dependence, domain variability, and limited generalisability [16]. CNN–LSTM models are still relevant because they can capture local textual patterns and sequential dependencies [17].

Although these developments have been made, there are still gaps in the current literature. This is because numerous studies use a single data set, use only a few models, or use a single performance metric. Compared to other studies, only a few studies compare classical machine-learning models, recurrent and attention-based neural architectures, and pretrained transformers in a similar multi-dataset protocol. Often computational requirements and interpretability are examined separately from predictive performance, although both are clearly related to the feasibility of use in practice.

To overcome these, four classical machine-learning algorithms and five deep-learning architectures are assessed in this study across five benchmark datasets. There are four domain evaluations and two cross domain transfer settings in the experimental design. Evaluation of the model performance is done based on accuracy, average F1-score, and cross-domain retention rate. Other factors include training time, latency of inference, dependency on GPUs, memory use, cost relative to the compute, and interpretability. The aims of the study are:

- To compare the within-domain performance of nine machine-learning and deep-learning models.
- To evaluate model generalisation across different news domains.

- To quantify the performance retained during source-to-target domain transfer.
- To compare the computational and memory requirements of the evaluated models.
- To examine trade-offs among accuracy, efficiency, and interpretability.
- To provide evidence-based guidance for selecting models under different deployment constraints.

II. METHODOLOGY

This paper is a comparative experimental study to assess the efficacy of classical machine-learning models, deep-learning models, and their practical applicability in detecting fake news. The methodology explores classification within the domain and generalisation across the domain, computational cost, efficiency of inference and interpretability. A total of five benchmark datasets and nine classification models were compared in a similar experimental set up.

Research Questions

The following research questions were used to design the experiments:

- **RQ1**: Which models using machine learning and deep learning gives the best classification result when trained and tested in the same news domain?
- **RQ2:** How well do the models generalise to other news domains when moved from political and world to celebrity and entertainment news?
- **RQ3:** What are trade-offs between predictive performance, computation, inference efficiency and interpretability, and how should these factors be considered for different deployment circumstances?

These questions consider predictiveness and operational suitability. A model that works well on one set of data may not perform the same on a different set of data. Likewise, in restricted environments, such as with limited resources of GPUs, memory, time to inference, or interpretability, the most accurate model may not be suitable.

Dataset Selection

In this study, five benchmark datasets were employed which were freely available. The numbers of reports indicate that a combined experimental corpus has been created, which includes 175,511 labelled instances. The datasets differ in terms of the domain, document size, class structure, annotation process, and data source.

Table 1. Summary of the benchmark datasets used in the study

Dataset	Records	Classes	Domain	Average length	Label source
LIAR [18]	12,836	Six labels converted to binary	Political statements	~17 words	PolitiFact verdicts
ISOT Fake News	44,898	Binary	litics and world news	~780 words	Reuters and fake-news websites
FakeNewsNet GossipCop [19]	22,141	Binary	Celebrity and entertainment	~420 words	Editorial credibility ratings
WELFake [20]	72,134	Binary	General web news	~400 words	Four merged sources
Fake and Real News Dataset	23,502	Binary	litics and world events	~850 words	Reuters and web-scraped sources

The LIAR dataset was created by Wang [18] and has 12,836 short political statements labeled according to six truthfulness categories based on PolitiFact verdicts. The labels for the six original classes collapsed into two classes: fake and real, consistent with the binary classification setting of the other data sets.

The ISOT Fake News Dataset was created within the Information Security and Object Technology Research Lab (ISOT) at the University of Victoria. It consists of truthful articles from reputable news outlets and fake articles from websites that have been flagged by fact checking organisations as not to be trusted [21].

FakeNewsNet is a news content database and contextual data that reflects information from fact-checking sites [19]. In the main cross domain evaluation, the GossipCop subset consisting of mainly celebrity and entertainment news was targeted.

The WELFake dataset is a collection of content from four different fake-news sources which results in 72,134 labelled news articles [20]. It is comparatively large and covers a wide range of web-news topics, which makes it applicable for

assessing the performance of the models over general web-news content.

The Fake and Real News Dataset was obtained from the Kaggle repository maintained by Bisaillon [22]. It was used as a binary political and world-news benchmark in both the within-domain evaluation and the secondary cross-domain experiment.

Four datasets—LIAR, ISOT Fake News, WELFake, and the Fake and Real News Dataset—were used for within-domain evaluation. FakeNewsNet GossipCop was reserved as the target dataset in the primary cross-domain experiment. Consequently, five datasets were included in the overall experimental framework, whereas within-domain results were reported for four datasets.

Text Preprocessing

Text preprocessing was adapted to the requirements of classical and deep-learning models. Initial normalisation included the removal or standardisation of HTML tags, URLs, usernames, special characters, and other non-textual elements. Tokenisation was performed using the spaCy English-language model.

For classical machine-learning models, stop words were removed and Porter stemming was applied before feature extraction. The processed documents were represented using TF-IDF-weighted unigram and bigram features. The vocabulary was restricted to a maximum of 50,000 features.

In models that relied on deep learning, stop words were kept in the text since the contextual embedding and pretrained language models require full extent of the text. The processed tokens were fed into the proper embedding or transformer input layers.

To overcome the class imbalance issue, the LIAR dataset has been converted into a binary version and then SMOTE has been applied. For the remainder of the data sets, stratified data splitting was applied to ensure the class distributions of the data sets were maintained across the data set partitions.

Experimental and Model Design

The experimental setup aimed at evaluating and comparing the classical machine-learning and deep-learning methods in the same fake-news classification pipeline. The models were categorized into two groups after the preparation of the data set and preprocessing of the texts. The first group contained Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Machine and Random Forest. The models were based on the TF-IDF weighed unigram and bigram representations, limited to the top 50,000 features.

The second group consisted of CNN, BiLSTM, BiLSTM with additive attention, BERT-base, and RoBERTa. Learned or pretrained word representations were used in the CNN and recurrent models, whereas BERT-base and RoBERTa employed contextual transformer-based representations. Both architectures gave binary predictions of fake or true news items.

All nine models were evaluated using the same general evaluation framework. The data used in domain experiments were drawn from the same data set in both the training and test sets. To evaluate the model transfer from a source dataset to a target dataset in a different news domain, cross-domain experiments are conducted. The outputs of the model were then compared in terms of accuracy, average F1 score, performance retention rate, training time, inference latency, memory usage, model GPU dependency, relative computational cost and interpretability.

This design allowed model evaluation and benchmarking of predictive performance and operational suitability without altering the configurations of models that were stated. The whole experimental procedure is shown in Fig. 1.

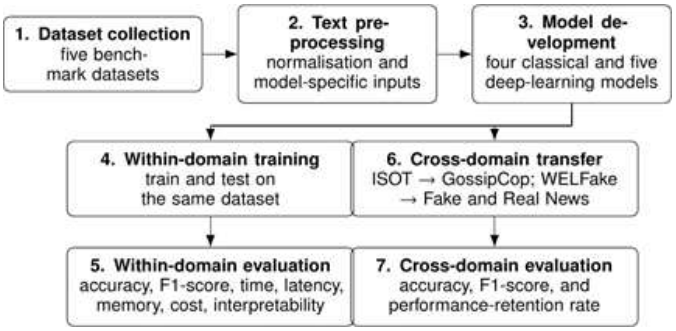

Fig. 1. Overall experimental workflow for within-domain and cross-domain fake-news detection.

Classical Machine-Learning Models

Four classical classifiers were implemented using scikit-learn version 1.3.2:

- Logistic Regression: L2 regularisation with C = 1.0.
- Multinomial Naive Bayes: Laplace smoothing was applied.

- Linear Support Vector Machine: A linear kernel with C = 1.0.
- Random Forest: Included as a nonlinear ensemble-learning baseline.

All four classical models used the same TF-IDF feature space to support a consistent comparison.

Deep-Learning Architectures

Five deep-learning architectures were evaluated:

- Convolutional Neural Network: Implemented using 256 filters with filter sizes of 2, 3, and 4 over 300-dimensional GloVe embeddings.
- Bidirectional LSTM: A two-layer BiLSTM containing 128 units in each direction.
- BiLSTM with Additive Attention: The BiLSTM representation was combined with an additive attention mechanism that assigned importance scores to individual tokens.
- BERT-base: The approximately 110-million-parameter model was fine-tuned for three epochs using AdamW with a learning rate of 2×10^{-5}.
- RoBERTa: Fine-tuned using the same principal training configuration as BERT-base.

The term additive attention is used consistently because it reflects the attention mechanism implemented in the BiLSTM architecture.

Experimental Protocol

The study used within-domain and cross-domain evaluation settings.

In the within-domain setting, each model was trained and tested using samples from the same dataset. Results were reported for LIAR, ISOT, WELFake, and the Fake and Real News Dataset.

The primary cross-domain experiment used ISOT as the source dataset and FakeNewsNet GossipCop as the target dataset:

$$\text{ISOT} \rightarrow \text{GossipCop} \quad (1)$$

This experiment evaluates transfer from political and world news to celebrity and entertainment content.

The secondary cross-domain experiment used WELFake as the source dataset and the Fake and Real News Dataset as the target:

$$\text{WELFake} \rightarrow \text{Fake and Real News} \quad (2)$$

The secondary experiment was included to examine whether the cross-domain performance pattern observed in the primary transfer remained consistent across another dataset pair.

Evaluation Criteria

Model performance was assessed using classification accuracy and average F1-score. Cross-domain robustness was additionally measured using the performance-retention rate:

$$\text{Retention Rate} = \text{Cross-domain Accuracy} / \text{Within-domain Accuracy} \times 100 \quad (3)$$

Operational suitability was assessed based on how long it takes to train the model, the number of GPUs used, the latency of inference per sample, the amount of RAM or VRAM required, the relative computational cost, and the interpretability of the model in a qualitative sense. These criteria allowed the comparison of predictive performance, as well as the suitability of the model under various resource, latency and accountability constraints.

III. RESULTS

We report the performance of the nine models evaluated on the within domain classification, as well as the cross-domain generalisation results and computational resources. These results are only discussed in the Discussion section.

Within-Domain Classification Performance

Table 2 presents the classification accuracy of all nine models on the four within-domain test datasets: ISOT, LIAR, WELFake, and the Fake and Real News Dataset. FakeNewsNet GossipCop was reserved for the primary cross-domain evaluation and is therefore not included in the within-domain results. The table also reports the average F1-score of each model.

Table 2. Within-domain classification accuracy and average F1-score across four benchmark datasets

Model	ISOT	LIAR	WELFake	Fake & Real	Avg. F1
Logistic Regression	91.2%	64.7%	88.1%	91.6%	0.836
Naive Bayes	89.4%	62.3%	85.9%	89.8%	0.818
Linear SVM	93.4%	67.8%	90.3%	93.1%	0.862
Random Forest	92.7%	66.1%	89.7%	92.4%	0.851

CNN	93.8%	69.4%	91.2%	93.6%	0.874
BiLSTM	95.3%	71.2%	93.1%	95.0%	0.898
BiLSTM + Additive Attention	96.1%	73.8%	94.4%	96.2%	0.912
BERT-base	97.1%	76.3%	96.2%	97.3%	0.934
RoBERTa	97.8%	77.9%	96.9%	97.9%	0.941

RoBERTa achieved the highest accuracy on all four within-domain datasets and obtained the highest average F1-score of 0.941. BERT-base produced the second-highest results, with an average F1-score of 0.934. Among the classical machine-learning models, Linear SVM achieved the highest average F1-score of 0.862.

The LIAR dataset produced lower classification accuracy for all models than the other datasets. Accuracy on LIAR ranged from 62.3% for Naive Bayes to 77.9% for RoBERTa. On the remaining datasets, the transformer models achieved accuracy above 96%.

Training Time and Interpretability

Table 3 reports the approximate training time and qualitative interpretability rating of each evaluated model.

Table 3. Training time and interpretability of the evaluated models

Model	Training time	Interpretability
Logistic Regression	<1 min	High
Naive Bayes	<1 min	High
Linear SVM	~8 min	Moderate
Random Forest	~5 min	Moderate
CNN	~8 min	Low
BiLSTM	~25 min	Low
BiLSTM + Additive Attention	~35 min	Moderate
BERT-base	~2 hrs	Low
RoBERTa	~3 hrs	Low

Logistic Regression and Naive Bayes required less than one minute of training and were rated as highly interpretable. Linear SVM and Random Forest required moderate training time and were classified as moderately interpretable. Among the neural models, BiLSTM with additive attention provided moderate interpretability, whereas CNN, BiLSTM, BERT-base, and RoBERTa were classified as having low interpretability.

Cross-Domain Generalisation Performance

Tables 4 and 5 present the cross-domain accuracy and performance-retention rates for the primary and secondary source-to-target transfer settings, respectively. In the primary experiment, the models were trained on ISOT and evaluated on FakeNewsNet GossipCop. In the secondary experiment, the models were trained on WELFake and evaluated on the Fake and Real News Dataset.

Table 4. Cross-domain performance for the ISOT-to-GossipCop transfer

Model	Accuracy	Retention rate
Logistic Regression	67.4%	73.9%
Naive Bayes	65.1%	72.8%
Linear SVM	71.2%	76.3%
Random Forest	68.4%	73.8%
CNN	75.1%	80.1%
BiLSTM	78.4%	82.3%
BiLSTM + Additive Attention	81.2%	84.5%
BERT-base	83.7%	86.2%
RoBERTa	84.9%	86.8%

Table 5. Cross-domain performance for the WELFake-to-Fake-and-Real transfer

Model	Accuracy	Retention rate
Logistic Regression	69.2%	75.3%
Naive Bayes	67.0%	74.6%
Linear SVM	72.8%	78.2%
Random Forest	70.1%	75.9%

Model	Accuracy	Retention rate
CNN	76.4%	81.6%
BiLSTM	79.9%	85.9%
BiLSTM + Additive Attention	82.7%	86.0%
BERT-base	85.1%	87.4%
RoBERTa	86.3%	88.2%

RoBERTa achieved the highest cross-domain accuracy and retention rate in both transfer settings. It obtained 84.9% accuracy with an 86.8% retention rate in the ISOT-to-GossipCop experiment and 86.3% accuracy with an 88.2% retention rate in the WELFake-to-Fake-and-Real experiment. BERT-base ranked second, followed by BiLSTM with additive attention. Linear SVM was the strongest classical model, achieving 71.2% and 72.8% accuracy in the two experiments.

Computational Requirements

Table 6 summarises the approximate computational requirements of the nine evaluated models. The reported measures include training time, GPU dependency, inference latency per sample, memory consumption, and relative computational cost. Relative cost was normalised using Logistic Regression as the baseline.

Table 6. Computational requirements and relative cost of the evaluated models

Model	Training time	GPU required	Latency (ms/sample)	Memory	Relative cost
Logistic Regression	<1 min	No	0.1	~200 MB	1×
Naive Bayes	<1 min	No	0.1	~100 MB	1×
Linear SVM	~8 min	No	0.3	~800 MB	8×
Random Forest	~5 min	No	2.0	~1.2 GB	5×
CNN	~8 min	Yes	1.2	~2 GB	30×
BiLSTM	~25 min	Yes	4.0	~3 GB	90×
BiLSTM + Additive Attention	~35 min	Yes	5.5	~4 GB	130×
BERT-base	~2 h	Yes	38	~10 GB	700×
RoBERTa	~3 h	Yes	42	~12 GB	1,000×

Logistic Regression and Naive Bayes exhibited the lowest computational requirements, with training times below one minute, inference latency of 0.1 ms per sample, and no GPU dependency. Linear SVM also remained computationally efficient, although its relative cost was eight times the Logistic Regression baseline.

Among the deep-learning models, CNN required the shortest training time and the lowest memory capacity. BiLSTM with additive attention fell somewhere in the middle, to be trained for about 35 minutes, using 4 GB of memory and taking 5.5 ms to infer per sample. The model with the highest compute needs were BERT-base and RoBERTa. RoBERTa took about three hours for training, 12 GB of memory, and 42 ms of time per sample, or a 1000-fold speedup over Logistic Regression.

IV. DISCUSSION

The results have shown a consistent performance hierarchy for both within domain and cross domain evaluation settings. RoBERTa outperformed all its peers on all the four within-domain datasets, achieving 97.8% accuracy on ISOT and 97.9% accuracy on the Fake and Real News Dataset, and the highest average F1-Score of 0.941. BERT-base was the second-best model, and the best classical machine-learning model was the Linear SVM. Short political statements and the reclassification of LIAR's original six level truthfulness classification scheme into a binary classification make it a more challenging classification problem than a full-length news article, as all models struggled with this classification.

Cross-domain evaluation reduced the performance of every model, confirming that strong within-domain accuracy does not guarantee reliable generalisation. In the ISOT-to-GossipCop transfer, RoBERTa achieved 84.9% accuracy, compared with 71.2% for Linear SVM, a difference of 13.7 percentage points. In the same transfer, WELFake-to-Fake-and-Real, RoBERTa can attain 86.3% accuracy. The results suggest that the proportion of performance that will be preserved by pretrained transformers will be larger when there is a domain shift. BiLSTM with additive attention, also achieved competitive accuracy rates in the two transfer settings (81.2% and 82.7%) with significantly lower computational resources compared to BERT and RoBERTa.

The results are consistent with the previous studies. Neural architecture is demonstrated to have the capacity to include richer textual and contextual information than traditional feature-based models and this is evidenced by the superior results of the BiLSTM-based models [1]. These cross-dataset evaluations previously showed significant degradation in performance when classifiers were transferred to a different, unseen domain, thus corroborating the decrease seen in the cross-domain evaluations [7]. The better performance of BERT-base and RoBERTa is also in line with the contextual representation power of bidirectional transformers [3] and the enhanced pretraining approach of RoBERTa [4].

The lesson to be learned is that the selection of models should not rely solely on the predictive value. RoBERTa is the best choice if maximum cross-domain performance is desired and enough GPU memory, or inference resources are available. Its reported relative cost, however, is ~1000 times Logistic Regression. For applications that require interpretability, CPU-only environments, and rapid prototyping, classical models continue to be a good choice. An intermediate solution is BiLSTM with additive attention, which is a compromise between the two extremes of cross-domain accuracy, computational efficiency and partial interpretability.

There are several limitations in the study. It uses only text data and ignores image data, propagation and user-level information. Only two pairs of datasets are available for cross domain evaluation, and all datasets are mostly English language. The statistical significance test, repeated experimental runs, adversarial paraphrasing and temporal concept drift were not considered. The results may also be affected by differences in the annotation procedures and artefacts inherent in the dataset.

Finally, future research directions include domain-adaptation and few-shot learning techniques, multimodal neural architectures using textual and visual data, distilled transformer architectures that can be made more efficient in terms of latency and memory usage, and continual-learning architectures for changing misinformation trends. In addition, there is a need for multilingual and multi-domain benchmarks for wider evaluation in linguistic and cultural contexts.

V. CONCLUSION

In this study, four classical machine learning models and five deep learning architectures are compared on five benchmark datasets for within or across domain fake-news detection. The findings indicate that cross-domain robustness, computation expense, inference needs, and interpretability are equally important factors to consider when selecting models besides classification accuracy. RoBERTa realized the highest accuracy of 97.8% in ISOT and 84.9% in the main ISOT-to-GossipCop cross-domain experiment. It beat the best classical model, Linear SVM, by 13.7 percentage point in this transfer setting.

The secondary WELfake-to-Fake-and-Real experiment also resulted in a similar ranking, verifying the better generalisation performance of transformer-based models. But none of the models is best in all deployments. Logistic Regression, Naive Bayes, and linear SVM have significantly lower training time, memory and computation requirements, and are well-suited for CPU-based, fast prototyping and interpretability-focused applications. In conclusion, BiLSTM with additive attention strikes a balance, achieving competitive cross-domain effectiveness while using fewer resources compared to BERT and RoBERTa.

The results overall show the limitations of relying on accuracy within a domain when assessing fake-news detection systems. When deploying a model, it is important to optimise between predictive performance and the operational constraints. Automated detection should be used alongside, and not be a replacement for, professional fact checking, editorial judgment, platform control and digital-literacy efforts.

REFERENCES

1. N. Ruchansky, S. Seo, Y. Liu, CSI: a hybrid deep model for fake news detection, in Proceedings of the 2017 ACM Conference on Information and Knowledge Management, 797–806 (2017). https://doi.org/10.1145/3132847.3132877
2. K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: a data mining perspective, SIGKDD Explor. Newsl. 19(1), 22–36 (2017). https://doi.org/10.1145/3137597.3137600
3. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in Proceedings of NAACL-HLT 2019, 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
4. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., RoBERTa: a robustly optimized BERT pretraining approach, arXiv:1907.11692 (2019). https://doi.org/10.48550/arXiv.1907.11692
5. X. Zhou, R. Zafarani, A survey of fake news: fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53(5), Article 109 (2020). https://doi.org/10.1145/3395046
6. H. Ahmed, I. Traore, S. Saad, Detection of online fake news using N-gram analysis and machine learning techniques, in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Lecture Notes in Computer Science 10618, 127–138 (Springer, Cham, 2017). https://doi.org/10.1007/978-3-319-69155-8_9
7. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in Proceedings of the 27th International Conference on Computational Linguistics, 3391–3401 (2018).
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30, 5998–6008 (2017).
9. K. Popat, S. Mukherjee, A. Yates, G. Weikum, DeClarE: debunking fake news and false claims using evidence-aware deep learning, in Proceedings of EMNLP 2018, 22–32 (2018). https://doi.org/10.18653/v1/D18-1003
10. M. Granik, V. Mesyura, Fake news detection using Naive Bayes classifier, in Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering, 900–903 (2017). https://doi.org/10.1109/UKRCON.2017.8100379
11. S. Kula, M. Choraś, R. Kozik, Application of the BERT-based architecture in fake news detection, in Advances in Intelligent Systems and Computing, 1267, 239–249 (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-57805-3_23
12. A. Silva, L. Luo, S. Karunasekera, C. Leckie, Embracing domain differences in fake news: cross-domain fake news detection using multimodal data, Proc. AAAI Conf. Artif. Intell. 35(1), 557–565 (2021). https://doi.org/10.1609/aaai.v35i1.16134
13. A. Choudhry, I. Khatri, A. Chakraborty, D.K. Vishwakarma, M. Prasad, Emotion-guided cross-domain fake news detection using adversarial domain adaptation, in Proceedings of the 19th International Conference on Natural Language Processing, 75–79 (2022).
14. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, et al., Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17(59), 1–35 (2016).
15. N. Capuano, G. Fenza, V. Loia, F.D. Nota, Content-based fake news detection with machine and deep learning: a systematic review, Neurocomputing 530, 91–103 (2023). https://doi.org/10.1016/j.neucom.2023.02.005
16. A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Inf. Sci. 497, 38–55 (2019). https://doi.org/10.1016/j.ins.2019.05.035
17. M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G.S. Choi, B.-W. On, Fake news stance detection using deep learning architecture (CNN–LSTM), IEEE Access 8, 156695–156706 (2020). https://doi.org/10.1109/ACCESS.2020.3019735
18. W.Y. Wang, “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, Vol. 2, 422–426 (2017). https://doi.org/10.18653/v1/P17-2067
19. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8(3), 171–188 (2020). https://doi.org/10.1089/big.2020.0062
20. P.K. Verma, P. Agrawal, I. Amorim, R. Prodan, WELFake: word embedding over linguistic features for

fake news detection, IEEE Trans. Comput. Soc. Syst. 8(4), 881–893 (2021). https://doi.org/10.1109/TCSS.2021.3068519

21. ISOT Research Lab, ISOT Fake News Dataset, University of Victoria (2017), accessed 10 June 2026. https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/
22. C. Bisaillon, Fake and Real News Dataset, Kaggle (accessed 10 June 2026). https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset