High Accuracy, Low Generalization: Structural Homogeneity and Cross-Dataset Evaluation in Fake-News Benchmarks

Hiram Calvo, Mayte H. Laureano
Center for Computing Research
Instituto Politécnico Nacional
hcalvo@cic.ipn.mx, mhernandezl2021@cic.ipn.mx

Abstract

State-of-the-art fake-news classifiers frequently report near-ceiling accuracy on widely used benchmarks such as ISOT, Misinfo, and WELFake. We argue that such results often reflect structural homogeneity and provenance-based separability rather than robust claim-level veracity inference. Anchored in the Information Disorder framework, we analyze how dataset construction operationalizes the notion of "fake" and how this shapes model behavior. We conduct systematic bidirectional cross-dataset experiments across six transfer directions and evaluate performance not only by mean accuracy, but also by variance and directional asymmetry. Results reveal substantial degradation under distribution shift and pronounced transfer asymmetries between dataset pairs. Although not always achieving the highest mean accuracy, affective augmentation combining dimensional (VAD) and categorical (Ekman) representations yields the lowest variance and smallest directional gap, indicating superior cross-domain stability. Our findings expose the disconnect between accuracy-driven benchmarking and construct-valid evaluation. We argue that progress in fake-news detection requires shifting from isolated in-domain optimization toward robustness-oriented, bidirectional, and distribution-aware assessment practices.

Keywords: fake news detection, information disorder, dataset bias, cross-dataset evaluation, distribution shift

1. Introduction

Fake-news classification has become a canonical supervised NLP task, frequently reporting accuracy values exceeding 98%—and in some cases approaching or surpassing 99%—on benchmark datasets. Systematic reviews confirm the prevalence of such near-ceiling results across a wide range of machine-learning architectures (Villela et al., 2023). At face value, these figures may suggest that automated veracity detection is largely solved.

At the same time, the notion of "fake news" is not a well-defined natural category. Prior work has emphasized its overlap with broader forms of information disorder (Lazer et al., 2018), the role of deliberate deception and imitation of journalistic style (Gelfert, 2018), and the diversity of phenomena grouped under the term, including satire, propaganda, and fabrication (Tandoc et al., 2018). These perspectives highlight that the target label in fake-news detection is conceptually heterogeneous and socially constructed.

In practice, however, machine-learning research depends heavily on publicly available supervised datasets. Among these, the ISOT dataset (Ahmed et al., 2017) has become particularly influential. The original ISOT study reported approximately 92% accuracy using n-gram features and classical classifiers, while more recent work using deep learning and transformer-based architectures reports near-ceiling performance, often exceeding 98% (Lawal and Abdulrauf, 2025; Villela et al., 2023). Such results raise a methodological question: what exactly is being learned? In ISOT, true articles are largely crawled from Reuters, whereas fake articles are collected from outlets flagged as unreliable. This provenance-based construction introduces systematic lexical, stylistic, and formatting differences between classes, allowing models to exploit ecosystem-level regularities rather than perform claim-level epistemic inference.

Three corpora—ISOT, Misinfo, and WELFake (Ahmed et al., 2017; Peutz, 2023; Verma et al., 2021)—are widely used due to their size and binary labeling schemes. However, each encodes different assumptions about the negative class and is constructed through distinct pipelines, potentially introducing spurious correlations between labels, sources, and stylistic patterns.

When class labels correlate strongly with source identity or stylistic conventions, classification may approximate corpus discrimination rather than epistemic reasoning. This raises a central validity question: what do near-ceiling benchmark results actually measure?

We argue that high intra-corpus accuracy often reflects structural separability and dataset homogeneity rather than transferable veracity reasoning. To examine this, we benchmark representative modeling setups on ISOT, Misinfo, and WELFake and complement them with controlled cross-dataset transfer experiments. The results reveal a substantial generalization gap: models that appear nearly perfect under random within-dataset splits degrade markedly when evaluated

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 25–33
12 May 2026.

across corpora. By situating our findings within the Information Disorder framework (Wardle and Derakhshan, 2017), we shift the focus from raw accuracy to construct validity and robustness under distributional shift.

2. Related Work

2.1. Fake-News Detection Models

Early work on fake-news detection relied on lexical features such as n-grams, TF–IDF representations, and classical classifiers including SVM, logistic regression, and decision trees (Ahmed et al., 2017). On datasets such as ISOT, these approaches already achieved accuracy above 90%, suggesting strong lexical separability between classes.

With the rise of deep learning, convolutional and recurrent architectures (CNN, LSTM, BiLSTM) became common (Kaliyar et al., 2020). Subsequent transformer-based approaches, particularly BERT and its variants, further increased reported performance (Kaliyar et al., 2021). Systematic reviews confirm that transformer-based architectures now dominate the field and frequently report accuracy exceeding 95%, with some studies approaching or surpassing 99% on benchmark corpora (Villela et al., 2023).

Hybrid models combining contextual embeddings with gradient boosting or ensemble methods also report near-ceiling performance. For example, BERT embeddings combined with LightGBM achieve 99% accuracy on ISOT under random splits (Essa et al., 2023). Similarly, recent BiLSTM-based approaches on ISOT report 98.98% accuracy and recall above 99% (Lawal and Abdulrauf, 2025). Such figures suggest that, at least under standard evaluation protocols, separability between classes is extremely high.

Beyond English-language corpora, multilingual and cross-lingual fake-news detection has also expanded. Large language models (LLMs) and instruction-tuned transformers have recently been explored for misinformation detection, often achieving strong zero-shot or few-shot results (Li et al., 2024; Koka et al., 2024). However, these studies typically evaluate within-dataset performance and rarely stress-test cross-domain transfer.

Our work differs from prior studies by explicitly quantifying directional asymmetry and demonstrating that near-ceiling performance can be reproduced by minimal lexical rules.

2.2. Dataset Bias, Distribution Shift, and Shortcut Learning

Parallel to advances in model architecture, research in NLP and machine learning has highlighted the risk of shortcut learning, whereby models exploit highly predictive but semantically shallow correlations present in the training data (Geirhos et al., 2020). In text classification, such shortcuts may include stylistic markers, formatting artifacts, or source-specific lexical cues.

In the context of fake-news detection, provenance-based labeling is particularly susceptible to shortcut exploitation. If true and fake classes are drawn from distinct outlet ecosystems, models may learn outlet identification rather than veracity inference. Surveys of misinformation detection acknowledge the presence of dataset bias and limited generalization, though systematic stress-testing remains rare (Islam et al., 2020; Villela et al., 2023).

Recent work has explicitly investigated cross-dataset generalization. Pszona et al. (2023) show that models trained on one fake-news corpus often degrade significantly when evaluated on another, suggesting that dataset-specific artifacts drive performance. Similarly, studies on domain adaptation for misinformation detection report substantial drops under distribution shift (Shu et al., 2022; Huang et al., 2021). These findings align with broader observations in NLP that benchmark performance does not necessarily imply robustness under temporal, topical, or source shift.

Despite these concerns, many papers reporting near-ceiling accuracy rely on random within-dataset splits without source-held-out or cross-dataset evaluation. As a result, systematic evidence about transferability remains limited relative to the volume of high-accuracy claims. Our work aims to address this gap by directly quantifying cross-dataset degradation across ISOT, Misinfo, and WELFake.

3. Dataset Semantics and Structural Homogeneity

3.1. ISOT

The ISOT Fake News Dataset (Ahmed et al., 2017) consists of two files: *True.csv* (21,417 articles) and *Fake.csv* (23,481 articles), primarily covering 2016–2017 political and world news. True articles are largely sourced from Reuters, while fake articles are collected from outlets flagged as unreliable.

This provenance-based construction introduces strong structural regularities. The true class reflects wire-service style and formatting, whereas the fake class aggregates politically charged content from heterogeneous sources. As a result, classification can rely on stylistic and source cues rather than claim-level reasoning.

3.2. Misinfo (79k)

The Misinfo dataset contains 78,617 articles (34,975 true; 43,642 fake/misinformation/propaganda). True articles originate from mainstream outlets such as Reuters, The New York Times, and The Washington Post, while the negative class aggregates diverse sources, including extremist websites and curated disinformation cases.

Although this mixture increases semantic heterogeneity, structural clustering remains: different source ecosystems exhibit consistent stylistic and rhetorical patterns. Models may therefore still exploit these cues instead of factual inconsistencies.

3.3. WELFake

WELFake contains 72,134 articles (35,028 real; 37,106 fake), constructed by merging multiple datasets (Kaggle, McIntire, Reuters, BuzzFeed) to increase diversity (Verma et al., 2021).

While this reduces reliance on a single source, binary provenance signals persist across merged corpora. Moreover, substantial textual overlap with ISOT is observed due to shared sources. This challenges dataset independence and motivates explicit deduplication in cross-dataset evaluation.

3.4. Structural Implications

Across ISOT, Misinfo, and WELFake, class labels correlate strongly with source provenance and topic distributions. Such correlations create conditions conducive to shortcut learning (Geirhos et al., 2020), where models exploit stable stylistic or ecosystem-level cues. High within-dataset accuracy under random splits may therefore reflect structural homogeneity rather than transferable epistemic reasoning.

4. Intra-corpus classification and shortcut analysis

We first evaluated intra-corpus classification on the ISOT dataset under a standard train/validation split (70/15/15, stratified by label). Models were trained and validated exclusively on ISOT, without cross-domain mixing. Performance was measured using accuracy, macro F1, and weighted F1.

We focus on ISOT as a maximal case of structural separability, where class labels are strongly aligned with source-specific lexical and stylistic cues. This makes it particularly suitable for analyzing the extent to which near-ceiling performance can be explained by shallow signals rather than semantic understanding.

4.1. Transformer-based models

Across configurations, transformer models achieved near-perfect validation performance:

- Accuracy: 0.99+
- Macro F1: 0.99+
- Weighted F1: 0.99+

These results are consistent with prior literature reporting $> 98\%$ accuracy on ISOT. However, such performance alone does not establish robust facticity discrimination, as intra-corpus evaluation may permit learning of dataset-specific regularities.

4.2. Shallow stylistic modeling

To assess whether intra-corpus performance on ISOT requires deep contextual representations, we trained a logistic regression classifier using only low-dimensional stylistic features extracted from raw text. No lexical n-grams or embeddings were used.

Feature extraction. For each document, we computed a set of normalized surface features designed to capture stylistic and provenance-related regularities. These included binary indicators such as `has_reuters` (presence of the token "(Reuters)") and `has_video` (presence of markers such as "[Video]" or "video"), as well as continuous features such as `upper_ratio` (proportion of uppercase letters over alphabetic characters), `qm_per_100` and `excl_per_100` (question and exclamation marks per 100 characters), `log_chars`, and `log_words`. Additional cues included punctuation density, quotation density, source-name matches, and counts of emphatic uppercase tokens. All continuous features were standardized before training. The classifier used L_2-regularized logistic regression with default scikit-learn settings.

Model	Accuracy	Macro F1
Logistic (features)	0.9960	0.9960
Logistic (w/o Reuters)	0.8935	0.8935

Table 1: Intra-corpus validation performance on ISOT using shallow stylistic features.

Using the full feature set, performance matches transformer-based models within rounding error, indicating that ISOT is nearly linearly separable in a low-dimensional stylistic space. To test whether this result is dominated by a single provenance marker, we performed an ablation in which the token "(Reuters)" was removed from the text before feature extraction, and the derived features `has_reuters` and `has_dateline` were excluded from the model. Even after this removal, the

classifier still achieved 0.8935 accuracy and 0.8935 Macro F1 on validation, showing that Reuters is the strongest cue but not the only source of separability.

Feature importance. Table 2 shows the largest-magnitude coefficients from the full logistic regression model (positive values predict *true*, negative values predict *fake*). As expected, `has_reuters` is by far the dominant predictor.

Feature	Coefficient
`has_reuters`	+6.36
`log_words`	+1.38
`upper_ratio`	-0.38
`has_video`	-1.49
`qm_per_100`	-1.37
`log_chars`	-1.59

Table 2: Selected coefficients from the full logistic regression model on ISOT. Positive values favor the *true* class; negative values favor the *fake* class.

Table 3 summarizes the most influential coefficients after removing Reuters-related cues. The remaining decision boundary is still driven by simple stylistic properties, especially document length, uppercase usage, question-mark density, exclamation frequency, and video markers.

Feature	Coefficient
Most predictive of fake	
`log_words`	-5.90
`upper_ratio`	-4.47
`qm_per_100`	-2.66
`excl_per_100`	-1.32
`has_video`	-1.17
Most predictive of true	
`log_chars`	+4.00
`quote_ratio`	+1.84
`n_chars`	+0.74
`punct_per_100`	+0.28

Table 3: Selected coefficients from the logistic regression model after removing Reuters-related cues. Positive values favor the *true* class; negative values favor the *fake* class.

These results show that the near-ceiling separability of ISOT is heavily driven by provenance markers, but also supported by additional shallow stylistic differences between classes. In other words, removing Reuters substantially weakens the shortcut, yet intra-corpus classification remains far above chance without requiring semantic representations.

4.3. Ablation and minimal baselines

To further analyze the source of near-ceiling performance in ISOT, we evaluated minimal classifiers based on shallow stylistic cues.

Building on the shallow stylistic model above, we further evaluated a one-rule classifier based on a single binary feature:

```
if has_reuters == 1 → TRUE
        else → FAKE
```

This classifier achieved 0.9963 accuracy on ISOT validation. Table 4 summarizes the intra-corpus performance of transformer-based models and minimal baselines. A single binary lexical cue is therefore sufficient to reproduce near-ceiling performance on ISOT.

Model	Acc.	M. F1	W. F1
BERT (baseline)	0.999+	0.999+	0.999+
Logistic (style features)	0.9960	0.9960	0.9960
1-rule (has_reuters)	0.9963	0.9963	0.9963

Table 4: Intra-corpus results on ISOT.

4.4. Temporal considerations.

The datasets used in this study span a relatively narrow and overlapping time period, primarily covering news events from 2016–2017 in the case of ISOT, with partially overlapping periods in Misinfo and WELFake due to shared or reused sources. Following common practice in the literature, we employed stratified random splits rather than temporal splits.

We acknowledge that the absence of temporal partitioning may allow models to be exposed to similar or related events across training and validation sets. Prior work has highlighted the importance of temporal separation in fact-checking and misinformation detection, arguing that models can otherwise exploit event-level overlap and lack of counter-evidence (Glockner et al., 2022).

However, the ablation results presented in this section indicate that near-ceiling performance can be achieved using shallow stylistic cues alone, including a single provenance marker. This suggests that high intra-corpus performance is not primarily driven by temporal memorization of events, but by structural regularities and source-specific artifacts embedded in the datasets.

Therefore, while temporal splits are important for evaluating robustness in realistic fact-checking scenarios, the evidence presented here indicates that the dominant factor behind near-perfect intra-corpus results in ISOT is dataset-level separability rather than event-level generalization.

5. Cross-Dataset Experiments

We evaluate several BERT-based configurations under cross-dataset transfer conditions. Models are trained on one dataset and evaluated on another without target fine-tuning.

To evaluate robustness under dataset shift, we conducted systematic cross-dataset transfer experiments using three widely adopted fake-news corpora: WELFake, Misinfo, and ISOT.

To assess whether affective enrichment improves cross-domain generalization, we additionally evaluate several BERT-based configurations, including variants augmented with VAD (Mohammad, 2018), Ekman emotion categories (Ekman, 1992), and SenticNet features (Cambria et al., 2018), which provide complementary affective signals at lexical and conceptual levels.

Affective feature integration. VAD features encode continuous affective dimensions (valence, arousal, dominance) at the lexical level. Ekman categories provide a discrete representation of basic emotions, capturing coarse affective states such as anger, fear, or joy. SenticNet supplies concept-level affective and semantic features derived from commonsense knowledge bases.

In all cases, these signals are computed from the input text and concatenated with the BERT representation prior to classification.

For each pair of datasets (A, B), we trained on A and evaluated on B, yielding six transfer directions:

1. WELFake $\rightarrow$ Misinfo
2. Misinfo $\rightarrow$ WELFake
3. ISOT $\rightarrow$ WELFake
4. WELFake $\rightarrow$ ISOT
5. ISOT $\rightarrow$ Misinfo
6. Misinfo $\rightarrow$ ISOT

All models were fine-tuned under identical conditions (BERT-base backbone, max length 128, batch size 128, three epochs, class weighting, mixed precision training). We report Accuracy and Macro F1 on the external test set in each direction. Macro F1 is considered as it captures class balance under distributional shift.

Full per-direction classification reports (including class-wise precision, recall, and F1 scores) are provided in the Appendix (Tables 8–13).

Table 5 summarizes Accuracy and Macro F1 for compact comparison across all six transfer directions. Across domains, performance drops relative to within-dataset evaluation, indicating limited transferability.

5.1. Duplicate Analysis and Data Leakage Control

Before conducting cross-dataset evaluation, we performed exact-match duplicate detection between corpora. Between WELFake (train) and ISOT (test), we identified 39,687 exact duplicates out of 44,898 ISOT instances (approximately 88.4%). To prevent leakage, all overlapping instances were removed prior to training.

This finding indicates that WELFake partially subsumes ISOT content, rendering naïve cross-dataset evaluation misleading unless strict deduplication is applied. The persistence of near-ceiling WELFake→ISOT performance after deduplication further suggests that ISOT remains structurally easy as a target domain.

6. Discussion

High intra-dataset accuracy does not imply generalization under distribution shift. When datasets are structurally homogeneous, classification approximates ecosystem discrimination.

Binary provenance-based labels risk conflating source identification with epistemic inference.

6.1. Cross-Domain Stability Analysis

To evaluate robustness under dataset shift, we computed the Macro F1 score across all six cross-dataset transfer directions.

For each model, we report the mean Macro F1 across directions and its standard deviation. Lower variance indicates higher cross-domain stability.

To quantify directional asymmetry, we compute the absolute Macro F1 difference between both transfer directions of each dataset pair (e.g., W→M vs M→W). Lower values indicate greater bidirectional consistency and thus stronger cross-domain robustness. Table 7 reports the directional gaps for the three dataset pairs, as well as the mean gap per model.

Overall, BERT+VAD+Ekman exhibits the lowest average directional gap, indicating the most symmetric transfer behaviour. Ekman alone ranks second. In contrast, SenticNet and VAD show the largest asymmetries, particularly in the ISOT–Misinfo pair.

6.2. Ranking of Cross-Domain Stability

Ranking models by increasing standard deviation (see Table 6) yields:

1. BERT + VAD + Ekman (most stable)
2. BERT + Ekman
3. BERT simple
4. BERT + SenticNet
5. BERT + VAD (least stable)

Model	W→M		M→W		I→W		W→I		I→M		M→I	
	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1
BERT	0.851	0.846	0.914	0.913	0.794	0.789	0.996	0.996	0.638	0.548	0.965	0.964
VAD	0.854	0.846	0.695	0.659	**0.812**	**0.810**	0.997	0.997	0.576	0.424	0.879	0.875
Ekman	0.846	0.842	0.952	0.952	0.745	0.733	**0.999**	**0.998**	**0.719**	**0.685**	0.669	0.614
VAD+Ekman	**0.863**	**0.858**	0.849	0.845	0.803	0.800	0.998	0.997	0.686	0.641	0.828	0.819
SenticNet	0.858	0.854	**0.958**	**0.958**	0.696	0.674	0.997	0.997	0.636	0.543	**0.987**	**0.987**

Table 5: Accuracy (Acc) and Macro F1 (MF1) across all six cross-dataset transfer directions. Best value per column is shown in bold.

Model	Std. Dev. (Macro F1)
BERT simple	0.1630
BERT + VAD	0.2012
BERT + Ekman	0.1526
BERT + VAD + Ekman	**0.1147**
BERT + SenticNet	0.1875

Table 6: Sample standard deviation of Macro F1 across six transfer directions. Lower values indicate greater cross-domain stability.

Model BERT+	W↔M	W↔I	I↔M	M.Gap
Base	0.0671	0.2073	0.4159	0.2301
VAD	0.1871	**0.1869**	0.4516	0.2752
Ekman	0.1104	0.2655	**0.0702**	0.1487
VAD+Ekman	**0.0138**	0.1973	0.1787	**0.1299**
SenticNet	0.1036	0.3231	0.4434	0.2900

Table 7: Directional Macro F1 gap (absolute difference between bidirectional transfers). Lower values indicate stronger symmetry and cross-domain robustness.

Several structural patterns emerge from the analysis.

First, raw in-dataset performance does not predict cross-domain robustness. Models achieving near-ceiling accuracy on certain datasets exhibit substantial degradation when evaluated under domain shift.

Second, affective enrichment behaves differently depending on the representation used. Ekman-based features consistently reduce performance variance across datasets. When combined with VAD, the resulting model achieves the lowest cross-domain fluctuation (Std. Dev. = 0.113), suggesting a regularization effect that stabilizes decision boundaries across stylistic and source-based shifts.

In contrast, SenticNet produces the highest peaks in some transfer directions (e.g., Misinfo → ISOT and WELFake → ISOT) but also sharp collapses in others (e.g., ISOT → Misinfo). This indicates high sensitivity to training-domain lexical-affective alignment.

VAD alone exhibits the highest instability (Std. Dev. = 0.207), suggesting that coarse affective dimensions without discrete emotional categories may insufficiently constrain cross-domain generalization.

Finally, the consistent asymmetry between directions (e.g., ISOT → Misinfo vs. Misinfo → ISOT) confirms that dataset provenance and stylistic priors strongly influence transferability. Training on more heterogeneous corpora appears to improve outward generalization, whereas training on more homogeneous datasets leads to brittle representations.

These observations align with the hypothesis that dataset structure plays a dominant role in model behavior. In particular, ISOT appears to be largely separable using source-specific lexical cues, such as the presence of the token "(Reuters)". This explains why both shallow and deep models achieve near-perfect intra-corpus performance, while cross-dataset generalization collapses. The phenomenon is consistent with shortcut learning and dataset bias, but here it emerges under minimal feature assumptions.

Taken together, these results indicate that cross-domain evaluation and variance analysis should complement standard accuracy reporting in fake-news classification. Stability under dataset inversion provides a stronger robustness criterion than isolated peak performance.

7. Conclusions

Fake-news benchmarks frequently report near-ceiling performance, yet cross-dataset evaluation reveals substantial and asymmetric generalization gaps. These results show that accuracy on isolated benchmarks is not a reliable indicator of robustness, but largely reflects dataset separability rather than genuine veracity understanding.

Across six cross-dataset transfer directions involving WELFake, Misinfo, and ISOT, we observe three systematic phenomena. First, average performance and stability are not equivalent objectives: in some cases, BERT+SenticNet achieves the highest mean Macro F1, yet exhibits considerable directional asymmetry and variance. Second, BERT+SenticNet attains near-ceiling results in cer-

tain directions but collapses under domain reversal, resulting in the highest mean directional gap. This pattern suggests strong sensitivity to corpus-specific stylistic and lexical distributions. Third, the combined VAD+Ekman representation yields the lowest standard deviation and the smallest bidirectional gap, indicating the most stable cross-domain behavior.

The superior robustness of VAD+Ekman may stem not only from feature concatenation, but from the complementarity of affective representations. VAD encodes continuous affective gradients (valence, arousal, dominance), providing smooth global signals, whereas Ekman categories introduce discrete emotional anchors. Their joint integration may provide an implicit regularizing effect: when distribution shift destabilizes one affective dimension, the other may preserve transferable structure. This hybrid affective embedding reduces directional asymmetry and mitigates collapse under domain shift.

These results suggest that robustness emerges not from maximizing in-domain accuracy, but from constraining models with semantically meaningful inductive biases. Emotional structure, when represented through complementary continuous and categorical spaces, appears to provide such a constraint. Importantly, robustness must be evaluated not only via mean performance, but through variance and directional gap metrics that explicitly capture asymmetry under transfer.

Progress in fake-news detection therefore requires a methodological shift: from accuracy maximization on isolated benchmarks toward cross-domain robustness, bidirectional evaluation, and construct-valid modeling. Without such a shift, near-ceiling scores risk reflecting dataset artifacts rather than genuine generalizable understanding.

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful discussion, and Instituto Politécnico Nacional (COFAA, SIP-IPN) and the Mexican Government (SECIHTI, SNII) for their financial support for this work.

A. Full Cross-Domain Classification Results

This appendix reports the complete cross-domain evaluation results for all six transfer directions. For each experiment, we provide Accuracy (Acc), Macro F1 (MF1), Weighted F1 (WF1), per-class F1 scores, and per-class Recall. The best value in each column is highlighted in bold.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138. Springer.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of AAAI Conference on Artificial Intelligence*, 32(1).

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ehab Essa, Karima Omar, and Ali Alqahtani. 2023. Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, 9(6):6581–6592.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Axel Gelfert. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117. Accessed 2026-02-25.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. Dafd: Domain adaptation framework for fake news detection. In *International conference on neural information processing*, pages 305–316. Springer.

Md. Rafiqul Islam, M. M. Islam, Md. Shafiqul Azad, M. S. Uddin, Kamal Daud, and M. A. Hossain. 2020. A survey on fake news detection using machine learning and deep learning techniques. *IEEE Access*, 8:178007–178025.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80:11765–11788.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.8509	0.8460	0.8489	0.8733	0.8188	0.9303	0.7528
VAD	0.8536	0.8462	0.8498	0.8799	0.8126	**0.9702**	0.7095
Ekman	0.8463	0.8419	0.8447	0.8683	0.8156	0.9167	0.7594
VAD+Ekman	**0.8625**	**0.8583**	**0.8608**	**0.8827**	**0.8338**	0.9363	0.7713
SenticNet	0.8578	0.8539	0.8564	0.8777	0.8301	0.9237	**0.7765**

Table 8: Full classification results for the transfer experiment WELFake → Misinfo. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9140	0.9131	0.9133	0.9221	0.9041	0.9951	0.8293
VAD	0.6949	0.6591	0.6616	0.7695	0.5487	0.9967	0.3794
Ekman	0.9524	0.9523	0.9524	0.9547	0.9500	0.9800	0.9236
VAD+Ekman	0.8489	0.8445	0.8451	0.8708	0.8182	0.9959	0.6953
SenticNet	**0.9577**	**0.9575**	**0.9576**	**0.9602**	**0.9549**	**0.9981**	**0.9154**

Table 9: Full classification results for the transfer experiment Misinfo → WELFake. Best values per column are shown in bold.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumya Sinha. 2020. Fndnet – a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Sahas Koka, Anthony Vuong, and Anish Kataria. 2024. Evaluating the efficacy of large language models in detecting fake news: a comparative analysis. *arXiv preprint arXiv:2406.06584*.

Maaruf Lawal and Abdulrashid Abdulrauf. 2025. Fake news detection using Bi-LSTM architecture: A deep learning approach on the isot dataset. *Journal of Computing Theories and Applications*, 3:104–114.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1146–1151. Accessed 2026-02-25.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–184.

Steven Peutz. 2023. Misinformation & fake news text dataset 79k. Dataset card describing 34,975 true and 43,642 fake/misinfo/propaganda items. Accessed 2026-02-25.

Maria Pszona, Maria Janicka, Grzegorz Wojdyga, and Aleksander Wawer. 2023. Towards universal methods for fake news detection. *Natural Language Engineering*, 29(4):1004–1042.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection*, pages 215–232. Springer.

Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893. Open full text circulated by authors; Accessed 2026-02-25.

H. F. Villela, F. Corrêa, J. S. D. A. N. Ribeiro, A. Rabelo, and D. B. F. Carvalho. 2023. Fake news detection: A systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1):47–58.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Report prepared for the Council of Europe. Accessed 2026-02-25.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.7935	0.7887	0.7877	0.8206	0.7568	0.9734	0.6240
VAD	**0.8115**	**0.8101**	**0.8096**	**0.8263**	**0.7939**	0.9244	**0.7051**
Ekman	0.7446	0.7329	0.7312	0.7888	0.6769	0.9831	0.5198
VAD+Ekman	0.8034	0.8001	0.7993	0.8257	0.7744	0.9603	0.6555
SenticNet	0.6961	0.6735	0.6710	0.7594	0.5877	**0.9884**	0.4207

Table 10: Full classification results for the transfer experiment ISOT → WELFake. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9962	0.9960	0.9962	0.9951	0.9968	0.9903	**1.0000**
VAD	0.9971	0.9970	0.9971	0.9964	0.9976	0.9927	**1.0000**
Ekman	**0.9985**	**0.9984**	**0.9985**	**0.9981**	**0.9987**	**0.9961**	**1.0000**
VAD+Ekman	0.9975	0.9974	0.9975	0.9968	0.9979	0.9937	**1.0000**
SenticNet	0.9967	0.9966	0.9967	0.9959	0.9973	0.9918	**1.0000**

Table 11: Full classification results for the transfer experiment WELFake → ISOT. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.6384	0.5484	0.5707	0.7500	0.3469	0.9766	0.2160
VAD	0.5764	0.4235	0.4563	0.7204	0.1265	**0.9827**	0.0690
Ekman	**0.7188**	**0.6846**	**0.6961**	**0.7884**	**0.5809**	0.9434	**0.4382**
VAD+Ekman	0.6857	0.6407	0.6548	0.7678	0.5136	0.9359	0.3732
SenticNet	0.6356	0.5431	0.5658	0.7487	0.3374	0.9776	0.2086

Table 12: Full classification results for the transfer experiment ISOT → Misinfo. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9645	0.9643	0.9644	0.9670	0.9615	0.9965	0.9294
VAD	0.8786	0.8751	0.8761	0.8958	0.8544	0.9985	0.7471
Ekman	0.6688	0.6144	0.6211	0.7592	0.4696	0.9983	0.3074
VAD+Ekman	0.8277	0.8194	0.8212	0.8582	0.7806	0.9967	0.6425
SenticNet	**0.9866**	**0.9865**	**0.9866**	**0.9873**	**0.9858**	**0.9989**	**0.9731**

Table 13: Full classification results for the transfer experiment Misinfo → ISOT. Best values per column are shown in bold.