IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

Zero-Short Fake News Detection Using Bert: A Review of Semantic Classification and Research Directions

Tarate Siddhant[1], Prof. M. N. Pimple[2], Prof. A. R. Kardile[3], Dr. H. B. Jadhav[4]
Student, Computer Engineering, Adsul Technical Campus, Ahilyanagar, India[1]
Guide, Computer Engineering, Adsul Technical Campus, Ahilyanagar, India[2]
Assistant Professor, Computer Engineering, Adsul Technical Campus, Ahilyanagar, India[3]
HOD, Computer Engineering, Adsul Technical Campus, Ahilyanagar, India[4]

Abstract : *The rapid circulation of unverified information across social-media platforms creates a need for automated systems that can screen large volumes of textual content. This paper reviews a zero-shot fake-news detection approach that combines contextual representations from Bidirectional Encoder Representations from Transformers (BERT) with a zero-shot learning (ZSL) classifier. Unlike conventional classifiers that depend strongly on task-specific labelled examples, zero-shot classification uses semantic relationships between an input representation and candidate class descriptions.*

The reviewed ZS-FND study uses the textual portion of the Fakeddit dataset and reports 98.39% accuracy, 97.33% precision, 95.67% recall, 96.49% F1-score and 0.0160 mean absolute error. These are published results, not experiments conducted for this review. The paper explains the processing pipeline, summarises the reported evaluation, and discusses topic shift, explainability, source verification and evaluation risks.

A practical research framework is outlined for testing generalisation across topics and datasets. The review concludes that zero-shot models are promising screening tools, but automated labels should not be treated as proof of factual accuracy without independent verification.

Keywords: Fake news detection, zero-shot learning, BERT, natural language processing, text classification, social media, misinformation

I. INTRODUCTION

Digital communication platforms allow news and opinions to reach large audiences within seconds. This speed is useful for sharing information, but it also allows misleading claims, manipulated headlines and fabricated stories to spread before they are checked. Engagement signals such as likes, shares and comments can increase visibility; however, popularity is not evidence that a claim is true. Manual review alone is difficult to scale because platforms process a continuous stream of user-generated content [1].

Automated fake-news detection applies natural language processing (NLP), machine learning (ML) and deep learning (DL) to identify patterns associated with unreliable content. Traditional approaches commonly use word counts, n-grams and term frequency–inverse document frequency (TF-IDF), with classifiers such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression and Random Forest. Neural methods, including convolutional neural networks (CNNs), recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, can learn more complex text patterns [1].

A continuing difficulty is that a classifier trained on one topic or dataset may perform poorly when new events, writing styles or claims appear. Zero-shot learning offers a possible way to reduce dependence on labelled examples for every

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

class. The reviewed ZS-FND approach uses BERT to encode contextual meaning and a zero-shot classifier to compare the resulting representation with candidate labels [1]. This paper explains the published method, identifies its strengths and limitations, and proposes additional research checks. It does not claim to reproduce the original experiment. semantic information from a pre-trained language model can reduce dependence on a large, domain-specific labelled dataset. A high score on one benchmark, however, does not guarantee reliability on live social-media content.

II. OBJECTIVES AND CONTRIBUTIONS

The objectives are: (1) explain the role of BERT embeddings in representing news text; (2) describe zero-shot classification and semantic label matching; (3) summarise the dataset and evaluation results reported in the source study; (4) compare the approach with selected ML and DL baselines; (5) identify risks involving topic shift, class imbalance, explainability and false predictions; and (6) propose a reproducible evaluation plan for future work.

The contribution is a consolidated, implementation-oriented explanation of the ZS-FND pipeline together with additional research considerations. It is a review and research framework, not a newly trained detection model.

III. RELATED WORK

Earlier fake-news detection research has used conventional ML classifiers with features derived from word frequency and text statistics. These approaches can be efficient, but performance depends on feature design, dataset coverage and the relationship between training and test samples. Ensemble methods and large-scale data processing have also been explored [1].

Deep-learning research has applied CNN, RNN and LSTM architectures to learn patterns from word sequences. Hybrid approaches combine these architectures or use dimensionality-reduction methods such as principal component analysis. Transformer-based systems represent words in context, allowing a word's representation to reflect its surrounding text rather than a fixed vocabulary entry [1].

BERT-based systems have been combined with ensemble learning and multimodal information. Other work has explored zero-shot learning for social-media news and prompt-based methods for low-resource classification. The reviewed article notes that performance varies with the dataset and task design [1]. Zero-shot learning does not mean that a system has no prior training: it uses knowledge acquired by a pre-trained model and semantic descriptions of candidate labels.

IV. PURPOSE OF THE SYSTEM

The purpose of the fake-news detection system is to support the initial screening of large volumes of online news text and help reviewers identify posts that may contain misleading information. The reviewed approach uses BERT to represent the contextual meaning of a news item and zero-shot semantic matching to compare that representation with candidate labels such as "real news" and "fake news".

The system is intended to reduce dependence on task-specific labelled examples for every new topic, provide a consistent first-level classification signal, and assist human fact-checkers in prioritising content for review. Its output is only a screening indication; independent evidence checking and human judgement remain necessary before a claim is treated as false.

EXISTING SYSTEM

Traditional fake-news detection systems commonly use manual review or machine-learning methods based on word counts, n-grams and TF-IDF features. Classifiers such as Naive Bayes, Support Vector Machine, Logistic Regression and Random Forest learn patterns from labelled examples. Later approaches use CNN, RNN and LSTM networks to learn more complex text patterns. These methods can be useful, but their performance depends on the quality and coverage of the training data and the selected features. A model trained on one topic may not work as well when

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

unfamiliar events, new writing styles or claims appear. Manual checking is also difficult to scale when content is published continuously.

EXISTING SYSTEM WORKFLOW DIAGRAM

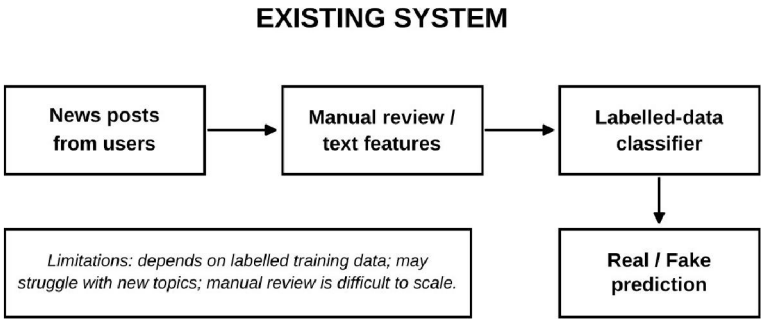


Figure 1. General workflow of conventional fake-news detection systems, summarised from the approaches discussed in the review paper.

Note: This diagram represents the general existing-system approach described in the paper; it is not a reproduction of a single evaluated system.

V. RESEARCH GAP AND QUESTIONS

Benchmark datasets may not represent the speed and diversity of live misinformation. Models can also learn unreliable correlations, such as sensational wording or recurring publisher names. Accuracy alone does not show whether rare types of misinformation are missed or legitimate reports are incorrectly flagged. Semantic similarity is not equivalent to evidence-based fact verification.

The following questions are suitable for future controlled study: RQ1: How well does BERT-based zero-shot classification transfer to topics absent from development data? RQ2: How does it compare with TF-IDF-based SVM, Logistic Regression and an LSTM baseline under identical data splits? RQ3: Which preprocessing decisions most affect precision, recall and false-positive rate? RQ4: How stable are predictions when headlines are paraphrased or new events emerge? RQ5: Can explanations and source context improve human review without overstating model certainty?

VI. DATASET AND PREPROCESSING

The source study uses the Fakeddit dataset, containing 166,740 news-related posts. It reports 54.31% of samples as fake and 45.69% as real. Although Fakeddit includes multimodal information such as text, images and metadata, the reviewed ZS-FND pipeline focuses on textual content [1]. A text-only system cannot directly evaluate image manipulation or visual context.

Preprocessing prepares raw posts for language-model input. The source describes tokenization, removal of unnecessary special characters and stop words, spelling correction, lowercase conversion and lemmatization. The text is then passed through the BERT tokenizer, which adds [CLS] and [SEP] tokens, maps tokens to vocabulary IDs and applies padding or truncation to a fixed sequence length. An attention mask distinguishes actual tokens from padding [1].

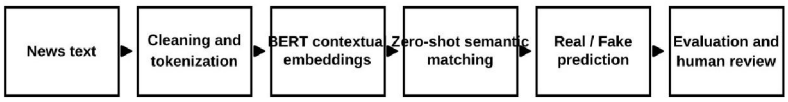
For a fair experiment, preprocessing rules should be documented and applied consistently. Operations that learn information from data—such as vocabulary fitting, feature selection or resampling—must be fitted using training data only. Duplicate and near-duplicate posts should also be checked because they can cause information leakage between training and test sets.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-41146

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

VII. PROPOSED SYSTEM AND METHODOLOGY

The reviewed process has six stages. First, textual posts are collected. Second, text is cleaned and tokenized. Third, BERT produces contextual embeddings. Fourth, the embeddings are mapped into a semantic attribute space. Fifth, the zero-shot classifier calculates similarity between the post representation and candidate class descriptions such as “real news” and “fake news”. Finally, the class with the highest similarity is returned [1].



A prediction is a screening signal and should be followed by independent verification.

Figure 2. Overview of the reviewed BERT and zero-shot classification workflow.

The input representation is written as h_i = BERT(x_i). A projection layer maps it to an attribute space: a_i = W h_i + b, where W and b are learned parameters. The classifier compares a_i with the semantic attribute vector A_y for each candidate label. A similarity measure such as cosine similarity is used, and the label with the highest score is selected [1]. The prediction can be written as y_hat = arg max over y of sim(a_i, A_y). This indicates semantic compatibility; it does not prove that a claim is factually correct.

VIII. EVALUATION METRICS AND REPORTED RESULTS

The source article evaluates the model using accuracy, precision, recall, F1-score and mean absolute error (MAE), and reports the use of 10-fold cross-validation. Accuracy measures the overall proportion of correct predictions. Precision measures the proportion of predicted fake items that are labelled fake; recall measures the proportion of actual fake items detected. F1-score balances precision and recall. MAE measures the average absolute difference between encoded predictions and labels [1].

Metric	Reported value	Meaning
Accuracy	98.39%	Overall proportion of correct classifications
Precision	97.33%	Correctness of predicted fake items
Recall	95.67%	Proportion of labelled fake items detected
F1-score	96.49%	Balance between precision and recall
MAE	0.0160	Mean absolute prediction error as reported

Table 1. Performance values reported for ZS-FND in the source article [1].

The source article reports that ZS-FND achieved higher values than the ELD-FN and FakeNED baseline approaches. It also reports higher accuracy, precision, recall and F1-score than selected traditional ML and DL classifiers, including LSTM, SVM, CNN, Bernoulli Naive Bayes, Logistic Regression and Multinomial Naive Bayes [1]. These comparisons should be interpreted in the context of the dataset, label definitions and experimental setup used in that study.

The article also examines preprocessing and resampling. Its discussion reports stronger overall performance when preprocessing is enabled and reports that the no-resampling setting performed best among the tested resampling conditions. Such findings should not be generalised to every dataset: class distributions and collection methods can change the outcome [1].

IX. DISCUSSION AND ADDITIONAL RESEARCH CONSIDERATIONS

Context and evidence: A sentence can resemble misinformation and still be true; a carefully written false claim can also appear credible. Future systems should consider evidence retrieval, publication date, source reliability and claim-level verification as complementary signals. These are recommendations in this review, not components claimed to have been evaluated in the cited ZS-FND study.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

Confidence and human review: A practical system should retain similarity scores for both candidate labels and route low-margin predictions for manual checking. The interface should use cautious wording such as "potentially misleading" rather than declaring a post false. Reviewers should be able to inspect the original post, its source and supporting evidence.

Robustness and fairness: Testing should include paraphrases, spelling variations, short headlines, multilingual posts and breaking-news topics. Results should be reported by topic and class, not only as one aggregate score. Researchers should check whether the model disproportionately flags particular publishers, communities or writing styles.

Reproducibility: A future implementation should publish the dataset version, cleaning rules, train/validation/test split, random seeds, model version, candidate label descriptions and hardware details. Near-duplicate detection and temporal splits can reduce leakage and provide a more realistic estimate of future performance.

X. CHALLENGES AND LIMITATIONS

Dataset dependence: Fakeddit is useful as a benchmark, but it cannot represent every platform, language, region or news event. Topic- and time-based testing is needed before generalising results.

Semantic ambiguity: Similar wording may appear in satire, opinion, quotation, reporting and deliberate misinformation. A two-class label can hide these distinctions.

Changing information: A statement that is unverified at one time may later be confirmed or disproved. Detection systems need timestamps and updated evidence.

False predictions: Incorrectly flagging legitimate reporting can damage trust, while missing fabricated content can allow harmful claims to spread. Both error types should be measured.

Explainability: Similarity scores alone may not provide a sufficient explanation. A deployment should retain decision logs and provide a route for human appeal.

XI. FUTURE SCOPE

Future research can extend the text-only pipeline with reliable source metadata and evidence-based claim verification. Multimodal analysis may be investigated for posts that combine text, images and video, with each modality evaluated separately as well as jointly. Multilingual testing is important because misinformation is not limited to English-language content.

A further direction is temporal evaluation: develop the system using earlier material and test it on later events. This can reveal whether the model adapts to emerging topics. Researchers may also compare candidate-label descriptions, calibrate confidence scores and evaluate lightweight deployment options for high-volume screening. All extensions should use transparent protocols and independent datasets.

XII. CONCLUSION

This paper reviewed a BERT-based zero-shot approach for fake-news detection and organised its main stages into an IJARSCT-style research-paper structure. The method uses text preprocessing, contextual embeddings and semantic similarity to classify posts into candidate categories. The source study reports 98.39% accuracy, 97.33% precision, 95.67% recall, 96.49% F1-score and 0.0160 MAE in its evaluation setting [1]. These are published results and have not been independently reproduced in this review.

The review identifies important next steps beyond benchmark classification: testing across time and topics, measuring false alarms, improving explanations, checking dataset leakage and incorporating independent evidence before a post is treated as false. Zero-shot classification can support large-scale screening, but it should assist—not replace—careful fact-checking and human judgement.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

The author acknowledges the Department of Computer Engineering, Adsul's Technical Campus, for academic support and the opportunity to study current research in natural language processing and misinformation detection.

REFERENCES

[1] R. Baashirah, "Zero-Shot Automated Detection of Fake News: An Innovative Approach (ZS-FND)," IEEE Access, vol. 12, pp. 182828–182840, 2024, doi: 10.1109/ACCESS.2024.3462151.
[2] K. Nakamura, S. Levy, and W. Yang Wang, "r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-Grained Fake News Detection," arXiv:1911.03854, 2019.
[3] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake news classification using transformer based enhanced LSTM and BERT," International Journal of Cognitive Computing Engineering, vol. 3, pp. 98–105, 2022.
[4] B. Palani, S. Elango, and V. Viswanathan, "CB-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT," Multimedia Tools and Applications, vol. 81, no. 4, pp. 5587–5620, 2022.
[5] A. J. Keya, M. A. H. Wadud, M. F. Mridha, M. Alatiyyah, and M. A. Hamid, "AugFake-BERT: Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification," Applied Sciences, vol. 12, no. 17, art. 8398, 2022.
[6] X. Xie, L. Zhang, and L. Li, "Zero-shot learning for text classification with pre-trained language models," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 3487–3497, 2022.
[7] H. Zhang, W. Yang, and Y. Sun, "Zero-shot text classification using transformer-based embeddings," Journal of Artificial Intelligence Research, vol. 72, pp. 183–200, 2023.
[8] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: Typology, time of detection, taxonomies," International Journal of Information Security, vol. 22, no. 1, pp. 177–212, 2023.