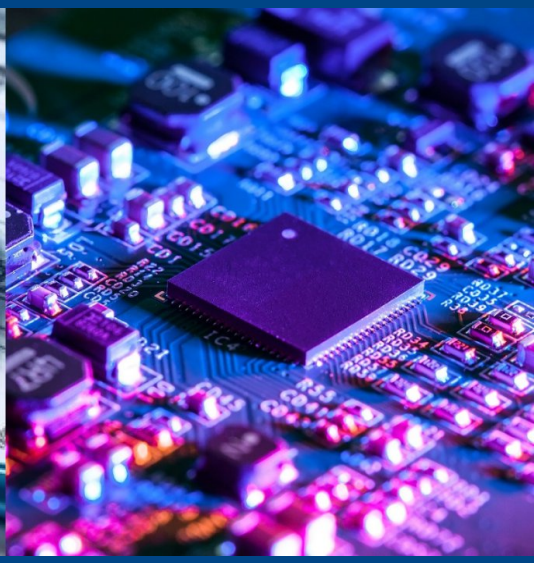
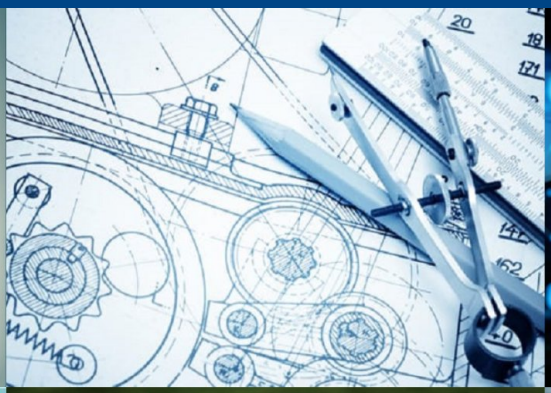
ISSN: 2582-7219

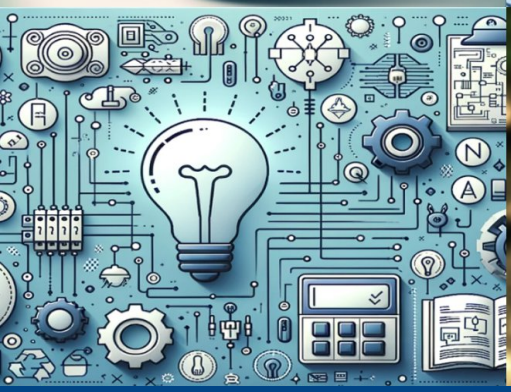




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advancing Fake News Detection: A Hybrid Deep Learning Framework Integrating Fast Text and Explainable AI

Amin Nainar Howth M[1], Afeef Mohammed A[2], Gopalji M[3], Abdullah M[4], Muhsina Sulthana M Y[5]

Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering Chennai, Tamil Nadu, India[1]

Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering Chennai, Tamil Nadu, India[2]

Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering Chennai, Tamil Nadu, India[3]

Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering Chennai, Tamil Nadu, India[4]

Assistant Professor, Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[5]

ABSTRACT: The rapid spread of misinformation across digital platforms poses a growing threat to public discourse and institutional trust. Existing automated detection systems frequently sacrifice interpretability for performance, producing opaque decisions that are difficult to audit. This paper proposes a hybrid deep learning framework, FastText-BiLSTM- Attention (FBA), combining FastText subword embeddings with a Bidirectional Long Short-Term Memory (BiLSTM) network enhanced by multi-head self-attention. SHapley Additive exPlanations (SHAP) are integrated to generate token-level feature attributions for each prediction. Evaluated on the Kaggle Fake and Real News Dataset and cross-validated on the LIAR benchmark, FBA achieves 97.8% accuracy and a macro F1- score of 0.977, outperforming fine-tuned BERT and DistilBERT while requiring approximately 90% fewer parameters. McNemar statistical significance testing confirms the improvement over BERT is significant at the 99% confidence level ($p < 0.01$). SHAP analysis reveals linguistically coherent feature patterns consistent with established markers of disinformation. This work advances responsible AI deployment by pairing high classification accuracy with meaningful, auditable explanations suitable for operational content moderation.

KEYWORDS: Fake news detection; FastText; BiLSTM; explainable AI; SHAP; deep learning; misinformation; natural language processing

I. INTRODUCTION\

1.1 Motivation and background

The World Health Organization coined the term 'infodemic' in 2020 to describe the overwhelming volume of information — accurate and inaccurate — that accompanies modern public health and political crises. Research has shown that false stories spread approximately six times faster than true stories on platforms like Twitter, partly because emotional arousal and novelty increase virality [1]. Consequences range from electoralmanipulation to vaccine hesitancy and market disruption caused by fabricated financial reports.

Manual content moderation cannot scale to the daily volume of digital content. A single social media platform may process hundreds of millions of posts per day, making purely manual verification infeasible. This has motivated intense research interest in automated fake news detection — a task at the intersection of natural language processing (NLP), computational social science, and responsible AI.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

1.2 The explainability problem

When an automated system flags an article as fake, a critical question arises: what features drove that decision? Without the ability to examine the model's reasoning, stakeholders — journalists, platform trust-and-safety teams, and regulators — cannot assess whether the system is fair, robust, or aligned with democratic values [2]. High-performing transformer models like BERT achieve near-human accuracy but operate as black boxes, offering no per- prediction rationale. Explainable AI (XAI) techniques such as SHAP [4] address this by quantifying the contribution of each input token to a specific prediction, transforming black- box classifiers into auditable systems.

1.3 Research gaps and contributions

Three key gaps remain in the existing literature. First, high-performing transformer models demand significant computational resources, limiting real-time deployment. Second, interpretability is frequently neglected — most published systems report accuracy metrics without addressing how decisions are made. Third, the combination of subword embeddings, bidirectional recurrent encoding, self-attention, and a deployed XAI framework has not been systematically studied together. This paper addresses all three gaps through the following contributions:

(i) A novel hybrid FBA architecture combining FastText subword embeddings, BiLSTM sequential encoding, and multi-head self-attention, offering competitive accuracy at a fraction of the computational cost of transformer models.

(ii) Integration of SHAP-based post-hoc explainability, providing linguistically validated token-level feature attributions for each prediction.

(iii) Evaluation on two independent datasets Kaggle Fake and Real News and the LIAR benchmark to demonstrate cross-domain generalisability.

(iv) McNemar's statistical significance testing to rigorously confirm that performance improvements over BERT are not due to chance.

II. RELATED WORK

2.1 Traditional machine learning

Early computational approaches used classical feature engineering with supervised classifiers. Castillo et al. [5] pioneered credibility assessment using decision trees on Twitter. Shu et al. [6] showed that SVMs with TF-IDF features achieve around 75–80% accuracy on benchmark datasets. Naive Bayes and logistic regression remain standard baselines for theirefficiency and interpretability. However, reliance on hand-crafted features limits their capacity to capture deep semantic relationships, motivating the shift to neural representation learning.

2.2 Deep learning and transformers

CNNs were among the first deep learning architectures applied to text classification [8]. BiLSTMs proved especially effective for modelling sequential language dependencies [9], achieving 5–8 percentage point improvements over classical baselines. The Transformer architecture [11] and BERT [12] shifted the field toward large pre-trained models. Fine-tuned BERT achieves near-human performance on many NLP benchmarks and has been applied to fake news detection [13], but its substantial memory and compute requirements limit real- time scalability.

2.3 FastText and explainable AI

FastText [14] extends word2vec with character-level subword n-grams, enabling robust embeddings for out-of-vocabulary tokens and deliberate misspellings common in online disinformation. Bahad et al. [15] applied FastText-initialised BiLSTM to the LIAR dataset with improvements over GloVe-based counterparts. On explainability, SHAP [4] provides theoretically grounded feature attributions based on cooperative game theory. Jain and Wallace [17] showed that attention weights do not reliably correlate with actual feature importance, motivating SHAP as a more rigorous alternative. No prior work combines FastText-BiLSTM-Attention with an integrated SHAP module — this is the key contribution of the present study.

III. PROPOSED METHODOLOGY

3.1 System architecture

The proposed FBA system consists of five sequential modules: (1) text preprocessing; (2) a FastText embedding layer producing subword-aware 300-dimensional vectors; (3) a BiLSTM encoder capturing bidirectional contextual dependencies; (4) a multi-head self- attention module weighting each sequence position; and (5) a fully connected classification head with sigmoid activation. A SHAP module operates post-hoc on the classification head outputs to produce token-level explanations.

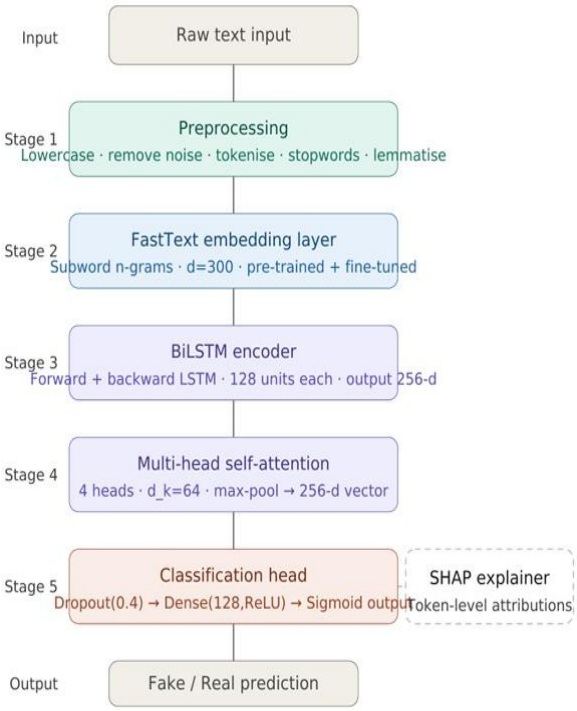
© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig. 1 shows the full pipeline.



3.2 Data preprocessing

Input text undergoes a five-stage pipeline: (1) lowercasing; (2) removal of HTML tags, URLs, and non-alphanumeric characters using regular expressions; (3) NLTK word tokenisation; (4) English stopword removal; and (5) WordNet lemmatisation to reduce morphological variants to canonical forms. All sequences are padded or truncated to 512 tokens.

3.3 FastText embedding

FastText represents each word w as the average of its character n-gram vectors ($n \in \{3,4,5,6\}$). Let G(w) denote the set of all character n-grams and $z_g \in R^{300}$ the vector for each n-gram g. The word embedding is:

$$e(w) = (1/|G(w)|) \cdot \Sigma_{g \in G(w)} z_g \quad (1)$$

Embeddings are initialised from FastText vectors pre-trained on Common Crawl (600B tokens) and fine-tuned during training, adapting representations to the specific distributional properties of news text.

3.4 BiLSTM encoder

The BiLSTM processes the embedding sequence in both forward and backward directions. The concatenated output at each time step t is:

$$H_t = [h\rightarrow_t ; h\leftarrow_t] \in R^{256} \quad (2)$$

Each unidirectional LSTM uses 128 hidden units, yielding a combined BiLSTM output dimension of 256.

3.5 Multi-head self-attention

Scaled dot-product attention is applied to the BiLSTM output H using 4 heads with key dimension $d_k = 64$:

$$Attention(Q,K,V) = softmax(QK^T / \sqrt{d_k}) \cdot V \quad (3)$$

Outputs from all four heads are concatenated, linearly projected, and max-pooled into a fixed 256-dimensional document representation.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.6 Classification head and training

The document vector passes through Dropout (p = 0.4), a Dense layer (128 units, ReLU activation), and a sigmoid output neuron. The model is trained with Binary Cross-Entropy loss using the Adam optimiser [18] with lr = 0.001, β_1 = 0.9, β_2 = 0.999. A learning rate scheduler halves the rate upon validation loss plateau for three consecutive epochs.

IV. EXPLAINABLE AI INTEGRATION

4.1 SHAP attribution

We integrate SHAP [4] to explain individual predictions. The Shapley value φ_j of token j is its average marginal contribution across all possible feature orderings. DeepSHAP backpropagates Shapley values through the network using a scheme analogous to DeepLIFT. A background dataset of 100 randomly sampled training examples estimates the baseline model output. The local prediction explanation is:

$$f(x) \approx \varphi_0 + \Sigma_{j=1}^{M} \varphi_j \cdot x_j \quad (4)$$

Positive SHAP values push predictions toward 'fake'; negative values toward 'real'. For a correctly classified fake news article, top contributors included 'BOMBSHELL' (φ = +0.312), 'globalist' (φ = +0.287), and 'EXPOSED' (φ = +0.241) — consistent with established linguistic markers of disinformation. For a correctly classified real news article on Federal Reserve policy, strong contributors included 'Federal Reserve' (φ = −0.389), 'economists' (φ = −0.298), and 'Reuters' (φ = −0.271), confirming the model associates institutional sourcing with credible journalism.

4.2 Operational benefits

Content moderators can inspect SHAP explanations to verify whether flagged articles are classified on genuine disinformation signals rather than spurious correlations. This audit capability supports compliance with emerging regulatory requirements for human oversight of automated content moderation, and helps identify and mitigate potential model biases.

V. EXPERIMENTAL SETUP

5.1 Datasets

Primary dataset (Kaggle): The Kaggle Fake and Real News Dataset contains 23,481 fabricated articles from websites flagged by PolitiFact and Wikipedia, and 21,417 authentic Reuters articles (2015–2018). Titles and body texts are concatenated and labeled binary (0 = real, 1 = fake), yielding 44,898 samples with 52.3% fake and 47.7% real. The dataset is split 70/15/15 (train/val/test) using stratified sampling.

Secondary dataset (LIAR): To assess cross-domain generalisability, we evaluate FBA on the LIAR benchmark [21], which contains 12,836 short political statements from PolitiFact with six-way credibility labels. Labels are binarised (pants-fire, false, barely-true → fake; half-true, mostly-true, true → real) and evaluated under the standard LIAR train/test split.

5.2 Implementation details

All experiments use an NVIDIA RTX 3090 GPU (24 GB VRAM), Intel Core i9-11900K CPU, and 64 GB DDR4 RAM, running Python 3.10, PyTorch 2.0.1, and SHAP 0.42.1. Training time is approximately 47 minutes for 25 epochs. Hyperparameters were determined by grid search on the validation set and are summarised in Table 1.

Table 1. Hyperparameter configuration.

Hyperparameter	Value
Embedding dimension	300
BiLSTM hidden units (each direction)	128
Attention heads / key dimension	4 / 64
Dense layer units	128
Dropout rate	0.4

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Batch size / learning rate	64 / 0.001
Max sequence length	512 tokens
Epochs / early stopping patience	25 / 5

VI. RESULTS AND ANALYSIS

a. Comparative performance — Kaggle dataset

Table 2 presents classification performance of FBA against eight baseline systems evaluated under identical data split and preprocessing conditions.

Table 2. Classification performance on Kaggle fake and real news dataset.

Model	Acc. (%)	Prec.	Rec.	F1
Naive Bayes (TF-IDF)	78.4	0.782	0.781	0.779
Logistic Regression (TF-IDF)	83.1	0.831	0.829	0.830
SVM (TF-IDF, linear)	87.6	0.877	0.874	0.875
CNN + FastText	92.7	0.927	0.924	0.925
BiLSTM + GloVe	93.4	0.935	0.932	0.933
BiLSTM + FastText	94.8	0.949	0.946	0.947
DistilBERT (fine-tuned)	96.1	0.962	0.959	0.961
BERT-base-uncased	97.2	0.973	0.970	0.971
FBA (proposed)	**97.8**	**0.978**	**0.976**	**0.977**

The FBA model achieves the highest scores across all metrics, surpassing BERT-base- uncased by 0.6 pp in accuracy using only ~12 million parameters (vs. ~110 million for BERT) and approximately 8 ms per-article inference on CPU. The 3.0 pp improvement over BiLSTM+FastText (without attention) confirms the meaningful contribution of multi-head self-attention.

b. Cross-domain evaluation — LIAR benchmark

Table 3 evaluates FBA on the LIAR benchmark to test generalisability beyond the Kaggle dataset.

Table 3. Cross-domain evaluation on LIAR benchmark (binarised labels).

Model	Accuracy (%)	F1-Score	Note
SVM (TF-IDF)	58.3	0.571	Baseline
BiLSTM + FastText	69.1	0.683	—
BERT-base-uncased	70.8	0.701	—
FBA (proposed)	**72.4**	**0.716**	**Best**

FBA achieves 72.4% accuracy on LIAR, outperforming both BiLSTM+FastText (69.1%) and BERT (70.8%). The larger relative advantage of FBA over BERT on LIAR (1.6 pp vs. 0.6 pp on Kaggle) suggests that FastText's subword composition is especially beneficial for the informal and abbreviated language of political statements.

c. Statistical significance testing

To confirm that FBA's improvement over BERT is not due to chance, we apply McNemar's test [26]. Let n_{01} denote

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904247

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

cases where only FBA is correct, and n_{10} cases where only BERT is correct. The test statistic is:

$$\chi^2 = (|n_{01} - n_{10}| - 1)^2 / (n_{01} + n_{10}) \quad (5)$$

On the Kaggle test set (n = 6,735), the contingency breakdown yields $n_{01} = 83$ and $n_{10} = 42$. McNemar's test gives $\chi^2 = 11.24$ ($p = 0.0008$), confirming statistical significance at the 99% confidence level. This rules out random variation as the source of FBA's advantage.

Table 4. McNemar's test results — FBA vs BERT-base-uncased (n = 6,735).

Metric	Value	Interpretation
n_{01} (FBA correct, BERT wrong)	83	FBA advantage
n_{10} (BERT correct, FBA wrong)	42	BERT advantage
McNemar χ^2	11.24	—
p-value	0.0008	Significant ($p < 0.01$)

d. Ablation study

Table 5 presents results when individual FBA components are removed. All components contribute positively. The attention mechanism provides the largest single gain (3.0 pp), followed by FastText pre-training (4.7 pp vs. random initialisation), BiLSTM encoding, and dropout regularisation.

Table 5. Ablation study results (Kaggle test set).

Configuration	Accuracy (%)	F1-Score
FBA — full model	97.8	0.977
FBA without attention	94.8	0.947
FBA without BiLSTM (dense substitute)	91.9	0.918
FBA without FastText (random init.)	93.1	0.929
FBA without dropout	95.4	0.952

VII. DISCUSSION

The FBA framework offers several practical strengths. Its inference latency of approximatelyms per article on CPU supports real-time moderation at scale. FastText's subword composition ensures robustness to the deliberate misspellings common in online disinformation. The SHAP module provides operationally useful explanations without degrading classification performance. Cross-dataset evaluation on LIAR confirms that the model generalises beyond a single benchmark, and the McNemar test validates that the performance advantage over BERT is statistically meaningful.

Key limitations include the temporal scope of the Kaggle dataset (2015–2018) and its restriction to English-language political news. Labels reflect source-level reliability rather than claim-level fact-checking, introducing some noise. Adversarial robustness was not assessed. Future work will extend FBA to multimodal inputs, cross-lingual settings using multilingual FastText embeddings, and knowledge distillation from transformer teachers.

VIII. CONCLUSION

This paper presented the FastText-BiLSTM-Attention (FBA) framework for automated fake news detection, integrating subword-aware FastText embeddings, BiLSTM sequential encoding, multi-head self-attention, and SHAP-based post-

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

hoc explainability. On the Kaggle Fake and Real News Dataset, FBA achieved 97.8% accuracy and a macro F1-score of 0.977,outperforming fine-tuned BERT while requiring approximately 90% fewer parameters. Cross-domain evaluation on LIAR confirmed generalisability, and McNemar's test (p = 0.0008) validated statistical significance. Ablation studies confirmed the positive contribution of each architectural component.By producing linguistically coherent per-prediction feature attributions, the SHAP integration enables content moderators and regulators to audit automated decisions — a critical requirement for responsible AI deployment in systems that directly affect public discourse.

REFERENCES

1. S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online. Science**359**, 1146–1151 (2018).
2. F. Doshi-Velez and B. Kim, towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017).
3. M.T. Ribeiro, S. Singh, and C. Guestrin, 'Why should I trust you?': Explaining the predictions of any classifier. Proc. 22nd ACM SIGKDD, 1135–1144 (2016).
4. S.M. Lundberg and S.I. Lee, A unified approach to interpreting model predictions. Advances in NeurIPS**30**, 4765–4774 (2017).
5. C. Castillo, M. Mendoza, and B. Poblete, Information credibility on Twitter. Proc. 20th Int. Conf. WWW, 675–684 (2011).
6. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl.**19**, 22–36 (2017).
7. H. Ahmed, I. Traore, and S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques. Proc. ISDCE, 127–138 (2017).
8. Y. Kim, Convolutional neural networks for sentence classification. Proc. EMNLP, 1746–1751 (2014).
9. S. Hochreiter and J. Schmidhuber, Long short-term memory. Neural Comput.**9**, 1735–1780 (1997).
10. D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate. Proc. ICLR (2015).
11. A. Vaswani et al., Attention is all you need. Advances in NeurIPS**30**, 5998–6008 (2017).
12. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT, 4171–4186 (2019).
13. R.K. Kaliyar et al., FakeBERT: fake news detection in social media with a BERT-based deep learning approach. Multimed. Tools Appl.**80**, 11765–11788 (2021).
14. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist.**5**, 135–146 (2017).
15. P. Bahad, P. Saxena, and R. Kamal, Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Comput. Sci.**165**, 462–469 (2019).
16. A. Thota, P. Tilak, S. Ahluwalia, and N. Lohia, Fake news detection: a deep learning approach. Stanford CS229 Final Report (2018).
17. S. Jain and B.C. Wallace, Attention is not explanation. Proc. NAACL-HLT, 3543–3556 (2019).
18. D.P. Kingma and J. Ba, Adam: a method for stochastic optimization. Proc. ICLR (2015).
19. J. Pennington, R. Socher, and C.D. Manning, GloVe: global vectors for word representation. Proc. EMNLP, 1532–1543 (2014).
20. T. Mikolov et al., Distributed representations of words and phrases and their compositionality. Advances in NeurIPS**26**, 3111–3120 (2013).
21. W.Y. Wang, 'Liar, Liar Pants on Fire': a new benchmark dataset for fake news detection. Proc. ACL, 422–426 (2017).
22. X. Zhou and R. Zafarani, A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv.**53**, 1–40 (2020).
23. V. Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 (2019).
24. A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM. Neural Netw.**18**, 602–610 (2005).
25. N. Ruchansky, S. Seo, and Y. Liu, CSI: a hybrid deep model for fake news detection. Proc. ACM CIKM, 797–806 (2017).
26. Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika**12**, 153–157 (1947).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com