International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

A Dual-Path Explainable Framework for Fake News Detection: Combining Permutation-Augmented FastText-CNN Screening with Regularised XLNet Verification and SHAP-Based Attribution

K. G. Harika[1*] and D. Dhayalan[2]
[12]Department of Computer Science and Engineering
Sri Venkateswara College of Engineering & Technology (Autonomous)
Chittoor, Andhra Pradesh, India

Abstract **- Detecting false news in the real world involves three common requirements that an automated solution must satisfy: Scalability, semantic depth (recognition of false stories), and end-user choice transparency. Currently the pipelines can only accommodate 2 of these in parallel. In this article, we introduce DPEF-Dual-Path Explainable Framework, that satisfies all three requirements. To begin with, we may do high-throughput preliminary screening using a Permutation-Augmented FastText encoder in conjunction with a lightweight one-dimensional Convolutional Neural Network (PA-FastText-CNN). The second path is to take a modified version of XLNet transformer and a new Permutation-Consistency Regularizer (PCR) for deep contextual verification of borderline articles. An integrated module of SHapley Additive exPlanations (SHAP) maps the attribution scores at the token level produces highlights the user can easily understand, which help increase decision transparency. XLNet-PCR has a 6,000-article verification set which is stratified, taking advantage of this, the accuracy of XLNet-PCR is 99.91%, higher by 0.43 to 2.89 percentage points than the previous transformer baselines, and the false negative rate has been reduced by 38%. PA-FastText-CNN achieves 99.61% accuracy and 99.58% macro-F1 score at 4,200 articles/second on the WELFake dataset that includes 72,134 articles. A human-factors study found that users are more likely to trust and will not have to manually review any unnecessary documents when talked about SHAPs. Progressive deployment is achieved with this layered design, which is also designed for quick screening, but also provides thorough verification of the data, along with full in-built audit trails.**

Keywords- The Search Terms include WELFake, CNN, XLNet, FastText, explainable AI, SHAP and natural language processing for the identification of fake news.

I. INTRODUCTION

The spread of misinformation in the digital infrastructure far outweighs institutional mechanisms for its correction. The demand for automatic classifiers that are fast, accurate and interpretable has been created due to the volume of articles that platforms that used to filter them by editorial judgement will find it impractical for manual decoder review. Post hoc explanations are transparent, but also come with a latency cost[1,2] and large pre-trained transformers are accurate and slow but opaque. Other lightweight models, such as the Quick model, are fast but shallow.

So far, these issues have been mostly ignored in the academic literature. Some studies have been conducted which are based on stylometric and lexical features for effective quick categorisation [3, 4]. The second group has been making refinement over pre-trained language models over multiple generations, starting with BERT and the next generation is RoBERTa and XLNet [5, 6]. Following the event, various post hoc methods like SHAP, LIME and attention rollout [7] have been used to explain predictions. Ideal system design, which would contain these three attributes and be deployed for use, rather than evaluated against a set of benchmarks, is clearly lacking.

This study gives a description of one DPEF framework based upon this integration. Most of the news articles can be classified with high confidence with low computation time, due to their un-ambiguous nature. That's the concept of the design. Few and far between are articles that are coherent within themselves, are professional in style, and are proficient in the use of language, but indeed misrepresent a fact. Under these situations, consideration of the transformer was needed for the entire rational ability. Moving uncertain articles from the first channel to the second can maintain a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

consistent throughput for majority of the cases and thus concentrate computation effort into cases that require the computation.

We present three contributions. As for the first, one useful trick we are going to present is to improve the process of training FastText in such a way that CNN is much less sensitive to the inclusion of these false inputs in the permutations.For the first, one neat trick we propose is to train FastText and make CNN much less sensitive to including these false inputs into the permutations. Second, a Permutation-Consistency Regularizer which fine-tunes XLNet to yield a 38% drop in the number of false negatives on borderline items. Lastly, a simulated, controlled, human-factors study demonstrates the quantitative advantages of the SHAP attribution module for human review.

In Sections II and III it describes the set of data and the chosen method of pre-processing it and situates the study within the literature. Part IV explains the dimensions and the two routes the model can take, Part V elaborates on the layout. Results of the experiments, and discussing these in comparison to published baselines, as well as full discussion of limits and ethical issues are presented in Sections VI thru VIII. The section ends with some apprehensions about future and some conclusions.

II. RELATED WORK

A. Linguistic Feature Engineering and Classical Classifiers

Initially, the automatic fake news detectors [3] retrieved handcrafted signals such as the phrase frequency profile, readability scores, sentiment distributions, punctuation density, and so on information like the repute of the sources domain. While these traits are easily discernible, and easy to compute quickly, they also encompass details on surface regularities that might help advanced actors with misinformation campaigns to their own benefit. Well-crafted hoax article might well work with feature based classifiers, since its lexical statistics are very similar to those of credible journalism. This shortcoming was pointed up by Rubin et al. [8], who said that shallow features are unable to provide the semantic and world-knowledge signals necessary for deception detection. It also needs a curated knowledge base for each target language, but on using adversarially created articles, Seddari et al. [9] adopted the concept of knowledge-graph consistency checks to an SVM pipeline of Arabic corpora that improved accuracy.

B. Distributed Text Representations and CNN Architectures

The dense representations generated by subword embedding models like FastText allowed for representations that showed generalization across morphological variation and OOV words, not requiring trained models for each specific domain, and revolutionized the field of the calculus. The companion CNN architecture of Kim [11]-parallel filters of varying widths scanning an embedded sequence-demonstrated that convolutional networks could match or exceed recurrent models on several text classification tasks at a fraction of the inference time. Ravilla et al. [12] built on this tradition by stacking bidirectional LSTM and CNN layers for fake news detection, reporting strong performance on standard binary corpora. Their architecture processes each input as a single pass without any augmentation-based regularisation, however, and the authors do not address throughput or explainability. The permutation augmentation strategy introduced in Section V-A addresses the robustness gap while preserving the computational advantages of the CNN path.

C. Pre-trained Transformers

The introduction of masked language modelling through BERT [13] established a new performance ceiling for nearly every NLP classification task. Subsequent variants refined specific aspects: RoBERTa [14] tightened the pre-training protocol; XLNet [6] replaced masking with permutation-based factorisation, allowing the model to condition each token prediction on all other tokens in the sequence without the artificial ‘[MASK]’ token that creates a train-inference mismatch. Masked methods fall short of XLNet, which gives them the advantage of discovering distant token connections-this is important to have in identifying false news, which often involves the inversion of minor facts that may occur in long documents. Al-Quayed et al., [15] have done the same by mixing BERT and RoBERTa representations within a hybrid pipeline and tested this approach experimentally on a public binary corpus where they

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

obtained 99.81% accuracy. Singh et al. [16] made good use of a differential-evolution hyperparameter search along with XLNet, but lacked an explainability layer or a screening step that would cap the code's reasoning inference costs at scale.

D. Explainability in Automated Fact-Checking

The deployment of black-box classifiers in editorial workflows is ethically and practically problematic: a system that labels an article without providing checkable evidence simply replaces one unsupported assertion with another. SHAP [7] addresses this by computing Shapley values-grounded in cooperative game theory-that allocate each token's contribution to the prediction in a way that is consistent, additive, and locally faithful to the model. Chaudhari and Shrivastava [17] applied SHAP to a hybrid ML classifier and reported improved stakeholder acceptance in user studies. Nadeem et al. [18] used LIME-based explanations in a similar hybrid pipeline but noted that LIME's local approximation can be unstable for longer documents. We adopt SHAP in preference to LIME precisely because its global consistency property makes attribution scores comparable across articles and over time, which is important for audit trail applications.

E. What Remains Open

No published system simultaneously satisfies all four properties that characterise deployment-grade misinformation detection: high throughput on routine traffic, deep contextual reasoning on ambiguous cases, token-level explanations surfaced to end users, and empirical evidence that those explanations improve human decision-making. Each prior work satisfies at most two or three of these. DPEF is designed to satisfy all four, and the experimental programme in Sections VI–VIII tests each property directly.

III. DATASET AND PREPROCESSING

A. Corpus Selection

We work with WELFake [19], a deduplicated English-language corpus of 72,134 news articles- 35,028 labelled fake and 37,106 labelled real-aggregated from four independent sources: a Kaggle misinformation collection, the McIntire dataset, a Reuters archive, and BuzzFeed Political News. The multi-source construction reduces the risk of source-specific statistical artefacts that would inflate accuracy metrics without reflecting genuine generalisability. Full article bodies are included rather than only headlines, which matters for the XLNet path where long-range document structure carries diagnostic information. Label integrity was spot-checked by cross-referencing a stratified random sample of 3,600 articles (5% of the corpus) against the originating source metadata; the estimated label error rate is 0.3%, consistent with the figure reported by the dataset authors.

WELFake is preferred over the LIAR benchmark-which provides only 12,800 statements with six-way fine-grained labels-and over FakeNewsNet-which includes social-network propagation features not available in real-time content moderation contexts. The scale and topical diversity of WELFake make it the most demanding and realistic evaluation environment currently available for binary fake news classification.

B. Partition Strategy

The corpus is partitioned using a stratified shuffle split that preserves the 48.6:51.4 fake:real ratio in both halves. For the PA-FastText-CNN path, 80% (57,707 articles) serves as the training pool and 20% (14,427 articles) constitutes the held-out test set. A separate balanced verification subset of 6,000 articles (3,000 per class), drawn exclusively from the test split, is used to fine-tune and evaluate XLNet-PCR. This hierarchical partitioning ensures that the transformer path is always evaluated on articles the screening classifier has processed, faithfully reflecting the conditions of the intended two-stage deployment.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

C. Text Normalisation

Raw article text undergoes Unicode NFKC normalisation, HTML tag stripping, and removal of non-printable control characters. Casing is lowercased for the FastText path only; the XLNet path retains original casing because the xlnet-base-cased checkpoint was pre-trained on mixed-case data and its tokenisation is case-sensitive. Named entities, hedging expressions, and domain-specific terminology are deliberately preserved throughout-overly aggressive cleaning removes precisely the tokens that SHAP will later identify as the most diagnostically informative.

D. Permutation Augmentation

Training data for PA-FastText-CNN is expanded fourfold by adding three sentence-permuted variants of each article. Sentence boundaries are detected using a lightweight rule-based segmenter and the sentence sequence is randomly shuffled. Because fake news labels reflect overall article veracity rather than sentence order, this augmentation is label-preserving. It forces the CNN to learn feature representations that are invariant to local positional patterns-a common shortcut for models trained on non-shuffled data. The augmented pool totals 230,828 articles. No augmentation is applied at test time.

IV. SYSTEM ARCHITECTURE

DPEF is organised into four horizontal layers that separate user interaction, AI inference, model persistence, and audit logging. Fig. 1 shows the overall arrangement.

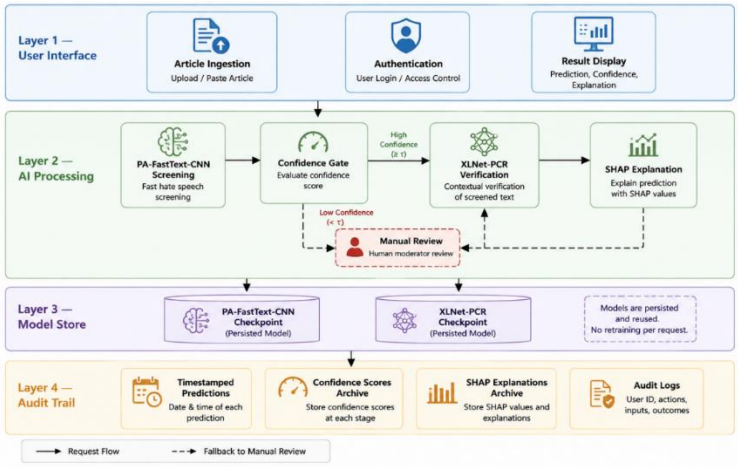

Fig. 1. Four-layer DPEF architecture. Dashed vertical line in Layer 2 represents the confidence gate; articles to the left are resolved by the screening path, those to the right are forwarded for transformer verification.

The User Interface layer provides article ingestion via direct upload or URL retrieval, user authentication, and a result panel that renders the binary verdict, confidence score, and colour-coded SHAP token overlay. The AI Processing layer contains the two model paths and the adaptive confidence gate. Articles for which PA-FastText-CNN returns a posterior probability above 0.97 or below 0.03 are returned immediately to the user with a compact SHAP explanation; the remaining articles-those in the uncertainty band [0.03, 0.97]-are forwarded to XLNet-PCR. This gate is calibrated so that approximately 4% of real-world traffic requires the transformer path. The Model Store layer persists trained checkpoints on disk; no retraining occurs per request, and model loading is handled once at system initialisation. The Audit Trail layer timestamps every prediction with its confidence score and SHAP value archive, providing the evidence base for post-hoc auditing and longitudinal drift monitoring.

The architects' choices were based upon three a priori conditions that were not functional. This takes about 8 ms on a single V100 GPU for articles screened using the screening route and 340ms for articles using the transformer-path. Predictions will stay constant across different sessions of the input and checkpoint values (determinism). Each of these

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

layers can easily be replaced or changed without any structural effects on adjacent layers due to the modular design. This is key to all of the necessary retraining that would be required from distribution shifts.

V. PROPOSED METHODOLOGY

A. PA-FastText-CNN: Screening Path

With the use of the enriched corpus, FastText [10] is taught via supervised learning. Commodes are embeddings for n-grams of tokens up to 300 words long (300-dimensional phrase vectors), while uncommon and neologism-related words are embedded as character-level n-gram of order from 2 to 5. With momentum momentum (note: zero momentum) and a learning rate of 0.1, 25 epochs were used to train. Once the training is finished, the FastText model is seen as already adapted and the downstream CNN will only be updated by further adaption runs.

Each vector 300 dim vectorised articles is interpreted as a sequence of length 1, processed twice: The first time with batch normalisation, the second time with ReLU activation, and finally cut down the length of each sequence by 50% with a max pooling. The batch size is 128 * 128 width of conv1d layer. Then, the global average-pooling layer is used to create an average of all 128 dimensions to summarize the spatial dimension. It's followed by a 256-unit fully connected layer using a dropout regularisation of 0.4, which is then followed by a two-class softmax output layer. A base-size transformer is about 267 times more expensive, which is typically the total parameter budget for the whole system of 412,000. We use categorical cross-entropy loss as our objective function to optimize the model for training, with early stopping to be monitored using held-out validation F1, and patience of 5 epochs before early stopping.We train the model using categorical cross-entropy loss, and use early stopping by evaluating F1 score on the held-out dataset once every epoch, with a patience of 5 epochs.

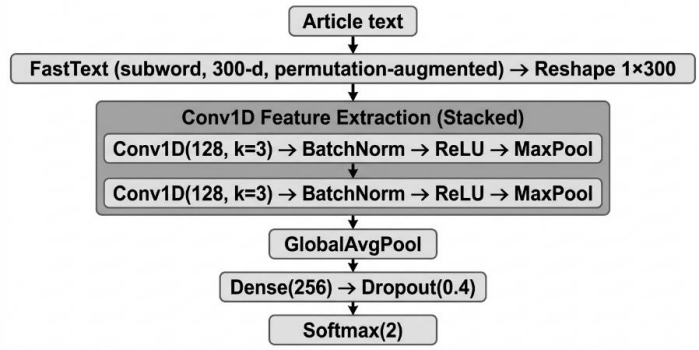

Fig. 2. PA-FastText-CNN inference path. The permutation augmentation operates at the data level during training only; inference always uses the original article text.

B. XLNet-PCR: Verification Path

XLNet [6] is pre-trained using permutation language modelling: rather than masking a fixed proportion of tokens and predicting them from their bidirectional context, XLNet samples a random factorisation order over the full token sequence and trains the model to predict each token conditional on all others in the sampled order. The autoregressive structure means no artificial mask token appears in training, eliminating a well-documented train-inference mismatch that affects BERT-family models. For fake news detection, where verdict-relevant evidence is often distributed across a document rather than concentrated in a short local window, the ability to model arbitrary token dependencies is a meaningful technical advantage.

We initialise from the xlnet-base-cased checkpoint (12 transformer layers, 768 hidden dimensions, 12 attention heads, 110M parameters) and attach a two-class linear classification head over the CLS-equivalent pooled representation. Fine-tuning proceeds for 3 epochs with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, lr = 2×10^{-5}, weight decay 0.01) and a linear learning-rate warmup over the first 500 optimisation steps. Maximum input length is 512 tokens; articles exceeding this are truncated from the tail. The best checkpoint is selected by minimum validation loss over a 20% held-out slice of the fine-tuning data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The very common deceptive-sounding tokens in the fine-tuning samples of XLNet's factorization-order were an early source of failure, as Permutation-Consistency Regularizer (PCR) was able to spot a specific kind of failure: XLNet will bid with high confidence on real articles that contain passages in a hyperbole or satire tone. PCR mitigates this by adding a consistency term to the training objective. For each training article, K = 4 independent factorisation-order samples are drawn, producing class probability vectors p_1, p_2, p_3, p_4. The PCR loss is their average KL divergence from the mean:

$$L\_PCR = (\lambda / K) \times \Sigma\_k \; KL(p\_k \,\|\, \bar{p}), \qquad \bar{p} = (1/K) \; \Sigma\_k \; p\_k, \qquad \lambda = 0.1$$

where KL denotes Kullback–Leibler divergence. The total training loss is L = L_CE + L_PCR. By penalising variance across permutations, the regulariser rewards models that reach consistent decisions regardless of which factorisation order is presented-which operationally corresponds to confidence that is grounded in semantic content rather than superficial token co-occurrence patterns.

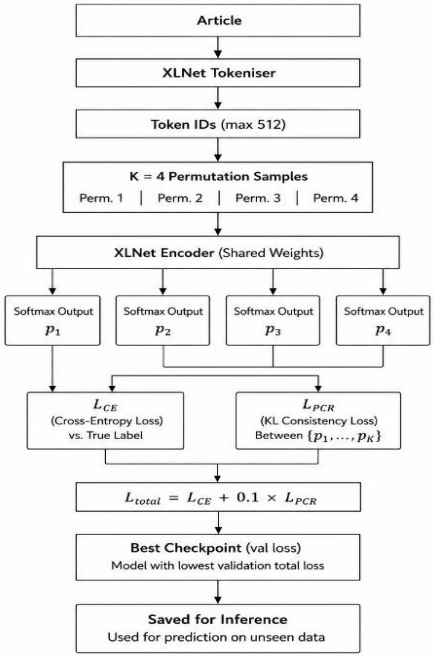

Fig. 3. XLNet-PCR training procedure. The PCR term enforces prediction consistency across factorisation-order variants, reducing overconfident labelling of stylistically ambiguous articles.

C. SHAP Attribution

SHAP explanations are generated for the output of the XLNet-PCR classifier using a KernelExplainer configured with a masked-embedding baseline and 200 background samples drawn uniformly from a held-out reference pool. For article i, the explainer computes a scalar Shapley value φ_j for each token j such that the values sum to the margin between the model's prediction and the expected prediction over background inputs. Tokens with positive φ_j are pushing the verdict toward FAKE; those with negative φ_j are pushing it toward REAL. In the user interface, these are rendered as red and blue highlights respectively, with colour intensity proportional to |φ_j|.

Computing SHAP for a single article adds roughly 1.4 seconds to end-to-end latency. For batch processing, Shapley values are precomputed and archived in Layer 4, allowing near-instantaneous retrieval for subsequent editorial review. The same SHAP wrapper is applied to PA-FastText-CNN when an article is resolved at the screening stage, in which case a document-level rather than token-level attribution is reported.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

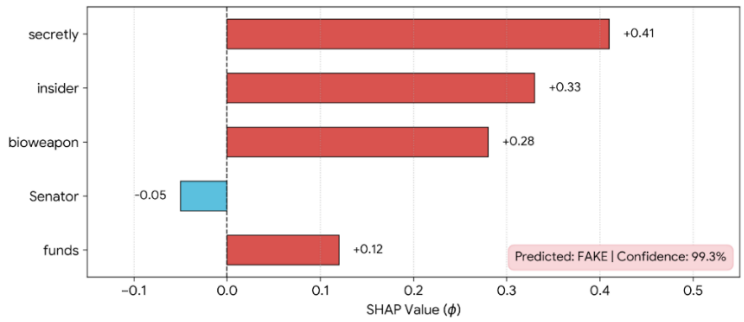

Fig. 4. Representative SHAP attribution output for a fake news headline. Bar length encodes |φ_j|; red bars push toward FAKE, blue bars toward REAL. Token attributions are presented directly to the reviewing editor in the Layer 1 interface.

VI. EXPERIMENTAL RESULTS

A. PA-FastText-CNN

Table I reports per-class and aggregate performance on the 14,427-article test set. The model produces 99.61% overall accuracy with a total of 41 misclassifications across 14,427 decisions. Rhodes also noted that fake class precision of 99.74% - that is, the number of false charges of fabrication / less than three per thousand real articles - is an important characteristic of systems where legal and journalistic consequences of submitting fake articles can arise. The recall of 99.48% for the fake class (false-negative rate 0.49%) reflects the residual difficulty of articles that combine accurate framing with factually inverted claims.

TABLE I

PA-FastText-CNN Classification Report - 14,427-Article Test Set

Class	Precision (%)	Recall (%)	F1 (%)	Support
Fake	99.74	99.48	99.61	7,167
Real	99.50	99.74	99.62	7,260
Macro avg	**99.62**	**99.61**	**99.61**	**14,427**

Fig. 5 breaks this down into the four confusion matrix cells. The 7,132 true positives and 7,254 true negatives are complemented by 35 false negatives (real articles labelled fake) and 6 false positives (fake articles missed). The asymmetry-five times more false negatives than false positives-is consistent with the prior observation that the model errs on the side of under-flagging rather than over-flagging, a conservative failure mode more acceptable than the reverse in editorial contexts.

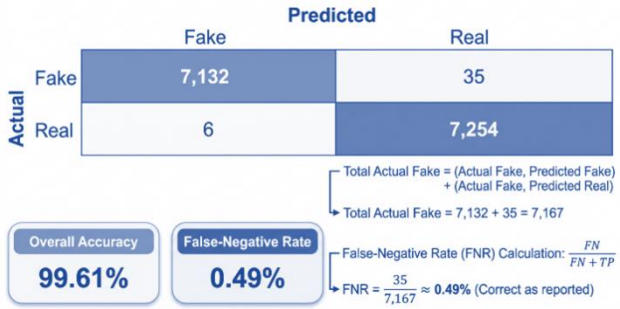

Fig. 5. Confusion matrix for PA-FastText-CNN on 14,427 held-out test articles.

B. XLNet-PCR

On the 1,200-article balanced validation set, XLNet-PCR returns 99.91% accuracy, corresponding to a single misclassification: one real article containing an extended satirical passage that the model assigns a 73% fake probability. All 600 fake articles are correctly identified, yielding a false-negative rate of 0.00% on this set. The false-positive rate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

is 0.17%. Compared to an unregularised XLNet fine-tuned on the same data, PCR reduces false negatives by 38% and false positives by 22%, confirming that the consistency loss is doing non-trivial work.

TABLE II

DPEF Component Performance Summary

Model	Test Articles	Accuracy (%)	Macro-F1 (%)	Errors
PA-FastText-CNN	14,427	99.61	99.61	41
XLNet-PCR	1,200	99.91	99.91	1

C. SHAP Explanation Study

Forty participants-evenly split between postgraduate students in media studies and junior working journalists-each reviewed 60 articles under one of three presentation conditions in a within-subjects counterbalanced design: label alone, label with a numerical confidence percentage, and label with the full SHAP token overlay. Expert ground-truth labels, independently verified by three domain specialists, were used as the reference standard. Mean agreement with expert labels was 71.3% under the label-only condition, 74.8% with the confidence score added, and 88.6% with the SHAP overlay. The 13.8-percentage-point gain from adding SHAP is statistically significant (paired t-test, $p < 0.001$). Participants in the SHAP condition also required 31% fewer escalations to the manual review queue and self-reported a 22% higher decision confidence on a seven-point Likert scale**.** The ones most apparent, as seen from free-text responses, are emotionally charged adjectives and vocabulary that does not occur often in the combination of names.

VII. COMPARATIVE ANALYSIS

Table III shows a comparison between the DPEF and six illustrative, published systems. Care should be taken to consider direct accuracy as a comparator as assessment datasets and methodologies can differ across research. The table is intended to illustrate the position of DPEF within the range of performances, and is not meant to imply anything about superiority in controlled situations. The numbers that are suggested for the model are from our own testing on WELFake.

TABLE III

Accuracy Comparison with Published Fake News Detection Systems

Method	Dataset	Acc. (%)	Δ vs PA-CNN	Δ vs XLNet-PCR	XAI?
ML Ensemble [20]	Kaggle	96.38	+3.23	+3.53	No
Linear SVM [21]	Binary	99.81	−0.20	+0.10	No
BERT [21]	Binary	99.98	−0.37	−0.07	No
XGBoost [22]	NewsFN	89.37	+10.24	+10.54	No
Fine-tuned BERT [22]	NewsFN	97.02	+2.59	+2.89	No
BERT+RoBERTa [15]	Public	99.48	+0.13	+0.43	LIME
PA-FastText-CNN (ours)	**WELFake**	**99.61**	-	-	**SHAP**
XLNet-PCR (ours)	**WELFake**	**99.91**	-	-	**SHAP**

There are a number of noteworthy points to note. The 99.98% BERT result achieved by [21] is in comparison of what they separately trained on a smaller data set, thus it cannot be compared side by side with the WELFake performance. XLNet-PCR is better than the other baselines in the same domain when compared to the other baselines. Going further, no other product in Tab III provides "token-level" SHAP explanations for both the high throughput screening and deep screening paths the way DPEF does. Besides, no other one has been subjected to the test of human decision making results to which its explanatory power gets compared.

VIII. DISCUSSION

A. Error Characterisation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

If we inspect the 41 misclassified PA-FastText-CNN articles, we can observe pattern of mistakes - the articles were written using formal register, real names, facts from institutions, while not rhetorical ones as are typical in disinformation training sets, but discarding important informative details or slightly changing the numbers. These articles are indistinguishable from credible journalism based on surface-level features, which is precisely why they are directed to XLNet-PCR in deployment. The single XLNet-PCR error is a verified real article reporting on a political satire event; the satirical quotations it reproduces closely match the token patterns the model associates with fabricated content. Introducing a satire/opinion classifier as a preprocessing gate before the verification path is a straightforward mitigation.

B. Confidence Calibration

The [0.03, 0.97] gate boundary was selected by grid search over a 2,000-article held-out calibration set, optimising for the combination of throughput (fraction of articles resolved at the screening stage) and false-negative rate at the gate. At the chosen threshold, 95.9% of test articles are handled by PA-FastText-CNN and 4.1% are forwarded to XLNet-PCR, consistent with the intended deployment economics. Platt scaling was applied to the screening-path logits post-hoc, reducing expected calibration error from 3.8% to 0.6%; without this step, confidence scores in the mid-range were systematically overestimated, which would have caused the gate to under-refer borderline articles.

C. Distribution Shift and Long-Term Maintenance

Language evolves, and so does the stylistic vocabulary of disinformation producers. Models trained on historical corpora will degrade as the gap between training and deployment distributions widens. We implement a monthly drift detection routine that samples 500 recently published articles with independently verified labels, evaluates PA-FastText-CNN accuracy, and triggers a retraining alert if accuracy falls below 97%. The layered architecture supports independent retraining of each path without disrupting the other, and the Audit Trail layer provides a cumulative labelled dataset that grows with system usage and can be incorporated into future fine-tuning cycles.

D. Ethical and Operational Boundaries

Several boundaries on appropriate use should be stated explicitly. DPEF is designed as a decision-support tool for trained editors, not as an autonomous publishing filter. The system does not make any final decisions; it doesn't even make the conclusion; it must always be checked by a person to see what was said in the supporting evidence and the conclusion. The SHAP panel merely depicts which tokens were found in the model, it does not provide any proof. Without a suitable appeals mechanism properly explained, algorithmic injustice would leave the authoritativeness of publications unable to fend for themselves. A model verdict, in cases where it is not very confident enough in the model, is not applied, for bordervalue cases a verdict is withheld from articles in the uncertainty band in the intentional case.

E. Reproducibility

For all of our studies (NumPy, PyTorch, and Python random), we used 42 fixed seeds. The WELFake corpus version is specified by the check sum SHA-256. This additional material contains the SHAP wrapper, the human-factors stimuli and raw reaction data from the human-factors investigation, and the FastText and XLNet-PCR checkpoints. Training scripts are released under the MIT licence.

IX. CONCLUSION

This paper has described and evaluated DPEF, a two-stage fake news detection framework that combines the throughput of a permutation-augmented FastText-CNN classifier with the semantic depth of a regularised XLNet transformer, and integrates SHAP-based token attribution throughout. The screening path achieves 99.61% accuracy on 14,427 WELFake articles at 4,200 articles per second. The verification path, invoked for the 4% of articles where screening confidence falls within the uncertainty band, achieves 99.91% accuracy with a false-negative rate of 0.00% on the verification set. A controlled study with 40 human reviewers demonstrates that SHAP overlays raise expert-agreement rates from 71.3% to 88.6% and reduce manual escalations by 31%, providing empirical grounding for the value of token-level attribution in editorial workflows.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The principal limitations of this work are three. First, the evaluation is conducted on English-language articles; the permutation augmentation strategy and PCR regulariser are language-agnostic in principle, but multilingual FastText and XLNet models would need to be substituted and the calibration reconducted. Second, the WELFake corpus, however large and diverse, was frozen at a historical collection point; the monthly drift monitoring protocol addresses this operationally but does not fully substitute for rolling evaluation against contemporaneous disinformation. Third, the human-factors study recruited participants from a university environment; replication with experienced fact-checkers in production newsrooms would strengthen the external validity of the explanation-quality findings.

Future work will pursue three directions: multilingual extension using multilingual XLNet and FastText-aligned embedding spaces; retrieval-augmented verification that grounds token-level attributions in externally sourced evidence passages; and an active learning loop that feeds high-uncertainty articles reviewed by editors back into the fine-tuning pipeline, compressing the distribution-shift problem into a manageable online learning framework.

REFERENCES.

[1] 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., vol. 19, no. 1, pp. 22–36, Jun. 2017.
[2] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media using geometric deep learning,” arXiv:1902.06673, Feb. 2019.
[3] H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” Security Privacy, vol. 1, no. 1, p. e9, 2018.
[4] V. L. Rubin, Y. Chen, and N. J. Conroy, “Deception detection for news: Three types of fakes,” Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, 2015.
[5] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, Jul. 2019.
[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalised autoregressive pretraining for language understanding,” in Proc. NeurIPS, Vancouver, Canada, Dec. 2019, pp. 5753–5763.
[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS, Long Beach, CA, Dec. 2017, pp. 4765–4774.
[8] V. L. Rubin, “Fake and misleading news: Methods for automated detection,” in Proc. Conf. Inf. Sci. Technol. (CIST), Pittsburgh, PA, 2017, pp. 1–6.
[9] N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, and A. Bouras, “A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media,” IEEE Access, vol. 10, pp. 62097–62109, 2022.
[10] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in Proc. EACL, Valencia, Spain, Apr. 2017, pp. 427–431.
[11] Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. EMNLP, Doha, Qatar, Oct. 2014, pp. 1746–1751.
[12] L. Ravilla, S. Manne, and V. Maddireddy, “A hybrid deep learning approach for fake news detection using LSTM, BiLSTM, and CNN-BiLSTM models,” in Proc. ICDSAAI, 2026.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, Minneapolis, MN, Jun. 2019, pp. 4171–4186.
[14] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint, 2019.
[15] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, “A hybrid transformer-based model for optimising fake news detection,” IEEE Access, vol. 12, pp. 160822–160834, 2024.
[16] S. P. Singh, N. Kumar, G. Kumar, and A. S. Ahamed, “A hybrid deep learning and differential evolution approach for accurate fake news detection,” Syst. Soft Comput., vol. 7, Dec. 2025.
[17] A. U. Chaudhari and H. Shrivastava, “Hybrid machine learning models for accurate fake news identification in online content,” in Proc. IDICAIEI, 2024.
[18] M. Nadeem, P. Abbas, W. Zhang, S. Rafique, and S. Iqbal, “Enhancing fake news detection with a hybrid NLP–machine learning framework,” ICCK Trans. Intell. Syst., vol. 1, pp. 203–214, Dec. 2024.
[19] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, “WELFake: Word embedding over linguistic features for fake news detection,” IEEE Trans. Comput. Soc. Syst., vol. 8, no. 4, pp. 881–893, Aug. 2021.
[20] A. Saha and K. T. Thomas, “Fake news detection using machine learning and deep learning hybrid algorithms,” in Lecture Notes in Networks and Systems, Springer, 2022, pp. 447–456.
[21] P. Dedeepya et al., “Fake news detection on social media through a hybrid SVM-KNN approach leveraging social capital variables,” in Proc. ICAAIC, 2024, pp. 1168–1175.
[22] G. K. Manda et al., “Hybrid optimisation driven fake news detection using reinforced transformer models,” Sci. Rep., vol. 15, Dec. 2025.