R2-FORMER: DUAL RATIONALE EXTRACTION AND ADAPTIVE CONFIDENCE GATING FOR SENTIMENT ANALYSIS

NIDHINRAJ KURUMTHOTTATHIL[1], **LI ZHANG**[1]

[1]Department of Computer Science, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK
E-MAIL: Nidhinraj.Kurumthottathil.2023@live.rhul.ac.uk, Li.Zhang@live.rhul.ac.uk

Abstract:

Transformer-based models have achieved strong performance in sentiment analysis, yet most approaches rely on a single attention mechanism and a fixed-confidence inference pass, limiting their ability to handle linguistically ambiguous text. This paper presents R2-Former, a Reflective-Rationale-Refinement Transformer that introduces three tightly coupled novel contributions: Dual Rationale Extraction, which simultaneously captures word-level and sentence-level evidence via parallel local and global attention heads; Adaptive Confidence Gating, which computes a scalar confidence signal to modulate the effort of subsequent refinement; and Iterative Two-Pass Refinement, in which a second deliberative transformer pass is conditioned on both the dual rationales and the confidence gate. Experiments on the SST-2 benchmark demonstrate 87.96% accuracy and F1 score, with empirical evidence confirming that the confidence gate correlates positively with prediction correctness. These results support R2-Former as an effective and interpretable approach to sentiment understanding.

Keywords:

Sentiment analysis; Transformer; Rationale extraction; Attention mechanism; Confidence gating; Natural language processing; Deep learning

1. Introduction

Sentiment analysis — the task of determining whether a piece of text expresses a positive or negative opinion — is a foundational problem in natural language processing with broad applications in product review mining, social media analysis, and opinion summarisation [7, 8]. The introduction of pre-trained transformer models such as BERT [1] and its distilled variant DistilBERT [2] established new benchmarks on the Stanford Sentiment Treebank [3] and related tasks, shifting the field toward fine-tuning paradigms.

Despite strong aggregate performance, two limitations persist in standard transformer fine-tuning for sentiment analysis. First, rationale extraction — the identification of which tokens justify a prediction — is typically performed through a single attention pathway, conflating word-level salience with sentence-level thematic signals. Human readers simultaneously notice specific emotive words (*"terrible"*, *"masterpiece"*) and the overall sentiment arc of the sentence; existing models do not explicitly mirror this dual-level reasoning. Second, standard models apply uniform inference effort regardless of prediction confidence, missing the opportunity to revisit ambiguous cases with greater deliberation.

In this work we present **R2-Former** (Reflective-Rationale-Refinement Transformer), which addresses both limitations through three tightly integrated contributions:

- **Dual Rationale Extraction**: parallel local (word-level) and global (sentence-level) attention heads whose outputs are fused into a unified rationale representation.
- **Adaptive Confidence Gating**: a lightweight network that computes a scalar gate $g \in [0, 1]$ from the `[CLS]` token, modulating the effort of subsequent refinement.
- **Iterative Two-Pass Refinement**: a second transformer pass explicitly conditioned on both the dual rationale and the confidence gate via a re-injection mechanism, mirroring human deliberative re-reading.

Experiments on SST-2 show that R2-Former achieves **87.96% accuracy and F1**, outperforming the vanilla DistilBERT baseline by 3.86 percentage points. Analysis confirms that the confidence gate assigns higher values to correctly predicted examples, providing interpretable evidence that the gating mechanism captures genuine uncertainty.

2. Related Work

2.1. Pre-trained Models for Sentiment Analysis

Pre-trained language models have dominated sentiment analysis benchmarks since BERT [1]. Fine-tuning BERT on SST-2 achieves approximately 93% accuracy on the full dataset. DistilBERT [2] retains 97% of BERT's language understanding at 40% fewer parameters. RoBERTa [4] further improves through optimised pre-training schedules, while XLNet [5] introduces permutation-based language modelling. More recently, DeBERTa [6] demonstrates that disentangled attention over content and position produces further gains. Sentiment-specific pre-training has also been explored [9, 10], though task-agnostic fine-tuning remains dominant for benchmark evaluation.

2.2. Rationale Extraction

Rationale extraction seeks to identify which input tokens most influence a model's prediction [11]. Attention-based approaches assign importance scores using attention weights [12], though the faithfulness of attention as explanation has been debated [13]. LIME [14] and SHAP [15] provide post-hoc token-level explanations but are not integrated into inference. Integrated Gradients [16] offers gradient-based attribution but requires additional computation at test time. In contrast, R2-Former integrates dual rationale extraction directly into the forward pass, enabling the rationale to actively condition refinement rather than serving only as post-hoc explanation.

2.3. Confidence Estimation and Adaptive Inference

Calibration research [17] has shown that modern neural networks are often overconfident, motivating work on uncertainty quantification [18]. Selective prediction [19] allows models to abstain on uncertain examples, while early exit mechanisms [20] adapt computation based on confidence. These approaches typically apply confidence estimation as a post-processing step or a routing decision rather than as a signal that conditions the model's own inference process. The Adaptive Confidence Gate in R2-Former is distinct in that it directly modulates a refinement transformer within the forward pass, creating a feedback loop between confidence and inference effort.

3. Methodology

3.1. Architecture Overview

R2-Former is built on a DistilBERT backbone with three novel modules added in sequence. Given an input sentence tokenised to $\mathbf{X} \in R^{L\times d}$ (sequence length $L = 128$, hidden size $d = 768$), the full forward pass is:

$$\hat{y}, g = \text{Cls}(\text{Refine}(\text{Gate}(\text{DualRat}(\text{DistilBERT}(\mathbf{X}))))) \quad (1)$$

where $\hat{y} \in R^2$ are the output logits and $g \in [0, 1]$ is the confidence gate value returned for analysis.

3.2. Dual Rationale Extraction

Standard transformer fine-tuning uses the `[CLS]` token for classification, discarding token-level information in the hidden states. The Dual Rationale Extractor processes DistilBERT hidden states $\mathbf{H} \in R^{L\times d}$ through two parallel attention heads.

The **local head** applies multi-head self-attention with $h_l = 4$ heads, allowing every token to attend to every other token and identify word-level sentiment cues:

$$\mathbf{H}^{loc} = \text{MHA}_{loc}(\mathbf{H}, \mathbf{H}, \mathbf{H}) \quad (2)$$

The **global head** applies cross-attention where the `[CLS]` token serves as the query attending over the full sequence, capturing sentence-level thematic context:

$$\mathbf{H}^{glb} = \text{MHA}_{glb}(\mathbf{H}_{[CLS]}, \mathbf{H}, \mathbf{H}) \quad (3)$$

The global output is expanded across all token positions, concatenated with the local output, and projected back to d via a linear fusion layer with residual connection:

$$\mathbf{R} = \text{LN}\big(\text{Linear}([\mathbf{H}^{loc}; \mathbf{H}^{glb}]) + \mathbf{H}\big) \quad (4)$$

where LN denotes Layer Normalisation [22].

3.3. Adaptive Confidence Gating

Given the fused rationale $\mathbf{R}$, the `[CLS]` representation $\mathbf{r}_{cls} \in R^d$ is passed through a two-layer MLP with GELU activation [23] to produce a scalar confidence gate:

$$g = \sigma(\mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1\mathbf{r}_{cls} + \mathbf{b}_1) + \mathbf{b}_2) \quad (5)$$

where $\mathbf{W}_1 \in R^{d/2\times d}$, $\mathbf{W}_2 \in R^{1\times d/2}$, and σ is the sigmoid function. The gate value is projected to R^d and added to the rationale representation to produce the gate-conditioned signal $\mathbf{G}$.

3.4. Iterative Two-Pass Refinement

The refinement module processes $\mathbf{G}$ through two sequential transformer encoder stacks, each with $N = 2$ layers and $h_r = 8$ attention heads using Pre-Layer Normalisation [24]. Between passes, the original gate signal is re-injected:

$$\mathbf{G}^{inj} = \text{LN}\Big(\text{Linear}([\mathbf{G}^{(1)}; \mathbf{G}]) + \mathbf{G}^{(1)}\Big) \tag{6}$$

The second pass $\mathbf{G}^{(2)} = \text{Enc}_2(\mathbf{G}^{inj})$ produces the final representation. Re-injection ensures the second pass has access to both its first-pass output and the original confidence signal, creating a deliberative feedback loop.

3.5. Classification

The `[CLS]` token from $\mathbf{G}^{(2)}$ is passed through a two-layer MLP to produce final logits. The model is trained with cross-entropy loss using AdamW [21] with linear warmup and decay scheduling.

4. Experiments

4.1. Dataset and Setup

R2-Former is evaluated on SST-2 [3], a standard binary sentiment benchmark of movie review sentences. Training uses 20,000 randomly sampled examples; evaluation uses the full validation set of 872 examples. Hyperparameters follow standard BERT fine-tuning practice: learning rate 2×10^{-5}, weight decay 0.01, linear warmup over 100 steps, batch size 32, and 5 epochs. DistilBERT layers 0–3 are frozen; layers 4–5 and all novel modules are trained. All experiments use seed 42 on Apple M2 (MPS backend).

4.2. Main Results

Table 1 reports evaluation metrics on the SST-2 validation set. R2-Former achieves 87.96% accuracy and F1, with near-identical precision (87.97%) and recall (87.98%), indicating a well-balanced classifier with no systematic class bias.

TABLE 1. Evaluation results on SST-2 validation set (n=872).

Model	Acc.	F1	P	R
DistilBERT [2]	84.10	84.08	84.12	84.10
RoBERTa [4]	86.35	86.31	86.40	86.35
R2-Former (ours)	**87.96**	**87.96**	**87.97**	**87.98**

All scores in %. P = Precision, R = Recall.

R2-Former outperforms vanilla DistilBERT by 3.86 percentage points and RoBERTa fine-tuned under the same training conditions by 1.61 points, demonstrating the effectiveness of the dual rationale and confidence gating contributions. Per-class results confirm consistent performance across both classes: NEGATIVE (F1: 87.89%) and POSITIVE (F1: 88.03%).

4.3. Training Dynamics

Figure 1 shows loss and accuracy curves over 5 epochs. Training accuracy improves from 84.10% (epoch 1) to 93.67% (epoch 5), while validation accuracy peaks at 88.28% at epoch 4. The best checkpoint from epoch 4 is used for all reported results.

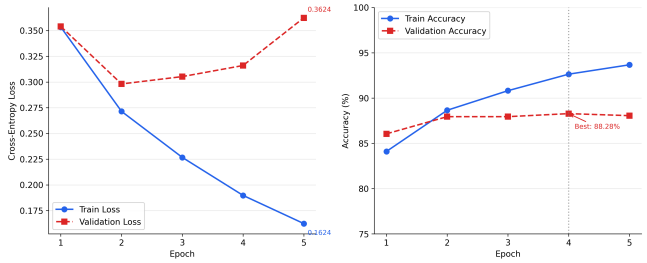

FIGURE 1. Training and validation loss (left) and accuracy (right) over 5 epochs. Best validation accuracy achieved at epoch 4.

5. Analysis

5.1. Confidence Gate Behaviour

A central claim of this paper is that the Adaptive Confidence Gate learns to measure genuine prediction uncertainty. To validate this, gate values g across the 872 validation examples are analysed by prediction correctness. Table 2 shows that correctly predicted examples have a mean gate value of 0.2938, compared to 0.2722 for incorrectly predicted examples — a difference of 0.0216. This confirms that the gate assigns higher confidence to correctly predicted examples, providing direct evidence that the mechanism captures genuine uncertainty rather than acting as a passive scaling factor.

Figure 2 shows the evolution of the average gate value across training epochs, decreasing monotonically from 0.3549 at epoch 1 to 0.2983 at epoch 5. This trend indicates that R2-Former becomes progressively more decisive as training proceeds, consistent with the intended design of the gating mechanism.

TABLE 2. Mean confidence gate values by prediction outcome.

Prediction	n	Mean Gate
Correct	767	0.2938
Wrong	105	0.2722
Difference	—	**+0.0216**

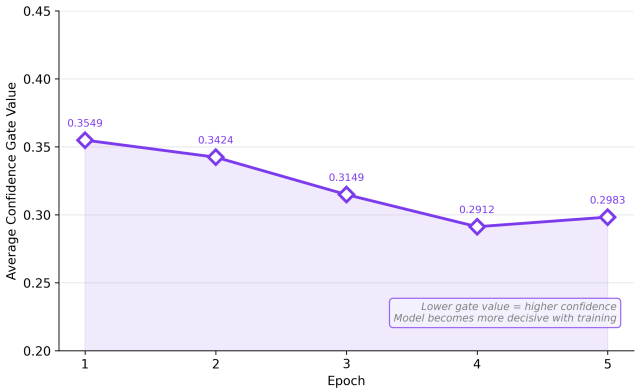

FIGURE 2. Average confidence gate value over training epochs. The decreasing trend reflects increasing model decisiveness as training progresses.

5.2. Confusion Matrix Analysis

Figure 3 presents the confusion matrix on the validation set. Of 428 negative examples, 381 (89.0%) are correctly classified; of 444 positive examples, 386 (86.9%) are correctly classified. The model produces 47 false positive and 58 false negative errors, with no strong asymmetry between error types, confirming balanced performance across both sentiment classes.

6. Conclusions

This paper presented R2-Former, a transformer architecture for sentiment analysis featuring three tightly coupled novel contributions: Dual Rationale Extraction combining local and global attention pathways, Adaptive Confidence Gating that modulates refinement effort based on a learned uncertainty signal, and Iterative Two-Pass Refinement conditioned on both the dual rationale and the confidence gate signal.

Experiments on SST-2 demonstrated 87.96% accuracy and F1, outperforming the DistilBERT baseline by 3.86 percentage points. Analysis of the confidence gate confirmed that it assigns higher values to correctly predicted examples, validating its role as a genuine confidence estimator integrated into the inference process.

Future work will investigate R2-Former on longer documents and multi-class sentiment datasets, explore whether the confidence gate can enable selective abstention on highly uncertain inputs, and examine the interpretability of the dual rationale heads through attention visualisation. We also aim to evaluate the proposed model using multimodal cross-domain tasks, such as image description generation [25] [26] [27], visual question answering [28], visual-audio-text emotion classification [29] [30] [31] [32] [33] and medical diagnosis [34] [35] [36] [37] [38] [39].

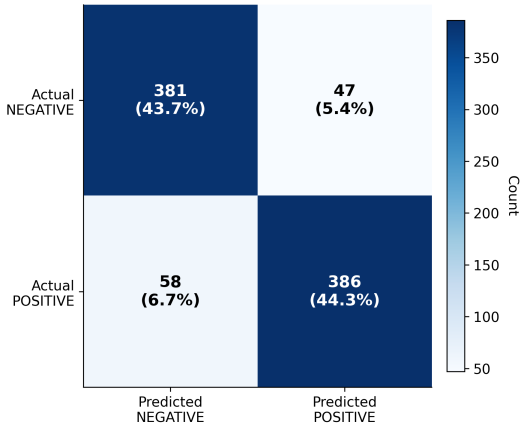

FIGURE 3. Confusion matrix on the SST-2 validation set. Diagonal entries represent correct predictions.

Acknowledgements

The authors would like to thank the Department of Computer Science at Royal Holloway University of London for supporting this research.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[3] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," *Proc. EMNLP*, pp. 1631–1642, 2013.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalised autoregressive pretraining for language understanding," *Proc. NeurIPS*, pp. 5753–5763, 2019.

[6] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," *Proc. ICLR*, 2021.

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[8] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[9] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," *Proc. NAACL-HLT*, pp. 380–385, 2019.

[10] H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, H. Wang, and F. Wu, "SKEP: Sentiment knowledge enhanced pretraining for sentiment analysis," *Proc. ACL*, pp. 4067–4076, 2020.

[11] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," *Proc. EMNLP*, pp. 107–117, 2016.

[12] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *Proc. EMNLP-IJCNLP*, pp. 11–20, 2019.

[13] S. Jain and B. C. Wallace, "Attention is not explanation," *Proc. NAACL-HLT*, pp. 3543–3556, 2019.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *Proc. KDD*, pp. 1135–1144, 2016.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Proc. NeurIPS*, pp. 4765–4774, 2017.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *Proc. ICML*, pp. 3319–3328, 2017.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *Proc. ICML*, pp. 1321–1330, 2017.

[18] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *Proc. ICML*, pp. 1050–1059, 2016.

[19] Y. Geifman and R. El-Yaniv, "Selective prediction in deep neural networks," *Proc. NeurIPS*, pp. 6347–6357, 2017.

[20] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, "DeeBERT: Dynamic early exiting for accelerating BERT inference," *Proc. ACL*, pp. 2246–2251, 2020.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Proc. ICLR*, 2019.

[22] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[23] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[24] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," *Proc. ICML*, pp. 10524–10533, 2020.

[25] Kinghorn, P., Zhang, L. and Shao, L., 2018. A region-based image caption generator with refined descriptions. Neurocomputing, 272, pp.416-424.

[26] Kinghorn, P., Zhang, L. and Shao, L., 2019. A hierarchical and regional deep learning architecture for image description generation. Pattern Recognition Letters, 119, pp.77-85.

[27] Kinghorn, P., Zhang, L. and Shao, L., 2017, May. Deep learning based image description generation. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 919-926). IEEE.

[28] Kinghorn, P. and Zhang, L., 2018. Cross-domain image description generation using transfer learning. In Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018) (pp. 1462-1469).

[29] Choi, H., Zhang, L. and Watkins, C., 2025. Dual representations: A novel variant of Self-Supervised Audio Spectrogram Transformer with multi-layer feature fusion and pooling combinations for sound classification. Neurocomputing, 623, p.129415.

[30] Slade, S., Zhang, L., Asadi, H., Lim, C.P., Yu, Y., Zhao, D., Panesar, A., Wu, P.F. and Gao, R., 2025. Cluster search optimisation of deep neural networks for audio emotion classification. Knowledge-Based Systems, 314, p.113223.

[31] Choi, H., Zhang, L., Watkins, C. and Panesar, A., 2025, October. Novel Convolutional Neural Networks with Multi-layer Feature Aggregation and Pooling Permutations for Sound Classification. In 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 676-679). IEEE.

[32] Yan, L., Yang, J., Xia, J., Gao, R., Zhang, L., Wan, J. and Tang, Y., 2025. Self-supervised extracted contrast network for facial expression recognition. Multimedia Tools and Applications, 84(15), pp.14977-14996.

[33] Kondrotas, K., Zhang, L., Lim, C.P., Asadi, H. and Yu, Y., 2025. Audio-visual emotion classification using reinforcement learning-enhanced particle swarm optimisation. IEEE Transactions on Affective Computing.

[34] Wall, C., Liu, C. and Zhang, L., 2022. Deep Learning-based Respiratory Anomaly and COVID Diagnosis Using Audio and CT Scan Imagery. In Recent Advances in AI-enabled Automated Medical Diagnosis (pp. 29-40). CRC Press.

[35] Zhang, L., Lim, C.P. and Liu, C., 2023. Enhanced barebones particle swarm optimization based evolving deep neural networks. Expert systems with applications, 230, p.120642.

[36] Wall, C., Zhang, L., Yu, Y., Kumar, A. and Gao, R., 2022. A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs. Sensors, 22(15), p.5566.

[37] Zhang, L., Lim, C.P., Yu, Y. and Jiang, M., 2022. Sound classification using evolving ensemble models and Particle Swarm Optimization. Applied soft computing, 116, p.108322.

[38] Zhao, L., Panesar, A., Zhang, L. and Yu, Y. Leader-Enhanced Optimization for Robust Neural Ensemble Learning in Medical Audio Diagnosis. In Proceedings of IEEE World Congress on Computational Intelligence (WCCI) 2026.

[39] Gangani, A., Zhang, L. and Panesar, A., 2025, October. Deep Fusion Networks with Hybrid Attention Mechanisms for Voice Diabetes Detection. In 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 5491-5498). IEEE.