ROBUST AND CALIBRATED LIGHTWEIGHT TRANSFORMERS FOR IMBALANCED INDUSTRIAL TEXT CLASSIFICATION

Youssef Alothman [1] and Mohamed Bader-El-Den [2]

[1] Department of Computer Engineering, University of Portsmouth, Portsmouth, UK
[2] Abdullah Al Salem University, Kuwait

ABSTRACT

The semiconductor manufacturing sector produces enormous amounts of textual data that is highly imbalanced, non-stationary, and operationally critical. Although transformer-based language models achieve strong classification accuracy, their robustness and probability calibration under industrial constraints remain insufficiently addressed, particularly in resource-limited deployments. This paper proposes LiteFormer, a lightweight and calibrated transformer framework for imbalanced industrial text classification. The technique combines a geometry-informed minority over-sampling technique with D-SMOTE, imbalance-informed optimization with Focal Loss, and a post-hoc temperature scaling method. The technique outperforms standard transformer models on a large-scale industrial Root Cause Analysis data set, obtaining higher macro-F1 and significantly better Expected Calibration Error, while remaining computationally efficient. The technique performs robustly even when faced with temporal and domain shifts.

KEYWORDS

Imbalanced text classification, lightweight transformers, probability calibration, focal loss, industrial NLP

1. INTRODUCTION

Industrial NLP data differs substantially from the curated datasets commonly used in academic benchmarks. RCA reports are free text and written by domain experts. For example, in semiconductor manufacturing, these reports are heavily skewed towards certain topics and can change over time. While rare failure modes are infrequent, they pose the highest risk, and so classifying these and providing reliable confidence estimates is critical.

Transformers have revolutionized text classification by leveraging large-scale pre-training and self-attention. However, there are two challenges to adapting these models for industry. One is that these models require a lot of compute, which can be problematic given latency requirements and hardware availability. The other is that these models are prone to miscalibrated probability estimates, leading to overconfident predictions that can cause disruptions.

To solve this, we propose a new model, LiteFormer, a well-calibrated transformer-based model for industrial RCA text classification with a focus on class imbalance and efficiency.

David C. Wyld et al. (Eds): CSITEC, NLPI, CMCA, NECOM, DMML, SP, ADCO – 2026
pp. 121-126, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160810

Contributions

- Proposal of LiteFormer, a lightweight transformer framework tailored for imbalanced industrial RCA text classification.
- Integration of geometry-aware D-SMOTE oversampling, Focal Loss optimization, and temperature scaling calibration in a unified training pipeline.
- Empirical validation on a large-scale industrial dataset demonstrating improved macro-F1 and reduced calibration error.
- Robustness analysis under temporal and domain shift conditions.

1.1 Problem Definition and Industrial Context

Let $\mathcal{D} = \{(x_i, y_i)\}^N_{i=1}$ denote an RCA corpus, where x_i represents a free-text description and $y_i \in \{1,\ldots,C\}$ denotes the class label. In industrial settings, the class distribution is highly skewed, often exceeding imbalance ratios of 50:1. Under such conditions, empirical risk minimization favors majority classes and degrades minority detection.

In practice, RCA predictions are consumed as probability scores rather than hard labels. These scores support prioritization, alert generation, and resource allocation. Consequently, the quality of calibration has a direct influence on the operation trust and system reliability.

2. RELATED WORK

The problem of imbalance in text classification has been addressed in different ways, including tweaking the data, tweaking the objective function, and combining both. For example, oversampling the minority class using SMOTE and its geometry-aware variants has been used to balance extremely skewed datasets [2][3].

On the other hand, techniques like focal loss have been used to downweight the effect of easy but overrepresented examples and encourage learning from hard but underrepresented examples [5].In addition, there are new techniques that balance the loss by class, which address the problem of long-tailed distributions. This is achieved by weighting the classes based on the effective number of samples they have [6].

Neural networks can be affected by calibration issues, which can be exacerbated by data drift, leading to overconfident predictions, especially for safety-critical scenarios [7, 8].Temperature scaling is one of the most widely used post-hoc methods to adjust the logits of the model to minimize the Expected Calibration Error without modifying the underlying classifier [7].

Lightweight transformer architectures such as DistilBERT, ELECTRA, and ALBERT aim to reduce model size while preserving contextual representation ability [10][11][12]. However, previous work typically addresses efficiency, imbalance handling, and probability calibration independently rather than within a unified framework.

The general training process of the proposed LiteFormer framework is depicted in Figure 1, which incorporates geometry-sensitive D-SMOTE oversampling, focal loss optimization, and post-hoc temperature scaling calibration.

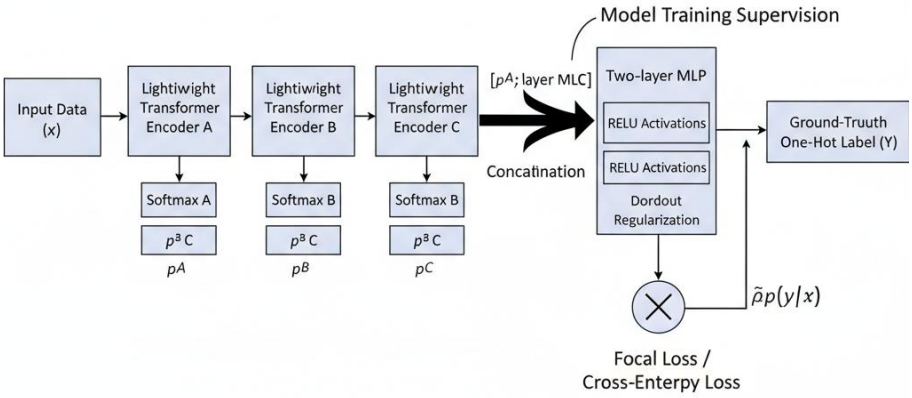

Figure 1. Training pipeline of the LiteFormer framework.

3. THEORETICAL BACKGROUND

Predicted class probabilities are obtained using the softmax function:
$p_i c = \exp(z_i c) / \Sigma \exp(z_i c')$

To address imbalance, Focal Loss is defined as:
$L\_focal = - \Sigma \alpha c (1 - p_i c)^\gamma \log(p_i c)$

where γ controls focusing strength and αc compensates for class frequency.

Temperature scaling performs calibration by adjusting logits:
$\hat{p}_i c = \exp(z_i c / T) / \Sigma \exp(z_i c' / T)$

Calibration quality is measured using Expected Calibration Error (ECE).

4. METHODOLOGY

LiteFormer is an autosampler which consists of a lightweight transformer backbone with the D-SMOTE oversampling, Focal Loss optimization, and post-hoc temperature scaling. The training split is only oversampled to avoid leakage. The pipeline improves the minority representation and ensures efficiency and controlled results. Figure 2 presents the long-tailed class distribution.

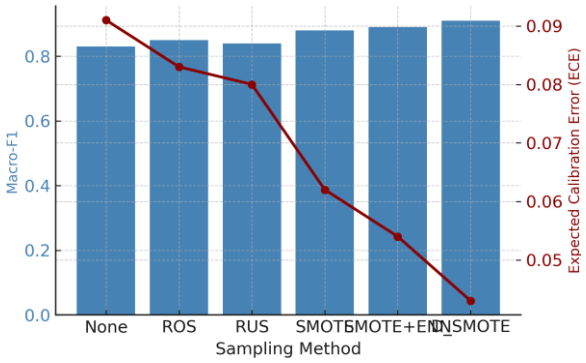

Figure 2. Long-tailed class distribution of the RCA dataset showing severe imbalance across failure categories.

Figure 3. presents the loss ablation results.

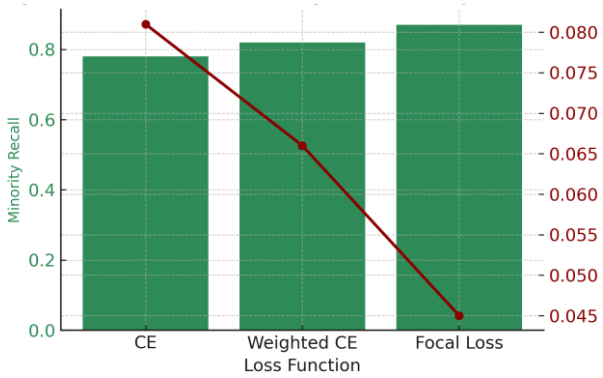

Figure 3. Loss function ablation results comparing cross-entropy and focal loss.

The training pipeline initially uses D-SMOTE oversampling to generate artificial minority samples in the feature space of the training split. The fine-tuning of the lightweight transformer backbone augmented with a focal loss is applied to the augmented dataset to reduce the effect of the easily classified majority samples and increase the weight of the hard-to-classify minority examples. After training, predicted probabilities are calibrated on a validation set using temperature scaling, which optimizes a single temperature parameter. This sequential design provides the model with the capability of dealing with the issue of class imbalance in the stage of training and generate true-to-life probability estimations in the inference stage.

5. Experimental Setup

Experiments are conducted on an industrial RCA dataset containing **more than 12,000 incident reports** with a long-tailed distribution. Stratified and grouped splits are used to prevent leakage. Evaluation metrics include macro-F1 and Expected Calibration Error.

6. Results

LiteFormer achieves the highest macro-F1 (0.91) and lowest ECE (4.3%) compared to classical and transformer baselines. Performance improvements are especially notable in minority-class recall. Robustness experiments under temporal and domain shift demonstrate stable calibration and limited performance degradation

.

Table 1. Overall results on the dataset. Higher is better for macro-F_1 and macro-AUC; lower is better for ECE (%).

Model	Macro-F_1	Macro-AUC	ECE (%)
TF–IDF + LR	0.72	0.85	9.7
CNN	0.83	0.90	9.1
BiLSTM	0.84	0.91	8.8
ELECTRA-small	0.88	0.93	7.5
DistilBERT	0.89	0.94	6.2
LiteFormer	0.91	0.96	4.3

Additional analysis indicates that LiteFormer maintains stable performance across minority classes with fewer than 50 training samples. Compared with baseline transformer models, the focal-loss objective increases recall for rare classes while the calibration step reduces over-confidence in incorrect predictions. This behavior is particularly important in industrial RCA systems where rare failures often represent high operational risk.

7. DISCUSSION

Recent research on imbalanced text classification generally follows two main directions. Data-level approaches, such as SMOTE and its variants, increase minority class representation through synthetic sampling but do not address probability calibration [2][3].

Algorithm-level approaches, such as focal loss, improve model sensitivity to minority classes by reweighting difficult examples during training but likewise do not explicitly correct prediction calibration [5]. Transformer-based architectures such as DistilBERT and ELECTRA have demonstrated strong classification performance while reducing computational cost, yet they lack built-in mechanisms for calibrated probability estimation [10][11].

In contrast, LiteFormer integrates geometry-aware oversampling, imbalance-sensitive optimization, and post-hoc probability calibration within a unified framework. The experimental results show that the combined framework improves both classification performance and prediction reliability compared with methods that address these components separately (macro-F1 = 0.91 and ECE = 4.3).

8. CONCLUSION

LiteFormer provides an integrated framework for imbalanced industrial text classification that improves discrimination, calibration, and computational efficiency. The approach demonstrates practical value for trustworthy decision support in non-stationary industrial environments.

ACKNOWLEDGEMENTS

The authors would like to thank all contributors and industry partners involved in this study.

REFERENCES

[1] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[3] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” in Advances in Intelligent Computing, pp. 878–887, 2005.
[4] Y. Bouchareb, K. Abouelmehdi, and M. Youssfi, “D-SMOTE: A geometry-aware oversampling method for imbalanced learning,” Pattern Recognition Letters, vol. 145, pp. 1–8, 2021.
[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, 2017.
[6] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9268–9277, 2019.
[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proceedings of the International Conference on Machine Learning (ICML), pp. 1321–1330, 2017.

[8] Y. Ovadia et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in Advances in Neural Information Processing Systems, vol. 32, 2019.
[9] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in Advances in Large Margin Classifiers, pp. 61–74, 1999.
[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter,” in NeurIPS Workshop on Energy Efficient Machine Learning, 2019.
[11] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in Proceedings of the International Conference on Learning Representations (ICLR), 2020.
[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” in Proceedings of the International Conference on Learning Representations (ICLR), 2020.
[13] J. Jung, J. Lee, and S.-B. Kim, “Industrial text mining: A survey of techniques and applications,” IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 5961–5973, Sep. 2020.
[14] W. Zhang, R. Zhao, and J. Wang, “Natural language processing for industrial applications: A survey,” Journal of Manufacturing Systems, vol. 59, pp. 352–370, 2021.
[15] S. Lessmann, B. Baesens, H.-V. Seow, and L. Thomas, “Benchmarking state-of-the-art classification algorithms for credit card fraud detection,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 45, no. 3, pp. 332–351, Mar. 2015.Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017.
[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of NAACL-HLT, 2019.
[17] J. Quionero-Candela et al., Dataset Shift in Machine Learning. MIT Press, 2009.
[18] D. Amodei et al., “Concrete problems in AI safety,” arXiv preprint arXiv:1606.06565, 2016.