

Beyond Global Calibration: Analyzing Class-Specific Miscalibration and High-Confidence Errors in Text Classification

Claudio M. V. de Andrade¹, Guilherme Fonseca¹, Celso França¹, Washington Cunha³

Adriana S. Pagano¹, Leonardo Rocha², Marcos André Gonçalves¹

claudio.valiense@dcc.ufmg.br, guilhermefonseca@dcc.ufmg.br, wcunha@unicamp.br

apagano@ufmg.br, lcrocha@ufs.edu.br, mgoncalv@dcc.ufmg.br

¹ Federal University of Minas Gerais, Brazil

² Federal University of São João Del-Rei, Brazil

³ University of Campinas, Brazil

Abstract

Confidence scores are increasingly used to assess uncertainty in text classification, yet aggregate calibration metrics may hide localized reliability failures. We investigate this issue across seven sentiment and topic classification datasets by comparing global and class-specific calibration and analyzing high-confidence errors. Our results show that low global calibration error can coexist with substantial class-specific variation, particularly in multi-class settings. We also find that high-confidence errors are not randomly distributed: they concentrate in specific true–predicted class confusions, with some incorrect predictions receiving confidence scores close to 1.0. These findings show that globally calibrated models can still exhibit systematic uncertainty failures, motivating class-aware and error-level analyses when confidence is used as an indicator of prediction reliability.¹

1 Introduction

Text classification (TC) is a core Natural Language Processing (NLP) task underlying applications such as information retrieval, content moderation, and digital archiving (Sebastiani, 2002; Kowsari et al., 2019). Recent pretrained Language Models (LMs) have substantially improved classification effectiveness (Devlin et al., 2019; Raffel et al., 2020; Bommasani et al., 2021; Zhang et al., 2025). However, in many applications, the predicted label is not enough: confidence scores are also used to decide whether a prediction should be trusted, flagged, or deferred. Assessing the reliability of such confidence estimates is therefore central to uncertainty-aware NLP.

¹The source code and documentation for reproducing our experiments are publicly available on GitHub: <https://github.com/claudiovaliense/emnlp-2026-Beyond-Global-Calibration->.

High predictive performance, however, does not imply reliable confidence. Neural classifiers may be poorly calibrated, overestimating or underestimating the empirical likelihood of correctness (Guo et al., 2017a; Desai and Durrett, 2020). Calibration is commonly evaluated with aggregate metrics such as Expected Calibration Error (ECE) (Naeini et al., 2015) and the Brier score (Guo et al., 2017a). Although useful, these metrics summarize reliability at the dataset level and may hide localized failures. In particular, calibration can vary across classes due to differences in frequency, ambiguity, semantic overlap, and separability (Panchenko et al., 2022; Ao et al., 2023).

A second limitation concerns *high-confidence errors*, i.e., incorrect predictions assigned strong confidence. These cases are especially problematic because they may bypass uncertainty-based safeguards that rely on low confidence to flag unreliable outputs. Moreover, they may concentrate in specific true–predicted class pairs, revealing systematic confusions that global calibration metrics cannot expose. We investigate these issues through two research questions:

- **RQ1:** How is miscalibration distributed across classes despite satisfactory global calibration in text classification?
- **RQ2:** Which true–predicted class confusions systematically concentrate high-confidence errors?

Across seven sentiment and topic classification datasets, we compare global and class-specific calibration and analyze the structure of high-confidence errors. Our results show that low global calibration error can coexist with substantial class-specific variation, particularly in multi-class settings. They also show that high-confidence errors concentrate in specific class confusions, with some incorrect predictions receiving confidence scores

close to 1.0. These findings demonstrate that aggregate calibration provides an incomplete view of uncertainty in text classification and motivate class-aware and error-level analyses when confidence is used as an indicator of prediction reliability.

2 Related Work

Neural text classifiers can achieve strong predictive performance while producing poorly calibrated confidence estimates (Guo et al., 2017a), including in pretrained language models (Desai and Durrett, 2020). Calibration is commonly assessed with aggregate metrics such as Expected Calibration Error (ECE) and the Brier score, which summarize the agreement between predicted confidence and empirical accuracy at the dataset level.

However, aggregate metrics may hide localized reliability failures, much as page-level search scores may overlook differences in the importance of individual page blocks (Fernandes et al., 2007). Prior work has shown that calibration can vary substantially across classes, motivating class-wise calibration analysis (Panchenko et al., 2022; Ao et al., 2023). Recent evidence in text classification further indicates that global calibration can mask severe miscalibration on incorrect predictions (Fonseca et al., 2026). This is particularly relevant for selective prediction and abstention, where confidence scores are used to decide whether a prediction should be trusted or deferred (Xin et al., 2021).

Building on this literature, we analyze the reliability of confidence beyond global calibration. Specifically, we examine whether globally calibrated models still exhibit class-specific miscalibration and whether high-confidence errors concentrate in systematic true–predicted class confusions.

3 Experimental Methodology and Setup

Figure 1 summarizes our experimental pipeline. We evaluate whether confidence reliability failures remain hidden under aggregate calibration by combining two complementary analyses. First, we compare global and class-specific calibration to address **RQ1**. Second, we analyze high-confidence errors according to their true–predicted class pairs to address **RQ2**. All analyses are performed on held-out test predictions, with calibration methods fitted only on validation predictions.

3.1 Datasets

We evaluate seven text classification datasets covering sentiment classification (PangMovie and SST)

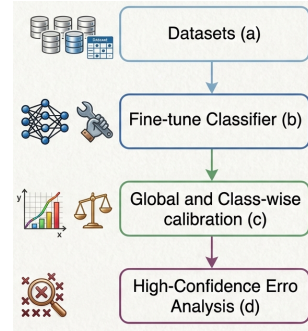


Figure 1: **Experimental methodology.** The pipeline comprises dataset partitioning, classifier training, post-hoc calibration, global and class-wise calibration analysis, and high-confidence error analysis.

and topic classification (AGNews, ACM Digital Library, DBLP, Reddit, and Twitter). These datasets vary in size, number of classes, class distribution, and textual context, including heterogeneous social-media discourse and topic-oriented semantic structure (Cunha et al., 2014; Viegas et al., 2020). This diversity allows us to examine whether calibration failures emerge under different classification settings, including binary and multi-class tasks. Dataset statistics are reported in Appendix A. For each dataset, we use stratified 5-fold cross-validation with a fixed random seed of 42, preserving class distributions across training, validation, and test partitions.

3.2 Classifier and Calibration

We use RoBERTa as the main classifier and additionally compare it with BERT and ModernBERT to verify whether the chosen base model provides a competitive calibration profile. For each fold, the classifier is fine-tuned on the training partition and then produces class probabilities for the validation and test sets. We define prediction confidence as the probability assigned to the predicted class.

Calibration is evaluated using Expected Calibration Error (ECE) and the Brier score. The Brier score (Equation 1) measures the accuracy of probabilistic predictions by computing the mean squared difference between the predicted probabilities and the true class labels. Lower Brier scores indicate better probabilistic predictions. ECE (Equation 2) measures calibration by computing a weighted average of the absolute difference between average accuracy (acc) and average confidence ($conf$) across confidence bins. Lower ECE values indicate better calibration (Pavlovic, 2025).

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (p_{ik} - y_{ik})^2, \quad (1)$$

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2)$$

We compare the original softmax probabilities against two post hoc calibration methods: Platt scaling (Platt, 1999) and isotonic regression (Zadrozny and Elkan, 2002). Calibrators are fitted exclusively on validation predictions and then applied to held-out test predictions. To address **RQ1**, we compute calibration both globally and separately for each class. Class-wise ECE is computed over instances sharing the same true label, allowing us to assess whether low aggregate calibration error hides class-specific reliability variation.

3.3 High-Confidence Error Analysis

To address **RQ2**, we examine whether confident failures are systematically structured across the label space. For each multiclass dataset, we select the 30% most confident test predictions and identify the instances that are incorrectly classified. Binary datasets are omitted, as they allow only two directed error pairs. We then group these high-confidence errors by their true–predicted class pairs and report both their frequency and mean confidence. This analysis reveals whether errors assigned high confidence are isolated cases or whether they concentrate in recurring class confusions.

3.4 Experimental Setup

We fine-tune classifiers using a learning rate of 2×10^{-5} , batch size 64, five epochs, and a maximum input length of 256 tokens (Cunha et al., 2023). Experiments are run on an NVIDIA Tesla P100 GPU with 32GB RAM. Pairwise comparisons are assessed with Wilcoxon signed-rank tests (Demšar, 2006). Results are aggregated across the five folds.

4 Experimental Results and Analyses

We organize the results around the two research questions introduced in Section 1. We first select the classifier and calibration method used in the subsequent analyses. We then examine how miscalibration is distributed across classes despite satisfactory global calibration (**RQ1**) and which true–predicted class confusions systematically concentrate high-confidence errors (**RQ2**).

4.1 Classifier and Calibration Method

Table 1 compares BERT, RoBERTa, and ModernBERT using their original softmax probabilities. Calibration behavior varies across datasets and metrics, with no single architecture consistently achieving the lowest Brier score or ECE. RoBERTa nevertheless provides a competitive calibration profile and better effectiveness (Table 6 in Appendix D). We therefore adopt RoBERTa as the base classifier for the fine-grained calibration and error analyses that follow. Additionally, Table 6 in Appendix D reports the classification effectiveness results, with RoBERTa achieving consistently strong performance across the datasets.

Dataset	BERT	RoBERTa	ModernBERT
Brier Score			
PangMovie	0.213±0.009	0.196±0.011	0.195±0.013
SST	0.236±0.003	0.218±0.019	0.221±0.016
ACM	0.311±0.008	0.334±0.008	0.353±0.014
AGNews	0.128±0.008	0.132±0.008	0.141±0.012
DBLP	0.248±0.012	0.262±0.010	0.274±0.012
Reddit	0.064±0.007	0.050±0.004	0.068±0.008
Twitter	0.191±0.008	0.178±0.008	0.223±0.019
ECE			
PangMovie	0.091±0.006	0.089±0.003	0.086±0.010
SST	0.100±0.001	0.099±0.010	0.097±0.011
ACM	0.051±0.003	0.112±0.002	0.113±0.011
AGNews	0.028±0.005	0.050±0.005	0.053±0.007
DBLP	0.050±0.007	0.092±0.008	0.102±0.006
Reddit	0.027±0.004	0.017±0.002	0.020±0.005
Twitter	0.022±0.007	0.064±0.007	0.079±0.011

Table 1: Brier Score and ECE for BERT, RoBERTa, and ModernBERT using their original softmax probabilities. Values are mean ± 95% confidence interval across five folds.

We next evaluate whether post-hoc calibration improves the reliability of RoBERTa confidence estimates. Table 2 compares the original softmax probabilities with Platt scaling and isotonic regression. Isotonic regression achieves the lowest Brier score and ECE across all seven datasets. The improvement is particularly pronounced for ECE: for example, ECE decreases from 0.112 to 0.032 on ACM, from 0.092 to 0.024 on DBLP, and from 0.064 to 0.026 on Twitter. We therefore use isotonic-calibrated probabilities in the subsequent analyses, allowing us to investigate whether reliability failures remain even after global calibration has been improved.

4.2 RQ1: How is miscalibration distributed across classes despite satisfactory global calibration in text classification?

Table 3 compares global ECE with class-specific ECE after isotonic calibration. Overall, global ECE remains low across all datasets, ranging from 0.014

Dataset	Original	Platt	Isotonic
Brier Score			
PangMovie	0.196±0.011	0.185±0.010	0.165±0.009
SST	0.218±0.019	0.205±0.014	0.184±0.013
ACM	0.334±0.008	0.340±0.008	0.315±0.009
AGNews	0.132±0.008	0.136±0.007	0.122±0.007
DBLP	0.262±0.010	0.272±0.009	0.245±0.007
Reddit	0.050±0.004	0.062±0.005	0.049±0.005
Twitter	0.178±0.008	0.185±0.007	0.168±0.005
ECE			
PangMovie	0.089±0.003	0.038±0.010	0.023±0.013
SST	0.099±0.010	0.027±0.015	0.022±0.004
ACM	0.112±0.002	0.078±0.013	0.032±0.010
AGNews	0.050±0.005	0.046±0.008	0.014±0.005
DBLP	0.092±0.008	0.083±0.004	0.024±0.006
Reddit	0.017±0.002	0.093±0.005	0.015±0.003
Twitter	0.064±0.007	0.077±0.007	0.026±0.009

Table 2: Brier Score and ECE for original, Platt-scaled, and isotonic-calibrated RoBERTa probabilities. Values are mean $\pm$ 95% confidence interval across five folds.

to 0.032. At first glance, these values suggest that the calibrated model provides reliable confidence estimates. However, the class-wise analysis reveals that this aggregate view is incomplete.

For the binary sentiment datasets, the gap between the best- and worst-calibrated classes is small: 0.002 for PangMovie and 0.001 for SST. In contrast, larger gaps appear in the multi-class topic datasets: 0.018 for ACM, 0.013 for AGNews, 0.012 for DBLP, 0.008 for Reddit, and 0.019 for Twitter. These differences indicate that calibration reliability is not uniformly distributed across the label space. A model may be globally well calibrated while still producing more reliable confidence estimates for some classes than for others.

Dataset	Global ECE	Best class-ECE	Worst class-ECE	Gap
PangMovie	0.023	0.021	0.023	0.002
SST	0.022	0.021	0.022	0.001
ACM	0.032	0.001	0.019	0.018
AGNews	0.014	0.005	0.018	0.013
DBLP	0.024	0.004	0.015	0.012
Reddit	0.015	0.001	0.009	0.008
Twitter	0.026	0.007	0.026	0.019

Table 3: **Comparison of global and class-wise calibration across datasets.** Gap denotes the difference between the worst- and best-calibrated classes.

Importantly, this result does not mean that class-wise ECE should be numerically higher than global ECE, since the two summarize calibration at different levels. The relevant observation is the variation across classes: global ECE captures average calibration behavior, but it does not show whether some classes are substantially less reliable than others. This pattern is especially relevant in multi-class settings, where classes may differ

in frequency, semantic overlap, ambiguity, and separability.

Global ECE and class-wise ECE are computed over different probability distributions and binning procedures. Global ECE is computed over all predictions jointly, whereas class-wise ECE is computed separately for each class. Consequently, the global value is not necessarily bounded by the minimum or maximum class-wise ECE. Differences in class frequencies, confidence distributions, and bin occupancy can lead to cases where the global ECE exceeds each individual class-wise value.

Answer to RQ1. Miscalibration is unevenly distributed across classes, particularly in multi-class datasets, even when global calibration error is low. Aggregate calibration therefore obscures substantial class-specific heterogeneity and provides an incomplete characterization of the reliability of confidence across the label space.

4.3 RQ2: Which true–predicted class confusions systematically concentrate high-confidence errors?

We next analyze whether confidence failures are structured at the level of concrete classification decisions. Table 4 reports the most frequent true–predicted class pairs among high-confidence errors in the multiclass datasets, together with their frequency and mean confidence.

Dataset	True	Pred.	N	Conf.
PangMovie	Positive	Negative	26	1.000
PangMovie	Negative	Positive	21	0.991
SST	Negative	Positive	34	0.990
SST	Positive	Negative	24	0.988
ACM	Comp. Meth.	Inf. Syst.	21	0.970
ACM	Inf. Syst.	Comp. Milieux	21	0.975
ACM	Comp. Syst. Org.	Software	20	0.969
ACM	Inf. Syst.	Comp. Meth.	20	0.970
AGNews	Sci./Tech.	Business	17	0.992
AGNews	Business	World	13	0.997
AGNews	World	Business	12	0.991
AGNews	Sci./Tech.	World	8	0.997
DBLP	Theory	Graphics	24	0.989
DBLP	Theory	Soft. Eng.	22	0.978
DBLP	Theory	Comp. Ling.	16	0.987
DBLP	Signal Proc.	Theory	15	0.990
Reddit	Gaming	Build a PC	5	1.000
Twitter	Sport/Gaming	Pop Culture	5	0.996

Table 4: **Most frequent high-confidence errors by true–predicted class pair.** Class names are abbreviated for compactness; N denotes the number of high-confidence errors observed within the selected top-30% and Conf. their mean confidence.

The results show that high-confidence errors are not uniformly distributed across classes. Instead, they recur in specific class confusions.

In ACM, for example, *Computing Methodologies* is predicted as *Information Systems* 21 times with mean confidence 0.970, while the reverse confusion also appears among the most frequent errors. In AGNews, *Sci./Tech.* is predicted as *Business* 17 times with mean confidence 0.992, and *Business*→*World* reaches mean confidence 0.997. DBLP exhibits a similar pattern, with *Theory* repeatedly confused with *Graphics*, *Software Engineering*, and *Computational Linguistics*, all with mean confidence above 0.978. To complement this analysis, Table 8 in Appendix C presents several examples of high-confidence misclassifications.

These findings expose a limitation of using confidence as a direct proxy for prediction reliability. Under a confidence-based criterion, these predictions would appear highly reliable, even though they are incorrect. As a result, mechanisms that flag uncertain outputs mainly through low confidence may fail to detect precisely these cases. Moreover, the recurrence of specific true–predicted class pairs shows that confident failures are systematic rather than isolated.

Together with the results for **RQ1**, this analysis shows that reliability failures can remain hidden at different levels. Global metrics may conceal class-specific variation, while error-level analysis reveals concrete high-confidence misclassifications that remain problematic even after calibration. Therefore, confidence reliability in text classification should be examined not only through aggregate calibration scores, but also through class-aware and error-level perspectives.

Answer to RQ2. High-confidence errors are unevenly distributed across true–predicted class pairs, revealing systematic and directional confusion patterns. Some errors occur at confidence levels close to 1.0, showing that high confidence does not necessarily indicate a reliable prediction. Consequently, confidence-based uncertainty mechanisms may fail to detect class-structured failure modes.

5 Conclusion and Future Work

This work examined whether aggregate calibration metrics are sufficient to assess the reliability of confidence in text classification. Across seven sentiment and topic classification datasets, we found that good overall calibration can still mask important class-level differences, especially in multiclass settings. We also observed recurring high-confidence errors in specific class confusions,

including incorrect predictions with confidence scores close to 1.0.

These findings are important when confidence is used to decide whether predictions should be trusted or flagged for review. They suggest that aggregate calibration metrics should be complemented with class-level and error-level analyses. Future work can extend this analysis to other model families and NLP tasks, error taxonomies focused on high-confidence incorrect predictions (Andrade et al., 2024), embedding-based analyses of class separability (Andrade et al., 2023), combining models to address the classes with the poorest calibration (Andrade et al., 2026), and investigate ways to reduce these class-specific and high-confidence errors.

Ethical Considerations

Although this work focuses on diagnostic analyses of model calibration and high-confidence errors, confidence estimates should not be interpreted as guarantees of correctness. In practical applications, overconfident predictions may lead users to place undue trust in automated systems, particularly in high-stakes domains. Our analysis therefore highlights the importance of evaluating both calibration and error patterns rather than relying solely on model confidence. The datasets considered in this work are publicly available benchmark datasets, and our experiments do not involve collecting or processing new personal or sensitive data.

Acknowledgements

This work was supported by CNPq, CAPES, INCT-TILDIAR (grant no. 408490/2024-1), CIIA-Saude (process No. PPE-00030-21), UNIMED-BH, FAPEMIG, AWS, NVIDIA, Google and FAPESP.

Limitations

This work has several limitations. First, our analysis is restricted to encoder-based pretrained language models (BERT, RoBERTa, and ModernBERT); decoder-only and generative LLMs may exhibit different calibration and high-confidence error patterns. Second, our high-confidence error analysis uses a fixed top-30% cutoff, and different cutoffs may yield different confusion patterns. Third, we consider two standard post hoc calibration methods, Platt scaling and isotonic regression; other calibration approaches may yield different class-wise reliability profiles. absence of temperature scaling (Guo et al., 2017b) is also a limitation of the current experimental comparison, and other calibration approaches may produce different class-wise reliability profiles. Finally, our experiments cover English sentiment and topic classification datasets. Whether the observed class-specific calibration heterogeneity and structured high-confidence errors generalize to other languages, domains, tasks, and model families remains an open question. Our analysis is primarily diagnostic: we characterize these failure modes but leave class-aware mitigation strategies for future work.

References

- Claudio Moisés Valiense De Andrade, Washington Cunha, Guilherme Fonseca, Ana Clara Souza Pagano, Luana De Castro Santos, Adriana Silvina Pagano, Leonardo Chaves Dutra Da Rocha, and Marcos André Gonçalves. 2024. Explaining the hardest errors of contextual embedding based classifiers. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 419–434, Miami, FL, USA. Association for Computational Linguistics.
- Claudio Moisés Valiense de Andrade, Washington Cunha, Davi Reis, Celso França, Wasterman Ávila Apolinário, Luana de Castro Santos, Adriana Silvina Pagano, Leonardo Chaves Dutra da Rocha, and Marcos André Gonçalves. 2026. “call my big sibling (cmbs)” – a confidence-based strategy leveraging instance selection to combine small and large language models for cost-effective text classification. *Journal of the Brazilian Computer Society*, 32(1):1233–1249.
- Claudio M.V. Andrade, Fabiano M. Belém, Washington Cunha, Celso França, Felipe Viegas, Leonardo Rocha, and Marcos André Gonçalves. 2023. On the class separability of contextual embeddings representations – or “the classifier does not matter when the (text) representation is so good!”. *Information Processing Management*, 60(4):103336.
- Shuang Ao, Stefan Rueger, and Advait Siddharthan. 2023. Two sides of miscalibration: identifying over and under-confidence prediction for network calibration. In *Uncertainty in Artificial Intelligence*, pages 77–87. PMLR.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Evandro Cunha, Gabriel Magno, Marcos André Gonçalves, César Cambraia, and Virgílio Almeida. 2014. He votes or she votes? female and male discursive strategies in twitter political hashtags. *PLOS ONE*, 9(1):e87041.
- Washington Cunha, Felipe Viegas, Celso França, Thierison Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2023. A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM CSUR*.
- Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 NAACL-HLT*, pages 4171–4186.
- David Fernandes, Edleno Silva de Moura, Berthier Ribeiro-Neto, Altigran Soares da Silva, and Marcos André Gonçalves. 2007. Computing block importance for searching on web sites. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management, CIKM ’07*, pages 165–174, New York, NY, USA. Association for Computing Machinery.
- Guilherme Fonseca, Gabriel Prenassi, Washington Cunha, Leonardo Chaves Dutra da Rocha, and Marcos André Gonçalves. 2026. When high accuracy hides poor calibration: Rethinking confidence evaluation in transformer-based text classification with balanced brier score. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45888–45900, San Diego, California, United States. Association for Computational Linguistics.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017a. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. [On calibration of modern neural networks](#).

Kamran Kowsari, Kian Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. [Predicting good probabilities with supervised learning](#). In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, page 625–632, New York, NY, USA. Association for Computing Machinery.

Michael Panchenko, Anes Benmerzoug, and Miguel de Benito Delgado. 2022. [Class-wise and reduced calibration methods](#).

Maja Pavlovic. 2025. [Understanding model calibration – a gentle introduction and visual exploration of calibration and the expected calibration error \(ece\)](#).

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Gonçalves. 2020. [Cluhtm - semantic hierarchical topic modeling based on cluwords](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150, Online. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. [The art of abstention: Selective prediction and error regularization for natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Bianca Zadrozny and Charles Elkan. 2002. [Transforming classifier scores into accurate multiclass probability estimates](#). In *Proceedings of the Eighth*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, page 694–699, New York, NY, USA. Association for Computing Machinery.

Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. 2025. Pushing the limit of llm capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528.

A Dataset Statistics

Table 5 summarizes the datasets used in our study. For sentiment analysis, we include two binary datasets, with varying document lengths and degrees of class balance. For topic classification, we rely on five datasets with diverse textual characteristics, ranging from short Twitter posts and news articles to longer academic titles and abstracts. Class distributions vary across datasets, with some exhibiting moderate imbalance, providing a realistic and challenging evaluation scenario.

Dataset	Domain	Docs	Avg Words	Class	Minor Class	Major Class
PangMovieSST	Movie	10662	21.02	2	5331	5331
	Movie	11841	19.18	2	5905	5936
ACM	Article	24897	63.52	11	63	6562
AGNews	News	12760	37.52	4	3190	3190
DBLP	Article	38128	141.43	10	1414	9746
Reddit	Reddit	10000	96.28	11	339	2346
Twitter	Twitter	6997	28.68	6	152	2738

Table 5: Datasets Statistics.

B Calibration

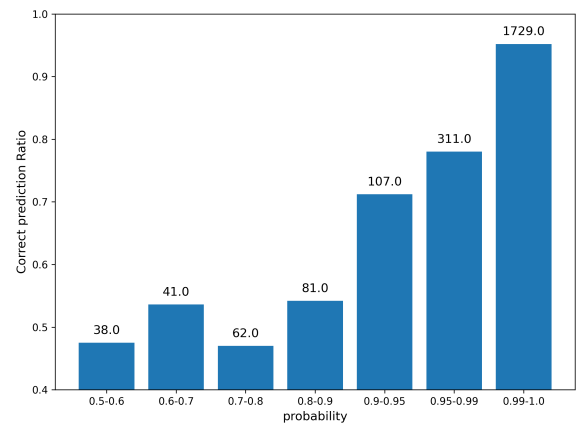


Figure 2: Softmax for SST dataset.

This section describes the calibration procedure applied to the model. First, we obtain the original class probabilities produced by the SLM via the softmax layer. Using the validation set probabilities, we train an isotonic regression calibrator,

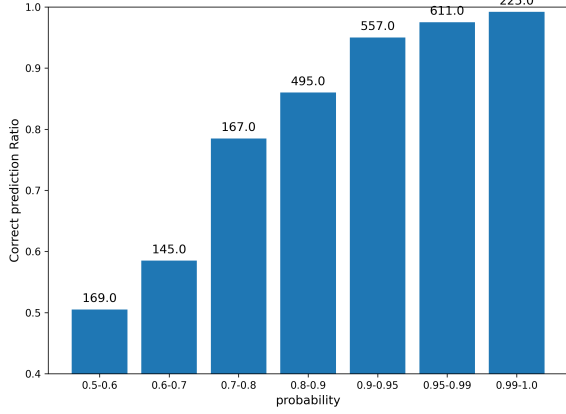


Figure 3: Isotonic for SST dataset.

which is then applied to the test set probabilities to more reliably identify hard test instances. Table 2 shows that isotonic calibration consistently improves the Brier score across datasets, indicating better probability calibration. In addition to quantitative results, we provide a visual analysis for the SST dataset: Figure 2 shows the softmax outputs, while Figure 3 presents the isotonic-calibrated probabilities. The latter exhibits a stronger alignment between predicted confidence (x-axis) and empirical accuracy, confirming the effectiveness of the calibration process.

C Examples of High-Confidence Misclassifications

Table 8 presents representative examples of high-confidence misclassifications identified among the 30% most confident predictions. The table reports the dataset, input text, true class, predicted class, and prediction confidence.

D Model Effectiveness

Table 6 presents the Macro-F1 effectiveness results for BERT, RoBERTa, and ModernBERT. Overall, RoBERTa achieves the best performance, motivating its selection as the base model for the subsequent experiments.

Dataset	BERT	RoBERTa	ModernBERT
PangMovie	87.5±0.7	88.7±0.9	88.6 ±0.7
SST	86.1±0.4	87.3±1.0	87.3±0.9
ACM	69.8±1.8	70.7±1.5	68.1±1.1
AGNews	91.8±0.6	92.2±0.4	91.3±0.6
DBLP	82.1±0.9	81.9±0.7	81.2±0.8
Reddit	94.3±1.1	96.0±0.5	81.2±0.8
Twitter	76.6±4.4	77.5±2.7	72.8±2.2

Table 6: Average Macro-F1 and 95% confidence interval for small language model.

E Class-wise Calibration Analysis

Table 7 reports class-wise ECE, Brier score, and log-loss (Niculescu-Mizil and Caruana, 2005) across ACM, AGNews, and DBLP. The results reveal substantial heterogeneity in calibration across classes within the same dataset. For example, ACM shows markedly higher Brier and log-loss values for Software and Information Systems than for Data and Mathematics of Computing, indicating that calibration quality can vary considerably across classes even under the same model and dataset. These results further support the need for class-wise calibration analysis rather than relying solely on aggregate metrics.

Table 7: Class-wise calibration metrics across datasets. Values are reported as mean (standard deviation).

Dataset	Class	ECE	Brier	Log-loss
ACM	General Literature	0.010 (0.003)	0.089 (0.019)	0.013 (0.001)
ACM	Hardware	0.006 (0.001)	0.112 (0.023)	0.023 (0.001)
ACM	Computer Systems Organization	0.008 (0.002)	0.139 (0.026)	0.031 (0.002)
ACM	Software	0.023 (0.004)	0.250 (0.015)	0.064 (0.003)
ACM	Data	0.005 (0.001)	0.027 (0.011)	0.002 (0.000)
ACM	Theory of Computation	0.006 (0.001)	0.102 (0.027)	0.021 (0.002)
ACM	Mathematics of Computing	0.004 (0.001)	0.083 (0.018)	0.015 (0.001)
ACM	Information Systems	0.016 (0.002)	0.229 (0.017)	0.051 (0.003)
ACM	Computing Methodologies	0.016 (0.003)	0.189 (0.022)	0.047 (0.003)
ACM	Computer Applications	0.006 (0.002)	0.092 (0.039)	0.011 (0.001)
ACM	Computing Milieux	0.012 (0.002)	0.148 (0.024)	0.037 (0.001)
AGNews	World	0.017 (0.002)	0.146 (0.022)	0.028 (0.004)
AGNews	Sports	0.009 (0.001)	0.067 (0.036)	0.009 (0.002)
AGNews	Business	0.014 (0.002)	0.190 (0.020)	0.044 (0.002)
AGNews	Science fiction or Technology	0.018 (0.004)	0.207 (0.051)	0.041 (0.001)
DBLP	Computer Vision	0.011 (0.001)	0.118 (0.023)	0.027 (0.002)
DBLP	Computational Linguistics	0.005 (0.001)	0.073 (0.008)	0.015 (0.001)
DBLP	Biomedical Engineering	0.006 (0.001)	0.092 (0.023)	0.017 (0.001)
DBLP	Software Engineering	0.006 (0.001)	0.091 (0.015)	0.021 (0.001)
DBLP	Graphics	0.006 (0.001)	0.092 (0.030)	0.017 (0.001)
DBLP	Data Mining	0.007 (0.002)	0.123 (0.012)	0.033 (0.001)
DBLP	Security and Cryptography	0.004 (0.001)	0.076 (0.027)	0.015 (0.001)
DBLP	Signal Processing	0.007 (0.002)	0.102 (0.006)	0.027 (0.001)
DBLP	Robotics	0.004 (0.001)	0.067 (0.031)	0.008 (0.000)
DBLP	Theory	0.013 (0.004)	0.260 (0.038)	0.066 (0.002)

Dataset	Text	True	Pred.	Conf.
ACM	an investigation on the use of machine learned models for estimating correction costs mauricio a. de almeida hakim lounis walcelio l. melo	Software	Computing Methodologies	0.974
ACM	formal language, grammar and set-constraint-based program analysis by abstract interpretation patrick cousot radhia cousot	Theory of Computation	Software	0.974
ACM	communicating with icons as computer commands philip rubens robert krull	Software	Information Systems	0.974
AGNews	UMC posts 160 surge in earnings", "TAIPEI United Microelectronics Corp., the world #39;s second-largest supplier of made-to-order semiconductors, said on Wednesday that its third-quarter profit rose 160 percent from a year earlier.	Sci./Tech.	Business	0.992
AGNews	"Chris Stone to leave Novell Inc.", "Vice chairman of the office of the chief executive officer in Novell Inc, Chris Stone is leaving the company to pursue other opportunities in life."	Sci./Tech.	Business	0.992
AGNews	SEAN recognizes China as market economy", "SHANGHAI : The Association of Southeast Asian Nations (ASEAN) has recognised China as a full market economy, Chinese state press reported following a meeting of ASEAN finance ministers in Indonesia.	World	Business	0.994
DBLP	model-based testing. "broadcasting" is an information dissemination process in which a member of a system generates a message communicated to all other members. we model this process by ordered rooted trees and investigate a special class of rooted trees allowing broadcasting from the root to all other vertices of the tree in the minimum time (over all rooted trees with n vertices). we characterize trees from this class ("mbt") and give an algorithm deciding membership in the class. we also present an algorithm to construct all mbt's with a given number of vertices and give a recursive formula to count these trees.	Software Engineering	Theory	0.989
DBLP	generating english summaries of time series data using the gricean maxims. we are developing technology for generating english textual summaries of time-series data, in three domains: weather forecasts, gas-turbine sensor readings, and hospital intensive care data. our weather-forecast generator is currently operational and being used daily by a meteorological company. we generate summaries in three steps: (a) selecting the most important trends and patterns to communicate; (b) mapping these patterns onto words and phrases; and (c) generating actual texts based on these words and phrases. in this paper we focus on the first step, (a), selecting the information to communicate, and describe how we perform this using modified versions of standard data analysis algorithms such as segmentation. the modifications arose out of empirical work with users and domain experts, and in fact can all be regarded as applications of the gricean maxims of quality, quantity, relevance, and manner, which describe how a cooperative speaker should behave in order to help a hearer correctly interpret a text. the gricean maxims are perhaps a key element of adapting data analysis algorithms for effective communication of information to human users, and should be considered by other researchers interested in communicating data to human users.	Data Mining	Computational linguistics	0.989

Table 8: Most frequent high-confidence errors by true–predicted class pair.