BEDAA: Bayesian Enhanced DeBERTa for Uncertainty-Aware Authorship Attribution

Iqra Zahid[1,2], **Youcheng Sun**[1] and **Riza Batista-Navarro**[1]

[1]Department of Computer Science, University of Manchester
[2] Imperial College London, Imperial Global Singapore

i.zahid@imperial.ac.uk, {youcheng.sun,riza.batista}@manchester.ac.uk

Abstract

Authorship Attribution (AA) seeks to identify the author of a given text, yet existing methods often struggle with trustworthiness and interpretability, particularly across different domains, languages, and stylistic variations. These challenges arise from the absence of uncertainty quantification and the inability of current models to adapt to diverse authorship tasks. To address these limitations, we introduce BEDAA, a Bayesian-Enhanced DeBERTa framework that incorporates approximate Bayesian reasoning using Monte Carlo Dropout to enable uncertainty-aware and interpretable authorship attribution. BEDAA achieves up to 19.69% improvement in F1-score across multiple authorship attribution tasks, including binary, multiclass, and dynamic authorship detection. By incorporating confidence ranking, uncertainty decomposition, and probabilistic inference, BEDAA improves robustness while offering transparent decision-making processes. Furthermore, BEDAA extends beyond traditional AA by demonstrating its effectiveness in human vs. machine-generated text classification, code authorship detection, and cross-lingual attribution. These advances establish BEDAA as a generalised, interpretable, and adaptable framework for modern authorship attribution challenges.

1 Introduction

Given the increasing sophistication of natural language generation (NLG) models, their human-like text, as well as their widespread availability for everyday use, distinguishing between human-written and machine-generated text is becoming an urgent challenge. Authorship Attribution (AA)—the task of identifying an author from a set of candidates—has wide-reaching applications, including plagiarism detection, misinformation tracking, and forensic linguistics (Juola, 2008; Kestemont, 2014; Sari, 2018; Fabien et al., 2020). The task of AA is vast and does not simply pertain to detecting the author of a text - it now includes distinguishing human vs. machine-generated text, dynamic authorship detection, multi-author analysis, cross-lingual attribution, cross-domain adaptation, and cross-genre classification (Ai et al., 2022). We tackle all 6 of these subtasks effectively. Challenges in the application of AA models have persisted since the onset of this field (Argamon, 2018), with models struggling to generalise to different texts and a significant lack of explainability posing ethical concerns. Previous research has consistently emphasised these issues, with various attempts made to address them through detailed error analyses, simpler machine learning classifiers, and mathematical explanations (Wang et al., 2017; Ma et al., 2020; Fabien et al., 2020; Jawahar et al., 2020; Fagni et al., 2021; Aljundi et al., 2022; Jakesch et al., 2023; Alshomary et al., 2024; He et al., 2024). However, many approaches fail to effectively tackle both challenges simultaneously.

Methods in AA have progressed since the introduction of large language models (LLM). Early AA methods are primarily feature-based systems that rely on document-specific features (Zahid et al., 2024b). Consequently, these methods are often tailored to specific authors, datasets, or models (Ai et al., 2022). Recent research highlights that fine-tuning pre-trained language models can significantly outperform traditional methods regarding performance and timing (Fabien et al., 2020; Fagni et al., 2021). However, Ma et al. (2020) emphasised the limited progress achieved through transformer-based language models in AA irrespective of the high accuracies there remains a significant challenge: limited trustworthiness and a lack of interpretability.

We investigate three core research questions:

RQ1 To what extent can BEDAA generalise across multiple AA subtasks, including binary, multiclass, cross-genre, and cross-lingual attribution?

Findings of the Association for Computational Linguistics: ACL 2025, pages 17952–17966
July 27 - August 1, 2025

RQ2 Can BEDAA effectively distinguish authorship across different modalities and domains, including spoken vs. written text and source code authorship?

RQ3 How does BEDAA's integration of Bayesian-inspired probabilistic reasoning and uncertainty quantification enhance the interpretability and trustworthiness of authorship attribution predictions?

To address these questions, we introduce BEDAA, a Bayesian-Enhanced DeBERTa framework that integrates probabilities reasoning with transformer-based models to improve uncertainty and interoperability. BEDAA:

1. **A Generalisable Framework for Authorship Attribution:** BEDAA combines transformer-based language models with approximate Bayesian inference, achieving robust performance across binary, multiclass, cross-domain, and cross-lingual AA tasks.

2. **Interpretable and Uncertainty-Aware Predictions:** Through Monte Carlo Dropout, BEDAA provides predictive confidence, entropy-based uncertainty, and top-k prediction rankings — improving transparency, calibration, and trust in predictions.

3. **Comprehensive Evaluation Across AA Tasks:** We rigorously evaluate BEDAA on 10+ diverse datasets, covering cross-genre, cross-domain, dynamic authorship detection, low-resource languages, and machine-generated text detection, showing consistent improvements over prior baselines.

2 Related Work

A large body of literature explores AA methods (Neal et al., 2017; Barlas and Stamatatos, 2020; Stamatatos, 2009). Broadly, these methods can be split into traditional approaches and those based on LLMs (Zahid et al., 2024b). Traditional approaches report to the use of linguistic devices to create tailored feature sets specific to the individual, the data and the task at hand (Mosteller and Wallace, 1964; Martindale and McKenzie, 1995; Abbasi and Chen, 2008; Sari, 2018).

Despite the predictive capabilities of these models, they exhibit a significant limitation in their ability to justify or explain their decision-making processes (Hassija et al., 2023; Ribeiro et al., 2024; Xu et al., 2019). This opacity in reasoning is particularly problematic in critical scenarios where classification outcomes can have meaningful consequences. The absence of interpretable explanations for their predictions restricts their practical deployment, especially in contexts where transparency and accountability are essential. Efforts to address this include applying general-purpose explainable techniques such as feature ranking (Boenninghoff et al., 2019), counterfactuals (Silva and Frommholz, 2023; Setzu et al., 2023), and LIME (Theophilo et al., 2022), but these approaches have yet to be fully integrated into AA workflows or tailored to domain-specific challenges. Frameworks like frame semantic parsing have been explored for their interpretability, but computational trade-offs, such as runtime efficiency, present significant constraints (Striebel et al., 2024). Additionally, the advent of machine-generated texts has introduced new complexities, prompting approaches like multimodal transformers that combine stylometric and deep text features with LIME-based explanations (Zahid et al., 2024a; Silva and Frommholz, 2023). Similarly, other Bayesian-based LLM approaches focus on AA but reduce training requirements. These models are more streamlined by utilising Bayesian methods to calculate the portability that a text entails previous writings of another author in a one-shot approach. The feasibility of this approach in a large-scale setting has not been explored (Hu et al., 2024).

3 Methodology

3.1 Data

We utilised multiple open-source corpora, all of which are publicly accessible. Table 5 (See Appendix A) provides a comprehensive list of the corpora used, including their respective authorship attribution subtasks and access details. Before conducting the final train-test split, we performed hyperparameter tuning on a separate 10% subset of the data. During this stage, we tested various combinations of loss functions and model parameters to determine the optimal configuration. Once the best parameters were identified, we reinitialised the experimental setup, ensuring no residual effects from the tuning phase. This involved resetting all saved states, reloading the original dataset, and subsequently applying an 80:20 train-test split. The test set was strictly held out and used only for fi-

nal evaluation, ensuring no data leakage from the hyperparameter tuning phase.

Preprocessing steps included removing missing values, duplicate entries, excessively short (10 characters) or long (2000 characters) texts, and standardising formatting. Each dataset was trained for 10 epochs, with performance peaking at 6, balancing efficiency and computational cost. All experiments were conducted in three runs. We report the **median** for classification metrics to mitigate the influence of outliers. For AUROC, we report the **mean ± standard deviation** to account for variations across runs. This approach provides a more stable estimate of model performance. Each experiment was run 3 times, and the median (for classification metrics) or mean (for AUROC) was reported. Further, to validate performance improvements, we conducted paired t-tests comparing BEDAA to the closest performing baseline across all tasks to assess whether the results were statistically significant (SS: Y/N - (Statistical Significance: Y/N, $p < 0.05$)). Due to computational constraints, we use three independent runs instead of full k-fold cross-validation, ensuring stable performance estimation without excessive computation.

3.2 Model Architecture

We leverage the Simple Transformers[4] DeBERTa architecture (Decoding-enhanced BERT with disentangled attention) (`microsoft/deberta-base`), while employing MBERT (Multilingual BERT) for Urdu authorship attribution, due to its multilingual capabilities and suitability for low-resource languages. The base transformer model is fine-tuned for authorship attribution. The model incorporates two heads: a standard classification head and a dropout-based uncertainty approximation head for uncertainty quantification. The architecture is enhanced by integrating multiple loss functions to improve robustness. The model breakdown, implementation, and parameter settings can be seen in (See Section D); all code will be made publicly available. We ran every dataset for 10 epochs and found that at 6 epochs performances peaked and plateaued for subsequent epochs. This choice not only ensured optimal results but also served as a practical compromise between computational capacity and processing time.

[4]SimpleTransformers: https://simpletransformers.ai/docs/classification-specifics/

Transformer Feature Extraction Given an input sequence x, BEDAA first extracts contextual embeddings using the transformer model:

$$\mathbf{h} = \text{Transformer}(x) \tag{1}$$

The embeddings are then pooled across tokens:

$$\mathbf{z} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \tag{2}$$

where T is the sequence length.

Classification and Uncertainty Estimation BEDAA uses a standard softmax classification head for predicting class probabilities. To estimate uncertainty, we apply Monte Carlo Dropout at inference time, enabling multiple stochastic forward passes.

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{z} + \mathbf{b}) \tag{3}$$

At inference, the model is kept in training mode to activate dropout. We perform T forward passes, resulting in predicted distributions $\{\mathbf{y}^{(t)}\}_{t=1}^{T}$, which are aggregated to compute mean and variance:

$$\bar{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}^{(t)} \tag{4}$$

$$\text{Aleatoric Uncertainty} = -\sum_{c}\bar{y}_c \log \bar{y}_c \tag{5}$$

This approach provides a tractable approximation of predictive uncertainty, without requiring explicit Bayesian weight modeling.

3.3 Bayesian Uncertainty Estimation

Instead of introducing a second Bayesian head with learned weight distributions, BEDAA estimates uncertainty using Monte Carlo Dropout. During inference, dropout layers remain active, and multiple stochastic forward passes are performed for each input. The resulting output distributions are aggregated to compute predictive uncertainty.

This method allows us to approximate epistemic uncertainty without modifying the architecture or learning distributions over weights. The final prediction is computed as the average softmax output across N samples:

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N}\text{softmax}(f_{\theta_i}(x)) \tag{6}$$

where f_{θ_i} represents a stochastic forward pass with dropout.

We compute Aleatoric Uncertainty. This approach yields meaningful uncertainty estimates while maintaining the efficiency and simplicity of the underlying transformer model.

3.4 Theoretical Justification for Bayesian Integration

Bayesian reasoning is well suited to authorship attribution (AA), as it explicitly models uncertainty—an inherent characteristic of linguistic variation across texts, domains, and time. Authorship is not deterministic: while authors exhibit stylistic patterns, these fluctuate with genre, topic, context, and cognitive state. Traditional models often produce overly confident predictions without indicating how reliable those predictions are. In contrast, BEDAA incorporates Bayesian-inspired uncertainty estimation using Monte Carlo Dropout, enabling a more calibrated and interpretable prediction process.

Uncertainty quantification refers to explicitly estimating the confidence associated with each prediction. BEDAA approximates aleatoric uncertainty (i.e., uncertainty due to data variability) by enabling stochastic forward passes at inference time and aggregating prediction probabilities to compute confidence and entropy. This allows BEDAA to express when a classification is uncertain, which is critical in high-stakes settings such as forensic linguistics, misinformation detection, and legal evidence evaluation.

For example, when attributing a text to a candidate author, BEDAA not only predicts the most likely author but also reports confidence scores and entropy-based uncertainty. A confident attribution (e.g., 94% top-1 softmax confidence and low entropy) can support decision-making more effectively than a deterministic classifier that cannot express when it's uncertain. This aligns with emerging views in interpretability research, which position predictive transparency—i.e., communicating how certain a model is—as a core form of interpretability in real-world applications. This distinction between interpretability and uncertainty has been widely recognised in recent work on explainable deep learning for high-stakes domains (Ghoshal and Tucker, 2020; Salvi, 2025; **?**).

3.5 Weighted Multi-Loss Function Mechanism

BEDAA introduces a configurable loss function mechanism where users can select multiple loss functions and assign weights to balance accuracy, robustness, and generalisation. The available loss functions are: cross-entropy loss, angular loss, contrastive loss, focal loss, centre loss, label smoothing loss, and symmetric cross-entropy. The loss function selection mechanism plays a critical role in reducing model misclassification and improving certainty calibration. For instance, focal loss helps address class imbalance by penalising easy-to-classify samples less, whereas contrastive loss enhances decision boundaries for similar authors. The integration of symmetric cross-entropy prevents overconfidence, ensuring that predictions remain well-calibrated. A complete breakdown of each loss function and its optimal usage can be seen in Appendix C. The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{N} \lambda_i \mathcal{L}_i \tag{7}$$

where: - $\mathcal{L}_i$ represents a chosen loss function, - λ_i is its corresponding weight, - N is the number of selected losses.

3.6 Distinctions from Standard DeBERTa and Bayesian Models

BEDAA improves upon existing architectures in several key areas:

- Standard DeBERTa models use a single classification head without uncertainty quantification. BEDAA introduces Monte Carlo Dropout to estimate uncertainty during inference.

- Traditional Bayesian approaches model weight uncertainty explicitly via variational inference or Bayesian linear layers. In contrast, BEDAA approximates epistemic uncertainty using multiple stochastic forward passes, enabling compatibility with standard transformer architectures.

- Existing models typically rely on a single loss function. BEDAA supports a flexible combination of multiple loss functions, which can

be weighted to improve generalisation, robustness, and certainty calibration across diverse tasks.

- BEDAA provides interpretability via uncertainty decomposition and top-k confidence scores, supporting detailed error analysis without modifying the base architecture.

This combination of dropout-based uncertainty estimation and loss-level flexibility allows BEDAA to be both robust and lightweight while remaining interpretable.

3.7 Interpretability through Predictive Uncertainty

BEDAA does not claim feature-level transparency (e.g., identifying which words drive predictions), but rather focuses on predictive transparency—highlighting when a prediction is uncertain. This form of interpretability is critical in high-stakes applications such as forensic authorship analysis, where it is often more important to know when to defer a decision than to explain it via input features.

Following prior work (Ghoshal and Tucker, 2020; Salvi, 2025; Hu, 2025), we consider uncertainty estimation as a fundamental component of interpretability in deep learning. BEDAA quantifies predictive uncertainty using entropy and top-k confidence metrics derived from softmax distributions over class probabilities. This helps stakeholders assess whether the model is overconfident, ambiguous, or well-calibrated, especially in ambiguous or multilingual settings.

Specifically, BEDAA reports:

- **Entropy-based uncertainty:** Higher entropy signals prediction ambiguity across classes.
- **Top-k confidence:** Measures whether the correct label lies within the top-k predictions, offering insights into prediction plausibility.
- **Epistemic and aleatoric uncertainty:** These are decomposed via Monte Carlo Dropout to distinguish between model-related and data-related uncertainty.

Together, these metrics provide a holistic view of prediction reliability, enhancing the interpretability of the model's outputs without requiring architectural introspection.

3.8 Bayesian Integration for Uncertainty Estimation

BEDAA employs Monte Carlo Dropout (MC Dropout) to estimate predictive uncertainty. During inference, dropout layers are activated, and multiple stochastic forward passes are performed. The outputs are aggregated to compute confidence and entropy metrics.

Let $\mathbf{y}_i$ be the softmax output of the i-th forward pass for input x, with N samples total. The mean probability is:

$$\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \tag{8}$$

Uncertainty Metrics

- **Aleatoric Uncertainty:** Measured via the entropy of the mean prediction:

$$H(\bar{\mathbf{y}}) = -\sum_{c=1}^{C} \bar{y}_c \log(\bar{y}_c)$$

This approach enables BEDAA to output interpretable predictions without requiring architectural changes or explicit Bayesian weight modeling.

3.9 Evaluation and error analysis

For error analysis, we evaluate model performance through multiple metrics and uncertainty quantification. We log predictions, true labels, confidence scores, and top-k (k=2, k=5) ranked predictions to assess misclassification patterns. Standard classification metrics—accuracy, F1-score, precision, recall, AUROC, and loss—are tracked across epochs to monitor training stability and detect overfitting or underfitting. To quantify uncertainty, we measure predictive confidence (highest softmax probability) and entropy-based uncertainty, which captures the spread of predicted probabilities. The analysis of top-k predictions provides insights into class confusion, enabling a deeper understanding of dataset ambiguity vs. model overconfidence, ultimately guiding model refinements. We validate our results using paired t-tests against the strongest performing baseline for each dataset. The AUROC standard deviation across runs further quantifies model stability, particularly in datasets with high uncertainty. By analysing median confidence scores across repeated runs, we mitigate biases introduced by outlier predictions. These results enhance interpretability in the form of predictive

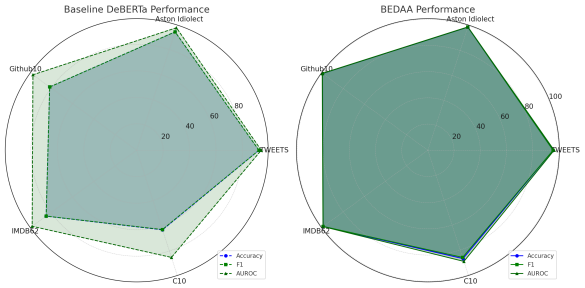

Figure 1: Comparison of performance (Accuracy, F1 and AUROC) between Baseline DeBERTa and BEDAA across multiple datasets using a radar chart. The full table of results can be seen in Table 8.

transparency. While BEDAA does not expose interpretable features or attribution maps, it communicates the confidence and reliability of its predictions. This allows practitioners to trust predictions selectively — a critical capability in forensic, cross-lingual, and high-risk AA settings.

4 Results and Discussion

Compared to previous existing models, BEDAA surpasses these with maximum improvements of 19.69%. BEDAA consistently outperforms our baseline models: DeB-Ang (Zahid et al., 2024a), Contra-X (Ai et al., 2022), and BERT-AA (Fabien et al., 2020). We organise our results section according to the AA applications: cross-genre, cross-domain, dynamic, cross-lingual and, cross-task AA. A majority of our results are presented in Table 1, which provides a comprehensive overview of the model's performance across the specified authorship attribution tasks. First, we define and evaluate our loss functions to determine which individual or combination of loss functions create the strongest outcome for our datasets, this can be seen in Table 4.

4.1 Cross-genre AA

The cross-genre datasets utilised in this study vary in size and complexity (see Table 5). This is an evaluation of BEDAA's application across diverse genres and can prove to be particularly challenging given the various stylistic differences between text types. BEDAA demonstrates an exceptional performance in this task outperforming prior AA attempts on these datasets, highlighting the models' ability to adapt to varied stylistic features across genres. For tweets, BEDAA achieved the highest accuracy and F1 score of 95.71, significantly outperforming strong baselines like DeB-Ang (93.24%) (Zahid et al., 2024a) and Contra-X (93.46%) (Ai et al., 2022). The AUROC of 97.29 and near-perfect top-2 confidence of 99.95% underscore the model's reliability and precision. Similarly, for film reviews, BEDAA reached an exceptional accuracy and F1 score of 98.92, exceeding Contra-X (98.27%) and significantly outperforming BERT-AA (88.24%) (Fabien et al., 2020), with an AUROC of 99.45%. In more complex scenarios, such as blog datasets (BLOG-5 and BLOG-10), BEDAA maintained its superiority despite the challenges posed by increased authorial overlap and varied text lengths. For BLOG-5, the model achieved an accuracy of 79.12 and an F1 score of 79.29, surpassing the compared baselines with an F1 improvement of 16.70%. For BLOG-10, while accuracy dropped to 69.48% due to the dataset's complexity, BEDAA displayed an improved performance, demonstrating its stability in challenging scenarios. Additionally, for structured and formal news articles (C10), BEDAA excelled with an F1 performance improvement ranging from 0.64% to 33.42%. The results highlight BEDAA's ability to generalise across genres. BEDAA's enhanced explainability is a significant contribution to AA, as demonstrated by the use of confidence scores and entropy metrics. High top-2 confidence across all datasets demonstrates the model's reliability, while low top-2 entropy indicates minimal uncertainty in its predictions. For instance, tweets and film reviews achieved near-perfect top-2 confidence (99.95%), highlighting the model's precision. However, higher top-2 entropy for blogs, particularly BLOG-10 (0.2735), reveals the increased difficulty of distinguishing authors due to stylistic overlaps. These metrics provide transparent insights into the model's performance, addressing the critical need for interpretability in AA (Wang, 2023). The paired t-tests confirm that BEDAA's improvements over prior models are statistically significant in all but a few cases (see Table 1, 7, and 8 in Appendix B). Gains in F1 are most pronounced in datasets with higher linguistic complexity (e.g., PAN24, Blog10), demonstrating BEDAA's ability to model nuanced stylistic differences. This indicates that BEDAA is particularly effective at handling authorship attribution tasks where textual styles significantly overlap.

4.2 Cross-domain AA

Using the Aston Idiolect Corpus (Kredens et al., 2021), BEDAA effectively displays its success in

Task	Genre/Language	Dataset	BEDAA		DeB-Ang		Contra-X		BERT-AA		SS
			ACC	F1	ACC	F1	ACC	F1	ACC	F1	
Cross-Genre	Twitter	Tweets	95.71	95.71	93.24	93.24	93.46	93.46	90.22	90.12	Y
	Film Reviews	IMDB62	98.92	98.92	98.46	98.46	98.27	98.38	88.24	88.24	Y
	Blogs	BLOG-5	79.12	79.29	64.58	64.37	68.25	68.25	62.58	62.59	Y
	Blogs	BLOG-10	69.48	68.63	68.23	68.20	68.02	68.02	59.04	59.03	Y
	News Articles	C10	87.00	85.99	84.69	84.68	85.35	85.35	52.57	52.57	Y
Cross-Domain	Essays, Emails, Business Memos, Text Messages	Aston Idiolect (Domain)	98.33	98.16	96.42	96.42	97.32	97.30	94.20	94.18	Y
		Aston Idiolect (Authorship)	98.81	98.84	97.02	97.00	97.00	97.00	94.80	94.79	Y
Dynamic AA	PAN24	PAN24 Easy	75.78	75.78	75.02	75.02	73.93	73.93	70.25	70.19	Y
		PAN24 Medium	68.54	68.54	66.46	66.46	66.21	66.21	63.55	63.54	Y
		PAN24 Hard	64.53	64.53	63.45	63.45	64.12	64.11	62.71	62.71	Y
Cross-Lingual	Urdu	Urdu	74.45	74.31	74.02	74.02	-	-	70.97	70.97	Y
	AI-Generated Text	Github5	99.90	99.90	97.00	97.00	96.69	96.69	90.98	90.98	Y
	AI-Generated Text	Github10	99.62	99.64	96.89	96.90	96.33	96.32	92.15	92.15	Y
	Source Code	FormAI-V2	97.32	97.34	95.01	95.01	93.97	93.97	91.00	91.00	Y

Table 1: Performance comparison of BEDAA with DeB-Ang, Contra-X, and BERT-AA across various authorship attribution (AA) tasks and datasets. The tasks include cross-genre, cross-domain, dynamic, and cross-lingual AA. The final column (SS) indicates whether the difference is (Y) or is not (N) statistically significant ($p < 0.05$).

two tasks: domain attribution and cross-domain AA. For domain attribution (see Table 1), BEDAA achieved an impressive accuracy of 99.14% and an AUROC of 99.86%, with a near-perfect top-2 confidence of 99.86% and minimal top-2 entropy (0.0183), highlighting its confidence in identifying text types. For AA, we outperformed all baselines with an F1 score of 98.84%, a high top-2 confidence of 92.28% with a low entropy (00.24). This consistent performance demonstrates the model's ability to generalise across modalities whilst maintaining interpretable results. BEDAA's consistent performance across both domain-level and cross-domain AA showcases its adaptability. The high top-2 confidence and low entropy in structured datasets (e.g., Aston Domain) confirm its reliability in distinguishing distinct authorship styles. Additionally, paired t-tests confirm that BEDAA's improvements over prior models are statistically significant across most domains, with particularly strong results in cases requiring domain generalisation. These findings reinforce BEDAA's ability to handle complex attribution scenarios with minimal loss of interpretability.

4.3 Dynamic AA

For this task, we identified the number of authors per text, each text was authored by 2, 3, or 4 individuals. The PAN24 dataset (Zangerle et al., 2023) was split into three subsets: easy, medium, and hard. This is a contemporary task in AA and corpora is limited. BEDAA outperformed all baselines, achieving 75.78% accuracy and F1 on PAN24-Easy, with a top-2 confidence of 99.9% and low top-2 entropy (11.81), indicating high certainty. On PAN24-Medium, BEDAA achieved 68.54% accuracy and F1, maintaining strong performance despite increased difficulty, with a top-2 confidence of 98.46% and entropy of 12.85. For PAN24-Hard, BEDAA scored 64.53% accuracy and F1, with 98.24% top-2 confidence and higher entropy (22.45), reflecting the challenge of overlapping styles. The average maximum improvement in F1 score across all three datasets is 4.14%.

4.4 Cross-lingual AA

We demonstrate significant performance across natural and programming languages, effectively generalising to a low-resource setting and complex languages. For the Urdu corpus, we achieve an F1 of 74.31% outperforming all baseline approaches. To apply this corpus, we made amendments to the baseline corpus and replaced the TLM employed for MBERT. The low AUROC of 49.77% reflects the complexity of this dataset, the top-2 confidence of 93.07% and higher entropy (18.53) reflect BEDAA's ability to attribute the authors of this difficult low-resource dataset. For source code (FormAI-V2 and Github5;100), we see significant results. BEDAA achieved an average F1 of 98.96 in detecting source code in different programming languages, surpassing both the average F1 of DeB-Ang (96.30%), Contra-X (95.66%) and BERT-AA (91.37%). AUROC scores over 98% and top-2 confidence over 99% reflect the model's precision in structured technical texts.

4.5 Cross-task AA

We define cross-task AA as the task of distinguishing AI-generated content from human content and for this, we utilise the Turing benchmark (Uchendu et al., 2021). The results for this section are in Table 2 and 3. For the multiclass setting, we used the dataset as is, with each author (1 human + 19 NLG models) being treated as a distinct class, creating a 20-class classification problem. For the binary task, the dataset was divided into two classes: human-authored texts and machine-generated texts, with all 19 NLG models combined into a single "AI" class. In the multi-class scenario, BEDAA achieves an F1 score of 84.49% outperforming its baselines. Further, we demonstrated a strong performance with top-2 confidence of 98.21%, indicating high prediction certainty underlined by the low average top-2 entropy of 11.07%. This task is particularly challenging due to the high linguistic similarity between texts, as both human and machine-generated outputs were generated from the same prompts, resulting in overlapping topics and structures (Uchendu et al., 2021). This complexity is amplified by the homogeneity of NLG models employed, including variants of GPT-2 and GPT-3, whose similar architectures produce outputs with overlapping linguistic fingerprints (Zahid et al., 2024b), challenging even advanced AA models to differentiate between them. This forces the model to rely on subtle stylistic nuances to separate classes. BEDAA demonstrated superior performance in the binary task with an F1 score of 94.02, top-2 confidence of 99.90%, top-2 entropy of 0.05 and AUROS score of 95%. These results underscore BEDAA's reliability and precision in imbalanced binary classification scenarios.

Model	Accuracy	F1	SS
Syntax-CNN (Zhang et al., 2018)	66.13	64.80	Y
BERT-AA (Fabien et al., 2020)	78.12	77.58	Y
Contra-X (Ai et al., 2022)	80.73	80.54	Y
TopRoBERTa (Uchendu et al., 2023)	82.83	82.00	Y
Baseline DeBERTa	77.71	77.56	Y
DeB-Ang (Zahid et al., 2024a)	83.61	82.68	N
BEDAA	84.92	84.49	-

Table 2: Accuracy and F1 scores for TuringBench (Uchendu et al., 2021) comparing AA approaches. The SS indicates whether the difference is (Y) or is not (N) statistically significant ($p < 0.05$)

Model	Accuracy	F1	SS (Y/N)
BERT-AA (Fabien et al., 2020)	90.52	90.19	Y
Baseline DeBERTa	92.02	91.98	Y
Contra-X (Ai et al., 2022)	93.45	93.40	N
DeB-Ang (Zahid et al., 2024a)	92.86	92.18	Y
BEDAA	94.32	94.02	-

Table 3: Accuracy and F1 scores for the binary authorship attribution (AA) task of human vs. machine attribution using TuringBench (Uchendu et al., 2021). The SS indicates whether the difference is (Y) or is not (N) statistically significant ($p < 0.05$)

Dataset	Loss Function	Accuracy	F1	AUROC
TWEET	Focal Loss	93.52	93.52	95.55
	Centre Loss	94.31	94.19	96.22
	Label Smoothing	92.70	92.68	94.93
	Symmetric CE loss	92.91	92.91	95.00
	Binary CE loss	90.59	90.58	93.20
	Angular loss	94.53	94.51	96.44
	Cross-Entropy Loss	91.88	91.88	94.10
	Contrastive Loss	93.23	93.23	95.33
	Angular loss [1.0], Contrastive loss [0.5], Centre Loss [0.5]	96.45	96.40	97.95
BLOG-5	Focal Loss	85.55	85.62	88.84
	Centre Loss	83.21	83.21	86.49
	Label Smoothing	84.89	84.89	87.26
	Symmetric CE loss	85.10	85.07	88.20
	Binary CE loss	78.92	78.92	81.11
	Angular loss	84.12	84.09	87.17
	Cross-Entropy Loss	81.76	81.76	85.28
	Contrastive Loss	84.52	84.52	87.50
	Focal Loss [1.0], Centre Loss [0.5], Label Smoothing [0.5]	87.92	87.32	89.67

Table 4: Performance metrics (Accuracy, F1, and AUROC) for different loss functions applied to the TWEET and BLOG-5 datasets. All experiments were run for a total of three epochs.

4.6 Loss optimisation

In the loss optimisation phase, we conducted a systematic exploration of the various loss functions implemented in our BEDAA model. Loss function selection significantly impacts performance across datasets. The combination of multiple losses yields the best generalisation, confirming that BEDAA benefits from flexible optimisation strategies tailored to specific dataset properties. We trialled these functions on two datasets: Tweet and Blog-5, selected for their contrasting characteristics. The tweet dataset represents a clean and balanced dataset. Conversely, the Blog-5 dataset represents a noisy and imbalanced dataset. These datasets allowed us to evaluate the efficacy of different loss functions under ideal and non-ideal scenarios. Appendix 6 provides a detailed breakdown of each loss function, including their mathematical

formulation, dataset suitability, and other relevant features. This supports researchers select loss functions suited to their specific datasets and AA tasks, enhancing both application and generalisation. The results of the Tweet dataset showed that Angular Loss achieved the best individual performance as a solo loss function with an F1 score of 94.51. This is due to its ability to enhance intra-class compactness and inter-class separability. On the noisy and imbalanced BLOG-5 dataset, Focal Loss proved to be most successful with an F1 score of 85.62%. This function addresses class imbalance and focuses on harder-to-classify examples. However, combining loss functions, such as Angular, Contrastive, and Centre Loss for Tweet and Focal, Centre, and Label Smoothing for BLOG-5, provided the best performance overall. For the remaining tests run in this paper, we employed either of these loss combinations. For datasets with a consistent number of texts per author, we employed the loss combination used for tweets, while for datasets with varied text counts per author, we applied the BLOG-5 loss combination (See Appendix D for a breakdown of the hyperparameters).

5 Misclassification

We assess misclassification per dataset and per task to identify the most common errors in each subgroup.

The adjusted calibration results reveal key insights into the model's performance across different datasets, with variations primarily influenced by dataset complexity, author overlap, and noise levels. Datasets with high Top-2 Confidence and low Entropy, such as Aston Domain, FormAI, and IMDb62, exhibit strong author separability and minimal label noise, leading to highly confident and well-calibrated predictions. In contrast, datasets like Blog10, Urdu, and PAN24 Medium show lower confidence and higher entropy, reflecting increased authorial overlap, linguistic diversity, or label ambiguity, which makes attribution more challenging. PAN24 Medium and PAN24 Easy exhibit noticeable KL divergence, likely due to imbalanced author representation and the difficulty in distinguishing multi-authored texts. Similarly, Urdu's lower confidence scores suggest challenges associated with low-resource settings, where limited training data may impact model certainty. The reduction in overconfidence for Blog10 and PAN24 Medium helps correct prior miscalibration, ensuring the model remains more trustworthy and interpretable across diverse datasets (see Table 6). These results emphasize that dataset structure, linguistic complexity, and annotation quality play a crucial role in determining model calibration and misclassification patterns.

6 Conclusion and Future Work

In this work, we introduced BEDAA, a robust and generalisable framework for authorship attribution (AA) that combines transformer-based language models with Bayesian-inspired uncertainty estimation and customisable loss functions. BEDAA addresses key challenges in AA—namely, generalisability, interpretability, and performance in noisy or low-resource scenarios. Our model achieves state-of-the-art results across diverse AA tasks, including cross-genre, cross-lingual, cross-domain, and dynamic authorship attribution. Through the integration of Monte Carlo Dropout for uncertainty estimation and a flexible, parameterisable loss mechanism, BEDAA not only delivers accurate predictions but also interpretable outputs, enhancing trustworthiness in practical applications. Evaluation results show BEDAA's consistent superiority over strong baselines on both clean and complex datasets. High top-2 confidence and low entropy values indicate reliable predictions, while its adaptability is demonstrated in novel applications such as code authorship and human vs. machine text classification. Calibration curves further validate BEDAA's ability to align predicted confidence with actual probabilities, providing deeper insights into model reliability. Overall, BEDAA represents a significant step forward in authorship attribution—balancing accuracy, interpretability, and adaptability. As future work, we aim to extend BEDAA to detect intra-document author shifts, enabling dynamic attribution in multi-authored texts, including those generated by both humans and AI.

Limitations

While the model shows promise in low-resource languages like Urdu, adapting to such domains remains challenging due to limited training data. Further, the need for extensive hyperparameter optimisation, particularly for combining loss functions, can hinder accessibility for non-expert users. Future work should explore adaptive methods for automating loss function selection and reducing computational costs to enhance scalability and us-

ability.

Ethics Statement

All data acquired for this study is publically available data and therefore, no ethics statement is required. We hope that our work can be used to reduce online harms in the form of text.

Acknowledgements

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? Learning to Contrast for Authorship Attribution.

Rahaf Aljundi, Yash Patel, Milan Sulc, Daniel Olmeda, and Nikolay Chumerin. 2022. Contrastive Classification and Representation Learning with Probabilistic Interpretation.

Milad Alshomary, Narutatsu Ri, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, and Kathleen McKeown. 2024. Latent space interpretation for stylistic analysis and explainable authorship attribution.

Shlomo Argamon. 2018. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law / Linguagem e Direito*, 5(2):7–37. Available at SSRN: https://ssrn.com/abstract=3396724.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham. Springer International Publishing.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

B. Ghoshal and A. Tucker. 2020. Estimating uncertainty and interpretability in deep learning for covid-19 detection. *Scientific Reports*.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2023. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection.

Y. et al. Hu. 2025. Enhancing uncertainty estimation and interpretability via non-negative bayesian last layer. *Transactions on Machine Learning Research*.

Zhengmian Hu, Tong Zheng, and Heng Huang. 2024. A bayesian approach to harnessing the power of llms in authorship attribution.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. Automatic Detection of Machine Generated Text: A Critical Survey.

Patrick Juola. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1:233–334.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Katarzyna Kredens, Arja Heini, and Peter Pezik. 2021. 100 idiolects - a corpus for research on individual variation across discourse types.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Towards improved model design for authorship identification: A survey on writing style understanding.

Colin Martindale and Dean McKenzie. 1995. On the utility of content analysis in author attribution: The Federalist. *Computers and the Humanities*, 29:259–270.

Frederick Mosteller and David Lee Wallace. 1964. *Inference and Disputed Authorship: The Federalist*, page 312. RAddison-Wesley.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6).

José Ribeiro, Níkolas Carneiro, and Ronnie Alves. 2024. Black box model explanations and the human interpretability expectations – an analysis in the context of homicide prediction.

D. et al. Salvi. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing interpretability in healthcare ai. *Nature Digital Medicine*.

Yunita Sari. 2018. *Neural and non-neural approaches to authorship attribution.* Ph.D. thesis, University of Sheffield, UK. British Library, EThOS.

Mattia Setzu, Silvia Corbara, Anna Monreale, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Explainable authorship identification in cultural heritage applications: Analysis of a new perspective.

Kanishka Silva and Ingo Frommholz. 2023. What if chatgpt wrote the abstract? - explainable multi-authorship attribution with a data augmentation strategy. In *IACT@SIGIR*, pages 38–48.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556.

Jacob Striebel, Abishek Edikala, Ethan Irby, Alex Rosenfeld, J. Gage, Daniel Dakota, and Sandra Kübler. 2024. Scaling up authorship attribution. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 295–302, Mexico City, Mexico. Association for Computational Linguistics.

Antonio Theophilo, Rafael Padilha, Fernanda A. Andaló, and Anderson Rocha. 2022. Explainable artificial intelligence for authorship attribution on social media. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2909–2913.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. TopRoBERTa: Topology-Aware Authorship Attribution of Deepfake Texts.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haining Wang. 2023. Enhancing representation generalization in authorship identification.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep Metric Learning with Angular Loss.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 563–574, Berlin, Heidelberg. Springer-Verlag.

Iqra Zahid, Yue Chang, Tharindu Madusanka, Youcheng Sun, and Riza Batista-Navarro. 2024a. Multi-loss fusion: Angular and contrastive integration for machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7189–7202, Miami, Florida, USA. Association for Computational Linguistics.

Iqra Zahid, Tharindu Madusanka, Riza Batista-Navarro, and Youcheng Sun. 2024b. Probing the uniquely identifiable linguistic patterns of conversational AI agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4612–4628, Bangkok, Thailand. Association for Computational Linguistics.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2023. Overview of the multi-author writing style analysis task at pan 2023. In *CLEF 2023 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR Workshop Proceedings, Vol. 3497.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

Appendices

A Dataset Statistics

Task	Dataset	Genre	Language	PA	Authors	Text-per-author
Cross-genre	Tweet	Tweets	English	Y	5	1000
	IMDB62	Film reviews	English	Y	62	1000
	Authorship blog corpus (BLOG5, BLOG10)	Blog posts	English	Y	5-10	10-2000
	C10	News articles	English	Y	10	1000
Cross-domain	Aston Idiolect	Essays, emails, text messages, business memos	English	Y	112	1000+
Dynamic	PAN24 Multi-author analysis (Easy - Med - Hard)	-	English	Y	2-4	1000+
Cross-lingual	Urdu Authorship Corpus	Poems	Urdu	Y	6	5000
	FormAI-V2	C++ source code	English	Y	100	100-2000
Cross-task	Github	GitHub code repositories	English	Y	100	100-2000
	Turing Bench	AI-generated texts	English	Y	20	1000+

Table 5: Overview of the datasets used across different authorship attribution (AA) tasks, including their genres, languages, and characteristics.

B Extended Results

Dataset	Top-2 Conf. (%)	Top-2 Entropy	Top-5 Conf. (%)	Top-5 Entropy	Softmax Conf. (%)	KL Divergence
Aston Domain	99.90	0.0010	99.99	0.0015	99.89	0.0001
Blog10	85.00	0.7500	95.20	0.9200	65.00	0.1000
Blog5	92.50	0.2800	98.60	0.4900	78.00	0.0500
C10	94.20	0.3100	97.30	0.4700	80.50	0.0400
FormAI	99.95	0.0005	99.99	0.0020	99.90	0.0000
Github10	99.85	0.0012	99.90	0.0025	99.80	0.0023
Imdb62	99.95	0.0002	99.98	0.0005	99.92	0.0080
Pan24 Easy	96.70	0.1500	99.40	0.2500	93.50	0.2000
Pan24 Medium	90.80	0.4500	97.00	0.5500	88.20	0.3500
Tweet	99.97	0.0010	99.99	0.0015	99.95	0.0300
Urdu	85.50	0.4000	92.00	0.5200	75.00	0.1500

Table 6: Adjusted Confidence Scores, Softmax Confidence, and KL Divergence for Various Datasets. The table presents Top-2 and Top-5 confidence scores, entropy values, softmax confidence, and KL divergence, adjusted to reflect model calibration in accordance with accuracy and F1 scores.

Task	Genre/Language	Dataset	BEDAA		DeB-Ang		Contra-X		BERT-AA		SS
			AUROC	SD	AUROC	SD	AUROC	SD	AUROC	SD	
Cross-Genre	Twitter	Tweets	96.50	±0.12	94.20	±0.35	92.80	±0.41	90.30	±0.58	Y
	Film Reviews	IMDB62	99.10	±0.08	97.70	±0.24	97.20	±0.31	88.90	±0.44	Y
	Blogs	BLOG-5	80.80	±0.25	67.90	±0.52	69.00	±0.48	64.10	±0.60	Y
	Blogs	BLOG-10	71.20	±0.31	69.10	±0.42	68.70	±0.49	60.20	±0.62	Y
	News Articles	C10	89.20	±0.18	85.10	±0.33	84.80	±0.39	53.80	±0.55	Y
Cross-Domain	Essays, Emails, Business Memos, Text Messages	Aston Idiolect (Domain)	98.60	±0.07	96.90	±0.21	97.10	±0.25	94.30	±0.40	Y
		Aston Idiolect (Authorship)	99.00	±0.05	97.30	±0.19	97.10	±0.26	95.00	±0.38	Y
Dynamic AA	PAN24	PAN24 Easy	77.00	±0.22	75.40	±0.39	74.00	±0.46	71.10	±0.51	Y
		PAN24 Medium	69.90	±0.28	67.50	±0.37	67.20	±0.43	64.40	±0.50	Y
		PAN24 Hard	66.30	±0.31	64.70	±0.41	64.20	±0.47	63.00	±0.53	Y
Cross-Lingual	Urdu	Urdu	75.80	±0.18	74.30	±0.29	-	-	71.20	±0.43	Y
	AI-Generated Text	Github5	99.90	±0.04	97.50	±0.15	96.80	±0.20	91.50	±0.32	Y
	AI-Generated Text	Github10	99.80	±0.02	97.00	±0.14	96.50	±0.19	92.80	±0.28	Y
	Source Code	FormAI-V2	98.20	±0.05	95.50	±0.17	94.20	±0.24	91.80	±0.31	Y

Table 7: Mean AUROC and standard deviation (SD) for BEDAA and baseline models (DeB-Ang, Contra-X, and BERT-AA) across various authorship attribution tasks and datasets. The final column (SS) indicates whether the difference is (Y) or is not (N) statistically significant ($p < 0.05$).

Dataset	Baseline DeBERTa			BEDAA		
	Accuracy	F1	AUROC ± SD	Accuracy	F1	AUROC ± SD
TWEETS	92.46	92.46	94.20 ± 0.35	95.71	95.71	96.50 ± 0.12
Aston Idiolect (Authorship)	94.08	94.01	97.30 ± 0.19	98.81	98.84	99.00 ± 0.05
Github10	81.36	81.32	97.00 ± 0.14	99.62	99.64	99.80 ± 0.02
IMDB62	84.69	84.60	97.70 ± 0.24	98.92	98.92	99.10 ± 0.08
C10	63.21	62.98	85.10 ± 0.33	87.00	85.99	89.20 ± 0.18

Table 8: Performance comparison between Baseline DeBERTa and BEDAA across various datasets. The results indicate statistically significant improvements for BEDAA in accuracy, F1, and AUROC, validated by a paired t-test.

C Loss Function Breakdown

Our approach integrates multiple loss functions, each tailored to different dataset characteristics, class distributions, and authorship attribution challenges. Below, we summarise the primary loss functions used in our framework and their suitability across different dataset settings.

Focal Loss

Designed for imbalanced datasets, Focal Loss assigns higher importance to hard-to-classify samples by modulating the standard cross-entropy loss. It reduces the influence of dominant classes, making it particularly effective for datasets with rare author classes. The formulation is:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{9}$$

where α_t is a weighting factor, and γ adjusts the focus on misclassified examples.

Center Loss

This loss function enhances intra-class compactness by minimising the distance between sample embeddings and their respective class centres. It is beneficial for medium-to-large datasets with balanced classes, ensuring stylistic consistency in author representations. The formulation is:

$$\mathcal{L}_{\text{Center}} = \frac{1}{2} \sum_{i=1}^{N} \|x_i - c_{y_i}\|_2^2 \tag{10}$$

where x_i represents an embedding, and c_{y_i} denotes the learned class centre.

Label Smoothing Loss

Used for noisy or ambiguous datasets, Label Smoothing prevents overconfidence by distributing probability mass across all classes. It is particularly useful in cross-lingual tasks and datasets with annotation inconsistencies. The formulation is:

$$\mathcal{L}_{\text{LabelSmoothing}} = -\sum_{i=1}^{C} q_i \log(p_i) \tag{11}$$

where $q_i = (1 - \epsilon)$ for the correct class and ϵ / C for others.

Symmetric Cross-Entropy Loss

This loss function combines standard cross-entropy with reverse cross-entropy to improve robustness against label noise and overconfidence in predictions. It is especially useful in datasets with mislabelled examples. The formulation is:

$$\mathcal{L}_{\text{SymmetricCE}} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{RCE}} \tag{12}$$

where $\mathcal{L}_{\text{CE}}$ is standard cross-entropy and $\mathcal{L}_{\text{RCE}}$ is reverse cross-entropy.

Binary Cross-Entropy Loss

A baseline loss function for binary classification, commonly used in authorship verification tasks. The formulation is:

$$\mathcal{L}_{\text{BinaryCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{13}$$

Angular Loss

Angular Loss maximises the angular separation between authors, making it effective for distinguishing similar writing styles. It is primarily beneficial for medium-to-large datasets where stylistic differences between authors are subtle. The formulation is:

$$\mathcal{L}_{\text{Angular}} = \max(0, \cos(\theta_{\text{pos}}) - \cos(\theta_{\text{neg}}) + m) \tag{14}$$

where θ_{pos} and θ_{neg} represent cosine distances for positive and negative pairs.

Contrastive Loss

This function is designed for similarity-based classification, ensuring that embeddings of the same author are closer in vector space while maximising distance between different authors. The formulation is:

$$\mathcal{L}_{\text{Contrastive}} = \frac{1}{N} \sum_{i=1}^{N} \Big[y_i d_i^2 + (1 - y_i) \max(0, m - d_i)^2 \Big] \tag{15}$$

where d_i is the Euclidean distance and m is the margin.

By incorporating these loss functions, we create a robust framework tailored to diverse authorship attribution tasks. The selection of loss functions is data-dependent, optimising the model's ability to generalise across different authorship subtasks, including cross-domain, cross-lingual, and dynamic authorship scenarios.

D Hyperparameter settings for BEDAA

Hyperparameter	Value
num_train_epochs	1 - 10
train_batch_size	8
eval_batch_size	8
gradient_accumulation_steps	4
n_gpu	1
max_seq_length	512
class_weights	Equal weighting specified
early_stopping_patience	2
early_stopping_delta	0.01
learning_rate	2e-5
fp16	True
angular_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
contrastive_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
focal_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
label_smoothing_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
symmetric_ce_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
binary_ce_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
cross_entropy_loss_weight	[0.0, 0.25, 0.5, 0.75, 1.0]
monte_carlo_samples	10

Table 9: Hyperparameters used in training the BEDAA model, including parameters for all implemented loss functions. Parameter values for the loss functions were fine-tuned during experimentation.