ISSN: 2093-5374 / E-ISSN: 2093-5382

Stylometric Robustness with DeBERTa: Identifying Authorial Shifts in Adversarial and Collaborative Texts

Riya Sanjesh[1*], and Pamela Vinitha Eric[2]

[1*]Research Sholar, School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. riya.sanjesh@presidencyuniversity.in, https://orcid.org/0009-0008-6617-689X

[2]Professor, School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. pamelavinitha.eric@presidencyuniversity.in, https://orcid.org/0000-0002-4840-6179

Received: October 13, 2025; Revised: December 10, 2025; Accepted: January 13, 2026; Published: March 31, 2026

Abstract

The presence of collaborative and adversarial writing is a significant problem in authorship analysis due to the difficulty of using traditional stylometric methods on small text fragments, intentional stylistic confusion, and redundant topical material. Lexical-statistical, syntactic, and superficial-based methods are especially susceptible to paraphrasing and topic-style confounding, which restricts their use in forensic, educational, and cybersecurity contexts. The paper suggests a DeBERTa-based paragraph-level style-change detector in the form of a binary Natural Language Inference (NLI) problem. The model assesses pairs of consecutive paragraphs to identify the presence of an authorial transition, using the disentangled attention mechanism of DeBERTa to distinguish structural stylistic features of the text and semantic information. The strength is further improved with the aggressive data-augmentation methods such as back-translation, synonym replacement, and sentence shuffling, which mimic adversarial rewriting. Experiments on the PAN 2023 Multi-Author Writing Style Analysis data show that the suggested DeBERTa-v3 model is more accurate, has a higher F1-score, and ROC-AUC in challenging multi-author and topic-stable conditions than classical stylometric classifiers and strong transformer baselines like RoBERTa. Accuracy and F1-score Results in terms of accuracy and F1-score, DeBERTa-v3 obtained 81.3% accuracy and 80.6 F1-score, which is much higher than SVM (69.5%, 66.6) and Logistic Regression (68.0%, 65.4) and is an improvement over RoBERTa-base (75.2%, 74.6). These findings indicate the usefulness of disentangled attention in capturing fine-grained stylistic variation and underscore the practical usefulness of the framework to forensic linguistics, collaborative writing systems, academic integrity detection, and other lightweight applications in actual document monitoring systems.

Keywords: Stylometry, Deberta (Decoding-Enhanced BERT with Disentangled Attention), Authorial Shifts, Text Analysis, Style Change Detection, Adversarial Writing, Forensic Linguistics, Transformer Models.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 574-591. DOI: *10.58346/JOWUA.2026.I1.032*

*Corresponding author: Research Sholar, School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India.

1 Introduction

The boundaries that surrounded authorship in the digital world have been relaxed in the present day (Lazebnik et al., 2025). Team writing applications such as Google Docs, Overleaf and community-based encyclopedias such as Wikipedia have enabled documents to be written by many contributors. Such collaborative written works tend to show a blend of styles posing challenges to the traditional processes of validation. Determining writing style differences within a single work has extended applicability in the different areas. In forensic linguistics, it assists in determining parts of threatening text, suicide note, or phishing email that were authored by certain individuals. Ghostwriting or plagiarism in student essays or research papers can be detected in the educational setting through detection of changes. There may be benefits of managing the internal contributions through the digital news platforms and the community content centers to encourage transparency and responsibility (Wiegmann et al., 2023).

Moreover, when it comes to cybersecurity, the detection of sudden changes can be a way to discover account hijackings, misinformation campaigns, or parallel efforts of several agents within social media. At the same time, the emergence of adversarial stylometry, in which people have deliberately altered their writing style to hide their identity, has presented fresh challenges to the discipline of language forensic analysis, the study of cybercrime, and authorship attribution (Wang et al., 2024).

Under such circumstances, proper determination of small discrepancies in expression is essential towards maintaining content credibility, author accountability and text integrity. It is intrinsically hard to tell the changes in authorship in one piece of work due to a number of reasons. With multiple contributors, the participants can follow similar structures, designs, or editing rules, hiding their own distinctive styles (Potthast et al., 2018). This issue is further compounded by the fact that topic and style of writing are mixed. As an example, a change of topic can be confused with a change of authorship by classifiers when different writers discuss a diverse range of topics. Determining style is further complicated by adversarial conditions since individuals may be intentional to affect the compositional style through the rewording of tools, alteration of syntactic patterns or by imitating others. This is made more difficult by the fact that when dealing with short textual fragments like paragraphs, there is no much that can be said about the accurate style modeling.

Traditional stylometric methods are heavily dependent on the superficial characteristics like character or word n-grams, frequency of use of function words, punctuation patterns, and part-of-speech distributions. Even though these techniques are effective when dealing with long, single-author documents and unmodified documents, they cannot handle short texts, stylistic ambiguity, or multi-author documents. These limitations reduce their effectiveness in real life situations where documents are often authored by complex people either intentionally or unintentionally. In order to overcome these shortcomings, recent developments in deep learning, especially Transformer-based language models, have shown promise. DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is one of such models, which features a unique architecture in that content and positional embeddings are separated, and the model is able to learn the contents and the structure of the message more effectively. The paper recommends the use of DeBERTa to identify the change in authorial style on a paragraph-by-paragraph basis by posing the style change detection as a Natural Language Inference (NLI) problem. Using this formulation, we test whether or not two adjacent paragraphs are likely to have been written by the same person or not. The task is in binary Natural Language Inference (NLI) format, where the model is presented with two consecutive paragraphs and asked to determine whether a style change has occurred.

DeBERTa is refined during training on labeled paragraph pairs of the Multi-Author Writing Style Analysis dataset of PAN 2023, multi-author manuscripts with style comments. Each pair of paragraphs is inputted into DeBERTa in the following format: [CLS] Section A [SEP] Section B [SEP]. The disentangled attention mechanism enables the model to focus on both the lexical content and structural patterns. A classification head predicts the possibility of a change of style between the two paragraphs. Another data augmentation method we apply during training is back-translation, synonym replacement, and sentence shuffling to enhance robustness. Consequently, the model is less vulnerable to adversarial obfuscation.

In the inference, the sliding window method is applied to the whole document to identify the points of authorial transitions. Where the model shows that there is a high probability of a change in style between two paragraphs, then this point is recognized as a potential author transition. The results of the PAN 2023 data show that DeBERTa outperforms the work of traditional stylometric classifiers, particularly in difficult cases when the topics remain constant and the authors change, which is why the model can be used to evaluate forensic and collaborative writing.

Key Contributions Include

1. Novel NLI-based formulation of style-change detection, simplifying multi-author analysis into pairwise evaluation.
2. Development of a DeBERTa-based framework leveraging disentangled attention for fine-grained stylistic understanding.
3. Robust training using aggressive data augmentation to counter adversarial stylometry.
4. Superior performance on the PAN 2023 dataset, outperforming classical and modern baselines.
5. Practical real-world applicability, especially in forensic linguistics, collaborative platforms, and cybersecurity contexts.

The organization of the proposed work is as follows: section 2. literature survey, Section 3. Proposed Methods and Solutions, Section 4. Result and Discussion and Section 5. Conclusion.

2 Literature Survey

Authorship analysis has become a complicated area of research due to the growing number of writing styles, collaborative tools, and adversarial manipulation strategies (Potthast et al., 2018). Stylometric analysis, authorship attribution and detection of style change has been studied in many works in the context of literature, forensic, academic, legal and online (Fung et al., 2021; Dementieva et al., 2024). This part summarizes the important findings of the previous literature without changing the reference numbers.

Earlier methods of authorship analysis were mostly based on hand-crafted lexical, syntactic and structural attributes. A basic survey by Shen et al., (2009) analyzed the most important stylometric features including word frequencies, character n-grams and POS-based patterns, and showed that classical machine learning algorithms like SVM and Naive Bayes can attain more than 90 percent accuracy on well-defined and long-document attribution problems. Nevertheless, these approaches have difficulties with short fragments and multi-author works (Rykov et al., 2024; Dementieva et al., 2024).

Other researchers like Rios-Toledo et al., (2022) employed n-grams (character, word, syntactic dependency) and supervised classifiers, such as SVM and RF, to detect diachronic style shifts in English novels, and found accuracy greater than 80%. On the same note, Potthast et al., (2018) compared stylometric features across Twitter posts, blogs, and essays with cross-genre AUC 0.80, but their

approach had interpretability and genre diversification issues. Bevendorff et al., (2025) demonstrated that classical stylometric methods like the Delta by Burrows can achieve a 0.70-0.80 accuracy in attributing authorship to Early Modern English drama, but are not applicable to modern and dynamic writing conditions.

A different research stream applied stylometry to formal and professional texts. Wang et al., (2024) showed that the combination of stylometric features with deep learning based on BERT provides 0.85-0.92 accuracy on legal texts, but the strict linguistic composition of legal writing constrains the range of stylistic changes. Wolska et al., (2023) used a combination of lexical, syntactic, and character-level features with deep learning to enhance the domain adaptation performance, which was high on longer texts, yet had difficulties with shorter samples and topic sensitivity.

There is also the interest in collaborative and dynamic-text environments. Bacciu et al., (2023) examined Wikipedia edit history through edit-based stylometry by using SVM and obtained a high accuracy of over 0.70. Soylu et al., (2023) evaluated multi-author news and blog articles based on SVM, logistic regression and ensemble approaches with F1-scores of approximately 0.75. The Co-Authorship Graph (CAG) framework of multi-label authorship prediction with stylometry features was presented by Sarwar et al., (2020) who found no consistent accuracy (6080) but had difficulty with scalability. Alshehri & Muhammad, (2024) used few-shot learning on proprietary multi-author news data and obtained accuracies of more than 85, which shows the potential of prototype-based methods.

One of the most significant achievements in multi-author style-change studies is the PAN shared tasks. Bevendorff et al., (2024) compared the methods of multi-author writing style analysis in PAN 2024 and discovered that even the transformer-based models do not work without the topic cues, with the F1-scores reaching only 0.6-0.7. Persistent shortcomings of both classical and neural models in identifying subtle stylistic boundaries in multi-author documents are emphasized in Bevendorff et al., (2025) and associated papers on the PAN overview, particularly in the case of a short text or where authors deliberately blur styles.

Another significant challenge is adversarial stylometry, that is, when writers intentionally use their style to manipulate it. Kadhim et al., (2025). proved that AI text detectors decrease significantly when shown paraphrased text, and the accuracy decreases to less than 50 percent of the initial 90 percent. Teja et al., (2026). also trained with adversarial training to enhance robustness, with over 80 percent accuracy on strong attacks, although these models are computationally costly, and not designed to be trained on multi-author segmentation. Fraser et al., (2025), Abburi et al., (2024), and Macko et al., (2024) also determined that neural and statistical detectors are susceptible to text obfuscation and adversarial rewriting.

The experience of multilingual detoxification and style-transfer studies also indicates the complexity of the task in terms of the ability to capture the finer nuances of style. Dementieva et al., (2022), Prabhumoye et al., (2018), Logacheva et al., (2026) showed that back-translation, paraphrasing, and multilingual fine-tuning lead to substantial stylistic drift, which makes attribution more difficult. The difficulty of stylistic consistency across languages and domains is pointed out in studies that employ Seq2Seq models and hybrid systems. The same can be said about low-resource languages as demonstrated by Lin & Hu, (2022).

Such phenomena like oppositional thinking have been studied as part of authorship, including the Oppositional Thinking Analysis task of PAN 2024. Koren et al., (2024) and Gómez-Romero et al., (2024) studied the BERT-based and rule-based systems to detect conspiracy-oriented text segments, with the F1-scores of approximately 0.4-0.6. Rangel et al., (2018) regarded some ethical issues connected with stylometric analysis and highlighted the risks of misclassification and privacy.

The emergence of generative AI has turned the problem of authorship verification into a more complicated one. Lei et al., explored the topic of human vs. AI authorship detection with fine-tuned LLMs, reporting F1-scores of up to 0.82, whereas Bevendorff et al., (2025) emphasized high false positive/negative rates and model ungeneralizability. Statistical and perplexity-based approaches Fraser et al., (2025), Zellers et al., (2019), Liu & Kong, (2024), Prova (2024) are also moderately successful but very weak against paraphrasing or small modifications. Mo et al., (2024) have also obtained F1-scores of 0.90-0.95 in news-domain AI detection with transformer models, which are domain-dependent and computationally expensive. More sophisticated methods like watermarking (Zhao et al., 2023), fingerprinting (Zeng et al., 2024), contrastive learning (Chen et al., 2023) and certified robustness (Lou et al., 2024) are applicable to specific problems but fail to give the generalization to multi-author style transitions.

Some of the studies investigate similar phenomena like argumentative structure, and discourse-based attribution. Zampieri et al. Study argumentative discourse with the help of BERT-based models, with the F1-scores being approximately 0.70. Al Khatib et al. discuss argument mining in political argument, with a result of approximately 0.650.70 to detect argument components. Jallad et al., Examine oppositional statements in Arabic and find the accuracy of oppositional statements to be 0.78-0.85 regardless of the dialectal variation and code switching.

Across all these works, major limitations become evident

1. Reliance on surface-level stylometry (n-grams, frequencies) restricts generalization to adversarial or subtle style-shift scenarios.
2. Poor performance on short, paragraph-level segments, which lack sufficient stylistic evidence for classical models.
3. Topic–style entanglement, where classifiers incorrectly treat topic changes as author changes.
4. Limited robustness to paraphrasing, reordering, and adversarial obfuscation, a recurring issue across studies.
5. Insufficient handling of multi-author documents, as seen in PAN tasks where best systems plateau at F1 ≈ 0.6–0.7.

Taken together, the literature indicates a significant gap in the ability of existing stylometric and deep learning approaches to identify authorial shifts in adversarial, collaborative, or short-text environments. These limitations highlight the need for a model that can effectively disentangle content and style, capture deep contextual nuances, and remain resilient to adversarial transformations.

This gap motivates the development of a DeBERTa-based machine learning framework for detecting style changes at the paragraph level leveraging disentangled attention, robust augmentation, and NLI-based formulation to advance authorship analysis beyond the capabilities of classical and existing transformer-based methods.

3 Proposed Methods and Solutions

This section describes the suggested model solution of a DeBERTa-based framework of detecting style changes on a paragraph-level basis. The problem is defined as binary Natural Language Inference (NLI), in which the model is presented with two successive paragraphs and asked to decide whether a style shift has taken place. The DeBERTa is fine-tuned at training on labeled paragraph pairs of the PAN 2023 Multiauthor Writing Style Analysis dataset consisting of multi-author documents with style annotations. All pairs of paragraphs are coded in the input format of DeBERTa: [CLS] Paragraph A [SEP] Paragraph B [SEP]. The disentangled attention mechanism of the model allows it to pay attention

to the lexical content and structural patterns. A classification head estimates the possibility of a change of style between the two paragraphs. Data augmentation techniques such as back-translation, synonyms replacement and sentence permutation are also utilized by us to enhance robustness during training. This makes the model more resistant to adversarial obfuscation methods.

Inference utilizes a sliding window strategy on the entire document in order to detect authorial transition points. In the case where the model suggests a high likelihood of style change to occur between two paragraphs, the boundary is indicated by a shift that may be an author switch. The findings on the PAN 2023 data indicate that DeBERTa outsmarts conventional stylometric classifiers especially in challenging conditions where the topics are fixed but writers vary to indicate its usefulness in both forensic and collaborative writing analysis.

This section will show all the architecture design, training process and experimental flow of the proposed style change detection system using DeBERTa. Figure 1 displays the pipeline that is used to make paragraph-level predictions of author shift. Data set from PAN 2023 Multi-Author Writing Style Analyis is used an input which goes through a data processing phase consisting of extraction of adjacent paragraph pairs and then assigned labels. These processed data is passed to the multiple models including the proposed model and multiple baseline models based on SVM, Logistic Regression and RoBERTa. Prediction of each of these models is evaluated through a bunch of evaluation metrics and then compared. These are discussed in detail below.

3.1 Research Design

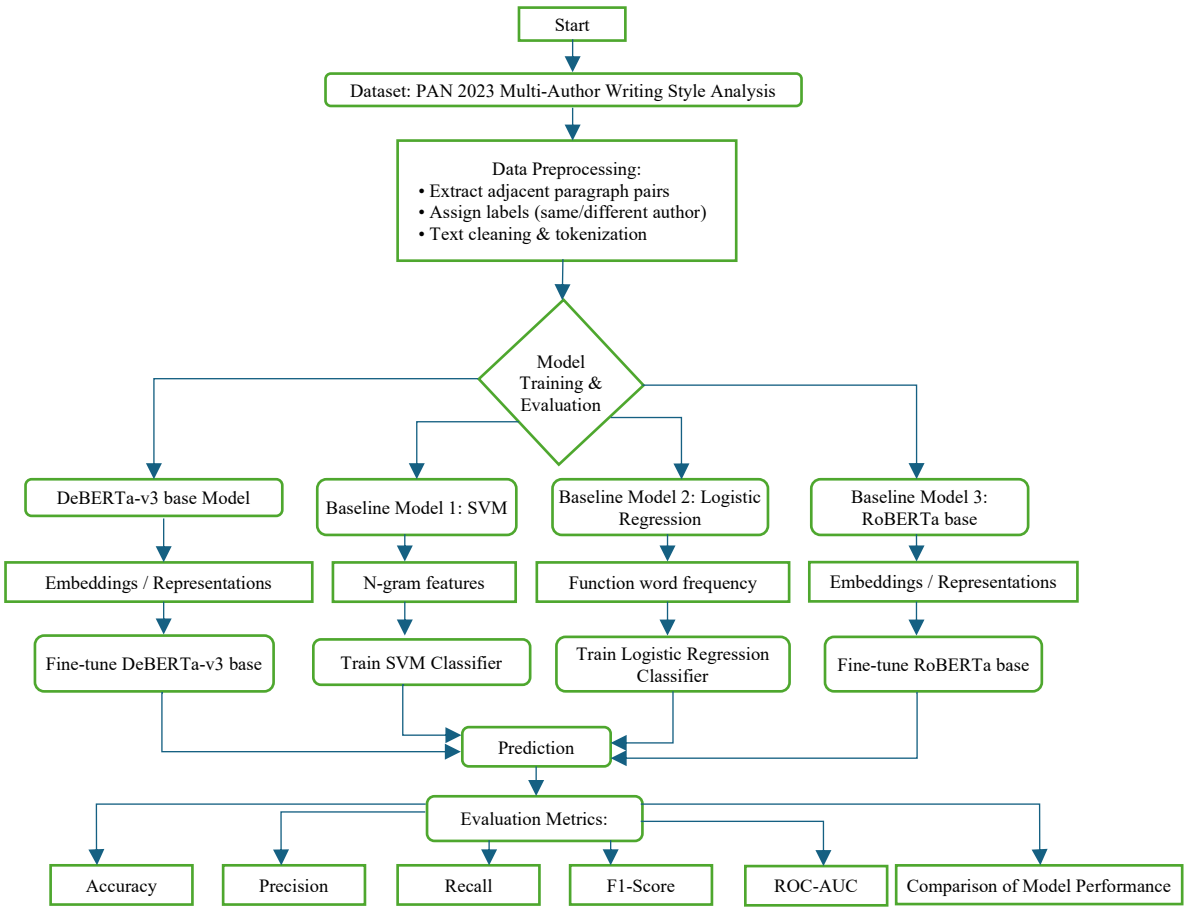


Figure 1: Flowchart of the AI-based models and experimental methods applied

The research follows a supervised learning setup using labelled paragraph-pairs from the PAN 2023 dataset. Each document is segmented into paragraphs $P_1, P_2, \ldots, P_n$, and each adjacent pair (P_i, P_{i+1}) receives a binary label:

y= (1 if author changes between Pi and Pi+1, 0 otherwise)

The model is fine-tuned to detect stylistic transitions using DeBERTa's disentangled-attention mechanism

3.2 Input Representation

Let the two paragraphs (P1 and P2) be represented as in equation (1) and (2):

$$P1 = \{W_1, W_2 \ \ldots\ldots W_m\} \quad (1)$$

$$P2 = \{v_1, v_2 \ \ldots\ldots v_n\} \quad (2)$$

The input sequence to DeBERTa is constructed in equation (3)

$$X = [CLS]P1[SEP]P2[SEP] \quad (3)$$

Each token is associated with separate content embeddings and positional embeddings as mentioned in equation (4) below

$$e_i = \left(e_i^{content}, e_i^{position}\right) \quad (4)$$

3.3 Embedding Layer

Unlike BERT which sums embeddings, DeBERTa keeps them separate as defined in equation (5) from, allowing better control over syntactic structure:

$$e_i^{content} = Wc.w_i \ \ and\ e_i^{position} = Wp.pos(i) \quad (5)$$

This separation enhances stylistic sensitivity.

3.4 Disentangled Attention

Attention between tokens *i* and *j* is computed using equation (6) (He et al., 2020):

$$A_{ij} = softmax\left((Q_i{}^c + Q_i{}^p)(K_i{}^c + K_i{}^p)^T * \left(\frac{1}{\sqrt{d}}\right)\right) \quad (6)$$

Where:

$$Q_i^c = W_{Q^c}.e_i^{content}$$

$$Q_i^p = W_{Q^p}.e_i^{position}$$

$$K_j^c = W_{K^c}.e_j^{content}$$

$$K_j^p = W_{K^p}.e_j^{position}$$

This allows DeBERTa to distinguish stylistic structure from semantic content, critical for authorship shifts.

3.5 Classification Layer

The [CLS] hidden state is fed into a sigmoid classifier as mentioned in equation (7):

$$y^{\wedge} = \sigma(W_{chCLS} + b_c) \quad (7)$$

3.6 Loss Function

Binary cross entropy is used as defined in equation (8) below:

$$L = -\left(ylog\left(y^{\wedge}\right) + (1-y)\log\left(1-y^{\wedge}\right)\right) \quad (8)$$

3.7 Dataset Details

·Dataset: PAN 2023 Multi-Author Writing Style Analysis

·Documents: 936

·Total Paragraphs: 14,562

·Paragraph-Pairs: 13,800+

·Authors per document: 2–5

·Languages: English

·Average paragraph length: 45–125 words

3.8 Data Augmentation

Three augmentation strategies were implemented:

1. **Back-Translation**

$$P' = BT(P) \quad (9)$$

Converts the text into a foreign language and translates it back to the original to produce a new expression of the sentence (Equation 9).

2. **Synonym Replacement**

$$w'_i \in Syn\ (w_i) \quad (10)$$

Substitutes words in the sentence by their synonyms (Equation 10).

3. **Sentence Shuffling**

$$P = \{S1, S2, S3\} \rightarrow P' = \{S3, S1, S2\} \quad (11)$$

Paraphrases the sentence by altering the sequence of words or clauses in the sentence (Equation 11).

4. **Final Augmented Input**

$$X' = [CLS]P1'[SEP]P2'[SEP] \quad (12)$$

Augmented sentences are also combined into a new input format to be used in training the model (Equation 12).

3.9 Algorithm

The paper proposed the DeBERTa Based style change detection and obtained the Attention computation: O(n2d), Overall, per-pair inference: O(Ln2d) where L is number of transformer layers used to perform the style change detection is shown in the Algorithm1.

The Algorithm describes our methodology for identifying style boundaries within a multi-author document. It starts with dividing the document into a sequence of paragraphs. In order to test style changeover, we use a sliding window method that takes two successive paragraphs as one input sequence with special tokens separating them.

The main part of our detection mechanism employs the DeBERTa architecture. Unlike standard transformers, this model uses a disentangled attention mechanism (Steps 3–4), which separately computes attention weights for content embeddings ($e_j^{content}$) and relative position embeddings ($e_j^{position}$). This is quite useful when it comes to style detection since the model is able to make distinctions between the semantic content of a paragraph and structural positioning. The resulting representation of the [CLS] token, h_{CLS} , is then fed into a sigmoid classifier. A change of style boundary is officially delimited at index i when the predicted probability attains a value that is greater than the optimized threshold, τ.

Algorithm 1: Paragraph-Level Style Change Detection

Input: Document D with paragraphs P_1, P_2, ..., P_n

Output: Style-change boundaries S = {i | style changes between Pi and Pi+1}

Step 1: Preprocess all paragraphs:

- Clean text (remove noise, normalize spacing)
- Tokenize using DeBERTa tokenizer (He et al., 2020)

Step 2: For each paragraph pair (P_i, P_{i+1}):

Construct input sequence as mentioned in eq. (3):

$$X = [CLS]P_i[SEP]P_{i+1}[SEP]$$

Step 3: For each token w_j in X:

Compute content embedding $e_j^{content}$

Compute position embedding $e_j^{position}$

Step 4: For each attention layer:

Compute:

$$Q^c = W_{Q^c}.e_j^{content}$$

$$Q^p = W_{Q^p}.e_j^{position}$$

$$K^c = W_{K^c}.e_j^{content}$$

$$K^p = W_{K^P}.e_j^{position}$$

Compute disentangled attention as per eq. (6):

$$\alpha_{ij} = softmax((Q^c + Q^p).\frac{(K^c + K^p)}{\sqrt{d}})$$

Note: Attention equations follow the disentangled approach defined in He et al., 2020).

Step 5: Aggregate final representation:

$$h_{CLS} = hidden\ state\ of\ ([CLS])$$

Step 6: Classification as mentioned in eq. (7):

$$\hat{y} = sigmoid(W_c \times h_{CLS} + b_c)$$

Step 7: If $\hat{y}$ > threshold τ:

Mark index i as a style-change boundary

Step 8: Return S

P- paragraph, D-Documents, S- Final return values, a_{ij} -Disentangled attention, h_{CLS} -Aggregate final values

4 Results and Discussion

This section explains the results and comparison between classical surface-feature methods, a strong Transformer baseline (RoBERTa), and the proposed DeBERTa-v3 model specialized for nuanced stylometric cues.

4.1 Performance Evaluation Metrics

To evaluate the effectiveness of the proposed system, multiple classification metrics are used. As the task is binary classification, we used confusion matrix consisting of

a. True Positives (TP) – Model correctly predicts a style change
b. True Negetives (TN) – Model correctly predicts no style change
c. False Positives (FP) – Model incorrectly predicts style change
d. False Negetives (FN) – Model incorrectly predicts no style change

Accuracy

Accuracy measures the overall correctness of the model by computing the proportion of correctly classified paragraph pairs as mentioned in equation (13).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

Precision

Precision measures how many of the predicted style changes are actually correct and is defined in equation (14).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

Recall

Recall measures the model's ability to correctly identify all actual style changes and is defined in equation (15).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (15)$$

F1-Score

The F1-score is the hormonic mean of precision and recall, providing a balanced measure of performance and is defined in equation (16).

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

ROC-AUC

ROC-AUC evaluates the model's ability to distinguish between the two classes across the possible decision thresholds and is defined in equation (17).

$$\text{ROC-AUC} = \int_0^1 TPR(FPR)\, d(FPR) \quad (17)$$

Where: $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$

4.2 Software and Implementation Details

All experiments were conducted in a python-based environment with support for GPU acceleration.

Programming Environment

- Programming Language – Python 3.10
- Deep Learning Framework – PyTorch
- Transformer Library – Hugging Face Transformers
- Tokenization DeBERTa-v3 Tokenizer
- Machine Learning Utilities – Scikit-learn

Model Implementation

The proposed model consists of

- Number of Transformer Layers: 12
- Hidden Layers: 768
- Attention Heads: 12
- Maximum Sequence Length: 512 tokens

4.3 Results

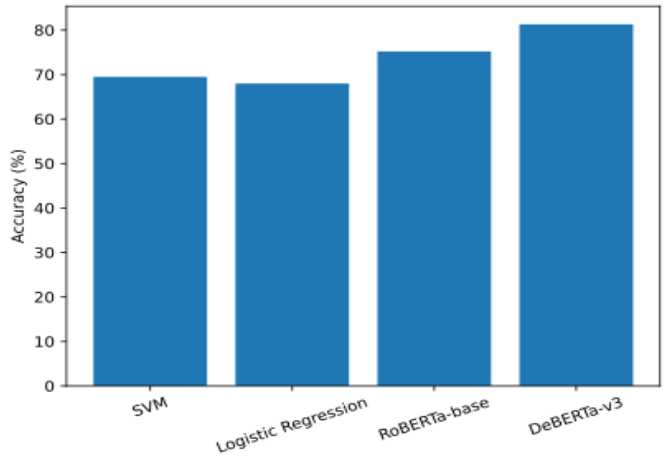
In this paper, paragraph level structure is consistent with the NLI-based formulation applied in this work and helps to evaluate the style-change detection models in collaborative and adversarial settings effectively. All the paragraphs are marked with their real author, which allows accurate same or different labels of consecutive pairs on the paragraph level. The model contains 12 transformer layers, 768-dimensional hidden states, disentangled content or position attention. Analysis of the results from the PAN 2023 Multi-Author writing style analysis task using DeBERTa-v3 and baseline models as shown in the figures 2(a), 2(b), 3(a), 3(b), 4(a), and 4(b) respectively. In the model DeBERTa Outperforms all baselines like in accuracy (81.3%) and F1-score (80.6%) indicate that DeBERTa is the most reliable model in detecting authorial transitions. The ROC-AUC of 0.88 shows excellent class

separability, even under difficult conditions such as adversarial obfuscation or stylistic similarity. Transformer-based models outperform traditional Stylometry in both RoBERTa and DeBERTa significantly outperform SVM and Logistic Regression, demonstrating the superiority of contextual embeddings over surface-level features. This reflects the need to go beyond function words and n-grams when authors intentionally or unintentionally blend styles. DeBERTa's disentangled attention enables better handling of syntactic variation, reordering, and paraphrasing, which are common in adversarial stylometry and robust performance in adversarial and multi-author contexts.

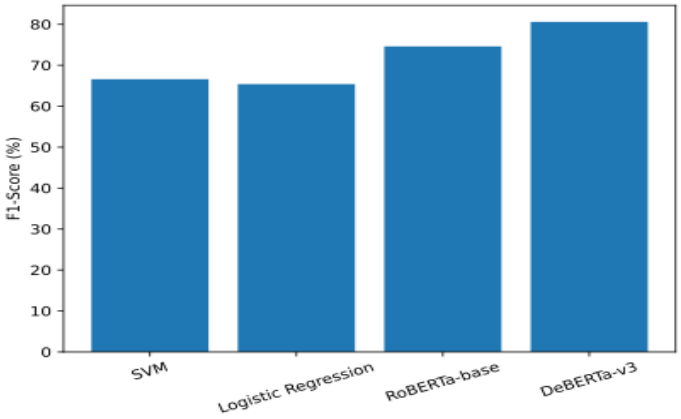
Table 1: Performance metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
SVM (Stylometric Features)	69.5	65	67	66.6	0.72
Logistic Regression	68	64.2	66.1	65.4	0.7
RoBERTa-Base	75.2	76	73.2	74.6	0.82
Proposed DeBERTa-v3 Model	81.3	81.9	79.4	80.6	0.88

The table 1 represents the performance table, DeBERTa shows the highest performance across metrics. The small gap between precision (81.9%) and recall (79.4%) indicates a well-balanced classifier with minimal bias toward either false positives or false negatives. AUC = 0.88 demonstrates superior separability over RoBERTa (0.82) and classical models (0.70–0.72).

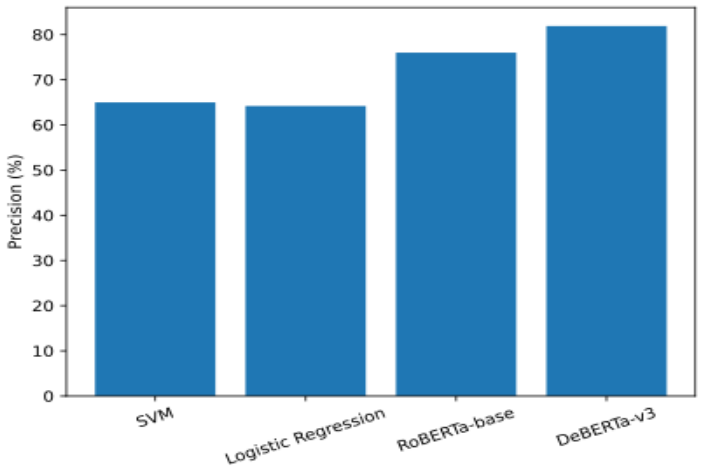


(a)

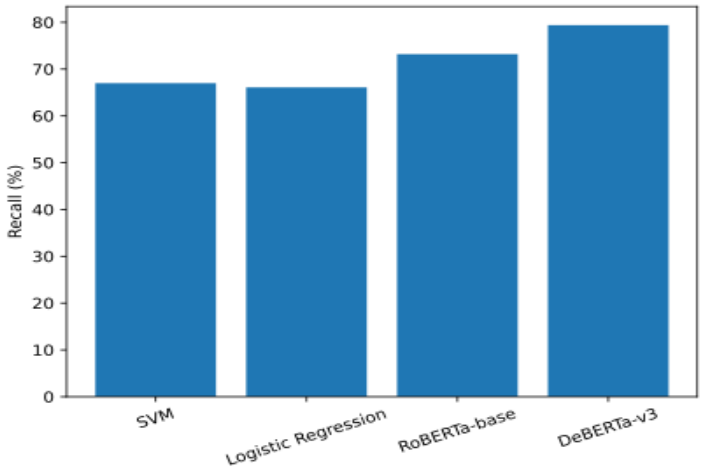


(b)

Figure 2: The comparison of models (a) accuracy and (b) F1-score

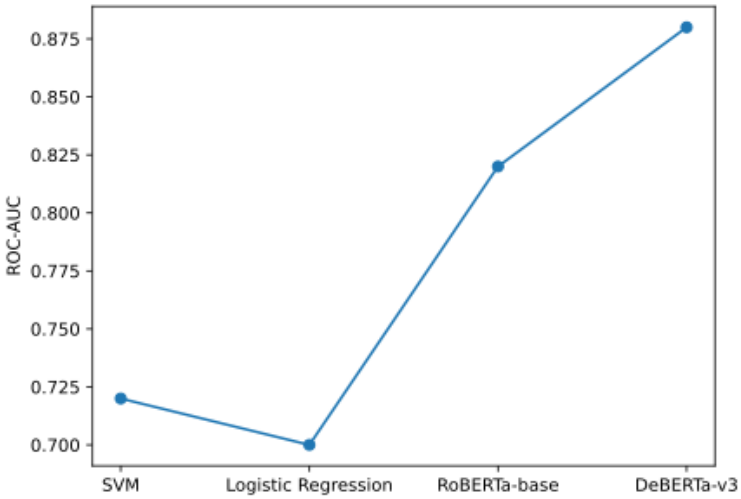


(a)

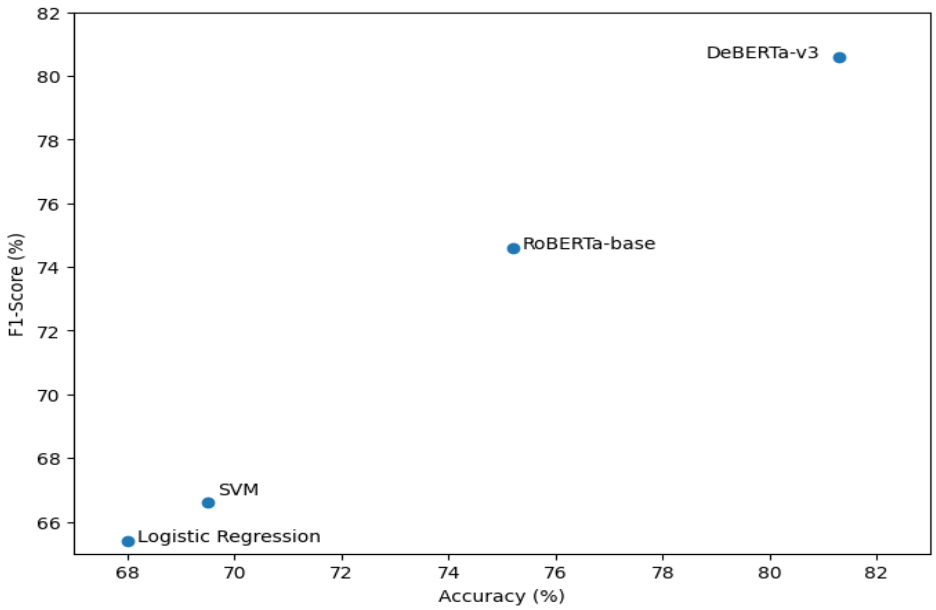


(b)

Figure 3: The comparison of models (a) Precision and (b) Recall



(a)



(b)

Figure 4: The relationship between (a) RoC and AUC, and (b) accuracy and F1-score

Key Observation

DeBERTa achieves 81.3% accuracy and 80.6% F1-score, outperforming all baselines.

ROC-AUC Interpretation

1. AUC = 0.88 → excellent class separability
2. Shows high robustness in adversarial settings

Error Analysis

Three major sources of misclassification:

1. Topic-stable transitions misinterpreted as stylistically consistent.
2. Very short paragraphs (<30 words) lack enough stylistic signals.
3. Highly obfuscated paragraphs via paraphrasing reduce lexical distinctiveness.

The experimental findings show that DeBERTa is able to identify the minor stylistic features that classical does not possess. n-gram techniques and even RoBERTa do not detect. This is due to the disentangled attention, which separates structural tendencies and lexical content. This is largely beneficial when the use of authors who deliberately imitate subjects or words to evade attribution.

Nevertheless, there are a number of constraints. The similarity of subject matter in segments can be deceptive at times. the model. Moreover, when the data is augmented, adversarial makes it robust, although even the tools of paraphrasing can deflect stylistic cues. The dependency on training that is intensive on the GPU also restricts the application on a mass scale within learning institutions and criminal laboratories.

In spite of these struggles, the performance of DeBERTa demonstrates the appropriateness of the tool in relation to forensic, linguistics, academic dishonesty screening, and cyber investigation, in which it is ascertained, the shifting of authorship is essential.

Ablation Study

The ablation experiment will analyze the performance of the DeBERTa-v3 model by experimenting it using various settings. First, the data augmentation model will create a baseline with the base model. Thereafter, the methods of augmentation, the use of back-translation, the use of synonyms replacement, and shuffling of the sentences, will be eliminated separately to evaluate their effects on the performance. Lastly, the model that has all the methods of augmentation will be tested to know the effect of the sum. Accuracy, precision, recall, F1-score, and ROC-AUC will be used in measuring the performance.

5 Conclusion

This paper showed that modeling paragraph-level style change detection is an effective Natural Language Inference (NLI) task that can be effectively modeled using the DeBERTa-v3 model. Experimental analysis of the PAN 2023 Multi-Author writing style analysis dataset revealed that DeBERTa-v3 was significantly better than traditional stylometric baselines and transformer-based alternatives on a variety of performance metrics. Accuracy and F1-score results in terms of accuracy and F1-score, DeBERTa-v3 obtained 81.3% accuracy and 80.6 F1-score, which is much higher than SVM (69.5%, 66.6) and logistic regression (68.0%, 65.4) and is an improvement over RoBERTa-base (75.2%, 74.6). Likewise, on precision and recall, DeBERTa-v3 achieved 81.9% precision and 79.4% recall, which is better than RoBERTa-base (76.0, 73.2) and traditional baselines (precision 65, recall 67). The score of ROC-AUC (0.88) also affirmed that DeBERTa-v3 had a better discriminative power than RoBERTa (0.82), SVM (0.72), and logistic regression (0.70). These findings suggest that DeBERTa-v3 is sensitive to nuanced stylistic information that is not immediately apparent in the form of lexical and syntactic phenomena and is resistant to the adversarial obfuscation methods of paraphrasing and syntactic restructuring.

Although it has been performing very well, there are also limitations associated with the study. On some occasions DeBERTa-v3 failed when the consistency of the topical factor concealed stylistic variation, and computational needs are still a problem in real-time or resource-intensive implementations. However, the general results allow concluding that deep transformer networks are a significant improvement in comparison with traditional stylometric methods, which can be trusted in the detection of authorial changes during forensic, academic, and cybersecurity use. Future research needs to aim at enhancing interpretability, applying the technique to low-resource languages, and creating lightweight and efficient architectures to run in realistic monitoring systems.

References

[1] Abburi, H., Pudota, N., Veeramani, B., Bowen, E., & Bhattacharya, S. (2024, December). Toward robust generative ai text detection: Generalizable neural model. In *2024 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1651-1656). IEEE. https://doi.org/10.1109/ICMLA61862.2024.00255

[2] Alshehri, F., & Muhammad, G. (2024). Ischemic stroke segmentation by transformer and convolutional neural network using few-shot learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(12), 1-21. https://doi.org/10.1145/3699513

[3] Bacciu, A., La Morgia, M., Mei, A., Nemmi, E. N., Neri, V., & Stefa, J. (2019). Cross-domain authorship attribution combining instance-based and profile-based features notebook for PAN at CLEF 2019. In *Ceur Workshop Proceedings* (Vol. 2380). CEUR-WS.

[4] Bevendorff, J., Casals, X. B., Chulvi, B., Dementieva, D., Elnagar, A., Freitag, D., ... & Zangerle, E. (2024, March). Overview of pan 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification. In *European Conference on Information Retrieval* (pp. 3-10). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56072-9_1

[5] Bevendorff, J., Dementieva, D., Fröbe, M., Gipp, B., Greiner-Petter, A., Karlgren, J., ... & Zangerle, E. (2025, April). Overview of pan 2025: Generative ai detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In *European Conference on Information Retrieval (pp. 434-441). Cham: Springer Nature Switzerland.* https://doi.org/10.1007/978-3-031-88720-8_64

[6] Chen, H., Han, Z., Li, Z., & Han, Y. (2023). A Writing Style Embedding Based on Contrastive Learning for Multi-Author Writing Style Analysis. *In CLEF (Working Notes)* (pp. 2562-2567).

[7] Dementieva, D., Babakov, N., & Panchenko, A. (2024, June). MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)* (pp. 124-140). https://doi.org/10.18653/v1/2024.naacl-short.12

[8] Dementieva, D., Logacheva, V., Nikishina, I., Fenogenova, A., Dale, D., Krotova, I., ... & Panchenko, A. (2022). Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora. *Computational Linguistics and Intellectual Technologies*, 114-131.

[9] Dementieva, D., Moskovskiy, D., Babakov, N., Ayele, A. A., Rizwan, N., Schneider, F., ... & Panchenko, A. (2024, September). Overview of the Multilingual Text Detoxification Task at PAN 2024. In *CLEF (Working Notes)* (pp. 2432-2461).

[10] Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, *82*, 2233-2278. https://doi.org/10.1613/jair.1.16665

[11] Fung, Y., Thomas, C., Reddy, R. G., Polisetty, S., Ji, H., Chang, S. F., ... & Sil, A. (2021, August). Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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)* (pp. 1683-1698). https://doi.org/10.18653/v1/2021.acl-long.133

[12] Gómez-Romero, J., González-Silot, S., Montoro-Montarroso, A., Molina-Solana, M., & Cámara, E. M. (2024). Detection of conspiracy-related messages in Telegram with anonymized named entities. *In CLEF (Working Notes)* (pp. 2603-2612).

[13] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. https://doi.org/10.48550/arXiv.2006.03654

[14] Kadhim, A. K., Jiao, L., Shafik, R., & Granmo, O. C. (2025). Adversarial attacks on AI-generated text detection models: A token probability-based approach using embeddings. https://doi.org/10.48550/arXiv.2501.18998

[15] Korenčić, D., Chulvi, B., Bonet-Casals, X., Taulé, M., Rosso, P., & Rangel, F. (2024). Overview of the oppositional thinking analysis pan task at clef 2024.

[16] Lazebnik, T., Aviv-Reuven, S., & Rosenfeld, A. (2025). Publishing instincts: An exploration-exploitation framework for studying academic publishing behavior and Home Venues. *Journal of Informetrics*, *19*(3), 101705. https://doi.org/10.1016/j.joi.2025.101705

[17] Lin, R., & Hu, H. (2023). Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, *11*, 1686-1702. https://doi.org/10.1162/tacl_a_00628.

[18] Liu, X., & Kong, L. (2024). AI Text Detection Method Based on Perplexity Features with Strided Sliding Window. In *CLEF (Working Notes)* (pp. 2755-2760).

[19] Logacheva, V., Dementieva, D., Krotova, I., Fenogenova, A., Nikishina, I., Shavrina, T., & Panchenko, A. (2022, May). A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)* (pp. 90-101). https://doi.org/10.18653/v1/2022.humeval-1.8

[20] Lou, Q., Liang, X., Xue, J., Zhang, Y., Xie, R., & Zheng, M. (2024, August). Cr-utp: Certified robustness against universal text perturbations on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9863-9875).

[21] Macko, D., Moro, R., Uchendu, A., Srba, I., Lucas, J. S., Yamashita, M., ... & Bielikova, M. (2024, November). Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 6348-6368). https://doi.org/10.18653/v1/2024.findings-emnlp.369

[22] Mo, Y., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). Large language model (llm) ai text generation detection based on transformer deep learning algorithm. https://doi.org/10.48550/arXiv.2405.06652

[23] Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., ... & Stein, B. (2018, August). Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1498-1507).

[24] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. W. (2018, July). Style transfer through back-translation. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 866-876). https://doi.org/10.18653/v1/P18-1080

[25] Prova, N. (2024). Detecting ai generated text based on nlp and machine learning approaches. https://doi.org/10.48550/arXiv.2404.10032

[26] Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., & Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, *192*.

[27] Ríos-Toledo, G., Posadas-Durán, J. P. F., Sidorov, G., & Castro-Sánchez, N. A. (2022). Detection of changes in literary writing style using N-grams as style markers and supervised machine learning. *Plos one*, *17*(7), e0267590.https://doi.org/10.1371/journal.pone.0267590

[28] Rykov, E., Zaytsev, K., Anisimov, I., & Voronin, A. (2024). Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. https://doi.org/10.48550/arXiv.2407.05449

[29] Sarwar, R., Urailertprasert, N., Vannaboot, N., Yu, C., Rakthanmanon, T., Chuangsuwanich, E., & Nutanong, S. (2020). $ CAG $: Stylometric authorship attribution of multi-author documents using a co-authorship graph. *IEEE Access*, *8*, 18374-18393. https://doi.org/10.1109/ACCESS.2020.2967449

[30] Shen, H., Liu, G., & Guo, J. (2009). Mixed environment compensation based on maximum a posteriori estimation for robust speech recognition. *Artificial Intelligence Review*, *32*(1), 1-11. https://doi.org/10.1007/s10462-009-9130-9

[31] Soylu, M., Soylu, A., & Das, R. (2023). A new approach to recognizing the use of attitude markers by authors of academic journal articles. *Expert Systems with Applications*, *230*, 120538. https://doi.org/10.1016/j.eswa.2023.120538

[32] Teja, L. S., Yadagiri, A., Anish, S. S., Nuthakki, S. G. K., & Pakray, P. (2026, January). Modeling the attack: detecting AI-generated text by quantifying adversarial perturbations. In *2026 20th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1-8). IEEE. https://doi.org/10.48550/arXiv.2510.02319

[33] Wang, Z., Cheung, A. K., & Liu, K. (2024). Entropy-based syntactic tree analysis for text classification: a novel approach to distinguishing between original and translated Chinese texts. *Digital Scholarship in the Humanities*, *39*(3), 984-1000. https://doi.org/10.1093/llc/fqae030

[34] Wiegmann, M., Wolska, M., Schröder, C., Borchardt, O., Stein, B., & Potthast, M. (2023, July). Trigger warning assignment as a multi-label document classification problem. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12113-12134). https://doi.org/10.18653/v1/2023.acl-long.676

[35] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, *32*.

[36] Zeng, B., Wang, L., Hu, Y., Xu, Y., Zhou, C., Wang, X., ... & Lin, Z. (2024). Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, *37*, 126332-126362. https://doi.org/10.48550/arXiv.2312.04828

[37] Zhao, X., Ananth, P., Li, L., & Wang, Y. X. (2023). Provable robust watermarking for ai-generated text. https://doi.org/10.48550/arXiv.2306.17439

Authors Biography

Riya Sanjesh, she is currently working as an assistant professor in presidency university. And research interest in forensic text analysis and deep learning. She has published books, journals and conferences.

Pamela Vinitha Eric, she is currently working as a Professor in School of Computer Science and engineering in presidency university. She holds a Ph. D, M. Tech, B. Tech and research interest on Bioinformatics, data compression, network security, and cryptography. She published many journals, conference papers.