# Bi-Directional Cross-Entropy Loss and Stylometric Feature Combined Classifier

Notebook for PAN and ELOQUENT at CLEF 2025

Yitao Sun[1], Svetlana Afanaseva[2], Kevin Stowe[3] and Kailash Patil[4]

[1]*Pindrop, New York, USA*
[2]*Pindrop, Seattle, USA*
[3]*Pindrop, Chicago, USA*
[4]*Pindrop, Atlanta, USA*

## Abstract

In the context of the PAN 2025 Voight-Kampff Generative AI Detection Task, Subtask 1[1], we present a hybrid method that leverages BiScope's bi-directional cross-entropy loss[2] alongside a suite of stylometric features to enhance detection performance. BiScope captures perplexity asymmetries between forward and backward language modeling, revealing latent inconsistencies characteristic of generated content. To complement this, we extract stylometric features—covering lexical diversity, syntactic complexity, and structural idiosyncrasies. Empirical results on the PAN 2025 benchmark datasets demonstrate that this integrated framework is a strong contender for effective generative AI detection.

## Keywords

PAN 2025, Voight-Kampff AI Detection Sensitivity Task, AI-generated text detection, Bidirectional cross-entropy loss, Stylometric analysis, Feature fusion

## 1. Introduction

The rise of large language models (LLMs) has made machine-generated text nearly indistinguishable from human writing, creating a pressing need for reliable detection methods. This challenge is central to the PAN 2025 Voight-Kampff Generative AI Detection Task, Subtask 1 [1], which focuses on identifying AI-generated content from a single text segment.

In response, we propose a hybrid detection framework that combines BiScope's bi-directional cross-entropy loss[2] with a rich set of stylometric features[3, 4]. BiScope captures asymmetries in token predictability from both forward and backward language models, revealing distributional irregularities often present in generated text. While effective, this approach alone may miss deeper stylistic cues that characterize human authorship.

To enhance detection accuracy, we integrate stylometric features—including lexical richness, syntactic patterns, and punctuation usage—that reflect consistent writing habits. This combination of low-level probabilistic signals and high-level stylistic markers provides a more holistic representation of authorship.

Our method is model-agnostic and domain-flexible. Experiments on the PAN 2025 dataset demonstrate that this dual-modality approach outperforms single-feature baselines, highlighting the value of combining linguistic signals for robust generative AI detection.

## 2. Background

Our approach is motivated by the NIST 2024 Generative AI (GenAI) Text-to-Text (T2T) Discriminator Task[5], which evaluated systems for distinguishing human-written from AI-generated summaries.

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*

ysun@pindrop.com (Y. Sun); safanaseva@pindrop.com (S. Afanaseva); kevin.stowe@pindrop.com (K. Stowe); kpatil@pindrop.com (K. Patil)
0009-0007-3743-9331 (Y. Sun)

We build on insights from the top-performing teams in the challenge: the first-place system employed BiScope's bi-directional cross-entropy loss to uncover token-level distributional anomalies[2], while the third-place system leveraged stylometric analysis to capture higher-level linguistic patterns such as lexical diversity and syntactic style. By combining these complementary strategies, we aim to enhance detection robustness and interpretability.

By integrating BiScope's probabilistic analysis with stylometric feature extraction, our method aims to leverage the strengths of both approaches. This hybrid framework is designed to enhance detection accuracy by capturing both low-level distributional irregularities and high-level stylistic nuances, providing a more robust solution for identifying AI-generated text.

# 3. System Overview

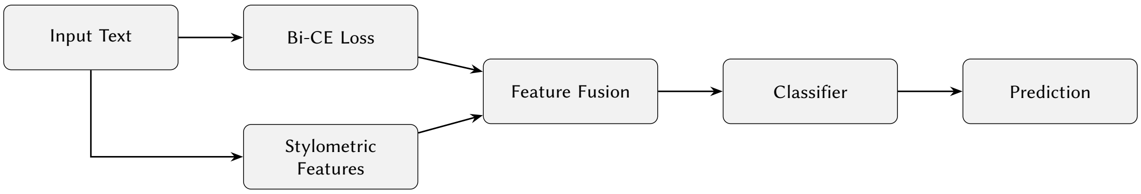

**Figure 1:** Model architecture

## 3.1. Stylometric Features

We tested a variety of linguistic and stylometric features. The features are largely based on previous work in AI-generated text detection [3, 4]. Additionally, we used a large language model (LLM) Claude [6] for suggestions of relevant features and implemented these. We broadly categorize these features into five different categories:

- **Character-level**: proportions of special characters, punctuation
- **Lexical**: unique words, abstract nouns
- **Syntactic**: part-of-speech-based features, multi-clause sentences
- **Structural**: total words, total sentences, sentence and paragraph length
- **Stylistic**: repetition, discourse markers, readability

A total of 101 features were initially generated and subsequently refined through univariate feature selection. We determined that selecting the top 25 most significant features produces optimal performance. The final set of these 25 features is listed in the Appendix 6.

## 3.2. Bi-directional Cross-entropy Loss Features

Bi-directional Cross-entropy (Bi-CE) loss is a method used to improve the detection of AI-generated text by measuring the consistency of token predictions in both forward and backward directions[2]. Traditional cross-entropy loss evaluates the likelihood of the next token given the previous context (left-to-right). Bi-CE extends this by also considering the reverse context (right-to-left), thus providing a more robust estimation of token likelihood.

Formally, the Bi-CE loss is computed as the sum of the forward and backward cross-entropy losses:

$$\mathcal{L}_{\text{Bi-CE}} = \mathcal{L}_{\text{forward}} + \mathcal{L}_{\text{backward}}, \tag{1}$$

where

$$\mathcal{L}_{\text{forward}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}),$$

$$\mathcal{L}_{\text{backward}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{>t}).$$

By capturing information from both directions, Bi-CE loss features provide a stronger signal for distinguishing human-written text from AI-generated content, as the latter tends to exhibit patterns that are less coherent when evaluated bidirectionally.

In our method, these features are extracted from a pre-trained language model and fed into downstream classifiers to enhance detection performance.

We transform a single text sample into a numerical feature vector by:

- Summarizing the text to create a prompt.
- Feeding prompt and text into a model. (**Llama2-7b**)
- Computing token-level forward and backward losses.
- Extracting statistical features over segments of the token losses. (**mean, max, min, and standard deviation** of both FCE and BCE losses)

We created 72 different statistical features of both FCE and BCE losses, similar to stylometric features, we then filtered these based on univariate feature selection. We reatined the 25 most important features yields the best results

### 3.3. Classifier

The proposed classifier is an ensemble model that combines five different machine learning algorithms. This architecture integrates probabilistic, boosting, and tree-based techniques using a soft voting scheme with tuned weights. The main components of the ensemble include:

- **Gaussian Naive Bayes**: A probabilistic classifier based on the assumption of Gaussian-distributed features, serving as a baseline model.
- **AdaBoost Classifier**: An adaptive boosting algorithm implemented with a fixed random seed for reproducibility.
- **LightGBM Classifier**: A gradient boosting model optimized for efficient parallel computation.
- **CatBoost Classifier**: A gradient boosting algorithm optimized for production environments.
- **Random Forest Classifier**: A bagging ensemble of 256 decision trees that provides diverse and robust predictions.

The classifier is trained on 50 retained Bi-CE Loss and Stylometric features extracted from the text dataset provided by the PAN competition for training[7].

## 4. Results

**Table 1**
Performance metrics for the Pindrop model.[8]

| Team | Software | ROC-AUC | Brier | C@1 | F1 | F0.5u | Mean | FPR | FNR |
|---|---|---|---|---|---|---|---|---|---|
| Pindrop | blistering-band | 0.903 | 0.877 | 0.843 | 0.883 | 0.933 | 0.890 | 0.087 | 0.152 |

## 5. Conclusion

In this work, we proposed a hybrid method for detecting AI-generated text that leverages both bidirectional cross-entropy (Bi-CE) loss and a comprehensive set of stylometric features. By combining statistical patterns captured from pre-trained language models with linguistic cues traditionally used in authorship analysis, our system offers a robust approach to distinguishing human-written from machine-generated content. Through univariate feature selection, we refined 173 initial features down to the most informative 50, balancing model complexity and performance. The final ensemble classifier, composed of five complementary algorithms, demonstrated strong predictive capability on the PAN 2025 testing dataset. Our findings underscore the effectiveness of combining intrinsic language model signals with surface-level stylistic features for advanced text forensics. Future work will explore model generalization across domains and further integration of semantic features.

## 6. Acknowledgments

This work was conducted as part of our research at Pindrop Secruity. We thank our colleagues across the Pindrop team for their support and contributions to the experiments and development efforts described in this paper.

## Declaration on Generative AI

During the preparation of this work, we used GPT-4 in order to conduct grammar and spelling check. In addition, we used GPT-4 for figures 1 in order to generate figure format. After using these tools, we reviewed and edited the content as needed and assume full responsibility for the content of the publication.

## References

[1] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[2] H. Guo, S. Cheng, X. Jin, Z. Zhang, K. Zhang, G. Tao, G. Shen, X. Zhang, Biscope: Ai-generated text detection by checking memorization of preceding tokens, in: Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2024. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc808cf2d2444b0abcceca366b771389-Abstract-Conference.html.

[3] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, H. Liu, Stylometric detection of ai-generated text in twitter timelines (2023). URL: https://arxiv.org/abs/2303.03697. arXiv:2303.03697.

[4] C. Opara, Styloai: Distinguishing ai-generated content with stylometric analysis, in: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Springer Nature Switzerland, 2024, pp. 105–114. URL: https://arxiv.org/abs/2405.10129.

[5] Y. Lee, G. Awad, A. Butt, L. Diduch, K. Peterson, S. Seo, I. Soboroff, H. Iyer, 2024 NIST Generative AI (GenAI): Evaluation Plan for Text-to-Text (T2T) Discriminators, Technical Report, National Institute of Standards and Technology, 2024. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957332.

[6] Anthropic, Claude llm (version 1.0), Large language model, 2023. URL: https://www.anthropic.com, accessed: Dec. 2024.

[7] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in:

J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[8] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.

[9] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, 2009. URL: https://www.nltk.org/.

[10] P. University, About wordnet, 2010. URL: https://wordnet.princeton.edu/.

[11] L. Shen, Lexicalrichness: A small module to compute textual lexical richness, 2022. URL: https://github.com/LSYS/lexicalrichness.

[12] A. Hahn, textstat: Text statistics for python, 2018. URL: https://github.com/shivam5992/textstat.

# Appendix A

Tables 2 below describes the 25 features used in our ensemble model. Part-of-speech and sentence-based features were generated via parsing with NLTK [9].

**Table 2**
Feature descriptions for the top 25 features

| Feature | Additional Notes | Type |
|---|---|---|
| Punctuation Count | Punctuation defined using Python's `string.punctuation` | Textual |
| Special Character Count | Special characters defined by regex | Textual |
| Hapax Legomenon Rate | Percentage of words that occur only once in the text | Lexical |
| Rare Verb Count | Number of verbs not in the most common 5000 words per WordNet[10] | Lexical |
| Stop Word Count | Stop words defined using NLTK's stopwords | Lexical |
| TfIdf Variance | Variance in term-frequency / document-frequency by sentence | Lexical |
| Type to Token Ratio | Number of unique words / number of total words | Lexical |
| Unique Bigram Count | Calculated with NLTK ngram | Lexical |
| Unique Trigram Count | Calculated with NLTK ngram | Lexical |
| Unique Word Count (regex) | Word count based on regular expression match | Lexical |
| Unique Word Count (LexicalRichness) | Unique word count provided by the LexicalRichness package [11] | Lexical |
| Unique Word Percentage | Unique Word Count (regex) / Word Count | Lexical |
| Word Count (regex) | Word count calculated by splitting text by spaces | Lexical |
| Word Count (LexicalRichness) | Word count provided by the LexicalRichness package [11] | Lexical |
| Flesch Reading Ease Score | Flesch Reading Ease scores calculated using the `textstat` package [12] | Semantic |
| Gunning Fog Index | Gunning Fog Index scores calculated using the `textstat` package [12] | Semantic |
| Adverb-like Count | Count of tags starting with 'RB' | Syntactic |
| Adverbs | Count of words tagged with specific 'RB' part of speech | Syntactic |
| Complex Sentence Count | Count of sentences that contain more than one verb phrase | Syntactic |
| Max Pattern Repetition | Occurrences of most common pattern / number of sentences | Syntactic |
| Token Count | Calculated using NLTK parse | Syntactic |
| Relationship Count | Total count of dependency relations | Syntactic |
| Sentence Count | Sentences split with NLTK `sent_tokenize` | Syntactic |
| Sentence Length Consistency | Std / mean of sentence lengths | Syntactic |
| Unique Pronouns | Words matching NLTK pronoun tag | Syntactic |
| Average Connection Strength | Cosine similarity between BERT sentence embeddings | Discourse |