# Modelling the Relative Contributions of Stylistic Features in Forensic Authorship Attribution

**G. Çağatay Sat** and **John Blake** and **Evgeny Pyshkin**
School of Computer Science and Engineering
University of Aizu
Aizuwakamatsu
Japan
s1312006,jblake,pyshe@u-aizu.ac.jp

## Abstract

This paper explores the extent to which stylistic features contribute to the task of authorship attribution in forensic contexts. Drawing on a filtered subset of the Enron email corpus, the study operationalizes stylistic indicators across four groups: lexical, syntactic, orthographic, and discoursal. Using R Programming Language for feature engineering and logistic regression modelling, we systematically assessed both the individual and interactive effects of these features on attribution accuracy. Results show that n-gram similarity consistently outperformed all other features, with the combined model of n-gram similarity and its interaction with other features achieving accuracy, precision and F1 scores of 91.6%, 93.3% and 91.7% respectively. The model was subsequently evaluated on a subset of the TEL corpus to assess its applicability in a forensic setting. The findings highlight the dominant role of lexical similarity and suggest that integrating interaction effects can yield further performance gains in forensic authorship analysis.

## 1 Introduction

Authorship analysis is a core area in forensic linguistics, particularly relevant in criminal investigations where authorship may be disputed or unknown (Coulthard and Johnson, 2000). Recent advances in computational linguistics enable more systematic modelling of authorial identity based on quantifiable linguistic patterns (Juola, 2006).

This study investigates the relative impact of different stylistic features on authorship attribution. Using the Enron email corpus, we identify stylistic variables, assess different combinations of them and their contributions to author identification. The study adopts a stepwise modelling approach using R (R Core Team, 2025), where each style variable is first computed separately, then combined to assess interactions and additive effects. The overarching aim is to determine which style variables are most influential and whether combining interactions of variables result in a more accurate model of authorship attribution.

## 2 Forensic Authorship Analysis

Authorship attribution involves identifying the likely author of a text based on linguistic features (Grant, 2013). In forensic settings, authorship analysis must meet evidentiary standards of reliability, validity, and interpretability (Coulthard and Johnson, 2013). Traditionally, this analysis has relied on qualitative judgements of stylistic idiosyncrasies. However, quantitative models grounded in corpus linguistics enable more rigorous and reproducible approaches (Koppel et al., 2009).

Despite the progress in computational methods, feature selection remains a major challenge in authorship analysis (Stamatatos, 2009). Different studies employ different sets of linguistic features, making it difficult to compare results and replicate findings. This study aims to investigate this issue by creating models using different combinations of style features and evaluating their relative contributions in a controlled setting.

## 3 Style

Stylistic features reflect habitual, often unconscious linguistic choices, that remain relatively stable across writing contexts (Argamon et al., 2003; Argamon, 2009). Style can be used to distinguish between human and AI-generated texts (Blake et al., 2025). Stylistic features capture individual expression patterns which may be independent of topic and genre. Distinctive authorship markers or features may be thought of as a “linguistic fingerprint” (Eder, 2011; Eder and Górski, 2023). Stylometric methods are able to identify authorship due to the slightly different registers each author adopts (Grieve, 2023), using measures such as function word frequency, syntactic preferences, and punctuation patterns (Argamon and Levitan, 2005; Torney et al., 2012).

Style, however, is a complex construct, subject to diverse theoretical interpretations and operational definitions. Stylistic features encompasses a broad spectrum of linguistic dimensions: lexical, syntactic, orthographic, pragmatic, and discoursal, each of which gives rise to multiple potential variables (Stamatatos, 2009; Grieve, 2007). Even canonical features, such as function words, admit varied representations. Relative frequency, positional distribution, or interactional patterns (Koppel et al., 2009) are cases in point.

The feature space of stylistic variables is effectively unbounded. Analysts may consider fine-grained markers such as POS bigrams, punctuation clusters, or spacing conventions, each encoding idiolectal signatures (Juola, 2006). Moreover, the combinatorial possibilities across features (e.g. grouped by linguistic level, communicative function, or statistical behaviour) are further multiplied through transformations, interactions, and composite indices. As Eder (Eder, 2011) demonstrates, even stylistic features with limited linguistic interpretability, such as letter n-grams, can result in high attribution accuracy, particularly in morphologically rich or inflectional languages. Nonetheless, the central premise of authorship attribution remains: that distinctive style, however instantiated, constitutes a stable, identifiable signal (Grant, 2013; Coulthard and Johnson, 2007). In forensic contexts, where attribution must be both accurate and explainable, modelling stylistic distinctiveness is indispensable (McMenamin, 2002; Turell, 2010).

# 4 Aim

In this study, we use a feature-based model for authorship attribution to compare the relative contributions of stylistic features in the attribution process. We conceptualize authorship identity as a function of style variables and operationalize these features using reproducible metrics. This approach allows forensic linguists to attribute authorship with greater clarity and explain the basis of attribution in clear linguistic terms.

We aim to answer two research questions (RQs), namely:

**RQ1:** Which stylistic features contribute most to accurate authorship attribution?

**RQ2:** What combinations of features yield optimal performance?

# 5 Method

## 5.1 Corpus Description

Two datasets were harnessed: the Enron Corpus (Hussain, 2020) and the Threatening English Language (TEL) Corpus (Gales et al., 2022). The Enron corpus consists of email communications from $20,328$ users, totaling over 500,000 messages. Due to the nature of the corpus there is a lot of repetitiveness (e.g. forwarded emails), thus to get a better grip of each author's profile these were filtered. In addition, outlying documents comprising too few or too many word tokens were removed. For this study, we used a tabular format of corpus available on Kaggle (Hussain, 2020). We selected 395 authors with at least 20 and at most 30 messages each and sampled $70\%$ for training, $28.9\%$ for validation and $1.1\%$ for testing. The TEL corpus, comprising 309 texts written by 203 authors, was also drawn upon for testing. Given the scarcity of publicly available corpora containing threatening language, the limited variety in the TEL corpus can be considered justifiable, as it remains an indispensable resource.

## 5.2 Features and their Extraction

A logistic regression model that consists of all features was created to assess the relative contribution of stylistic features. Categories include lexical (e.g. function words, lexical diversity), syntactic (e.g. sentence types, POS distribution), orthographic (e.g. punctuation, spelling), and discourse-pragmatic features (e.g. discourse markers). These initial results identify the most distinctive variables, which are then combined to evaluate their impact on classification accuracy and interaction effects.

Preprocessing involved tokenization, part-of-speech tagging using `spacyr` (Benoit and Matsuo, 2023), and conversion to lowercase. Function words, punctuation, sentence length, and other metrics were extracted using `quanteda` (Benoit et al., 2018) and custom R scripts.

All features were normalised using z-scores based on the training set to account for scale differences across the data. Features used are shown in Table 1. The relative features 1 through 25 are computed as $\frac{\text{Frequency}}{T}$, where $T$ is the total number of non-numerical tokens, the relative feature 26

(RelCase) – as $\frac{\text{Frequency}}{K}$, where $K$ is the total number of letters, while the relative features 27 through 29 – as $\frac{\text{Frequency}}{N}$, where $N$ is the total number of word tokens.

Feature 30 (CTTR) is computed as $\frac{\text{Type}}{\sqrt{2N}}$, where $N$ is the total number of word tokens, while for the feature 31, the standard cosine measure is used (see equation (2)). Then, the feature differences for each author-document combination were calculated using the Manhattan distance 1.

$$\textit{Manhattan Distance} = |x_1 - x_2| \tag{1}$$

The n-gram similarity score did not require such a step, as it was computed directly using cosine similarity between each document-author combination using 2, where $p$ is the number of n-grams, $x_{i,k}$ is the $k$th n-gram in the author profile vector, and $x_{d,k}$ is the $k$th n-gram in the document vector.

$$\begin{matrix}\textit{Cosine} \\ \textit{Similarity}\end{matrix} = \frac{\sum_{k=1}^{p} x_{i,k} x_{d,k}}{\sqrt{\sum_{k=1}^{p} x_{i,k}^2}\sqrt{\sum_{k=1}^{p} x_{d,k}^2}} \tag{2}$$

Furthermore, correct combinations of author-document pairs were labeled 1, while others 0. This made it possible to use the resulting sets in a binomial model construction, focusing only on the difference between the suspected authors and the documents.

### 5.3 Model Construction

We trained models for stylistic features using logistic regression with the `glmnet` package (Tay et al., 2023). During model construction, the training set was fitted using elastic net regularisation, and the resulting model was evaluated on a separate validation set. Validation performance was used to identify the optimal regularisation parameter $\lambda$, which was then used to refit the model on the training data. This final model was subsequently applied to the test set. For each document, the model output a score for every candidate author, and the author with the highest score was selected as the predicted label. Insights from this initial modeling phase informed further training runs, incorporating only the most discriminative features and their combinations. After several iterations of feature selection and model refinement, n-gram similarity remained the sole consistently predictive feature. Other features were either assigned zero coefficients by the model or contributed negligibly or negatively to predictive performance. As a result, we retained n-gram similarity as the core feature and reintroduced interactions between it and selected stylistic features.

The model created from this process gave the best accuracy and Macro-F1 score among the tested models.

### 5.4 Evaluation Metrics

We evaluated classification performance using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score as defined in equations 3–6.

These metrics were computed by comparing the predicted author (i.e. the author with the highest score assigned by the model) to the true author for each document. Macro-averaging treats each author equally, regardless of frequency.

$$\textit{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of documents}} \tag{3}$$

$$\textit{Macro-Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$\textit{Macro-Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{5}$$

$$\textit{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6}$$

Here, $TP_i$, $FP_i$, and $FN_i$ refer to the true positives, false positives, and false negatives for author $i$, and $N$ is the number of authors.

Statistical robustness was assessed using ANOVA and pairwise t-tests on classification scores.

The R scripts used in this study are available at: `https://github.com/gcs0/ModellingAA`.

## 6 Results

### 6.1 Initial Model Performance

The initial model, which included all features, produced the lowest performance, with both accuracy and macro-precision at 89.9Table 2 shows each model and its associated coefficient ($\beta$). In logistic regression, each coefficient ($\beta$) represents the estimated change in the log-odds of the outcome

Table 1: Feature Means and Standard Deviations of Features-per-Document Basis

| No. | Feature code | Description | Mean | SD |
|---|---|---|---|---|
| 1 | RelADJ | Relative Adjective Count | 0.0341 | 0.0200 |
| 2 | RelADP | Relative Adposition Count | 0.0726 | 0.0287 |
| 3 | RelADV | Relative Adverb Count | 0.0196 | 0.0160 |
| 4 | RelAUX | Relative Auxiliary Verb Count | 0.0406 | 0.0232 |
| 5 | RelCCONJ | Relative Coordinating Conjunction Count | 0.0202 | 0.0133 |
| 6 | RelDET | Relative Determiner Count | 0.0543 | 0.0280 |
| 7 | RelINTJ | Relative Interjection Count | 0.0051 | 0.0071 |
| 8 | RelNOUN | Relative Noun Count | 0.1389 | 0.0576 |
| 9 | RelPART | Relative Particle Count | 0.0190 | 0.0140 |
| 10 | RelPRON | Relative Pronoun Count | 0.0536 | 0.0365 |
| 11 | RelPROPN | Relative Proper Noun Count | 0.1343 | 0.0825 |
| 12 | RelPUNCT | Relative Punctuation Count | 0.1383 | 0.0890 |
| 13 | RelSPACE | Relative Space Character Count | 0.1148 | 0.0476 |
| 14 | RelVERB | Relative Verb Count | 0.0806 | 0.0331 |
| 15 | RelNUM | Relative Numeral Count | 0.0387 | 0.0387 |
| 16 | RelSCONJ | Relative Subordinating Conjunction Count | 0.0132 | 0.0109 |
| 17 | RelSYM | Relative Symbol Count | 0.0093 | 0.0139 |
| 18 | RelDot | Relative Dot Count | 0.0666 | 0.0746 |
| 19 | RelGS | Relative Genitive/Saxon Genitive Count | 0.0098 | 0.0395 |
| 20 | RelSlash | Relative Slash Count | 0.0165 | 0.0583 |
| 21 | RelSQuote | Relative Single Quote Count | 0.0080 | 0.0142 |
| 22 | RelDash | Relative Dash Count | 0.0476 | 0.0836 |
| 23 | RelEMark | Relative Exclamation Mark Count | 0.0027 | 0.0084 |
| 24 | RelColon | Relative Colon Count | 0.0188 | 0.0275 |
| 25 | RelX | Relative Other Count | 0.0126 | 0.0341 |
| 26 | RelCase | Relative Uppercase Count | 0.0877 | 0.0717 |
| 27 | RelFunction | Relative Function Word Count | 0.4055 | 0.1215 |
| 28 | RelAWL | Relative Academic Word Count | 0.0363 | 0.0281 |
| 29 | RelMisspelled | Relative Misspelled Word Count | 0.0007 | 0.0037 |
| 30 | CTTR | Corrected Type-Token Ratio | 5.5692 | 1.4447 |
| 31 | ngram_sim | | 0.3310 | 0.1631 |

**Note.** `ngram_sim` metrics are calculated from author-document pairs.

(i.e. correct authorship attribution) associated with a one-unit increase in the corresponding predictor variable, holding all other variables constant; positive coefficients indicate an increased likelihood of a correct match, while negative coefficients indicate a decreased likelihood.

Table 2: Initial Model Coefficients

| Model | $\beta$ | Model | $\beta$ |
|---|---|---|---|
| (intercept) | -4.70 | RelCase | 0.06 |
| RelFunction | 0.69 | RelAWL | -0.17 |
| CTTR | -0.03 | RelDot | 0.17 |
| RelGS | 0.26 | RelSlash | -0.09 |
| RelSQuote | -0.24 | RelDash | -0.21 |
| RelEMark | -0.13 | RelColon | 0.12 |
| RelADJ | -0.27 | RelADV | -0.08 |
| RelAUX | -0.19 | RelCCONJ | -0.05 |
| RelDET | -0.42 | RelINTJ | -0.06 |
| RelNOUN | -0.27 | RelPART | -0.16 |
| RelPRON | -0.65 | RelPROPN | -0.38 |
| RelPUNCT | -0.47 | RelSPACE | -0.35 |
| RelVERB | 0.35 | RelNUM | -0.05 |
| RelSCONJ | -0.16 | RelSYM | 0.01 |
| RelX | -0.25 | | |
| ngram_sim | 1.56 | | |

### 6.2 N-gram Only Performance

Taking n-gram score solely increased the accuracy to 90.8% and macro-precision to 92.8%.

### 6.3 N-gram and Interaction of Other Features Performance

Combining n-gram similarity scores with their interaction with other features raised accuracy to 91.6% and macro-precision to 93.3%.

Table 3 displays the coefficients of the final model, incorporating interaction terms between n-gram similarity and selected stylistic features. These interactions produced modest but consistent improvements in predictive performance.

### 6.4 Model Comparison

As can be seen in Table 4, the initial model gave an accuracy rating of 89.9%, while the final Model, which depends on n-gram similarity and its interactions with nine other features, gave a rating of 93.3%. The rising trend can also be seen on other metrics with Macro-Recall increasing from 89.5% to 91.1%, Macro-F1 from 89.6% to 91.7%, and most substantially Macro-Precision from 89.9% to

Table 3: Interaction Model Coefficients

| Model | $\beta$ |
|---|---|
| (Intercept) | -6.64 |
| ngram_sim | 2.03 |
| ngram_sim:RelSQuote | -0.06 |
| ngram_sim:RelDash | -0.03 |
| ngram_sim:RelADJ | -0.05 |
| ngram_sim:RelDET | -0.10 |
| ngram_sim:RelNOUN | -0.06 |
| ngram_sim:RelPRON | -0.19 |
| ngram_sim:RelPROPN | -0.09 |
| ngram_sim:RelPUNCT | -0.10 |
| ngram_sim:RelSPACE | -0.11 |

93.3%. The delta scores ($\Delta$) show the difference between the initial and final values.

Table 4: Comparison of Initial and Final Model Performance Metrics (in %)

| Metric | Initial | Final | $\Delta$ score |
|---|---|---|---|
| Macro-Precision | 89.9 | 93.3 | +3.4 |
| Macro-Recall | 89.5 | 91.1 | +1.6 |
| Accuracy | 89.9 | 91.5 | +1.6 |
| Macro-F1 | 89.6 | 91.7 | +2.1 |

Figure 1 visually summarises the performance differences between the initial and final models across all evaluation metrics presented in Table 4.

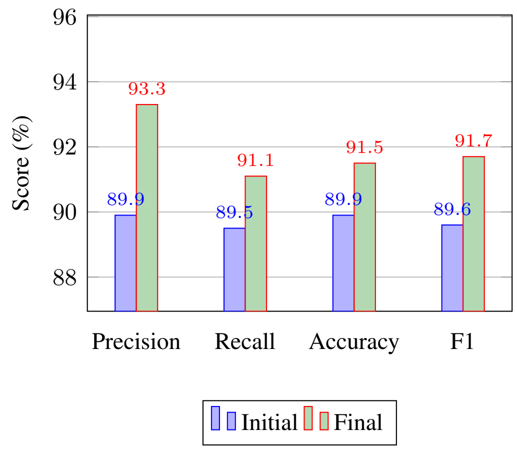


Figure 1: Performance comparison of initial and final models.

### 6.5 Confusion Matrix

The final model achieved over 90% on all macro-level test metrics mentioned. The confusion matrix heat map for all author accuracies is presented in Figure 2.

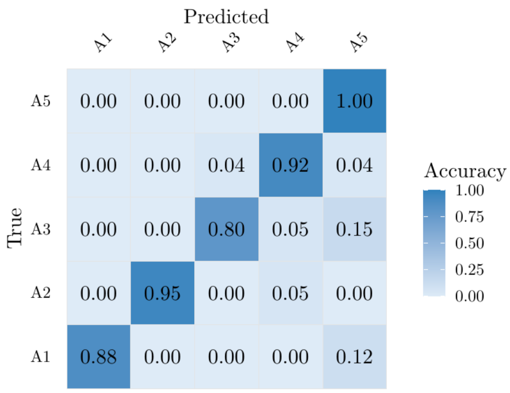


Figure 2: Confusion Matrix Heat Map for 5 Authors

Figure 3 shows a confusion matrix heat map for the final model, tested on a random sample of 10 authors and 226 documents in Enron Dataset. The model's overall accuracy was 80%, while macro-precision was 84%. Increasing the author number by double the original caused a considerable decrease in accuracy and macro-precision.

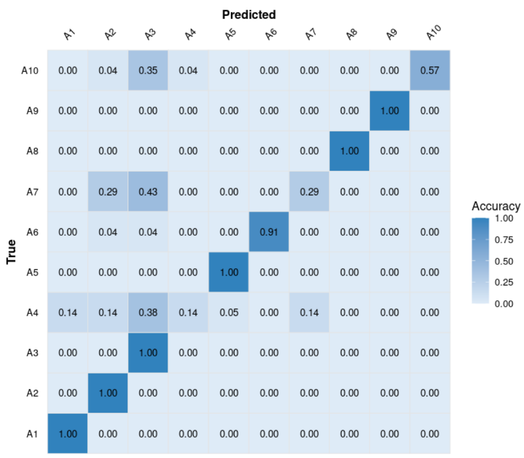


Figure 3: Confusion Matrix Heat Map For 10 Authors

When the same 10-author set is used with the final model set to apply an arbitrary confidence threshold of 0.002—classifying predictions below this threshold as Unclassified—the accuracy drops to 78.3%, while the macro-precision increases to 88%. The heat map corresponding to this approach is shown in Figure 4.

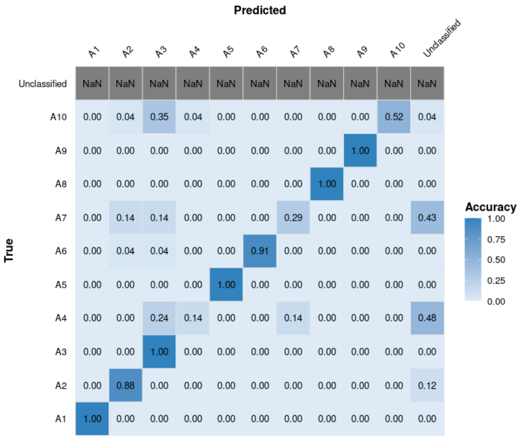


Figure 4: Confusion Matrix Heat Map with Unclassified

### 6.6 Performance on TEL Corpus

The TEL Corpus (Gales et al., 2023) consists of 309 documents attributed to over 200 authors. However, since many documents are written by unknown individuals and a significant number of authors do not meet the minimum document count required for analysis, we selected a subset of 24 documents authored by three individuals. When tested, the resulting accuracy was 79.17%, the macro-F1 score was 81.42%, and the macro-precision was 81.67%. This performance decline is likely due to the nature of the evaluation set: all selected authors had nine or fewer documents, which limited the reliability of author-level feature estimates. In addition, some documents contained relatively few tokens, further impacting classification performance. The confusion matrix heat map for the test can be seen in Figure 5.

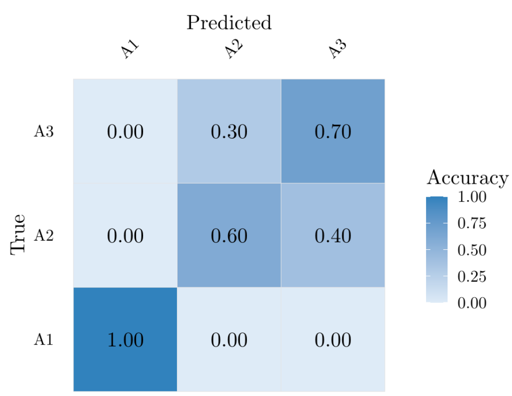


Figure 5: Confusion Matrix Heat Map for TEL Corpus

## 7 Discussion

The results of this study indicate that lexical patterns, particularly those captured through n-gram similarity, play a dominant role in authorship attribution. This is most likely because n-grams can encode stylistic patterns across multiple linguistic levels, including lexical choice, syntactic sequencing, and local discourse structure (Ríos-Toledo et al., 2022). Unlike individual stylistic markers, which may be sparse or vary in discriminatory power across documents, n-gram patterns provide dense, composite representations of authorial style. As such, they are well-suited to capturing the habitual linguistic choices that form the basis of authorship profiling. The consistent outperformance of n-gram similarity across all tested models highlights its robustness and interpretive value within forensic contexts.

Although other features such as punctuation, function words, and POS distributions are traditionally recognised as useful stylistic indicators, their standalone predictive power was limited in this study. However, interaction effects between n-gram similarity and selected stylistic variables produced modest but measurable attribution accuracy and precision gains. This suggests that while these features may not independently improve model performance, they may still contribute complementary information when combined with lexical similarity measures.

The findings of this study are consistent with prior work that emphasises the centrality of function words and other low-level features in authorship analysis (Argamon and Levitan, 2005). Earlier studies have demonstrated that function words are less topic-sensitive and more reflective of unconscious authorial habits, making them reliable markers in stylometric applications. Our results support this view, though they also reveal that such features, when used in isolation, are outperformed by broader lexical patterning captured through n-gram similarity.

The study also aligns with the conclusions of Stamatatos (2009), who advocated for combining multiple stylistic feature groups to improve attribution robustness. However, our work extends this line of inquiry by modelling not only additive but also interactive effects between variables. Rather than assuming that each feature contributes independently, we tested how their combinations influence the predictive performance of the model. This interaction-based approach revealed that some features contribute value only in conjunction with others, a finding that refines existing assumptions about feature selection in forensic authorship analysis.

This study is limited by its reliance on a formal, topic-consistent email corpus, which may not reflect the stylistic variability of real-world forensic texts. Although tested on a small subset of the TEL corpus, reduced performance underscores the need for broader genre and context validation. Additionally, while logistic regression offers interpretability, it may overlook nonlinear or complex feature interactions. Future work should evaluate more expressive models, such as SVMs or neural networks, provided their outputs remain forensically explainable.

Forensic linguists and practitioners involved in authorship attribution must balance predictive performance with explainability. This study demonstrates that models built on n-gram similarity, combined with interpretable stylistic features, can offer both high accuracy and plausible linguistic explanations. By quantifying the contribution of individual features and their interactions, the model facilitates transparent reporting in forensic contexts, where methods must be open to scrutiny in legal proceedings.

The findings also suggest that feature interactions should be considered when constructing attribution models. Features that do not independently yield predictive value may still enhance performance when combined with more robust indicators. This insight encourages the use of layered modelling strategies that foreground explainability while retaining methodological rigour.

## 8 Conclusion

This study investigated the relative contributions of stylistic features to forensic authorship attribution using a filtered subset of the Enron email corpus. Lexical, syntactic, orthographic, and discoursal features were extracted and modelled using logistic regression with elastic net regularisation. N-gram similarity emerged as the most effective predictor, outperforming all other features. Incorporating interaction terms with selected stylistic variables further improved performance.

Although most standalone features contributed minimally, their utility may vary with genre, corpus type, or attribution task. Future work should examine their relevance in more diverse, emotionally or deceptively charged texts.

The findings support using interpretable, feature-based models in forensic contexts, offering transparent, quantifiable evidence for author identification. Ongoing research will extend this model to genre-diverse corpora and integrate content and genre-sensitive features to enhance forensic validity.

## References


Shlomo Argamon. 2009. Automatically profiling the author of an anonymous text. In John Olsson, editor, *Crime and Language: The Linguistic Description of Criminal Language*, pages 69–76. Continuum.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Arye R Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–417.

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7.

Kenneth Benoit and Akitaka Matsuo. 2023. *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 1.3.0.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

John Blake, Abu Saleh Musa Miah, Krzysztof Kredens, and Jungpil Shin. 2025. Detection of ai-generated texts: A bi-lstm and attention-based approach. *IEEE Access*.

Malcolm Coulthard and Alison Johnson. 2000. *Forensic Linguistics: An Introduction to Language in the Justice System*. Routledge.

Malcolm Coulthard and Alison Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge.

Malcolm Coulthard and Alison Johnson. 2013. *Authorial and Temporal Attribution: Linguistic, Stylometric and Forensic Approaches*. Routledge.

Maciej Eder. 2011. Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6:99–114.

Maciej Eder and Rafał L Górski. 2023. Stylistic fingerprints, pos-tags, and inflected languages: A case study in polish. *Journal of Quantitative Linguistics*, 30(1):86–103.

Tammy Gales, Andrea Nini, and Ellen Symonds. 2022. The threatening english language (tel) corpus.

Tammy Gales, Andrea Nini, and Ellen Symonds. 2023. The threatening english language (tel) corpus. https://research.manchester.ac.uk/en/datasets/the-threatening-english-language-tel-corpus. Accessed April 2025.

Tim Grant. 2013. TXT 4N6: Method, consistency, and distinctiveness in the analysis of sms text messages. *Journal of Law and Policy*, 21(2):467–494.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Jack Grieve. 2023. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Javed Hussain. 2020. Enron email dataset.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Gerald R McMenamin. 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Germán Ríos-Toledo, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. Detection of changes in literary writing style using n-grams as style markers and supervised machine learning. *Plos one*, 17(7):e0267590.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. 2023. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31.

Kerrie Torney, Philip M McCarthy, and Danielle S McNamara. 2012. Using linguistic features to assess writers' similarity across genres. In *Proceedings of the 25th International FLAIRS Conference*.

Maria Teresa Turell. 2010. The use of linguistic evidence in forensic linguistics: Points of convergence and divergence between linguists and lawyers. In Malcolm Coulthard and Alison Johnson, editors, *The Routledge Handbook of Forensic Linguistics*, pages 517–530. Routledge.