applied sciences

Article

Authorship Attribution in Less-Resourced Languages: A Hybrid Transformer Approach for Romanian

Melania Nitu [1] and Mihai Dascalu [1,2,3,*]

[1] Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania; suzana_melania.nitu@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr.3, 050044 Bucharest, Romania
[3] The "G. Călinescu" Institute of Literary History and Theory, Romanian Academy, Calea 13 Septembrie, 050711 Bucharest, Romania
* Correspondence: mihai.dascalu@upb.ro

Abstract: Authorship attribution for less-resourced languages like Romanian, characterized by the scarcity of large, annotated datasets and the limited number of available NLP tools, poses unique challenges. This study focuses on a hybrid Transformer combining handcrafted linguistic features, ranging from surface indices like word frequencies to syntax, semantics, and discourse markers, with contextualized embeddings from a Romanian BERT encoder. The methodology involves extracting contextualized representations from a pre-trained Romanian BERT model and concatenating them with linguistic features, selected using the Kruskal–Wallis mean rank, to create a hybrid input vector for a classification layer. We compare this approach with a baseline ensemble of seven machine learning classifiers for authorship attribution employing majority soft voting. We conduct studies on both long texts (full texts) and short texts (paragraphs), with 19 authors and a subset of 10. Our hybrid Transformer outperforms existing methods, achieving an F1 score of 0.87 on the full dataset of the 19-author set (an 11% enhancement) and an F1 score of 0.95 on the 10-author subset (an increase of 10% over previous research studies). We conduct linguistic analysis leveraging textual complexity indices and employ McNemar and Cochran's Q statistical tests to evaluate the performance evolution across the best three models, while highlighting patterns in misclassifications. Our research contributes to diversifying methodologies for effective authorship attribution in resource-constrained linguistic environments. Furthermore, we publicly release the full dataset and the codebase associated with this study to encourage further exploration and development in this field.

Keywords: authorship attribution; linguistic features; contextualized embeddings; hybrid Transformer; ensemble learning; linguistic analysis; natural language processing

check for updates

Citation: Nitu, M.; Dascalu, M. Authorship Attribution in Less-Resourced Languages: A Hybrid Transformer Approach for Romanian. *Appl. Sci.* **2024**, *14*, 2700. https://doi.org/10.3390/app14072700

Academic Editors: Andrea Omicini, Maja Matetic and Marija Brkić Bakarić

Received: 18 February 2024
Revised: 14 March 2024
Accepted: 21 March 2024
Published: 23 March 2024

1. Introduction

Authorship attribution is the process of attributing a given text to its rightful author based on linguistic and stylistic characteristics. It involves extracting and analyzing distinctive patterns in writing styles, ranging from syntactic structures to lexical choices [1]. The significance of authorship attribution extends across various domains, namely, plagiarism or fake news detection [2], forensic linguistics [3], or literary studies [4]. Beyond cybersecurity or fake news detection, authorship attribution holds considerable value in the educational domain, particularly in digital libraries [5–7]. Moreover, identifying the authorship of academic works, research papers, and literary contributions can assist in organizing and categorizing content, as it can facilitate tracking academic contributions and ensuring proper acknowledgment of one's work.

Within the domain of authorship attribution, Misini et al. [8] identify three distinct tasks, namely, (1) authorship identification [9,10], which aims to identify the author of a given work; (2) authorship profiling or characterization [11–13], which explores the demographic traits such as age, gender, or educational level associated with the author;

and (3) authorship verification or similarity detection [14], which seeks to establish whether the presumed author aligns with the actual writer of a given document. All these tasks can be formulated as detection problems, where the goal is to determine the degree of similarity between texts by comparing their writing styles. Understanding the unique features that distinguish one author from another aids in solving practical problems and highlights the complexity of human writing. The methodology for authorship attribution typically involves a combination of computational and statistical techniques.

This study focuses on enhancing authorship attribution, specifically for the Romanian language. While existing methodologies have shown promising results in English and other widely studied languages, the linguistic complexity of Romanian presents unique challenges for which out-of-the-box solutions do not work. Even though the immediate focus is on Romanian, the methodologies applied in this study can easily be extended and applied to other limited or less-resourced languages. This study proposes a hybrid Transformer strategy to address these challenges, combining handcrafted linguistic features with contextualized embeddings. This approach is compared with a baseline ensemble of seven machine learning classifiers considering majority soft voting. In addition to proposing an enhanced authorship attribution method, this research includes a comparative evaluation against existing studies for Romanian and performs an analysis of the top discriminative features and authors' writing styles. The aim is to assess the efficiency of the proposed strategies when compared to previously established methodologies.

Current Study Contributions

Outlined below are the main contributions of our research:

- We release a publicly available corpus of Romanian stories comprising 1263 texts and 12,516 paragraphs written by 19 authors. The dataset can be accessed at https://huggingface.co/datasets/readerbench/ro-stories (accessed 12 February 2024). Additionally, this study examines existing methodologies and limitations in authorship attribution, with an emphasis on languages with limited resources, such as Romanian.
- We introduce a hybrid Transformer model that achieves state-of-the-art performance for authorship attribution for Romanian. This model, available at https://github.com/readerbench/ro-auth-detect (accessed 12 February 2024), combines the top predictive linguistic features selected using the Kruskal–Wallis mean rank and contextualized embeddings from a Romanian BERT model to predict the authors. A comparative analysis is conducted with ensemble learning, incorporating seven traditional machine learning classification algorithms based on linguistic features.
- We present a detailed analysis of the distinctions in the authors' writing styles based on the top discriminative linguistic features.

2. Related Work on Authorship Attribution

Over the years, various methodologies have been employed for authorship attribution. Traditional approaches often rely on statistical measures and handcrafted features, while newer methods leverage machine learning (ML)/deep learning (DL) models based on Transformer architectures. Feature-based models, stylometric analyses, and linguistic profiling are among the common methodologies employed. These techniques investigate aspects such as word frequency, sentence length, syntactic structures, and other linguistic attributes to create models capable of differentiating individual writing styles.

2.1. Linguistic Feature Engineering

In authorship attribution, the effective selection of features plays an important role in enhancing the performance of prediction models. Table 1 summarizes the main types of linguistic features generally employed in author attribution studies [15–17].

These feature categories collectively contribute to a comprehensive analysis of the linguistic and stylistic aspects corresponding to written texts. The selection and combination of these features depend on the specific goals of the authorship attribution task and the characteristics of the textual data being examined.

Table 1. Classification of linguistic features integrated in subsequent experiments.

Category	Metrics
Stylometric/Surface Features	Average sentence length Word usage patterns Punctuation distribution Vocabulary richness
Lexical Features	Word frequencies distribution Vocabulary size Unique words usage N-gram frequency
Syntactic Features	Part-of-speech distribution Syntactic tree structure Grammatical patterns Sentence structure complexity Dependency relations
Structural Features	Document organization Paragraph and sentence structure Heading usage Document length Formatting style
Content-Specific Features	Genre-specific keywords Theme-related phrases Domain-specific terminology Named entity recognition
Semantic Features	Word semantics from WordNet Pronoun usage & Co-reference patterns Cohesion Semantic similarity measures using topic modeling (e.g., LDA), static embeddings (e.g., Word2Vec and GloVe), or contextualized embeddings (e.g., BERT)
Discourse Features	Coherence measures Connective words usage
Time-Based Features	Temporal patterns of word usage Writing style evolution over time Date and time of document creation Writing trends Time-sensitive vocabulary

A common approach is to extract features with the highest discriminatory capabilities [15,16]. The selection of an appropriate set of features should consider several factors, including language, literary style (prose or poetry), text domain, text length, quantity of samples, and the number of considered features. The vast array of potential features necessitates a careful selection process to optimize the model performance. Feature selection involves identifying and keeping the most relevant and discriminative attributes while discarding the redundant or the less informative ones. This process mitigates the risk of overfitting, especially for the high-dimensional feature sets, and streamlines the computational complexity of the attribution model.

2.2. Classic Machine Learning Models

Diverse machine learning (ML) algorithms have been applied to analyze and interpret linguistic and stylistic features extracted from the written texts. Techniques such as Support Vector Machines (SVMs), Random Forests (RF), or neural networks (NNs) learn patterns and associations within the data, enabling models to discern the unique writing styles of different authors. The success of ML-based authorship attribution predominantly relies on the careful selection and engineering of features.

Statistical methods play a central role in quantitatively distinguishing an author's writing style. Computational stylometry employs methodologies such as word length, sentence length, vocabulary richness, and frequency analysis to analyze texts. Each author possesses a distinct writing style, which can be quantitatively differentiated using statistical techniques. The process involves text pre-processing, feature extraction, classification, and author attribution, with various classifiers utilized for this purpose. Elayidom et al. [18] leveraged a fuzzy learning classifier and an SVM, with the SVM model achieving higher accuracy in author identification. Combining these classifiers resulted in improved accuracy compared to individual SVM and fuzzy classifier approaches.

Diverse ML classifiers have been employed across various authorship analysis tasks. Chanchal et al. [19] introduced a stacked classifier for source code authorship attribution, aiming to identify the author of a given code in the context of open-source repositories like GitHub, Codalab, Kaggle, and Codeforces online judge. With the increasing number of software submissions, there is a growing concern about code plagiarism. The authors used a TF-IDF-based mechanism to represent the source code, utilizing word and character n-grams to generate code vectors. These vectors were then fed to various ML classifiers for prediction. Their approach achieved an accuracy of 82.95% on the test data. The practical application of their method extends to detecting plagiarized code and helping prevent legal issues in software development.

Alsmearat et al. [20] employed several classifier combinations, including Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVMs), and k-Nearest Neighbor (k-NN), for gender identification problems. SVMs and heuristics-based classifiers, coupled with stylometric features, yielded the best results. Similarly, Khdr et al. [11] applied NB, SVM, and DT (J48) for author profiling, specifically for gender and age prediction in SMS messages; the J48 Decision Tree exhibited higher accuracy in gender prediction. A distinct study [9] explored SVM and k-NN algorithms for author attribution, experimenting with varying instances and k values; in their research, SVM outperformed other methods, and for k values, 1-NN and 5-NN produced superior results. Abuhammad et al. [21] compared their results with the literature's best outcomes, conducting experiments with diverse feature combinations and data-preprocessing methods. In their tweet data analysis, they applied the chunking technique, achieving a high accuracy across varying chunk sizes.

A popular approach is to combine several ML models and train an ensemble of classifiers. Abbasi et al. [22] proposed employing ensemble learning and traditional ML techniques for authorship attribution. Their method extracted valuable author characteristics using a count vectorizer and bi-gram term frequency-inverse document frequency (TF-IDF). The experiment involved a comprehensive news dataset divided into three subsets. This twofold study focused on selecting 10 authors in the first fold and 20 authors in the second fold for a comparative baseline. Results indicated that the proposed ensemble learning outperformed previous state-of-the-art studies, achieving accuracy gains of 3.14% and 2.44% for the first scope, while the second scope exhibited 5.25% and 7.17% improvement.

2.3. Deep Learning Architectures

Deep learning (DL)-based authorship attribution considers leveraging NNs to discern patterns in written texts. This approach explores the hierarchical and complex relationships within language, allowing for a better understanding of authorial styles. DL architectures such as Recurrent Neural Network (RNN) and Transformer-based models excel at capturing long-range dependencies and contextual information, which are essential for authorship identification. By ingesting large amounts of textual data, these models autonomously learn the features that characterize an author's writing style, mitigating the need for manual feature engineering.

Research has shown the effectiveness of NN models in achieving precise author identification. Qian et al. [23] reached a high performance using a Gated Recurrent Unit (GRU), a type of RNN, on a dataset sourced from the Gutenberg Project (https://www.gutenberg.org/, accessed 12 February 2024). Their model sequentially represented words using GloVe pre-trained embeddings [24], employing average pooling and incorporating another GRU layer to capture the sequence of sentences within an article. Among the tested models, the article-level GRU achieved the highest accuracy, reaching 69.1% on the C50 dataset and 89.2% on the Gutenberg dataset. The study also explored authorship verification, where a Siamese network-based model achieved exceptional performance, achieving 99.8% accuracy on both the C50 and Gutenberg datasets. In a related exploration, Vaz [25] explored textual recommendations based on a limited set of stylometric features. Additionally, Pera [26] considered the author's writing style, yet the model learned from the perspective of reviewers rather than automatically extracting information from the content of each book; however, relying on user reviews may introduce bias, requiring widely read and reviewed books in the systems. Extending the scope, Zhang et al. [27] leveraged the entire e-book's content to organize the authors hierarchically. The study leveraged a multilayer self-organizing map (MLSOM), where authors were represented using a hierarchical tree structure, encompassing features such as biography, books, pages, and paragraphs. The benefit of employing the MLSOM algorithm stood in its capacity to efficiently handle complex representations, thus clustering e-books and authors.

Gupta et al. [28] employed two RNNs, namely Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The study evaluated the performance of index-based word embeddings against pre-trained embeddings on two separate datasets. When combined with word–index embeddings, the GRU model achieved a high accuracy in one of the datasets.

Authorship attribution for short online text snippets, such as emails, blogs, or forum posts, involves identifying the original author using stylometric traits as discussed in the research of Modupe et al. [29]. Nonetheless, accurately capturing an author's unique writing style presents challenges due to the complexity of the text. The authors proposed a model architecture that combines a convolutional neural network (CNN) with a bidirectional Long Short-Term Memory (LSTM) encoder. The framework extracted lexical stylometric characteristics, which were then processed through a bidirectional encoder to generate the features' vector representation. The feature array underwent normalization using distributed high networks to improve generalization and reduce errors. Subsequently, the bidirectional decoder analyzed the feature vector to extract the distinct writing styles of individual authors. The final step involved predicting the author of a text snippet using a classification layer. Results showed that their approach outperformed previous strategies.

Another study by Škorić et al. [30] investigated the use of different stylometric feature embeddings for finding the authors of multilingual literary texts. Four standalone embedding methods were leveraged, namely, word based, PoS based, lemma based, and PoS mask based, and a combined embedding method (composition based) was introduced. To form the combined embedding method, the standalone models were combined by various operations at the matrix level, such as the mean, product of norm of the matrices. The authors compared the results with multilingual BERT embeddings. The composition-based embedding method demonstrated improved performance compared to the baseline standalone methods or mBERT.

Uchendu et al. [31] proposed a method formulated as an authorship attribution task designed to identify the automatically generated texts and conduct performance evaluation between human- and machine-generated texts via the Turing Test. The authors leveraged linguistic features alongside NN architectures like RNN and CNN, and the analysis argued that the automatically generated texts presented significant or non-significant differences compared to human texts, depending on the generation method.

Romanov et al. [32] investigated the identification of authors of Russian texts using SVM and deep NN architectures, including LSTM, CNN with attention, and Transformer.

The findings suggested that while all methods were effective, SVM achieved the highest accuracy at 96%, attributed to the optimal parametrization and feature spaces. Deep NNs tended to be less effective, with 93% accuracy. Experiments revealed the susceptibility of SVM to deliberate text anonymization, resulting in a higher accuracy loss compared to deep NN, which experienced up to a 20% decrease in accuracy. The Transformer architecture proved the most effective for anonymized texts, achieving an accuracy of 81%. Another study [33] in the same publication series targeting Russian authorship attribution for literary texts leveraged FastText and SVM for feature extraction and selection, and introduced regularization methods. Their approach achieved an average of 83% in terms of accuracy, while FastText slightly outperformed the previous results, scoring an 84% accuracy.

Moreover, a different study [34] introduced a technique for author identification that combined chaos game representation with deep learning methodologies. Working at the character level, their method employed chaos game representation to transform documents into visual representations. Following this, a CNN algorithm was used for classification, reaching an average accuracy of 85.52%.

In their recent research, He et al. [35] provided a comprehensive examination of the methods, models, datasets, feature types, and evaluation metrics employed in author attribution studies conducted for both source code and English text. The survey included two deep learning studies [36,37] focused on source code author attribution. These studies introduce interpretable models and introduce the concept of saliency maps to enhance model interpretability. The initial model [36] achieved a 42% accuracy, utilizing an NN for embedding projection, alongside tSNE for visualization and the KNN classification algorithm for predictions. Similarly, the second study [37] utilized a CNN for embedding generation and, like its predecessor, employed a KNN classification method, resulting in a 70% accuracy rate. For authorship identification on text, various approaches including multiheaded RNNs [38] have shown promising results. Multiheaded RNNs [38] were leveraged to process multiple authors simultaneously and were optimized for small datasets to prevent overfitting. This method achieved an AUC average score of 80%. Other CNN-based approaches demonstrated elevated performance on short texts [39–44]. The survey [35] also outlined several challenges and constraints. Such limitations for deep learning techniques include considerations of the number of authors, i.e., the higher the variety of authors, the lower the accuracy. This tendency was observed for both ML and DL models, with DL scoring above ML in most of the cases.

2.4. Transformer-Based Models

Transformer-based models, exemplified by architectures like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), excel in contextualizing language features over longer contexts. The application of Transformer-based models in authorship attribution involves fine-tuning the pre-trained models on labeled datasets, enabling them to discern uniqueness in the writing styles of different authors. This approach mitigates the need for extensive feature engineering and enhances the model's adaptability across diverse genres, languages, and writing styles.

AlZahrani et al. [45] presented an authorship attribution use case in the domain of Islamic law, investigating the performance of recent Arabic pre-trained Transformer-based models, such as AraBERT, AraELECTRA, ARBERT, and MARBERT, in the context of legal texts. Given the absence of a dedicated dataset, the authors built their corpus using digital resources from Islamic law. Experimental results revealed that ARBERT and AraELECTRA outperform other models, achieving an impressive 96% accuracy. The study concluded that pre-trained Transformer models, when fine-tuned, yield the best results.

A recent research [46] used the authors' distinctive writing patterns for profiling purposes. Existing methods employing handcrafted features for classification tasks have shown limitations, particularly in out-of-domain scenarios. To overcome this, the paper introduced PART (Pre-trained Authorship Representation Transformer), a trained model focused on learning authorship embeddings instead of semantics. By leveraging pairs of text fragments written by the same author, the model determined authorship through cosine similarity evaluation, allowing for zero-shot generalization. The proposed model based on a pre-trained Transformer with an LSTM head was trained on diverse datasets incorporating literature, anonymous blog posts, and corporate emails. Evaluation results showcased zero-shot accuracies of 72.39% and 86.73% in determining authorship from a set of 250 authors. The paper concluded with qualitative assessments of the model's representations through data visualizations, providing insights into profiling features, such as book types, gender, age, and author occupation.

A different study [47] investigated the application of Transformer-based fine-tuned models. Because these models lack the transparency of traditional methods that rely on stylometric features to quantify style and maximize the distance between texts, the authors implemented a BERT-based model for authorship verification and evaluated its predictions using an adapted LIME explainer and an attention-based feature extraction procedure. A comparative analysis between traditional and Transformer-based approaches was conducted, emphasizing explainability through input alteration to verify the retrieval of influential features.

Huang et al. [48] proposed a study focusing on short texts, having limited information about the author. Leveraging pre-trained language models, the study proposed a model that combines BERTweet, a pre-trained language model for English tweets, with the capsule network. This combination proved effective in capturing deep features of sentence representations, leading to improvements in authorship attribution tasks for tweets. The model also incorporated user writing styles, achieving state-of-the-art results on a known tweet dataset.

Bauersfeld et al. [49] introduced a Transformer-based NN architecture designed to attribute anonymous manuscripts to their authors using only the text body and author names in the bibliography. In order to develop and evaluate the method, the researchers compiled the biggest authorship identification dataset to date, incorporating over 2M research papers publicly available on arXiv. In the arXiv subsets containing up to 2K different authors, the method achieved an impressive authorship attribution accuracy, correctly attributing up to 95% of the papers. The study enabled the accurate prediction of authorship and highlighted the weaknesses in the double-blind review process by identifying key aspects that make a paper attributable. The authors open-sourced the necessary tools to replicate the experiments, providing insights to support an unbiased double-blind review process.

Another study [50] introduced the C-Transformer framework, representing a mix between DL and Transformer for the author classification of Chinese poetry texts. The framework leveraged CNNs to learn contextual information, also comprising multi-head attention and Transformer layers, which capture detailed semantics, while the Latent Dirichlet Allocation (LDA) model extracted topical features from the poems. Extensive experiments across four datasets, each featuring a different number of poets, were performed to evaluate six baseline models. The results highlighted the superior performance of the proposed C-Transformer framework, showcasing an enhanced accuracy attributed to the incorporation of poetry topics.

2.5. Authorship Attribution in Romanian

Authorship attribution in less-resourced languages, such as Romanian, presents challenges due to the scarcity of available studies and limited datasets. While research in this area has primarily focused on well-resourced languages, such as English, the specific linguistic characteristics of Romanian texts remain relatively understudied.

In a recent study by Avram et al. [17], various Artificial Intelligence ML techniques were compared for the authorship classification of Romanian literary texts written by multiple authors, focusing on a limited number of speech parts (prepositions, adverbs, and conjunctions). The study introduced a new dataset of Romanian texts by different authors, incorporating diverse texts in terms of length, sources, or time periods. Three distinct feature lists were constructed based on the Inflexible Parts-of-Speech (IPoS), generating numerical representations of the texts. Five ML standalone models were employed, including Artificial Neural Networks (ANNs), Multi-Expression Programming (MEP), k-NN, SVM, and DT with C5.0. Experimental results indicated that the MEP method had the lowest overall error rate of 20.40%. Further analysis revealed that the dataset with prepositions and adverbs performed best. Several limitations were identified despite employing multiple methods. Determining the author of a text remained challenging, with only a few algorithms showcasing acceptable error rates on the test set. According to the authors, the dataset was also heterogeneous in various aspects, including text length, sources, time periods, and writing types, which may introduce complexity and variability into the analysis. Moreover, the choice to focus on a limited number of speech parts (prepositions, adverbs, and conjunctions) as features may overlook other relevant linguistic aspects that contribute to authorship attribution. While the MEP method achieved the best overall results with a 20.40% error rate, other methods' performance and generalizability to different datasets may vary, suggesting that the approach may not consistently achieve optimal results across all scenarios.

A second study by Avram [51] focused on authorship attribution using pre-trained language models, particularly BERT, to detect the authorship of Romanian texts. Similar to the previous study, this research used the same dataset, which is highly unbalanced in terms of the number of texts per author, source, time period, and writing type. Results showed a macro-accuracy exceeding 87%. Building upon previous work, the study leveraged shorter texts (200 tokens/words) for improved dataset balance. While previous methods achieved a maximum accuracy of 80.94%, BERT-based approaches yielded an 85.90% accuracy. However, a direct comparison with the previous method was challenging due to dataset modifications for the BERT processing requirements. Despite the promising results achieved with the BERT approach and shorter text segments, there were limitations to consider. While dividing texts into 200-word segments improves the dataset balance, it disregards the integrity of sentences. This approach may affect the context and semantic meaning of the text, which could influence the accuracy of authorship attribution. Therefore, further investigation is needed to assess the impact of maintaining sentence integrity and to understand the potential trade-offs between segmenting texts and preserving their original structure.

3. Method

Our approach introduces a hybrid Transformer and a baseline ensemble learning approach as illustrated in Figure 1. Both methods were evaluated on two sets of texts with different lengths: full texts (FT) and paragraphs (PP) containing around 200 words each, preserving paragraph integrity. In the hybrid Transformer method, the input text is tokenized and then processed by a pre-trained BERT model to generate text embeddings. Concurrently, linguistic features are extracted from the texts to serve as numerical input features. Feature selection is carried out using the Kruskal–Wallis mean rank, retaining the top features for each set. Subsequently, the text embeddings are combined with the numerical features to form a hybrid vector, which is afterward fed to a classification layer for authorship prediction. In contrast, the classical ML method leverages the Readerbench framework to extract linguistic features, which are then employed as features for a set of 7 machine learning models in an ensemble learning approach with a soft voting technique for authorship identification.

The selection of our models was motivated by the objective of establishing a baseline for comparison with prior research in Romanian authorship attribution. Considering the limited number of studies (only two known to date), our goal was to fill this gap by presenting a comprehensive analysis. Unlike previous approaches that primarily relied on standalone machine learning (ML) models and a basic BERT-based approach for authorship attribution, we introduced novel methodologies. These methods combine the strength of seven ML models and introduce a hybrid BERT-based model. Our hybrid model not only uses text input but also incorporates linguistic features for predictions. As a result, our proposed approach surpasses the state-of-the-art performance of prior studies.

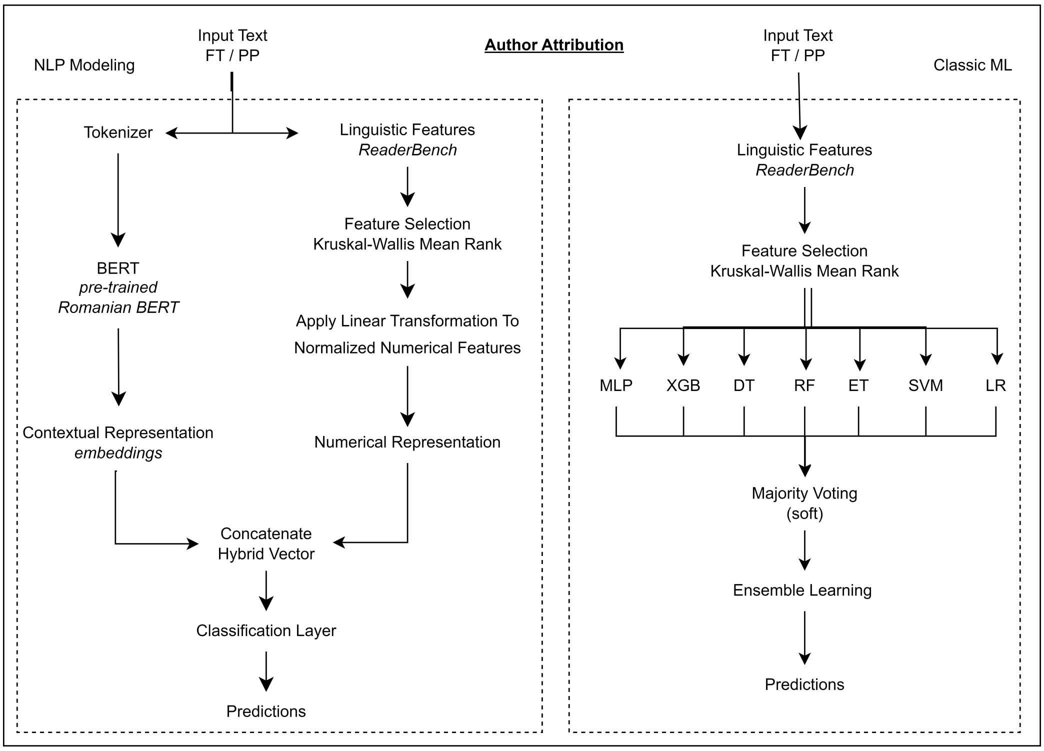

Figure 1. Authorship attribution methods.

3.1. Corpus

Our corpus consists of texts written by Romanian authors between the 19th century and the present, representing stories, short stories, fairy tales, and sketches. The corpus is presented in two versions: (a) the full version with texts from 19 authors, comprising 1263 full texts (FT) and 12,516 paragraphs (PP); (b) a subset with 10 authors, consisting of 250 full texts (FT) and 3021 paragraphs (PP). This dataset represents an extension of the ROST dataset (https://www.kaggle.com/datasets/sandamariaavram/rost-romanian-stories-and-other-texts, accessed 12 February 2024), which contains only 400 texts. We added as many other relevant authors and corresponding texts as we could find without having copyright issues. The smaller version was chosen to establish an equitable comparison with two prior Romanian studies, which had almost identical sets of authors (i.e., 9 out of 10). Paragraphs are around 200 words each for both sets and preserve paragraph integrity. FT and PP word distribution is presented in Figure 2.

In all experiments, we adopted an 80/20 train/test split. For the PP dataset, precautions were taken to prevent paragraphs from the same book or story from appearing in both the train and test sets. This is important for preventing the model from memorizing specific patterns of individual texts. If the model is exposed to identical or highly similar passages during both training and testing, it might learn to recognize these specific fragments rather than generalizing patterns that indicate an author's style. By including paragraphs from different books and stories in the training and testing datasets, the model learns broader, more general features of writing style that are more likely to be applicable across various texts and authors, thereby enhancing its capability to accurately attribute authorship.

Table 2 provides an overview of the dataset used in the study. It includes various statistics for each author, such as the number of full texts (FT) and paragraphs (PP), as well as the mean (M) and standard deviation (SD) of the number of words, unique words, and type–token ratio (TTR) for both full texts and paragraphs.

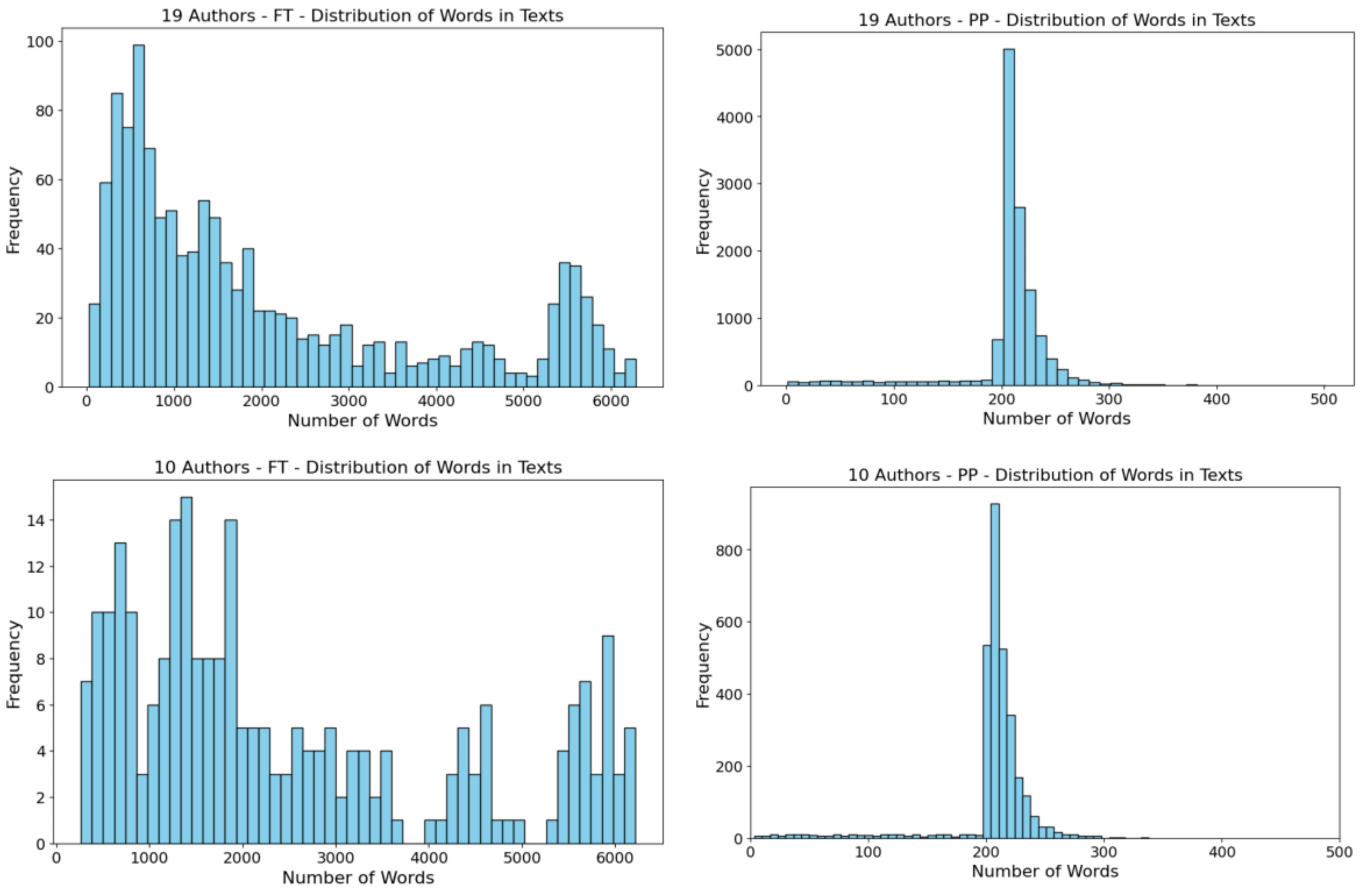

Figure 2. Word distribution for FT versus PP.

Table 2. Dataset overview (authors marked in bold are also included in the subset with 10 authors).

Author	FT	PP	M(SD) FT Words	M(SD) FT Unique Words	M(SD) FT TTR
Alexandru Vlahuta	96	647	1629.16 (1341.48)	735.19 (462.04)	0.5110 (0.0844)
Anton Bacalbasa	132	485	808.17 (720.04)	392.20 (244.57)	0.5256 (0.0660)
Barbu St. Delavrancea	47	747	4015.40 (2224.96)	1391.72 (658.60)	0.3730 (0.0599)
Costache Negruzzi	24	343	3482.62 (2253.38)	1236.46 (694.14)	0.4027 (0.0883)
Emil Garleanu	55	353	1533.58 (1582.43)	609.09 (449.03)	0.4649 (0.0767)
Emilia Plugaru	41	382	2176.71 (1705.21)	792.00 (454.83)	0.4091 (0.0702)
George Toparceanu	46	331	1689.11 (1246.86)	711.00 (412.92)	0.4728 (0.0815)
Ioan Slavici	89	1716	4692.76 (2156.69)	1306.64 (485.87)	0.3043 (0.0665)
Ion Creanga	45	424	2291.13 (2328.91)	720.96 (554.58)	0.4420 (0.1537)
Ion Luca Caragiale	60	585	2444.30 (1541.96)	895.13 (466.55)	0.3832 (0.0485)
Liviu Rebreanu	59	619	2544.49 (1770.39)	969.80 (518.88)	0.4165 (0.0654)
Mihai Eminescu	27	405	3642.78 (2167.54)	1284.67 (674.06)	0.3834 (0.0767)
Mihai Oltean	32	68	409.62 (394.16)	216.28 (174.42)	0.5938 (0.1093)
Mihail Sebastian	46	658	3478.37 (1826.51)	1234.85 (472.30)	0.3803 (0.0532)
Nicolae Filimon	35	375	2606.57 (1701.70)	998.20 (540.52)	0.4173 (0.0781)
Nicolae Iorga	306	2982	2437.67 (2215.16)	970.28 (741.50)	0.4834 (0.1054)
Panait Istrati	20	499	6299.85 (1202.32)	2177.75 (369.46)	0.3494 (0.0240)
Petre Ispirescu	40	630	3768.72 (1614.16)	1126.40 (359.51)	0.3183 (0.0517)
Traian Demetrescu	63	267	976.13 (581.40)	472.32 (234.24)	0.5279 (0.0845)
Aggregate	1263	12,516			

The dataset incorporates a diverse range of authors, each contributing with different numbers of texts and paragraphs. We can identify patterns and trends in their writing styles, vocabulary usage, and text lengths by comparing statistical measures across authors. To understand the statistical differences between the authors, we further explore surface metrics like the number of words, unique words, or lexical diversity (TTR score).

The number of words indicates the average length of texts authored by each writer—i.e., authors with higher mean values produce longer texts on average. Additionally, examining the standard deviation values enables us to understand the variability in text length within each author's work.

The unique words feature measures the richness of vocabulary used by each author. Authors with higher mean values for unique words tend to use a more diverse range of vocabulary in their writing. Comparing the mean and standard deviation values for unique words can reveal authors who consistently employ a wide range of vocabulary versus those who exhibit less variability in their word choices.

The type–token ratio (TTR) quantifies the lexical diversity of a text by calculating the ratio of unique words to total words. A higher TTR indicates greater lexical diversity, while a lower TTR suggests repetitive word usage. Authors with consistently high TTR values may exhibit more varied and sophisticated language use, while those with lower TTR values may prefer simpler or more repetitive language.

Discussing potential biases within our dataset and their impact on the model's predictions is essential. Our dataset exhibits an important degree of imbalance in the distribution of texts, with certain authors being disproportionately represented. For instance, we observe in the PP set 2982 paragraphs for Nicolae Iorga compared to only 68 paragraphs for Mihai Oltean. Similarly, there are only 20 stories from Panait Istrati compared to 132 from Anton Bacalbasa in the FT set. Such an imbalance can lead to skewed outcomes or diminished accuracy, potentially resulting in discrimination against specific authors. To mitigate this bias, our methodologies incorporate various techniques, including weighted loss and stratified labels for the data split.

We publicly release our extended dataset for future reference (https://huggingface.co/datasets/readerbench/ro-stories, accessed 12 February 2024).

3.2. Linguistic Features

In our study, we integrate a selection of features derived from the texts, serving as inputs for a classification model. This strategy aims to effectively capture elements of the author's writing style.

3.2.1. Feature Extraction

For the feature extraction step, we leverage the Readerbech framework [52] that offers various text analysis modules in multiple languages, such as English, French, Romanian, and Dutch. To our knowledge, Readerbench is the only open-source multilingual framework available also for Romanian, granting access to over 200 linguistic features (https://github.com/readerbench/Readerbench/wiki/Textual-Complexity-Indices, accessed 12 February 2024). These indices cover various factors, such as surface, syntactic, morphological, semantic, and discourse-specific elements, as well as cohesion metrics derived from specialized lexicalized ontologies and semantic models.

Surface indices focus on measures like sentence length, word length, the number of unique words, and word entropy. These indices operate on the premise that more complex texts contain a greater abundance of information and a wider range of concepts. Word complexity indices explore deeper into the complexity of individual words, considering factors such as syllable count, morphological complexity, and the richness of word meanings derived from sources like WordNet. Syntactic and morphological indices evaluate sentence-level features, including parts of speech and syntactic dependencies, providing insights into a text's structural complexity. Semantic cohesion indices assess the connectedness of ideas within a text, leveraging Cohesion Network Analysis (CNA) [53] and semantic models to measure both local and global cohesion. Discourse structure indices analyze discourse connectives and metrics derived from models of discourse, targeting the organization and elaboration of a text's content. Overall, these diverse ReaderBench indices (RBI) provide comprehensive insights into the complexity and structure of the text.

3.2.2. Feature Selection

The choice of features is key in authorship attribution, as it helps improve the efficiency of classification models by focusing on the most relevant and discriminative features. Since most features are non-normally distributed, we employ the Kruskal–Wallis mean rank [54], a non-parametric statistical method commonly used for feature selection in various ML tasks. This approach evaluates the relevance of different features by comparing their mean ranks across multiple groups or classes. In the context of authorship attribution, Kruskal–Wallis mean rank analysis helps identify the most discriminative features that contribute to distinguishing between authors' writing styles. By considering the mean ranks of features across authors on the train documents, this method enables the selection of those features that exhibit the greatest effect sizes, thus characterizing authorial styles.

Therefore, we used the Kruskal–Wallis method to select the top features, and we conducted experiments with varying numbers of features for each text set, namely FT and PP. The fransformer model achieved optimal performance with the top 100 features, whereas for the ML-based approach, we employed different numbers of features ranging between 50 and 500 as illustrated in Table 3.

Table 3. Number of top features selected via Kruskal–Wallis for different input corpora.

Corpus	No. Docs	No. Authors	Feature Selection	Model	No. Features
FT	250	10 authors	Kruskal–Wallis	ML Based	Top 50
				Transformer Based	Top 100
	1263	19 authors	Kruskal–Wallis	ML Based	Top 100
				Transformer Based	Top 100
PP	3021	10 authors	Kruskal–Wallis	ML Based	Top 300
				Transformer Based	Top 100
	12516	19 authors	Kruskal–Wallis	ML Based	Top 500
				Transformer Based	Top 100

3.3. Prediction Models

3.3.1. Baseline Ensemble Learning

We explore the efficiency of traditional ML models via an ensemble learning technique, a powerful method that combines the predictions of multiple individual classifiers to improve the overall performance, thus leveraging the diversity of different models to achieve better generalization. We employ a variety of classification methods as base learners, including Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB), Decision Trees (DT), Random Forests (RF), Extra Trees (ET), Support Vector Machines (SVMs), and Logistic Regression (LR).

To combine the predictions of these diverse classifiers, we employ an ensemble learning technique known as majority soft voting. In this approach, each base learner independently predicts the class probabilities for a given input, and the final decision is made by choosing the class with the highest average probability across all classifiers. Unlike simple majority voting, soft voting considers the confidence level of each classifier's prediction, enhancing the decision-making process. This ensemble approach takes advantage of the strengths of each individual model while surpassing their weaknesses, resulting in improved overall performance.

Compared to the study of Abbasi et al. [22], we present an expanded methodological framework designed for the Romanian language. While the initial study employed three machine learning models (XGB, MLP, and RF), our study incorporates a broader array of machine learning techniques, including seven distinct models. This expansion allows for a more comprehensive exploration of the classification, capturing diverse patterns within the data that may be better suited for specific algorithms. Furthermore, another distinction from the reference study lies in the feature extraction and selection process. While the initial study used TF-IDF and CountVectorizer for feature extraction, our study employs different linguistic features generated using RederBench. Additionally, the feature selection methodology diverges significantly; whereas the reference study did not employ any feature selection method, our approach incorporates Kruskal–Wallis mean rank. Kruskal–Wallis aids in identifying the most informative features by assessing the statistical differences among groups, thereby improving the model's discriminatory power. These methodological differences broaden the scope of our investigation and improve the baseline methodology.

3.3.2. Hybrid Transformer

The hybrid Transformer approach leverages a pre-trained Romanian BERT model, specifically RoBERT [55], to extract contextual representations from input data. These embeddings are then combined with numerical representations from linguistic features, which, after extraction and selection, are first normalized and undergo a linear transformation before being concatenated with the BERT embeddings as illustrated in Figure 3.

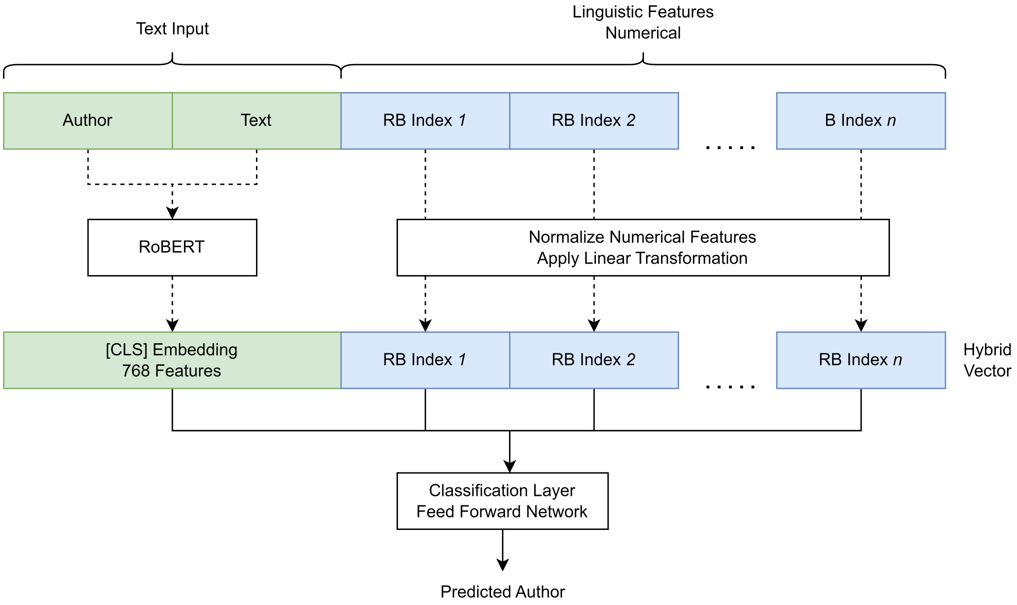

Figure 3. Hybrid Transformer model.

Formally, the problem can be defined as follows: let X represent the input data comprising textual content and associated numerical features, where $X = \{x_1, x_2, \ldots, x_n\}$ and n denotes the number of samples. Each x_i consists of a text sequence x_{text} and a corresponding set of numerical features x_{num}. The task is to predict the author label y for each input text. Given a dataset $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where X_i represents the features of the i-th sample and y_i denotes its corresponding author label, the objective is to learn a mapping function $f(X)$ that predicts the author label y accurately for unseen text data.

In the proposed approach, the textual content undergoes preprocessing to obtain token IDs, input masks, and segment IDs. These inputs are then fed into the encoder part of the Transformer architecture, consisting of multiple Transformer layers. Each Transformer layer comprises a self-attention mechanism followed by a feed-forward neural network layer. This process is repeated for the predetermined number of encoder layers, typically 12 in the BERT model, to produce context-based textual representations. Simultaneously, the numerical features are processed and incorporated into the hybrid feature vector. Finally, the feature vector is passed through a classification layer, which in our case is represented by a feed-forward neural network with softmax activation, to predict the class label of each text sample based on the probability distribution over author labels.

3.4. Parameter Tuning

The hyperparameter tuning using GridSearch generated optimal configurations for the ML-based approach models. On the PP dataset with 19 authors, the best hyperparameters for the MLP classifier included an activation function of relu, a regularization parameter (alpha) of 0.001, a single hidden layer with 100 neurons, and a maximum of 100 iterations. Similarly, for the XGB model, the best setting comprised a learning rate of 0.2, a maximum tree depth of 5200 estimators, a subsample rate of 0.8, and a colsample_bytree of 0.8. The parameter colsample_bytree determines the fraction of features (columns) to be randomly sampled and included when constructing each Decision Tree during the boosting process. A value of 0.8 means that each tree in the ensemble will randomly sample 80% of the features/columns when considering which features to split on. This parameter helps introduce randomness and reduce overfitting by preventing individual trees from becoming too specialized in specific features, thereby improving the model's generalization. The SVM model performed best with a regularization parameter (C) of 0.1, a linear kernel, and an automatic gamma scaling. LR achieved optimal results with a regularization parameter (C) of 1, a penalty of l1, and the liblinear solver. RF and DT models both had the same best parameters, including unlimited tree depth, automatic selection of the maximum number of features, and minimum leaf samples of 1, with slight variations in the minimum samples required for a split. The ET classifier's most effective setup involved a maximum depth of 20, automatic feature selection, minimum leaf samples of 1, minimum samples required for a split of 2, and 200 estimators. Similarly, for the FT dataset with 19 authors, the XGB model performed optimally with a configuration similar to that obtained for the PP dataset, while RF, LR, DT, and ET achieved their best performance with identical hyperparameters as in the PP dataset. The SVM model had a different optimal configuration, with a polynomial kernel of degree 2 instead of a linear kernel. These findings provide insights into the parameter settings that generated the highest classification accuracy for the ML-based approach on different corpora, facilitating the development of a robust attribution model.

Table 4 presents the results of the grid search for the best hyperparameters of ML models trained on the two corpora: paragraphs (PP) and full texts (FT). Each row corresponds to a specific model, and the columns represent the corpus, model type, and the best hyperparameters found during the grid search. Overall, the grid search results provide insights into the optimal configurations for training ML models on texts of different lengths, which can guide further experimentation and model refinement.

The tuning process for the hybrid Transformer involves optimizing various hyperparameters using GridSearch with 5- to 10-fold cross-validations, depending on the dataset size (FT or PP). Each model was trained for 5 to 10 epochs, allowing it to iteratively learn from the training data over multiple passes through the dataset, using a batch size of 32, which specifies the number of samples processed before updating the model's parameters. During training, the AdamW optimizer was employed with a learning rate of 1×10^{-5} and a weight decay of 0.01 to prevent overfitting and encourage stable convergence. Additionally, a learning rate scheduler was used to dynamically adjust the learning rate by a factor of 0.5 if the validation loss did not improve for a certain number of epochs. Moreover, the model leveraged a weighted loss function to address the class imbalance. By assigning higher weights to minority classes, the model is penalized more for misclassifying instances from these classes, encouraging it to learn better representations for these classes and improving its ability to generalize to unseen data. Furthermore, a dropout rate of 0.2 was applied after concatenating the BERT embeddings and the linguistic features (i.e., on the hybrid vector of 768 + 100 dimensions) and before the final classification layer to prevent overfitting. These hyperparameters were chosen based on their capability to optimize model performance and classification accuracy on the test dataset, ensuring efficient training, effective regularization, and robust predictive performance for authorship attribution tasks.

In terms of required computational resources, all models were trained on GPU machines, namely A100 and V100, using Google Colab Pro+ (https://research.google.com/colaboratory/faq.html, accessed 8 March 2024). GPUs offer advanced computing capabilities, enabling fast and efficient model training through hardware accelerators.

Table 4. GridSearch best parameters for ML models on the 19-author set.

Corpus	Model	Best Hyperparameters
FT	MLP	activation: relu, alpha: 0.0001, hidden_layer_sizes: (200,100), max_iter: 200
	XGB	colsample_bytree: 0.8, learning_rate: 0.2, max_depth: 5, n_estimators: 100, subsample: 0.8
	SVM	C: 0.1, degree: 2, gamma: scale, kernel: polynomial
	LR	C: 100, penalty: l1, solver: liblinear
	RF	max_depth: None, max_features: auto, min_samples_leaf: 1, min_samples_split: 10, n_estimators: 100
	DT	max_depth: None, max_features: auto, min_samples_leaf: 1, min_samples_split: 5
	ET	max_depth: None, max_features: auto, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200
PP	MLP	activation: relu, alpha: 0.001, hidden_layer_size: (100), max_iter: 100
	XGB	colsample_bytree: 0.8, learning_rate: 0.2, max_depth: 5, n_estimators: 200, subsample: 0.8
	SVM	C: 0.1, gamma: scale, kernel: linear
	LR	C: 1, penalty: l1, solver: liblinear
	RF	max_depth: None, max_features: auto, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
	DT	max_depth: 20, max_features: auto, min_samples_leaf: 1, min_samples_split: 5
	ET	max_depth: 20, max_features: auto, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200

4. Results

The results on the full dataset with 19 authors are presented in Tables 5 and 6. In the full text (FT) corpus, employing only RoBERT embeddings resulted in an F1 score of 0.85, while integrating BERT embeddings with linguistic features (hybrid RoBERT) exhibited the best performance with an F1 score of 0.87. Additionally, we observed that SVM achieved the highest F1 score of 0.83 among standalone models, while ensemble learning surpassed all standalone models with an F1 score of 0.84.

Correspondingly, in the paragraphs (PP) analysis, leveraging RoBERT resulted in an F1 score of 0.73, while hybrid RoBERT achieved the highest performance with an F1 score of 0.77. Furthermore, LR achieved the highest F1 score of 0.61. However, ensemble learning outperformed all standalone models with an F1 score of 0.69, indicating the advantage of combining predictions from multiple models.

Table 5. The 19-author full text (FT) results (bold represents the best results from each category).

Corpus	Input	Top Features	Model	F1	Error
FT	RBI	KW Top 100	Decision Trees	0.50	0.49
			Extra Trees	0.74	0.23
			Logistic Regression	0.72	0.25
			MLP	0.66	0.34
			Random Forest	0.69	0.28
			SVM	0.83	0.17
			XGBoost	0.80	0.18
			Ensemble Learning	**0.84**	**0.15**
	BERT Embeddings		RoBERT	0.85	0.14
	BERT Embeddings + RBI	KW Top 100	**Hybrid RoBERT**	**0.87**	**0.12**

Table 6. The 19-author paragraphs (PP) results (bold represents the best results from each category).

Corpus	Input	Top Features	Model	F1	Error
PP	RBI	KW Top 500	Decision Trees	0.36	0.63
			Extra Trees	0.44	0.47
			Logistic Regression	0.61	0.37
			MLP	0.58	0.40
			Random Forest	0.46	0.46
			SVM	0.60	0.39
			XGBoost	0.56	0.40
			Ensemble Learning	**0.69**	**0.28**
	BERT Embeddings		RoBERT	0.73	0.22
	BERT Embeddings + RBI	KW Top 100	**Hybrid RoBERT**	**0.77**	**0.23**

The results for authorship attribution on the subset with 10 authors are presented in Tables 7 and 8. For the full text (FT) corpus, employing RoBERT embeddings alone resulted in a competitive F1 score of 0.88, while integrating BERT embeddings with linguistic features (i.e., hybrid RoBERT) exhibited the best performance with an F1 score of 0.94, showcasing the advantage of leveraging a hybrid approach. Furthermore, employing various standalone ML models yielded diverse performance outcomes. RF and XGB achieved the highest F1 score of 0.81, closely followed by MLP and Extra Trees, with an F1 score of 0.78 and 0.77. However, the ensemble learning technique surpassed all standalone models, achieving an F1 score of 0.82, indicating the effectiveness of combining multiple models for improved performance.

Table 7. The 10-author full text (FT) results (bold represents the best results from each category).

Corpus	Input	Top Features	Model	F1	Error
FT	RBI	KW Top 50	Decision Trees	0.69	0.32
			Extra Trees	0.77	0.22
			Logistic Regression	0.75	0.24
			MLP	0.78	0.20
			Random Forest	0.81	0.18
			SVM	0.31	0.56
			XGBoost	0.81	0.18
			Ensemble Learning	**0.82**	**0.14**
	BERT Embeddings		RoBERT	0.88	0.11
	BERT Embeddings + RBI	KW Top 100	**Hybrid RoBERT**	**0.94**	**0.05**

Table 8. The 10-author paragraphs (PP) results (bold represents the best results from each category).

Corpus	Input	Top Features	Model	F1	Error
PP	RBI	KW Top 300	Decision Trees	0.51	0.48
			Extra Trees	0.62	0.37
			Logistic Regression	0.74	0.26
			MLP	0.66	0.34
			Random Forest	0.65	0.35
			SVM	0.72	0.28
			XGBoost	0.74	0.26
			Ensemble Learning	**0.78**	**0.20**
	BERT Embeddings		RoBERT	0.94	0.05
	BERT Embeddings + RBI	KW Top 100	**Hybrid RoBERT**	**0.95**	**0.04**

Similarly, for the paragraph (PP) analysis, employing RoBERT embeddings alone achieved a superior F1 score of 0.94, while the integration of BERT embeddings with linguistic features (i.e., hybrid RoBERT) achieved the highest performance with an F1 score of 0.95. These results underscore the relevance of leveraging hybrid features to enhance authorship attribution accuracy, particularly in datasets with few authors. Furthermore, standalone models exhibited varying performance levels. LR and XGB achieved the highest F1 scores of 0.74, followed closely by SVM, with an F1 score of 0.72. However, ensemble learning outperformed all standalone models with an F1 score of 0.78, highlighting its efficacy in aggregating predictions from diverse models.

We further compare the best three models on the 10-author set, namely, ensemble learning (EL), the BERT-based, and the hybrid Transformer models. We choose to present the results on the subset of 10 authors for an easier correlation with prior existing methods in the field, linguistic analysis, and trends in misclassifications, which are analyzed in the next section. First, we consider the performance of the three models on both the FT and PP 10-author test sets (see Table 9).

Table 9. Performance of the 3 classification models for the 10-author FT and PP test sets.

Classification Model	FP Correct Classification		PP Correct Classification	
	Yes	No	Yes	No
Ensemble Learning (EL)	41	9	542	152
RoBERT	44	6	654	40
Hybrid RoBERT	47	3	656	38

Second, we conduct Cochran's Q test [56] to determine whether there is a significant difference in proportions among our models' performance. Cochran's Q test helps assess whether there is a consistent pattern of differences in proportions across multiple groups (i.e., prediction models). The test compares the proportion of successes or positive outcomes (i.e., correct predictions) within each group, considering the repeated measures or dependent nature of the data. A significant result (i.e., a small p-value) from Cochran's Q test suggests that at least one of the groups differs significantly from the others in terms of the proportion of successes.

Third, we leverage pairwise McNemar's Test [57] to compare the performance of the three best models (i.e., ensemble-based, BERT- based, and hybrid Transformer) in our multiclass classification scenario. This comprehensive analysis allows us to evaluate the strengths and weaknesses of the classification models. McNemar's Test is a statistical significance test specifically designed for analyzing paired nominal data, making it suitable for comparing the performance of two classifiers on the same test dataset.

Tables 10 and 11 present the results of both statistical tests (i.e., Cochran's Q test and McNemar's pairwise tests) for the 10-author FT and PP test sets. For the 10-author FT dataset (see Table 10), Cochran's Q test ($\chi^2(2)$ = 6.75, p = 0.034) was first employed to assess the overall differences in the proportions of correct predictions across the models. Subsequently, pairwise McNemar tests were conducted to compare the performance of individual models, and the only statistically significant difference was observed between the EL and the hybrid Transformer model. However, it is important to note that the adjusted p-value for this comparison is 0.0939, indicating a potential increased risk of Type I error. No significant differences were observed between the EL and BERT models, or between the BERT and hybrid models.

Table 10. Results of Cochran's Q test and pairwise McNemar tests for 10-author FT test set.

Test	Value	*p*-Value	Adjusted *p*-Value
Cochran's Q	6.75	0.03422	-
McNemar (EL—BERT)	-	0.3750	0.3750
McNemar (EL—Hybrid)	-	0.0313	0.0939
McNemar (BERT—Hybrid)	-	0.3750	0.3750

For the 10-author PP dataset (see Table 11), Cochran's Q test ($\chi^2(2)$ = 201.1, $p < 0.001$) highlights significant differences in the prediction performance within the models. Subsequently, pairwise McNemar tests highlighted substantial variations in prediction performance among the models, with significant differences observed between EL and both BERT and hybrid Transformer models, while no significant difference was observed between the BERT and the hybrid RoBERT models, which had similar predictions with a slight improvement of the hybrid Transformer model.

Table 11. Results of Cochran's Q test and pairwise McNemar tests for 10-author PP test set.

Test	Value	*p*-Value	Adjusted *p*-Value
Cochran's Q	201.1	$<2.2 \times 10^{-16}$	-
McNemar (EL–BERT)	-	2.11×10^{-27}	3.16×10^{-27}
McNemar (EL–Hybrid	-	1.22×10^{-28}	3.66×10^{-28}
McNemar (BERT–Hybrid)	-	0.5	0.5

Overall, the BERT and hybrid Transformer models exhibit better performance than the EL model, with the hybrid architecture having the best performance.

5. Discussion

In this section, we discuss the results achieved by our proposed methods in comparison to existing approaches for authorship attribution involving datasets with varying numbers of authors, and we look at the top important features per author for the ML-based approach to understand which features most influenced the decision of the classifier. Moreover, the authors' writing characteristics are extracted from the analysis and interpreted in detail. Nevertheless, it is important to recognize the limitations of our study, which are also detailed below.

5.1. Comparison with Existing Methods

To ensure a fair comparison, we exclusively assessed existing methodologies in Romanian studies on authorship attribution. Given the varying linguistic complexities and model usage across languages, we focused entirely on Romanian studies. To our knowledge, only two existing studies exist, which are also outlined in Section 2 of this paper.

Focusing on datasets with 10 authors (see Table 12), we observe that our RoBERT model outperformed the existing approaches [22,51] for both FT and PP corpora. Additionally, our hybrid RoBERT model, which incorporates RBI features, achieved the highest F1 score of 0.95 for the PP corpus, indicating the effectiveness of leveraging both textual and numerical features for AA tasks. Moreover, our standalone ML-based approach, leveraging the RBI features with Kruskal–Wallis feature selection, achieved competitive results compared to the existing methods. Our method outperformed the previous studies [17,22] in terms of F1 score (0.81 versus 0.79) and error rate (0.18 versus 0.20) for the full text (FT) corpus. Moving on to ensemble learning methods, our approach achieved competitive performance compared to the existing ensemble approach [22] for both FT and PP corpora with 10 authors. Specifically, our ensemble learning method achieved an F1 score of 0.82 for the FT corpus and 0.78 for the PP corpus, arguing for the robustness of our approach in leveraging multiple classifiers to improve classification accuracy.

Regarding the full dataset with 19 authors, our hybrid RoBERT model achieved the highest F1 score of 0.87 for the FT corpus, highlighting the strength of leveraging both textual and numerical features for AA tasks involving larger author sets. Furthermore, our standalone ML-based approach with RBI features again achieved competing performance compared to existing methods, achieving the highest F1 score of 0.84 for the FT corpus with 19 authors. Our results are consistent and surpass the performance reported by Abbasi et al. [22], who included 20 authors for an English corpus. The authors leveraged DistilBERT and scored an F1 of 0.76, with an error rate of 0.23. The same study employed an ensemble learning model with TF-IDF as input features, scoring a maximum of 0.74 F1 and 0.25 error rate.

In analyzing the results, it is noteworthy that the rank and performance of the models vary across different datasets and input representations. The rank, denoted by Roman numerals, reflects the comparative performance of each model in terms of F1 score and error rate. Across both dataset versions, with 10 authors and 19 authors, our proposed method consistently achieved either the first or second rank. Particularly, for the 10-author version, the top three approaches include standalone models using RBI features with KW top features, ensemble learning with BERT embeddings, and our proposed hybrid RoBERT method. These methods showcased competitive performance, with our hybrid RoBERT outperforming the standalone models and closely matching the performance of the ensemble learning approach. Similarly, in the dataset with 19 authors, our method using hybrid RoBERT combining BERT embeddings and RBI features was top-ranked.

Table 12. Comparison with existing methods for 10 authors (bold marks the best result).

Corpus	No. Docs	No. Authors	Input	Model	Method	F1	Error	Rank
FT	400	10 authors	IPoS	Standalone	[17]	0.79	0.20	II
FT	250	10 authors	RBI (KW 50)	Standalone	our method	**0.81**	**0.18**	I
PP	3021	10 authors	RBI (KW 300)	Standalone	our method	0.74	0.26	III
FT	250	10 authors	RBI (KW 50)	ensemble learning	our method	**0.82**	**0.14**	I
PP	3021	10 authors	RBI (KW 300)	ensemble learning	our method	0.78	0.20	II
PP	6832	10 authors	BERT Embeddings	BERT-base-ro	[51]	0.85	0.14	
PP	3021	10 authors	BERT Embeddings	RoBERT	our method	0.94	0.05	II
FT	250	10 authors	BERT Embeddings	RoBERT	our method	0.88	0.11	
PP	3021	10 authors	BERT Embeddings + RBI	Hybrid RoBERT	our method	**0.95**	**0.04**	I
FT	250	10 authors	BERT Embeddings + RBI	Hybrid RoBERT	our method	0.94	0.05	III

Overall, our results underpin the importance of feature selection and modeling techniques in enhancing AA performance. The choice of the feature selection method and model architecture can impact the outcome. Leveraging feature selection techniques such as Kruskal–Wallis enabled us to identify the most informative features, leading to more effective model training and improved generalization. By incorporating numerical features such as RBI with textual embeddings from pre-trained BERT models, we captured richer feature representations, leading to improved classification accuracy. Additionally, the ensemble learning approach allowed us to combine the collective power of multiple classifiers, resulting in more robust and accurate predictions.

5.2. Analysis of Authors' Writing Styles

The top three most discriminative features per author for the top-performing models on the 10-author set, on both FT and PP, are outlined in Table 13. In this context, the polarity refers to whether the classification model is positively or negatively influenced by the respective feature (+ denotes positive polarity, while − denotes negative correlations). The importance of each feature is reflected in its corresponding coefficient from a simple Logistic Regression. These top features highlight distinct linguistic traits and syntactic structures that play an important role in identifying authorship within the provided texts. Further exploration and interpretation can provide relevant insights into and an understanding of the authors' writing styles and textual patterns. By considering the top features and their polarity in the classifier's decision-making process, we can redefine the distinctions of author writing styles and explore specific traits in greater detail.

For an overview of the authors writing styles based on the most discriminative features in their texts and lexical statistics, we perform a cross-correlation between the data in Tables 2 and 13. Based on this correlation, Barbu Delavrancea tends to use a high number of connector links and repetitions, along with a rich vocabulary of unique nouns, which are negatively correlated with the number of words and unique words. This suggests a concise writing style with frequent repetitions. These features, particularly the frequent repetitions and unique nouns, play an important role in distinguishing Delavrancea's texts from those of other authors. Nevertheless, the syntactic complexity represented by connector conjunctions and variations in dependency cases may pose challenges for classification models.

Table 13. Top discriminative features per author on the top-10 author subset (where Pos and Neg are polarities).

Author	FT		PP	
	Index	**Polarity (+/—) & Importance**	**Index**	**Polarity (+/—) & Importance**
B. Delavrancea	M(Connector_link/Par)	−0.71	Max(Connector_conj/Sent)	−0.33
	M(Repetitions/Par)	+0.67	SD(Dep_case/Sent)	−0.24
	M(UnqPOS_noun/Par)	+0.64	Max(POS_adv/Sent)	+0.19
Emil Garleanu	M(Punct/Par)	+1.22	SD(NmdEnt_person/Sent)	−0.32
	M(Commas/Par)	+0.74	M(NmtEnt_person/Sent)	−0.30
	M(Pron_third/Par)	+0.70	Max(Repetitions/Sent)	−0.22
Emilia Plugaru	M(UnqWd/Par)	+1.25	M(Connector_disj/Par)	−0.26
	M(Repetitions/Par)	−0.84	M(Connector_disj/Doc)	−0.26
	M(Connector_link/Par)	−0.83	Max(Connector_disj/Par)	−0.27
Ioan Slavici	M(Dep_nsubj/Par)	+0.93	M(Dep_nsubj/Par)	+0.32
	M(Commas/Par)	−0.80	M(Polysemy/Word)	+0.30
	M(Pron_third/Par)	+0.77	M(Dep_obj/Sent)	−0.28
Ion Creanga	M(Pron_third/Par)	−1.40	M(UnqWd/Sent)	+0.23
	M(Commas/Par)	+0.86	M(UnqPOS_verb/Sent)	−0.22
	M(Connector_disj/Par)	−0.63	Max(Dep_advcl/Sent)	+0.22
I.L. Caragiale	M(Pron_third/Par)	−1.71	M(ParseDepth/Sent)	−0.22
	M(Punct/Par)	−1.62	Max(POS_verb/Sent)	−0.22
	SD(POS_noun/Sent)	+1.05	Max(Connector_conj/Sent)	+0.21
Liviu Rebreanu	M(Connector_disj/Par)	+1.01	Max(Pron_third/Sent)	−0.23
	M(Dep_cc/Par)	+0.95	M(ParseDepth/Sent)	+0.21
	M(Connector_link/Par)	+0.87	Max(Dep_conj/Sent)	−0.20
Mihai Eminescu	M(Connector_link/Par)	−0.59	Max(Repetitions/Sent)	−0.21
	M(Wd/Par)	+0.48	M(Chars/Word)	−0.17
	M(UnqWd/Par)	−0.47	M(Commas/Par)	−0.17
Nicolae Filimon	M(Dep_cc/Par)	+0.87	Max(Repetitions/Sent)	+0.36
	M(Dep_conj/Par)	+0.81	M(Connector_temporal/Par)	−0.27
	M(Dep_det/Par)	+0.72	M(Connector_temporal/Doc)	−0.27
Petre Ispirescu	M(Pron_third/Par)	+1.63	Max(Dep_mark/Sent)	−0.25
	M(Dep_cc/Par)	−1.06	M(NmdEnt_person/Sent)	+0.24
	M(Dep_case/Par)	+0.90	SD(Polysemy/Word)	+0.24

+/−—positive/negative polarities; M—mean; SD—standard deviation; Max—maximum; /Doc—global count at document level; /Par—normalized counts at paragraph level; /Sent—local values at sentence leves; /Word—feature computed at word level; Dep_ —specific syntactic dependency (advcl—adverbial clause modifier; case—case marking; cc—coordination; conj—conjunct; det—determiner; disj—disjunct; mark—marker; nsubj—nominal subject); Unq—unique; Wd—Word; Pron—pronoun; NmdEnt—named entity from NER.

Emil Garleanu's texts are characterized by high usage of punctuation and commas, indicating complex sentence structures and suggesting a focus on rhythm. Third-person pronouns and repetitions indicate a narrative style characterized by vivid descriptions and storytelling. These features are essential for identifying Garleanu's authorship, providing distinct markers of their writing style.

Emilia Plugaru tends to use a variety of unique words per paragraph, which suggests a diverse vocabulary. Yet, repetitive connectors and disjunctions indicate a potential lack of coherence in her texts as suggested by the negative polarity associated with these features.

Ioan Slavici's writing style is characterized by a high usage of pronouns and nominal subjects per paragraph, which may indicate a narrative style focusing on characters and their actions. His writing style is distinguished by a balance between syntactic complexity, represented by the presence of commas and third-person pronouns, and narrative clarity, indicated by the use of third-person pronouns and the absence of repetitions. These features highlight both the structural and narrative aspects of his writing.

Ion Creanga exhibits a balanced use of commas and third-person pronouns per paragraph, suggesting moderate complexity in sentence structure and character interaction and reflecting a balance between syntactic complexity and narrative coherence.

I.L. Caragiale's writing style is marked by a high level of variability in sentence structure, indicated by a high standard deviation of parts of speech per sentence. This may suggest a dynamic and varied narrative style. The focus on syntactic structure is evidenced by the presence of punctuation and parse depth.

Liviu Rebreanu shows a preference for using a variety of connectors and links per paragraph, indicating a cohesive and logically structured narrative. His writing style is characterized by the presence of connector disjunctions and coherence markers, indicating a balance between structural complexity and narrative clarity and providing unique signatures of his style.

Eminescu's prose features many connector links but with a negative polarity, suggesting a less cohesive narrative style with disjointed elements. This implies a narrative style that may lack cohesion, featuring disjointed elements. Additionally, his texts are distinguished by frequent repetitions, indicating an intentional emphasis on certain aspects within the narrative.

Nicolae Filimon's writing is characterized by a high number of dependent clauses and conjunctions per paragraph, indicating syntactic complexity and a structured narrative style. Moreover, the positive polarities associated with repetitions and temporal connectors suggest a narrative style characterized by expressive descriptions and storytelling. These features collectively contribute to the identification of Filimon's authorship.

Petre Ispirescu exhibits a balanced use of third-person pronouns and case dependencies per paragraph, as well as connector disjunctions, suggesting a moderate level of narrative complexity and a balance between structural complexity and narrative coherence. Furthermore, the positive polarities associated with dependency markers and polysemy suggest a narrative style characterized by rich descriptions.

Summing up, each author's writing style exhibits unique linguistic features that are important for authorship attribution. Understanding the importance of these features and correlating them with other indicators, such as the number of words, unique words, and lexical diversity, helps in accurately identifying the authorship of literary texts, contributing to the field of literary analysis and AA.

5.3. Trends in Misclassifications

In this section, we investigate the trends related to misclassifications observed in our best-performing model (i.e., the hybrid Transformer) with the aim of enhancing our understanding of the limitations and challenges encountered by the model. Our objective is to reveal patterns within these misclassifications, identifying correlations between the types of errors produced by the model and the distinctive linguistic features associated with each author (presented in Table 13). To conduct this analysis, we refer to the confusion matrices (see Figures 4 and 5) that provide insights into the distribution of misclassifications across different authors and categories.

Analyzing the FT 10-author set, we observe three misclassifications. Ioan Slavici, for instance, was misclassified once with Emil Garleanu. The features associated with these errors include similar comma usage and third-person pronoun frequency. Liviu Rebreanu encountered one misclassification as Ion Luca Caragiale; however, the linguistic analysis did not include common characteristics in their top three discriminative features. Lastly, Mihai Eminescu was misclassified as Ion Creanga, and similar to the previous misclassification, correlating the two authors' profiles, there is no overlap in the top features of their writing style. From this analysis, it appears that misclassifications occur even when the top features of the misclassified authors do not overlap. This suggests that the misclassification is not attributed to specific linguistic features but rather contextual similarities.

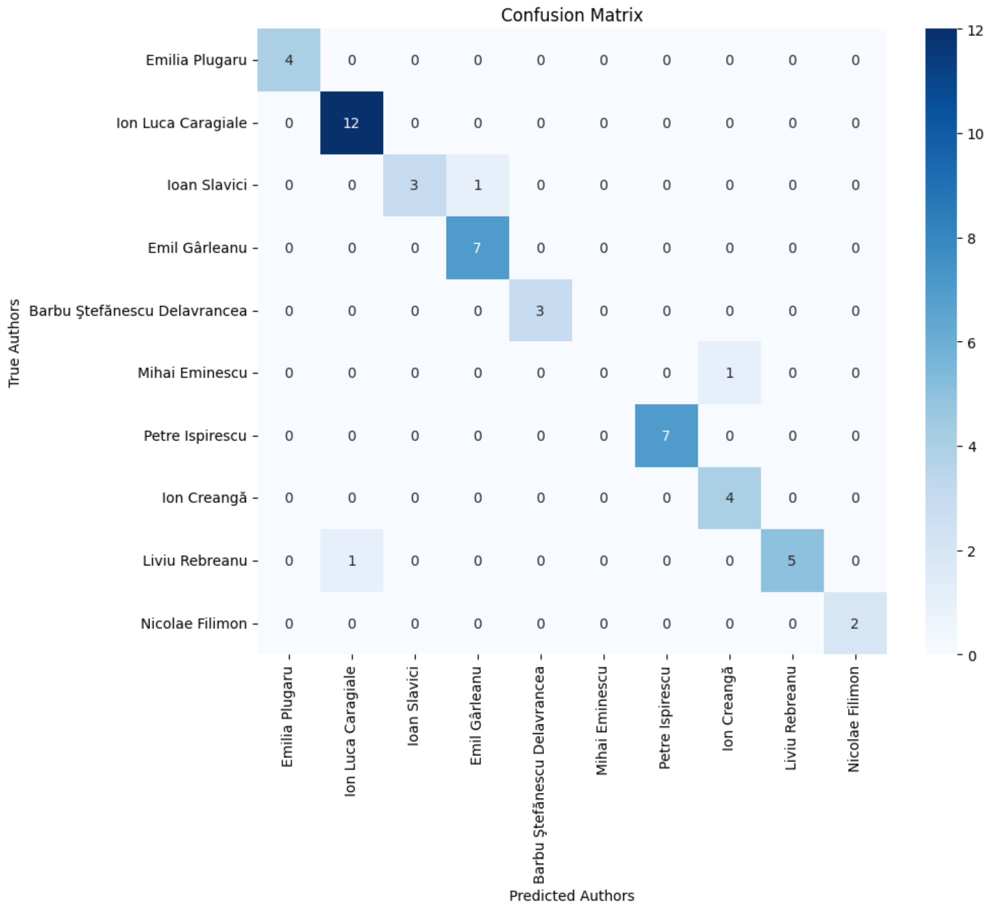

Figure 4. Hybrid Transformer: confusion matrix for FT 10-authors.

Further analysis extends to the PP 10-author set, uncovering additional misclassification patterns. Ioan Slavici faced the most misclassification, being mistaken three times for Liviu Rebreanu, eight times for Ion Luca Caragiale, and three times for Emil Garleanu. Features correlated with these errors comprise disjunction, third-person pronoun usage, and sentence-level syntactic dependency frequencies. Petre Ispirescu, for instance, was misclassified twice, and key features associated with these errors include third-person pronoun, coordination dependency, and named entity frequencies. Liviu Rebreanu experienced two misclassifications as Ion Luca Caragiale due to linguistic features like connector disjunction, coordination dependency, and connector frequencies. Barbu Stefanescu Delavrancea was misclassified as Ion Creanga, with common features comprising connector and sentence-level syntactic dependencies, and as Ioan Slavici, with similar writing at the sentence dependencies level. Ion Luca Caragiale was also misclassified as Slavici, and Rebreanu, due to features reflecting parse depth or punctuation usage. Furthermore, Emil Garleanu was misclassified as Liviu Rebreanu, with top features comprising punctuation frequency, named entity recognition, and third-person pronoun usage. Lastly, Ion Creanga experienced misclassifications, being mistaken for Petre Ispirescu or for Delavrancea and once for Mihai Eminescu. The linguistic features correlated with these errors involve third-person pronoun frequency, comma usage, and connector disjunction frequency.

Based on the information presented in this section, several patterns in misclassifications can be identified. Certain authors, such as Ioan Slavici and Emil Garleanu or Petre Ispirescu and Ion Creanga, are confused with each other, implying contextual similarities or thematic overlaps. Moreover, common linguistic features such as comma usage, third-person pronoun frequency, and syntactic dependencies are consistently associated with misclassifications across different authors. This underscores the complexity of author attribution, especially in distinguishing between authors with similar writing styles or thematic content. Despite the robustness of the hybrid model, certain authors consistently display misclassifications, suggesting the presence of confounding factors. Hence, there emerges a clear need for the continuous refinement of the methodologies, with a focus on feature selection and model adaptation to better accommodate individual author styles and minimize misclassifications.

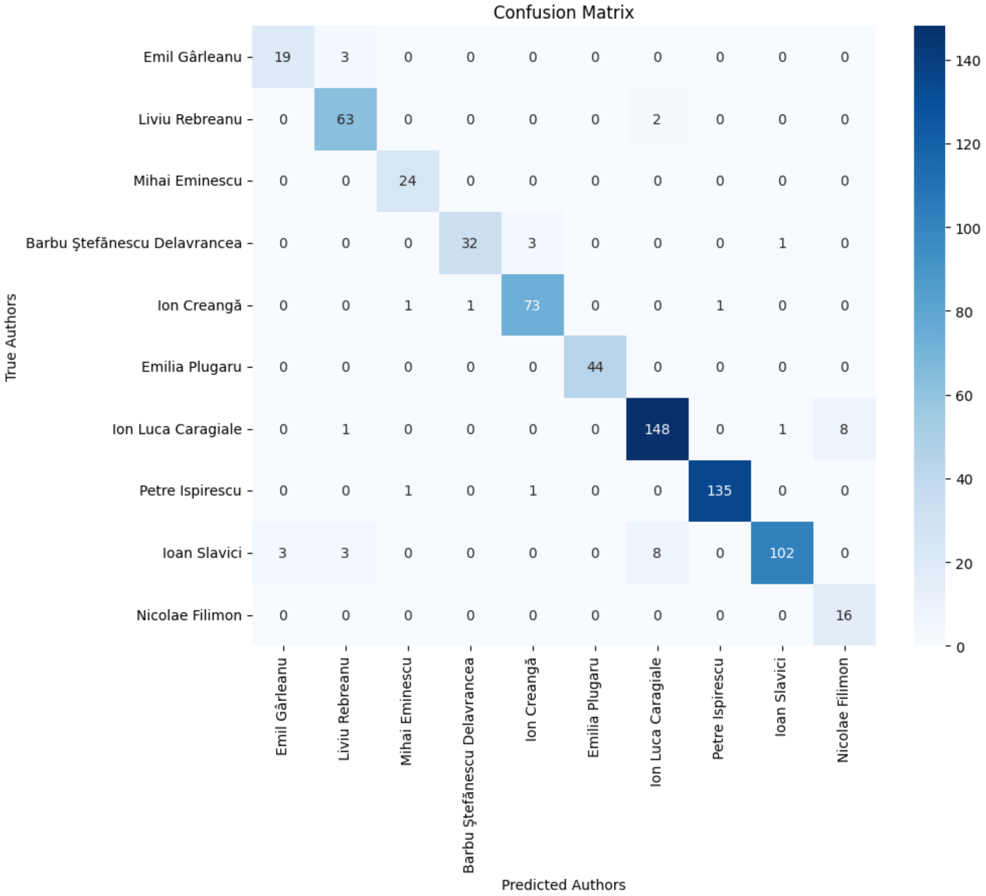

Figure 5. Hybrid Transformer: confusion matrix for PP 10 authors.

5.4. Limitations

In discussing the outcomes of our research, it is important to acknowledge certain limitations in our study. First, our evaluation exclusively targeted the performance of the proposed methodologies for Romanian texts. While our findings show promising results, the generalization of these approaches to other languages remains to be explored. The efficiency of the proposed models may vary when applied to languages with different linguistic characteristics. Second, our analysis depends on the quality of the dataset. While efforts were made to curate the corpus of Romanian stories, certain biases caused by an unbalanced dataset could potentially impact the robustness of our findings. Future research could benefit from including a larger and more diverse dataset, including a wider range of authors and writing styles.

6. Conclusions

The current study contributes to the field of authorship attribution in the context of Romanian, a less-resourced language. We have released an open-source corpus representing an extension of the existing ROST dataset consisting of Romanian stories, providing valuable resources for further research and development in this area. Moreover, we conducted an exploration of existing methodologies and their limitations for authorship attribution, setting the context for the proposed solutions.

One of the key contributions of our study is the introduction of a hybrid Transformer model tailored for Romanian authorship attribution, which is evaluated against a baseline ensemble of seven machine learning classifiers leveraging majority soft voting. The hybrid Transformer incorporates both numerical features from linguistic ReaderBench indices and textual features from a BERT encoder for author prediction. Furthermore, we consider feature selection via Kruskal–Wallis's non-parametric statistical test, enhancing the model's capability to identify relevant features for prediction. This methodology represents an innovative contribution to the field, offering a more refined and efficient approach to feature selection in authorship attribution tasks. Additionally, the proposed baseline ensemble model effectively improves the predictive performance compared to traditional standalone models or other existing ML methods. Furthermore, we compare the performance of our proposed models with similar existing methods. Results argue that our models consistently outperform or match the performance of existing approaches across various datasets and input representations.

Our hybrid RoBERT model achieved an F1-score of 0.95 and an error rate of 0.04 on the dataset with 10 authors when considering paragraphs as input. Additionally, when considering the dataset with 19 authors, our hybrid Transformer method reached an F1 score of 0.87 and an error rate of 0.12 on full-text data. These results underpin the effectiveness of our proposed method, outperforming the baseline ensemble learning approach, which achieved an F1-score of 0.82 with an error rate of 0.14 on the 10-author dataset and an F1-score of 0.84 with an error rate of 0.15 on the 19-author dataset.

In addition, the author′s writing styles were defined based on the most discriminative features in their texts, cross-correlated with lexical statistics. Our analysis provided insights into the linguistic patterns employed by the selected Romanian authors. Moreover, we conducted a linguistic analysis leveraging textual complexity indices and leveraged McNemar′s and Cochran′s Q statistical tests to evaluate the performance evolution across the best three models while also highlighting patterns in misclassifications.

In this paper, we focused on the challenges of authorship attribution in the context of Romanian texts, highlighting the unique linguistic complexities that pose problems to conventional solutions. While our discussion has primarily centered on developing methodologies for the Romanian language, we acknowledge the broader implications of our research beyond the immediate scope of our study. By addressing the problem of authorship attribution in Romanian, our work may contribute to a deeper understanding of computational linguistics and text analysis in multilingual environments. Moreover, the methods presented in this paper have diverse applications in diverse research areas. For instance, our methods can assist in analyzing disputed documents and legal texts in forensic linguistics. Additionally, our research may be used in historical document analysis to authenticate and attribute authorship to archival materials. Additionally, our research can enhance the exploration of authorial style in literary studies. Based on these potential applications, we underpin the significance of our research and its potential to extend beyond the limits of our immediate research context.

Directions for future research include expanding the scope of the corpus, refining the feature selection process, or exploring additional hybrid techniques. Additionally, investigating the applicability of our proposed models to other languages and domains would be an interesting avenue for exploration. Overall, our research contributes to a deeper understanding of authorship linguistic patterns in Romanian literature and provides a foundation for further research in the field of authorship attribution.

Author Contributions: Conceptualization, M.N. and M.D.; methodology, M.N. and M.D.; software, M.N.; validation, M.N.; formal analysis, M.N.; investigation, M.N.; resources, M.N.; data curation, M.N.; writing—original draft preparation, M.N.; writing—review and editing, M.D.; visualization, M.N.; supervision, M.D.; project administration, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS—UEFISCDI, project number PN-III-P4-PCE-2021-0300/PCE 127 from 29/06/2022, INTELLIV—"Soluții digitale inteligente pentru cercetarea și diseminarea literaturii române vechi și premoderne".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We release as open-source the dataset (https://huggingface.co/datasets/readerbench/ro-stories, accessed 12 February 2024), and the codebase along with statistical tests results (https://github.com/readerbench/ro-auth-detect, accessed on 12 February 2024). Additional information is available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AA	Authorship Attribution
AI	Artificial Intelligence
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
CNA	Cohesion Network Analysis
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Trees
EL	Ensemble Learning
FT	Full Text
GPT	Generative Pre-trained Transformer
GloVe	Global Vectors for Word Representation
GRU	Gated Recurrent Unit
IPoS	Inflexible Part of Speech
k-NN	k-nearest neighbor
KW	Kruskal–Wallis
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSTM	Long Short-Term Memory
MEP	Multi-Expression Programming
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLSOM	Multilayer Self-Organizing Map
NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
PP	Paragraphs
RF	Random Forests
RB	Readerbench
RBI	Readerbench Indices
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TF-IDF	Term-Frequency Inverse Document Frequency
TTR	Type–Token Ratio

References

1. De Oliveira, W.A., Jr.; Justino, E.; de Oliveira, L.S. Comparing compression models for authorship attribution. *Forensic Sci. Int.* **2013**, *228*, 100–104. [CrossRef]
2. Canhasi, E.; Shijaku, R.; Berisha, E. Albanian Fake News Detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 86. [CrossRef]

3. Belvisi, N.M.; Muhammad, N.; Alonso-Fernandez, F. Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features. In Proceedings of the 2020 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal, 29–30 April 2020; pp. 1–6.
4. Varela, P.; Justino, E.; Britto, A.; Bortolozzi, F. A computational approach for authorship attribution of literary texts using sintatic features. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4835–4842
5. Gasaway, L.N. Libraries, Users, and the Porblems of Authorship in the Digital Age. *DePaul L. Rev.* **2002**, *52*, 1193–1228. Available online: https://via.library.depaul.edu/law-review/vol52/iss4/7 (accessed on 12 February 2024).
6. Pandey, S.; Sahoo, S. Research Collaboration and Authorship Pattern in the field of Semantic Digital Libraries. *DESIDOC J. Libr. Inf. Technol.* **2020**, *40*, 375–381. [CrossRef]
7. Kim, J. Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics* **2018**, *116*, 1867–1886. [CrossRef]
8. Misini, A.; Kadriu, A.;Canhasi, E. A Survey on Authorship Analysis Tasks and Techniques. *SEEU Rev.* **2022**, *17*, 153–167. [CrossRef]
9. Ramnial, H.; Panchoo, S.; Pudaruth, S. Authorship Attribution Using Stylometry and Machine Learning Techniques. *Adv. Intell. Syst. Comput.* **2016**, *384*, 247–257.
10. Hossain, M.R.; Hoque, M.M.; Dewan, M.A.A.; Siddique, N.; Islam, M.N.; Sarker, I.H. Authorship Classification in a Resource Constraint Language Using Convolutional Neural Networks. *IEEE Access* **2021**, *9*, 100319–100338. [CrossRef]
11. Khdr, A.J.; Varol, C. Age and Gender Identification by SMS Text Messages. In Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 28–30 September 2018; pp. 1–5.
12. Deutsch, C.; Paraboni, I. Authorship attribution using author profiling classifiers. *Nat. Lang. Eng.* **2023**, *29*, 110–137. [CrossRef]
13. Suman, C.; Naman, A.; Saha, S.; Bhattacharyya, P. A Multimodal Author Profiling System for Tweets. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 1407–14162. [CrossRef]
14. Potha, N.; Stamatatos, E. A Profile-Based Method for Authorship Verification. In *Hellenic Conference on Artificial Intelligence*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2014; pp. 313–326.
15. Savoy, J. Feature selections for authorship attribution. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13), Coimbra, Portugal, 18–22 March 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 939–941.
16. Neocleous, A.; Loizides, A. Machine Learning and Feature Selection for Authorship Attribution: The Case of Mill, Taylor Mill and Taylor, in the Nineteenth Century. *IEEE Access* **2021**, *9*, 7143–7151. [CrossRef]
17. Avram, S.-M.; Oltean, M. A. Comparison of Several AI Techniques for Authorship Attribution on Romanian Texts. *Mathematics* **2022**, *10*, 4589. [CrossRef]
18. Elayidom, M.S.; Jose, C.; Puthussery, A.; Sasi, N.K. Text Classification For Authorship Attribution Analysis. *arXiv* **2013**, arXiv:1310.4909.
19. Suman, C.; Raj, A.; Saha, S.; Bhattacharyya, P. Source Code Authorship Attribution using Stacked classifier. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, Virtual, 13–17 December 2021; pp. 732–737.
20. Alsmearat, K.; Al-Ayyoub, M.; Al-Shalabi, R.; Kanaan, G.G. Author gender identification from Arabic text. *J. Inf. Secur. Appl.* **2017**, *35*, 85–95. [CrossRef]
21. Abuhammad, Y.; Addabe, Y.; Ayyad, N.; Yahya, A. Authorship Attribution of Modern Standard Arabic Short Texts. In Proceedings of the 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research (ArabWIC 2021), Sharjah, United Arab Emirates, 25–26 August 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–6.
22. Abbasi, A.; Javed, A.R.; Iqbal, F.; Jalil, Z.; Gadekallu, T.R.; Kryvinska, N. Authorship identification using ensemble learning. *Sci. Rep.* **2022**, *12*, 9537. [CrossRef]
23. Qian, C.; He, T.; Zhang, R. *Deep Learning based Authorship Identification*; Stanford Department of Electrical Engineering: Stanford, CA, USA, 2017.
24. Pennington, J.; Socher, R.; Manning, M. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Kerrville, TX, USA, 2014; pp. 1532–1543.
25. Vaz, P.C.; Martins de Matos, D.; Martins, B. Stylometric relevance-feedback towards a hybrid book recommendation algorithm. In Proceedings of the Workshop on Research Advances in Large Digital Book Repositories, Maui, HI, USA, 29 October 2012.
26. Pera, M.A.; Ng, Y.K. Analyzing Book-Related Features to Recommend Books for Emergent Readers. In Proceedings of the 26th ACM Conference on Hypertext and Social Media, Guzelyurt, Cyprus, 1–4 September 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 221–230.
27. Zhang, H.; Chow, T.; Wu, Q. Organizing Books and Authors by Multilayer SOM. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 2537–2550. [CrossRef]

28. Gupta, S.T.P.; Sahoo, J.K.; Roul, R.K. Authorship Identification using Recurrent Neural Networks. In Proceedings of the 2019 3rd International Conference on Information System and Data Mining (ICISDM '19), Houston, TX, USA, 6–8 April 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 133–137.
29. Modupe, A.; Celik, T.; Marivate, V.; Olugbara, O.O. Post-Authorship Attribution Using Regularized Deep Neural Network. *Appl. Sci.* **2022**, *12*, 7518. [CrossRef]
30. Škorić, M.; Stanković, R.; Ikonić Nešić, M.; Byszuk, J.; Eder, M. Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics* **2022**, *10*, 838. [CrossRef]
31. Uchendu, A.; Le, T.; Shu, K.; Lee, D. Authorship Attribution for Neural Text Generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020.
32. Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. [CrossRef]
33. Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Digital Authorship Attribution in Russian-Language Fanfiction and Classical Literature. *Algorithms* **2023**, *16*, 13. [CrossRef]
34. Stoean, C.; Lichtblau, D. Author Identification Using Chaos Game Representation and Deep Learning. *Mathematics* **2020**, *8*, 1933. [CrossRef]
35. He, X.; Lashkari, A.H.; Vombatkere, N.; Sharma, D.P. Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey. *Information* **2024**, *15*, 131. [CrossRef]
36. Bogdanova, A. Source Code Authorship Attribution Using File Embeddings. In Proceedings of the 2021 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity, Chicago, IL, USA, 17–22 October 2021; pp. 31–33.
37. Bogdanova, A.; Romanov, V. Explainable source code authorship attribution algorithm. *J. Phys. Conf. Ser.* **2021**, *2134*, 012011. [CrossRef]
38. Bagnall, D. Author identification using multi-headed recurrent neural networks. *arXiv* **2015**, arXiv:1506.04891.
39. Ruder, S.; Ghaffari, P.; Breslin, J.G. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv* **2016**, arXiv:1609.06686.
40. Shrestha, P.; Sierra, S.; González, F.A.; Montes-y Gómez, M.; Rosso, P.; Solorio, T. Convolutional Neural Networks for Authorship Attribution of Short Texts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 669–674.
41. Ferracane, E.; Wang, S.; Mooney, R. Leveraging discourse information effectively for authorship attribution. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 1 December 2017; Volume 1, pp. 584–593.
42. Hitschler, J.; Van den Berg, E.; Rehbein, I. Authorship attribution with convolutional neural networks and pos-eliding. In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 8 September 2017; pp. 53–58.
43. Boumber, D.; Zhang, Y.; Mukherjee, A. Experiments with convolutional neural networks for multi-label authorship attribution. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
44. Alsulami, B.; Dauber, E.; Harang, R.; Mancoridis, S.; Greenstadt, R. Source code authorship attribution using long shortterm memory based networks. In Proceedings of the European Symposium on Research in Computer Security, Oslo, Norway, 11–15 September 2017; Springer: Cham, Switzerland, 2017; pp. 65–82.
45. AlZahrani, F.M.; Al-Yahya, M. A Transformer-Based Approach to Authorship Attribution in Classical Arabic Texts. *Appl. Sci.* **2023**, *13*, 7255. [CrossRef]
46. Huertas-Tato, J.; Huertas-García, Á.; Martín, A.; Camacho, D. PART: Pre-trained Authorship Representation Transformer. *arXiv* **2022**, arXiv:2209.15373.
47. Kondyurin, I. Explainability of Transformers for Authorship Attribution. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 29 July 2022.
48. Huang, Z.; Iwaihara, M. Capsule Network Over Pre-Trained Language Model and User Writing Styles for Authorship Attribution on Short Texts. In Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System (CCRIS '22), Virtual, 26–28 August 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 104–110.
49. Bauersfeld, L.; Romero, A.; Muglikar, M.; Scaramuzza, D. Cracking double-blind review: Authorship attribution with deep learning. *PLoS ONE* **2023**, *18*, e0287611. [CrossRef]
50. Zhou, A.; Zhang, Y.; Lu, M. C-Transformer Model in Chinese Poetry Authorship Attribution. *Int. J. Innov. Comput. Inf. Control* **2022**, *18*, 901–916.
51. Avram, S.M. BERT-based Authorship Attribution on the Romanian Dataset called ROST. *arXiv* **2023**, arXiv:2301.12500.
52. Dascalu, M.; Gutu, G.; Ruseti, S.; Paraschiv, I.C.; Dessus, P.; McNamara, D.S.; Crossley, S.A.; Trausan-Matu, S. Readerbench: A Multi-lingual Framework for Analyzing Text Complexity. In *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, 12–15 September 2017*; Springer: Cham, Switzerland, 2017; pp. 606–609.
53. Dascalu, M.; McNamara, D. S.; Trausan-Matu, S.; Allen, L. K. Cohesion Network Analysis of CSCL Participation. *Behav. Res. Methods* **2018**, *50*, 604–619. [CrossRef] [PubMed]
54. McKight, P.; Najab, J.Kruskal-Wallis Test. In *The Concise Encyclopedia of Statistics*; Springer: New York, NY, USA, 2008.

55. Masala, M.; Ruseti, S.; Dascalu, M. RoBERT—A Romanian BERT Model. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 6626–6637.
56. Cochran, W.G. The Comparison of Percentages in Matched Samples. *Biometrika* **1950**, *37*, 256–266. [CrossRef]
57. Mcnemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.