

WORD AND PUNCTUATION N-GRAM FEATURES IN ROMANIAN AUTHORSHIP ATTRIBUTION

DANA LUPŞA AND RADU LUPŞA

ABSTRACT. This study addresses the problem of authorship attribution for Romanian texts, focusing on the use of N-gram features with an emphasis on semantic-independent representations. While character N-grams have been previously studied, this work extends the exploration to word and part-of-speech (POS) N-grams, as well as combinations involving punctuation, closed-class words, and filtered content words. Using the ROST corpus, we evaluate six supervised learning algorithms, with results averaged over multiple runs to ensure robustness. Our experiments show that Artificial Neural Networks (ANN) consistently achieve the highest performance, with word-based unigrams enhanced by punctuation reaching an average macro-accuracy of 0.93. Importantly, semantically independent features, such as closed-class words and POS replacements for nouns and verbs, yield small further improvements. These findings highlight the effectiveness of carefully designed N-gram features for Romanian AA and suggest that semantic-independent representations can complement traditional lexical approaches.

1. INTRODUCTION

Authorship attribution (**AA**) is the task of determining the author of a text based on its linguistic and stylistic properties. It has applications in fields such as digital humanities, plagiarism detection, forensic linguistics, and information security. A central challenge in AA is to identify features that capture the stylistic signature of an author while minimizing the influence of content and topic. Among the most widely used features are N-grams, which

Received by the editors: 20 September 2025.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* code [**Artificial Intelligence**]: Natural Language Processing – *Text Analysis*; code [**Artificial Intelligence**]: Learning – *Induction*.

Key words and phrases. Authorship Attribution, Machine Learning, N-gram features.

© Studia UBB Informatica. Published by Babeş-Bolyai University



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence.

encode recurring patterns at the level of characters, words, or part-of-speech (POS) tags.

While N-grams have been extensively studied in high-resource languages such as English, their use in Romanian AA remains relatively underexplored. The Romanian language presents additional challenges due to its rich morphology and limited availability of annotated corpora. Existing research on the ROST dataset has shown promising results with both traditional machine learning methods and modern neural approaches.

In this study, we provide a systematic evaluation of word and POS N-gram features, including their integration with punctuation, on a Romanian corpus. We introduce semantic-independent N-gram representations, showing how filtering out or replacing content words can enhance attribution performance. We demonstrate that ANN models achieve strong results, even with simple word unigrams plus punctuation, while semantically independent bigram features provide additional improvements. By focusing on both classical and semantic-independent feature sets, this study highlights effective strategies for authorship attribution in Romanian.

2. N-GRAMS FOR AUTHORSHIP ATTRIBUTION

In authorship attribution, **N-grams** are widely used features due to their ability to capture stylistic and syntactic patterns unique to each author [10, 2, 16, 15, 17, 9, 7]. The most common types of N-gram features used in this task are character N-grams, word N-grams and POS N-grams. It is considered that for **short texts**, character N-grams are typically more effective, while character and POS N-grams offer higher robustness. Although N-grams are widely used in languages like English, it's been relatively understudied for Romanian texts. An initial exploratory study, based on character N-grams, was presented in [11].

Character N-grams are sequences of N consecutive characters. They are expected to capture style irrespective of word boundaries, by identifying spelling habits, punctuation, and affixation patterns. Most important advantages are that they are language-independent and robust to noise.

Word N-grams are sequences of N consecutive words. They capture lexical collocations; can reveal frequent phrases and idioms, but they are sensitive to topic variation. They can identify an author based on the themes he approaches (semantic preference), while the aim is to find features of a specific style.

POS tag N-grams are sequences of part-of-speech (POS) tags. They focus on (syntactic) text structure and try to capture morpho-syntactic habits. A disadvantage is that, if POS tagging is performed automatically, it can introduce errors.

3. FEATURES WITH NO SEMANTIC INFORMATION FOR AUTHORSHIP ATTRIBUTION

Excluding semantic information in authorship attribution is often desirable, because we want to capture the stylistic signature of the author rather than the topic or meaning of the text. If semantic content dominates, the model may just learn *what* the author writes about, not *how* they write. Strategies to exclude semantic information include:

- Use function words or closed-class words
- Use POS tag sequences, that is: convert text into part-of-speech (POS) sequences, then extract N-grams. It is an attempt to reflect syntactic structure, not semantics.
- Mask or replace content words: replace all nouns, verbs with placeholders, and keep only structural patterns (POS or function words).

3.1. Function words and closed class words for semantic information of a text. Function words are words that have grammatical or relational meaning rather than lexical meaning. Unlike content words, which reflect the topic of a text, they are not topic-specific. Even if two texts discuss different subjects, the distribution of function words tends to reflect the writer's unconscious style. Because authors typically pay little conscious attention to how they use these words, their distributional patterns tend to be stable across different texts by the same writer, making them reliable stylistic markers.

By definition, closed-class words are word categories that form a fixed set in a language. While function words are closed-class, certain closed-class words are not function words, for example quantifiers like *some* or *many*. Also pronouns are excluded from the function word category because their primary role is to serve as nominal arguments. Closed-class words do not carry significant semantic information on their own. It's also worth noting that in some studies, the terms closed-class words and function words are treated interchangeably, as are open-class words and content words [4, 5, 13].

3.2. Nouns and verbs for semantic information of a text. In opposition with closed-class words, open-class words are nouns, verbs, adjectives, and adverbs. In general, verbs and nouns contribute more directly and specifically to the semantic theme of a text, than adjectives and adverbs. Nouns name entities, concepts, objects, people; they support the core ideas of a text. Verbs express actions, states, or processes; they give action and direction to the theme. Nouns and verbs are essential for understanding what the text is about. On the other hand, adverbs and adjectives are not so specific to the message of the text. The adjectives add details about qualities or characteristics of nouns

and they don't define the theme, they add nuance to it. Similarly, adverbs modify verbs, adjectives, or other adverbs, indicating manner, time, place, etc.

In this study, we extend the investigation of semantic-independent features by filtering out only nouns and verbs during feature extraction. To the best of our knowledge, this represents the first attempt to use only non-noun and non-verb words for Romanian authorship attribution.

4. ISSUES IN ROMANIAN AUTHORSHIP ATTRIBUTION

Authorship attribution for Romanian texts presents a unique set of challenges, largely due to the language's specific characteristics and the general limitations faced by research in less-resourced languages.

One of the most significant challenges is the lack of large, high-quality, and publicly available annotated datasets. While resources like the ROST corpus exist [2, 14], they are often smaller than what is available for high-resource languages like English, which can limit the effectiveness of data-hungry deep learning models. The ROST corpus is highly unbalanced, with a significant variation in the number of texts per author (ranging from 27 to 60) and a broad historical period (1850-2023). It can contain texts from various genres (e.g., novels, articles, short stories) exposing substantial variation in text length, from short texts with of 90 words to some exceeding 39000 words. These factors introduce significant challenges for model performance on unseen data.

Also, the high-quality Natural Language Processing (NLP) set of tools for Romanian is not as extensive as for major languages. This can hinder research that relies on detailed linguistic feature extraction.

Despite these challenges, recent research [1, 2, 3, 11, 14] has made significant progress in the field, by exploring approaches that combine traditional stylistic features with modern learning techniques, including fine-tuned BERT models for Romanian and achieving state-of-the-art performance.

5. THE EXPERIMENTS: MATERIALS AND METHODS

5.1. The dataset. The ROST (ROmanian Stories and other Texts) dataset is a significant and publicly available corpus for the Romanian language, specifically designed for Authorship Attribution (AA). It contains 400 documents authored by 10 writers, but it is important to know that the dataset is highly unbalanced, and the texts are diverse in both genre and length.

Several research papers have used the ROST dataset to benchmark authorship attribution models, using different approaches, from traditional machine learning to modern deep learning models. Table 1 provides a short overview of the results, showing the results and improvements researchers made in Romanian AA over time.

Study	Methods / Features	Evaluation	Notes
Avram et al. [2] (2022)	ANN, MEP, k-NN, SVM, DT; inflexible parts of speech	MAcc: 80.94%	Introduced ROST dataset
Avram [1] (2023)	Romanian-specific BERT	MAcc: 87.40%	
Nitu et al. [14] (2024)	Hybrid Transformers; RoBERT	F1: 95% , Error: 4%	
Lupsa et al. [11] (2025)	SVM, LR, k-NN, DT, RF, ANN; character N-gram	MAcc: 95.1 %	several cases with perfect classification

TABLE 1. Short overview of previous studies on the ROST dataset.

5.1.1. *Data Preprocessing.* Following the work in [11], we used most of the normalization reported there, in order to work with standardized data. We made the diacritic standardization and punctuation unification; we replaced all digits with a special character (@). We also marked the beginning and ending of paragraphs with the dollar sign (\$); this allows us to treat paragraph breaks as distinct tokens in the N-gram generation process. We also performed all experiments on both lowercase and original-cased text; while some differences were observed in the results, they were small and do not justify a separate analysis, as in [11].

5.1.2. *Features extraction.* To build the representations of the texts, we used word and Part-of-Speech (POS) N-grams. We extracted these features with the help of `spacy`'s `ro_core_news_sm`. This is a pretrained `spacy` model designed for Romanian language processing, that can split text into token words, punctuation, and symbols. The model also assigns POS tags, such as NOUN and VERB, which were needed for our analysis.

5.2. **Classification algorithms.** Authorship attribution has been addressed using numerous classification algorithms [15, 12, 2, 6]. Continuing the work in [11], our analysis utilized a suite of six supervised learning algorithms, implemented via the `scikit-learn` library: Decision Trees (DT), Random Forest (RF), k-Nearest Neighbor (k-NN), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Networks (ANN).

Similar to the methodology used in [11], we repeated the model training and evaluation five times, with a train-test split of 80 to 20, varying the random seed of the ML models for each run, to reduce the impact of stochastic variation on our findings. This approach ensures that our findings are robust and not dependent on a specific random seed. We also did the experiments for lowercase and original-cased letters, in order to cover both methodologies.

For data processing and manipulation, we used `NumPy` and `pandas`, and `scikit-learn` for model training and evaluation, while the textual data was transformed into a numerical vector format using `TfidfVectorizer`, which is part of the `scikit-learn` library.

In this study, accuracy and macro-accuracy are the primary metrics used to assess performance. Accuracy (**Acc**) is the simplest metric and it is generally used for a quick, general view in case of balanced datasets. Macro-Accuracy (**MAcc**) is the average of the accuracy achieved on each class independently, and it should be used for imbalanced classes. All reported Acc and MAcc values represent averages over multiple runs for a specific selection of experiment set, and they are presented in a concise way, in order to highlight the most important of our findings.

6. OVERALL RESULTS AND DISCUSSION

6.1. One-type feature for N-grams.

6.1.1. *Word N-grams.* To evaluate the impact of sequential word patterns for authorship attribution, we conducted experiments using word-based N-grams as features. In this setting, contiguous sequences of words of length N (with N ranging from 1 to 5) were extracted from the text corpus and used to construct the feature space.

A visual representation of the results (mean macro-accuracy scores) is provided in Figure 1A. Despite differences in how classifier performance changes with N , the figure indicates that the highest scores are generally obtained at $N = 1$, with a maximum value of 0.9377 achieved by ANN and $N = 1$. The results also show that, in four of the six classifiers, accuracy decreases as N grows.

6.1.2. *POS N-grams.* In addition to word-sequence N-grams representations, we conducted experiments using Part-of-Speech (POS) N-grams as features. In this approach, the original text was first tagged with POS labels, after which contiguous sequences of tags of length N (with N ranging from 1 to 5) were extracted to construct the feature space. By systematically varying the N-gram length, we aimed to assess how increasingly complex syntactic sequences influence classification performance. Comparatively, while word N-grams tend to capture richer lexical information, POS N-grams offer a complementary perspective by abstracting away from vocabulary-specific features.

A visual representation of these results is shown in Figure 1B. The best performance is observed for $N = 2$ to $N = 4$; however, all macro-accuracy values remain below 0.9. The figure also shows that the majority of classifiers achieved their best performance when considering sequences of length greater

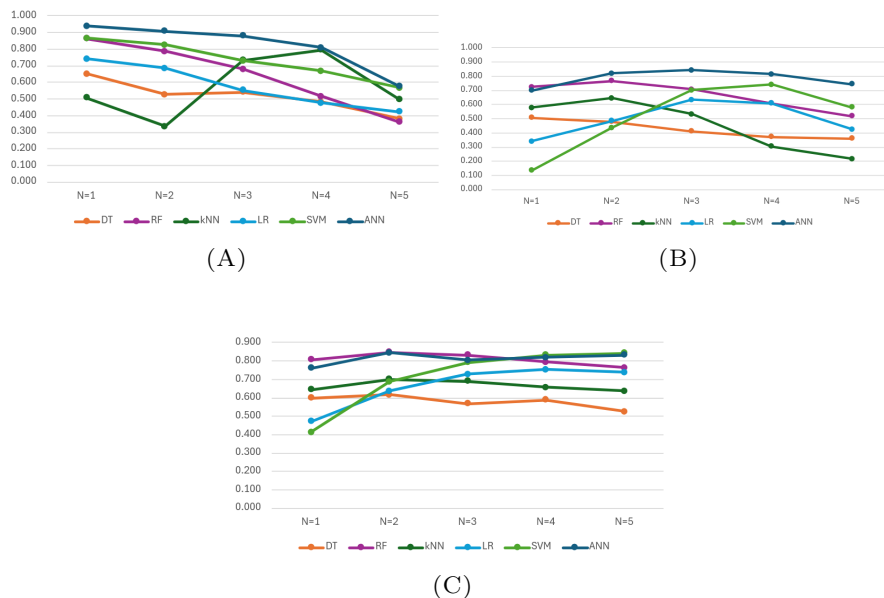


FIGURE 1. Results of classification experiments using exactly one feature type contributing to an N-gram. Used features are: only words in (A), only POS (part of speech) in (B), and only punctuation in (C). The figure shows the performance for each of the six classification methods: DT, RF, k-NN, LR, SVM, and ANN, when N varies from $N = 1$ to 5.

than 2, suggesting that contextual information beyond unigrams (syntactic structures and grammatical regularities) contributes significantly to classification accuracy.

6.1.3. *Punctuation N-grams.* We also conducted experiments in which punctuation marks were used exclusively as the basis for feature construction. In this setting, from the text corpus were extracted only punctuation symbols, and contiguous sequences of length N (with N ranging from 1 to 5) were extracted to form N-gram representations.

A visual representation of these results is presented in Figure 1C. Despite relying exclusively on punctuation-based features, the classification models achieve unexpectedly strong performance in certain cases, with maximum macro-accuracy values of 0.8453 for ANN at $N = 2$ and 0.8483 for RF at $N = 2$. While these findings highlight the potential discriminative power of punctuation, the achieved performance levels remain well below state-of-the-art performance levels.

6.2. Inclusion of punctuation in feature sets. Punctuation can encode subtle stylistic patterns [8, 9]. We evaluated the contribution of punctuation in context by adding it first to the basic word feature set and then to the POS feature set.

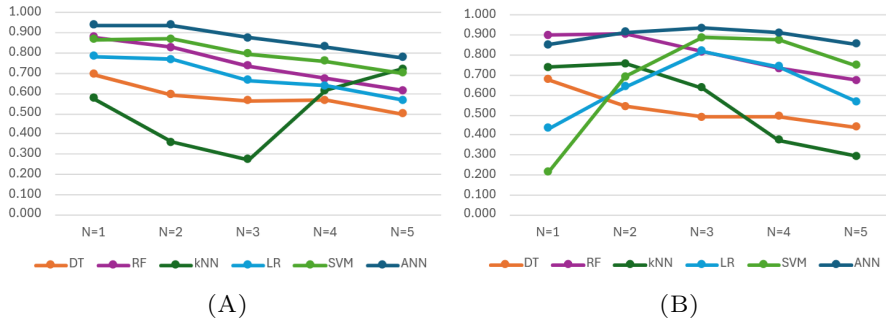


FIGURE 2. Results of classification experiments when punctuation is added to N-grams. Used features are: (A) Words and punctuation; (B) POS and punctuation. The figure shows the performance for each of the six classification methods: DT, RF, k-NN, LR, SVM, and ANN, when N varies from $N = 1$ to 5.

Results for features combining words with punctuation are shown in figure 2A, while the corresponding results for word-punctuation features are presented in figure 2B. Reported values represent the mean macro-accuracy scores obtained across multiple experimental runs. We can see that the highest average macro-accuracy scores, slightly exceeding 0.93, are obtained with the ANN classifier. Specifically, values of 0.9392 and 0.9379 are achieved using word-punctuation features at $N = 1$ and $N = 2$, respectively, while a score of 0.9357 is obtained with POS-punctuation features at $N = 3$. We can also remark that, for word-punctuation features, most classifiers show no improvement for $N > 2$.

6.3. Semantic-Independent N-gram Features. Another series of experiments was made in order to evaluate the results obtained by using semantically independent features. We explored the effect of using closed-class words in authorship attribution, and also experimented with filtering out nouns and verbs.

6.3.1. Closed-class words. In our experiments, we did not use a predefined list of closed-class words, but we considered in this category all the words with POS being different than nouns, verbs, adjectives and adverbs. Non-word items, like punctuation and other symbols, are included in the feature set, for

this processing. We experimented with two distinct set of features based on closed-class words. First, we eliminated all open-class words from the feature set. After that, open-class words were replaced with their corresponding part-of-speech annotation.

An overview of the results is presented in the figures 3A and 3B. Reported values represent the mean macro-accuracy scores obtained across multiple experimental runs. We can see that the ANN classifier consistently achieved the highest performance across all values of N (from 1 to 5). The maximum performance achieved with closed-class features is obtained in the second configuration, and is reached with ANN and $N = 2$, having average MAcc value of 0.9446 and average Acc value of 0.9449. It is worth noting that there is one case for each of the two feature sets (see Table 2, sections named *closedclass* and *closedclass-pos*) in the classification experiments in which the ANN model attains perfect accuracy, meaning that all texts are correctly assigned to their corresponding authors.

6.3.2. Filtering out nouns and verbs. Nouns and verbs are generally regarded as the most informative lexical categories in a text. To investigate their role in authorship attribution, we designed two distinct feature sets based on their manipulation. In the first case, all nouns and verbs were entirely removed from the feature space. In the second case, nouns and verbs were replaced by their corresponding POS annotation, thereby abstracting their lexical identity while preserving structural information.

An overview of the results is presented in the figures 3C and 3D. We can see that ANN obtained the best results, for all values of N (from 1 to 5), as in the majority of the other experiments. The maximum performance achieved with filtering out the nouns and verbs is obtained in the second configuration, with ANN and $N = 2$, having average MAcc value of 0.9494 and Acc value of 0.9504. Results obtained by using ANN are also presented in Table 2, sections named *nonnounverbs* and *nounverb-pos*. We should also mention that there is one case for each of the two feature sets (*nonnounverbs* and *nounverb-pos*) in the classification experiments in which all the texts are correctly assigned to their authors (i.e. the accuracy is 1).

6.4. Analysis of the results. The results indicate that improvements in accuracy do not follow a correlated pattern across the six classifiers, suggesting that each model responds differently to the feature representations employed.

ANN consistently achieves the best results across the majority of experiments. Therefore, in the following, we focus on comparing the results obtained with ANN. A detailed summary of the results, including average, standard deviation, maximum values for accuracy and macro-accuracy values obtained,

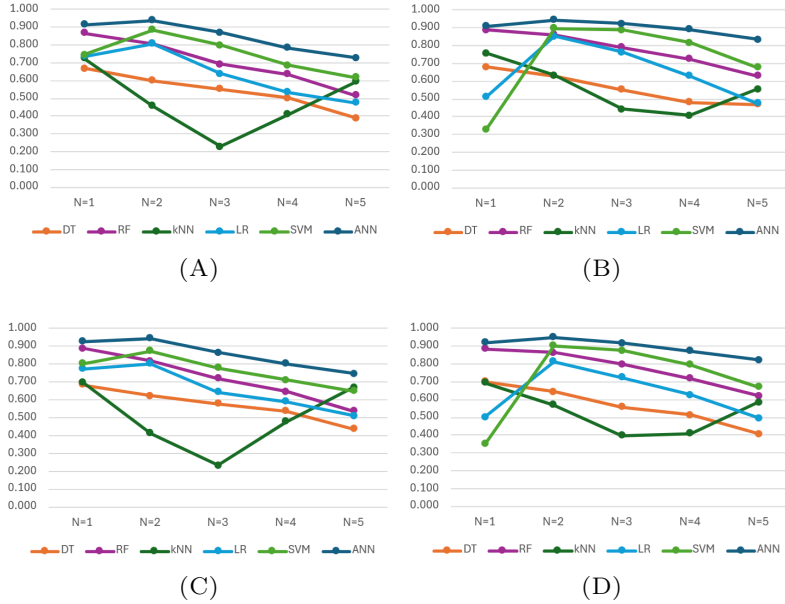


FIGURE 3. Results of classification experiments for N-grams formed by semantic-independent features. Used features are: (A) Closed class words and punctuation; (B) Closed class words, punctuation and POS tag for open class words. (C) Punctuations and all words different than nouns and verbs; (D) POS annotation for nouns and verbs and all other tokens (words and punctuation). The figure shows the performance for each of the six classification methods: DT, RF, k-NN, LR, SVM, and ANN, when N varies from $N = 1$ to 5.

is presented in Table 2. A brief visual summary of these results is presented in fig. 4.

The results obtained using the simple, classical word-based unigram features, including punctuation (see Section 6.2) yield an average macro-accuracy value of 0.9392. In our study on N-gram usage, if N-grams of order $N > 1$ fail to achieve better performance, it suggests that extending the classification to higher-order N-grams provides no additional benefit.

In Table 2, the MAcc value corresponding to word unigram features is marked with green, and all MAcc values exceeding this benchmark are highlighted in yellow for clarity. As shown in the table, approaches that rely on semantically independent features achieve higher macro-accuracy (MAcc) values. An interesting observation is that these improvements occur primarily at $N = 2$, whereas for $N = 1$ the MAcc values remain below 0.93—though still

Features		Acc			MAcc		
		mean	max	std	mean	max	std
<i>word</i>	N=1	0.9432	0.9753	0.0217	0.9392	0.9762	0.0249
	N=2	0.9442	0.9877	0.0195	0.9380	0.9929	0.0225
	N=3	0.8919	0.9383	0.0275	0.8772	0.9548	0.0315
	N=4	0.8506	0.9259	0.0280	0.8306	0.9025	0.0362
	N=5	0.7988	0.8395	0.0228	0.7789	0.8462	0.0338
<i>closedclass</i>	N=1	0.9170	0.9630	0.0246	0.9146	0.9735	0.0251
	N=2	0.9390	1.0000	0.0236	0.9381	1.0	0.0253
	N=3	0.8857	0.9383	0.0213	0.8698	0.9348	0.0301
	N=4	0.8067	0.8642	0.0355	0.7852	0.8624	0.0440
	N=5	0.7514	0.8519	0.0426	0.7265	0.8413	0.0433
<i>closedclass-pos</i>	N=1	0.9146	0.9630	0.0239	0.9100	0.9774	0.0298
	N=2	0.9449	1.0000	0.0261	0.9446	1.0	0.0278
	N=3	0.9272	0.9630	0.0257	0.9253	0.9786	0.0307
	N=4	0.8978	0.9630	0.0247	0.8908	0.9532	0.0274
	N=5	0.8556	0.9383	0.0405	0.8353	0.9276	0.0399
<i>nonounverb</i>	N=1	0.9314	0.9753	0.0227	0.9265	0.9762	0.0249
	N=2	0.9484	1.0000	0.0221	0.9457	1.0	0.0240
	N=3	0.8864	0.9259	0.0253	0.8657	0.9314	0.0358
	N=4	0.8306	0.8889	0.0233	0.8026	0.8869	0.0391
	N=5	0.7674	0.8519	0.0402	0.7457	0.8425	0.0486
<i>nounverb-pos</i>	N=1	0.9230	0.9630	0.0261	0.9204	0.9690	0.0307
	N=2	0.9504	1.0000	0.0244	0.9494	1.0	0.0263
	N=3	0.9215	0.9506	0.0255	0.9178	0.9607	0.0302
	N=4	0.8874	0.9506	0.0372	0.8746	0.9431	0.0416
	N=5	0.8442	0.9383	0.0450	0.8233	0.9122	0.0423
<i>pos</i>	N=1	0.8491	0.9259	0.0459	0.8514	0.9358	0.0485
	N=2	0.9217	0.9877	0.0268	0.9138	0.9917	0.0356
	N=3	0.9380	0.9753	0.0272	0.9357	0.9750	0.0292
	N=4	0.9168	0.9753	0.0377	0.9127	0.9774	0.0409
	N=5	0.8723	0.9630	0.0458	0.8552	0.9556	0.0433

TABLE 2. Performance results obtained by ANN for different N-gram feature configurations grouped by feature type. *word* and *pos* feature denote word and POS N-grams, including punctuation (as described in Section 6.2). *closedclass* refers to features restricted to closed-class words (as defined in Section 6.3.1), and *closedclass-pos* is for closed-class words combined with POS tags for all remaining items. *nonounverb* refers to experiments where all nouns and verbs were removed (see Section 6.3.2), while *nounverb-pos* denotes experiments in which nouns and verbs were retained but replaced with their corresponding POS tags.

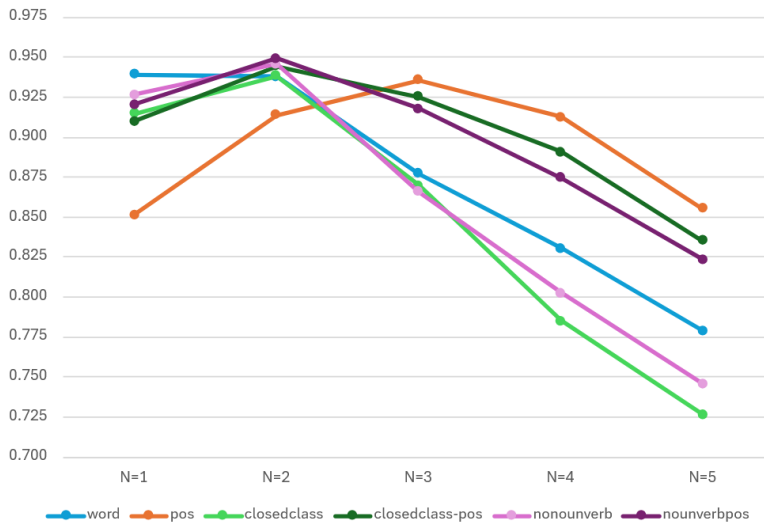


FIGURE 4. Macro-Accuracy averages for each N-gram size ($N = 1$ to 5) obtained by ANN, for the six feature selection methods: words with punctuation, part-of-speech (pos) with punctuation, features based on closed class words and features with no-nouns and no-verbs. The full set of results is reported in Table 2.

relatively high, exceeding 0.92. Also, the highest obtained value (0.9494) is close to the MAcc value of 0.951 reported in [11] (see Table 1).

Another interesting aspect is that, when the N-gram length increases beyond $N = 2$, i.e., for $N \geq 3$, the corresponding values consistently decrease, indicating reduced effectiveness of higher-order N-grams for most of feature selection cases. But feature representations using part-of-speech with punctuation information achieve better performance than the other feature selection when $N \geq 3$. This approach reaches a score of 0.9357 average MAcc for $N = 3$, and then decreases for higher values of N .

7. CONCLUSIONS

In this paper, we continued our investigation started in [11], addressing the authorship attribution problem on a Romanian dataset. To the best of our knowledge, this is the first study on Romanian that explores both word N-grams and POS N-grams, and also some of their combinations, focusing on N-grams of semantic independent features, as presented in section 6.3.

Our experiments show that Artificial Neural Networks (ANN) consistently achieve the highest performance. More than that, simple word-based unigram

features, enhanced with punctuation, reaching a macro-accuracy of 0.9392. Compared to this, feature representations designed to be semantically independent provide small improvements, particularly at the bigram level, with incorporating part-of-speech information for eliminated items, where they surpass the simple unigram scores and outperform the rest of results in our experiments. Higher-order N-grams ($N \geq 3$) show declining performance for most of the features. These findings highlight the effectiveness of carefully designed N-gram features for Romanian AA and suggest that semantic-independent representations can complement traditional lexical approaches.

REFERENCES

- [1] AVRAM, S.-M. Bert-based authorship attribution on the romanian dataset called rost. *arXiv preprint arXiv:2301.12500* (2023).
- [2] AVRAM, S.-M., AND OLTEAN, M. A comparison of several ai techniques for authorship attribution on romanian texts. *Mathematics* 10, 23 (2022), 4589.
- [3] BRICIU, A., CZIBULA, G., AND LUPEA, M. **AutoAt**: A deep autoencoder-based classification model for supervised authorship attribution. *Procedia Computer Science* 192 (10 2021), 397–406.
- [4] DE MARNEFFE, M.-C., NIVRE, J., AND ZEMAN, D. Function words in universal dependencies. *Linguistic Analysis* 43, 3–4 (2024), 549–588.
- [5] DREXLER, E. Qnrs: Toward language for intelligent machines, 2021.
- [6] HE, X., LASHKARI, A. H., VOMBATKERE, N., AND SHARMA, D. P. Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information* 15, 3 (2024).
- [7] HOUVARDAS, J., AND STAMATATOS, E. N-gram feature selection for authorship identification. In *Artificial Intelligence: Methodology, Systems, Applications* (2006).
- [8] HOWEDI, F., AND MOHD, M. Text classification for authorship attribution using naive bayes classifier with limited training data. *Computer Engineering and Intelligent Systems* 5 (2014), 48–56.
- [9] KOPPEL, M., SCHLER, J., AND ARGAMON, S. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology* 60, 1 (2009), 9–26.
- [10] LÓPEZ-ANGUITA, R., MONTEJO-RÁEZ, A., AND DÍAZ-GALIANO, M. C. Complexity measures and pos n-grams for author identification in several languages: Sinai at pan@clef 2018. In *Conference and Labs of the Evaluation Forum* (2018).
- [11] LUPSA, D., AVRAM, S.-M., AND LUPSA, R. Oldies but goldies: The potential of character n-grams for romanian texts. *Studia Universitatis Babeş-Bolyai Informatica* 70, 1-2 (2025), 25–42.
- [12] MISINI, A., KADRIU, A., AND CANHASI, E. A survey on authorship analysis tasks and techniques. *SEEU Review* 17 (12 2022), 153–167.
- [13] NICULESCU, O., AND VASILEANU, M. Prolongation in Romanian. In *Interspeech 2025* (2025), pp. 379–383.
- [14] NITU, M., AND DASCALU, M. Authorship attribution in less-resourced languages: A hybrid transformer approach for romanian. *Applied Sciences* 14, 7 (2024), 2700.
- [15] STAMATATOS, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60, 3 (2009), 538–556.

- [16] STAMATATOS, E. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21, 2 (01 2013), 421–439.
- [17] WANWAN, Z., AND JIN, M. A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics* 15 (04 2022).

DEPARTMENT OF COMPUTER SCIENCE, BABEŞ-BOLYAI UNIVERSITY, CLUJ-NAPOCA,
ROMANIA

Email address: `dana.lupsa@ubbcluj.ro`

Email address: `radu.lupsa@ubbcluj.ro`