AUTHORSHIP CLASSIFICATION TECHNIQUES: BRIDGING TEXTUAL DOMAINS AND LANGUAGES

*Arta Misini (1), Arbana Kadriu (1), Ercan Canhasi (2)**

(1) South East European University, Tetovo; (2) University "Ukshin Hoti" Prizren
[1] North Macedonia ; [2] Kosova

* Corresponding Author, e-mail: ercan.canhasi@uni-prizren.com

Abstract: Authorship classification analyzes an author's prior work to identify their writing style, a unique trait of each language and individual author. This research aims to conduct a thorough comparative analysis of various methods for classifying authorship. The study leverages two corpora: AAALitCorpus of Albanian literary texts and CCAT10 of English columns. We evaluate model-generated features across different configurations. The richness of the features and the breadth of the analysis provide a significant understanding of the problem, setting a new standard for comprehensive linguistic investigations across multiple languages. The study indicates that machine learning algorithms accurately discern authorial writing styles, highlighting the complexities of classifying authorship in a cross-linguistic context.

Key words: natural language processing, authorship classification, textual data, feature space, multiclass classification.

1. INTRODUCTION

Authorship classification (AA) is a decade-old issue in natural language processing (NLP). Its core objective lies in identifying a text's writer by analyzing the distinctive attributes of their previous works to ascertain the authorship of the text. Feature extraction is pivotal in this process, empowering classifiers to determine the author's identity based on the extracted attributes from a piece of writing.

Examining literary writings with disputed authorship has significantly shaped the evolution of AA over time. This interest in contested authorship in literature can be traced to studies dating back to the 19th century. These studies employed statistical analyses of writing styles, focusing on works such as Shakespeare's plays and the Federalist Papers. In today's digital age [1], the Internet has expanded anonymous content, making AA an increasingly concern. This issue carries substantial implications across various domains, including literature [2–4], journalism [5–8], and forensics [9].

AA aims to compile a collection of attributes that consistently characterize the writing style across multiple works attributed to the same author. Each author has a unique pattern in constructing sentences and selecting words. Consequently, most of methods rely on a feature set to identify the authorship of a given text [10]. These

features are subsequently employed by classifiers [11] to deduce the identity of the anonymous author.

This study utilizes an Albanian literary corpus and the CCAT10 English dataset [6, 12, 13] collected through web scraping [14]. A series of experiments is conducted employing model-generated features and machine learning (ML) methods for thorough analysis. This work serves as a significant contribution to the field of AA, offering a methodology that can be readily adapted for other languages.

The paper's structure follows: Section 2 overviews prior research on AA in various languages. In Section 3, we present the data collection process and corpus creation. Section 4 details the data and methods employed for implementing the AA models. In Section 5, we report the results of the conducted experiments. Section 6 presents the conclusion of the paper.

2. RELATED WORK

In this section, our objective is to present an extensive exploration of the research body regarding AA across various languages. Our examination is structured around three elements: feature extraction, classification methods, and datasets.

Feature extraction. Feature extraction, an essential element of AA, is pivotal in discerning the distinctive writing styles employed by different authors. Researchers have employed an array of features to tackle AA challenges. These features span lexical, morphological, syntactic, semantic, and structural dimensions, offering insights into a text's content, syntax, grammar, and vocabulary. When employed collectively, linguistic features form a comprehensive feature set for AA and other stylistic tasks [6, 7]. They enable the analysis of language and writing styles within texts, facilitating the identification of distinctive authorship characteristics. Among these features, n-grams, stand out as widely used elements within different tasks related to authorship. Extracting n-grams applies to any language and can effectively encapsulate stylistic nuances on lexical, syntactic, or structural dimensions. Table 1 provides a comprehensive summary of the most frequently utilized style markers offering insight into the core elements to analyze authorship.

Table 1. Stylometric features used in the authorship attribution task

Feature set	***Reference***
Lexical	*[2, 3, 15]*
Morphological	*[3, 15]*
Syntactic	*[3, 6, 7]*
Structural	*[3, 15]*

Authorship classification methods. Numerous studies have applied different methodologies across various languages. ML techniques have garnered attention for their efficacy in automated classification, allowing researchers to concentrate on automatic AA methods. A diverse array of ML-based algorithms has been utilized to categorize text samples according to their authorship. These algorithms undergo training on a corpus of known texts, enabling them to understand the distinctive attributes related

to different authors. Subsequently, these models are deployed to classify new, previously unseen texts. Furthermore, contemporary research has embraced deep learning techniques in AA. Deep learning (DL) [16, 17] has gained prominence across various domains. It adopts an end-to-end approach, eliminating the need for manual feature engineering and providing an attractive choice for AA. Table 2 is a repository of AA methodologies. This exploration aims to provide a comprehensive understanding of the methods that have shaped the AA landscape. Misini et al.'s research [18] offers an extensive exploration of the techniques applied in tasks related to authorship.

Table 2. Machine- and deep-learning methods used in authorship attribution studies

Technique	***Method***	***Reference***
Machine learning	*Naïve Bayes*	*[19]*
	Support Vector Machine	*[2, 15, 19]*
	Decision Tree	*[3]*
	k-Nearest Neighbor	*[15, 19]*
Deep learning	*Long Short-Term Memory*	*[2]*
	Convolutional Neural Network	*[2]*

Dataset repositories. Authorship attribution relies significantly on using text collections for evaluating various methods. This practice has been adopted across diverse linguistic contexts, encompassing languages such as English [20], Russian [21], Urdu [5], Bengali [2]. Benchmark datasets have been crucial in linguistics as the foundation for numerous AA studies. While many studies rely on established benchmark datasets like Enron [20], C50 [7], PAN [22], IMDb62 [6, 21] and others [9], the scarcity of standard datasets, particularly for low-resource languages, presents a unique challenge. Creating specialized corpora has paved the way for promising advancements in the field, demonstrated by projects like UNAAC [5], BAAD [2], UrduCorpus [5], A3C Corpus [8, 25], and more [4]. These corpora tailored for AA contribute significantly to the field, expanding its resources. AA research encompasses a diverse array of data, ranging from literary works [2–4], news articles [5, 8, 23, 24], tweets [9, 19], and blog posts [21]. Albanian, being resource-constrained, has garnered limited attention in NLP. To tackle this issue, we have crafted a novel corpus tailored to AA in Albanian. Detailed insights into the corpus creation process are provided in the subsequent section.

3. AAALitCorpus

Constructing a high-quality dataset is fundamental for successfully training and evaluating ML-based models in the field of AA. The creation of the dataset for this research involved a multi-step procedure to create a representative collection of literary texts authored by prominent Albanian writers. The goal was to develop a high-quality dataset for subsequent authorship analysis within Albanian literature. Our previous research, as detailed in [8, 25], investigated the utilization of newsroom columns for AA.

The dataset creation begins with an extensive review of the books in the University of Prizren library. This step was essential in identifying potential sources for the dataset, ensuring a diverse selection of texts. The criteria for selection included single-author

works and a diverse representation of literary genres. The selected literary works are then subjected to book scanning using a Viisan scanner, a specialized tool designed for digitizing physical texts. The CamBook app captures book pages precisely, ensuring the original content's preservation. Special attention was paid to copyright considerations in building this corpus, which led to the strategic scanning of book excerpts. The scanned pages are saved as image files. To enable text analysis, Optical Character Recognition (OCR) technology was employed to convert the scanned images into editable text content. Integrated Albanian text recognition software was employed for this purpose. Figure 1 visually represents the corpus creation process.

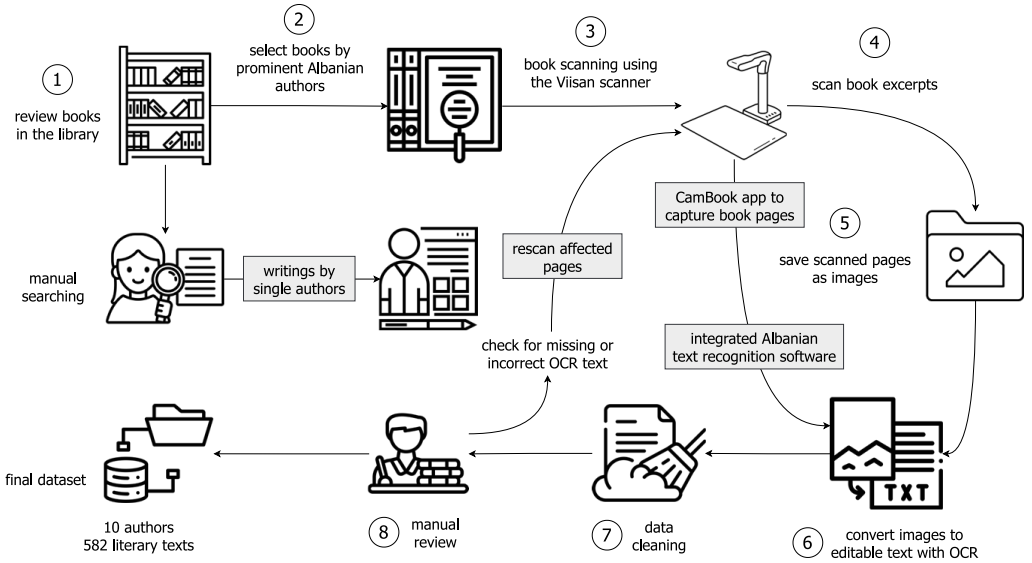

Figure 1. The process of creating AALitCorpus

A manual review was conducted to identify any instances of missing or incorrect OCR text. The affected pages were rescanned to ensure the dataset's completeness and accuracy. The process results in a final dataset of literary texts from ten Albanian authors. The proposed AAALitCorpus is a comprehensive and diverse collection of literary texts. Table 3 details the dataset's specifications and summary statistics.

Table 3. Statistical overview of the corpora

Feature	***AAALitCorpus***	***CCAT10***
total number of authors	*10*	*10*
total number of samples	*582*	*556*
average number of words	*1,444*	*1,099*
average number of unique words	*123*	*36*
average word length (in characters)	*5.3*	*5.8*
average sentence length (in words)	*15.9*	*26.1*
average number of words per author	*144*	*109*

The dataset's diversity, quality, and comprehensiveness are essential in ensuring the validity and reliability of subsequent analyses and experiments.

4. EXPERIMENTS

This section aims to assess the potential of various methods in the context of cross-linguistic AA through a series of experiments.

Data preparation. The dataset utilized for our experiments comprised a collection of Albanian literary texts and English newsroom columns. Importantly, we retained the original appearance of lowercase letters, punctuation, and special characters to preserve the unique variations in authors' handwriting styles.

Text representation models. In AA, a fundamental challenge lies in characterizing an author's style by extracting distinctive attributes from preprocessed data. Two prevalent models, TF-IDF and CountVectorizer, are employed for feature extraction in this context. The TF-IDF model captures the relative importance of terms and patterns in an author's writing. On the other hand, CountVectorizer provides insights into the frequency of specific words an author uses. Both models are experimented with, considering word- and character-level representations. We configured the n-gram range to encompass uni-, bi-, and tri-grams while limiting the maximum number of features to 1000. The aim is to identify the most suitable text modeling for the specific task of AA.

Classification methods. This subsection introduces the methods employed for authorship detection in our experiments, explicitly focusing on AA within a multiclass classification context. Multiclass classification deals with scenarios where each sample can be assigned to one of several distinct classes. We utilized a diverse array of ML-based classification algorithms. From an initial assessment of 30 ML-based models applied to different feature settings, we identified the top seven models for our experiments. The complete list of the employed algorithms in our investigation is as follows: eXtreme Gradient Boosting (XGB), Random Forest (RF), Logistic Regression (LR), Passive Aggressive Classifier (PAC), Linear Support Vector Classification (SVC), Multinomial Naïve Bayes (MNB), Multi-Layer Perceptron (MLP). Furthermore, we incorporated the fastText model into our study for comparative analysis. We employed default parameter values to provide a standardized assessment, allowing for a focused performance evaluation. The performance assessment of algorithms included the execution of a 5-fold cross-validation. We used the F1 score as an evaluation metric, considering its ability to effectively balance precision and recall performance. The subsequent section comprehensively presents and analyzes the experimental results.

5. RESULTS

This section reports our extensive range of experiments on the recently introduced dataset and CCAT10 corpus. We aim to delve into the efficacy of text representation models when analyzing authorship across textual varieties. For this purpose, we have employed various ML-based algorithms to uncover the underlying patterns in the data that enable insights into unique authorial characteristics. Table 4 presents the outcomes of the initial experiments conducted on the AAALitCorpus.

Table 4. Results of classification algorithms, text modeling techniques, and their configurations, with the best scores highlighted in bold (AAALitCorpus)

Features / Algorithms	*XGB*	*RF*	*LR*	*PAC*	*SVC*	*MNB*	*MLP*
character-level							
TF-IDF	*0.943*	***0.964***	*0.744*	*0.870*	*0.860*	*0.562*	*0.621*
CountVec	*0.913*	*0.908*	*0.923*	*0.795*	***0.936***	*0.881*	*0.862*
word-level							
TF-IDF	*0.976*	*0.979*	*0.976*	*0.997*	***0.998***	*0.991*	*0.997*
TF-IDF (SWR)	*0.935*	*0.970*	*0.993*	*0.997*	*0.995*	*0.997*	*0.997*
TF-IDF (CWR)	*0.938*	*0.966*	*0.965*	*0.993*	*0.995*	*0.988*	*0.995*
TF-IDF (SWR, CWR)	*0.887*	*0.926*	*0.993*	*0.991*	*0.995*	*0.988*	*0.995*
CountVec	*0.955*	*0.974*	*0.986*	*0.990*	*0.993*	*0.991*	*0.983*
CountVec (SWR)	*0.939*	*0.954*	*0.995*	*0.995*	***0.997***	*0.993*	*0.982*
CountVec (CWR)	*0.931*	*0.930*	*0.978*	*0.979*	*0.981*	*0.988*	*0.978*
CountVec (SWR, CWR)	*0.895*	*0.912*	*0.986*	*0.983*	*0.990*	*0.984*	*0.979*

SWR = Stop-Words Removal; **CWR** = Capitalized Words Removal.

We evaluate and compare the F1 scores of multiple classification algorithms across different configurations. Specifically, we explored the effectiveness of TF-IDF and CountVectorizer models in two settings: one with the inclusion of stopwords and the other without. Stopwords are common words often excluded from text analysis to focus on more meaningful content words. However, their existence can influence how the model perceives the text, especially in AA, where even function and common words can contribute to an author's distinctive voice.

Alternatively, capitalized terms, such as characters' names or important entities, can carry significant stylistic and contextual information. Therefore, we conducted experiments to assess how these text preprocessing variations impact our AA models' performance. This analysis allowed us to understand the role of text modeling techniques in capturing the distinctive features of authors' writing styles.

The most notable performance was attained through the SVC algorithm when using features extracted from the TF-IDF model. However, we encountered some intriguing scenarios in various configurations. In certain instances, excluding stopwords led to a decline in performance, while in others, it exhibited an improvement. Surprisingly, for some algorithms, the presence or absence of stopwords had minimal impact on the results. A similar pattern emerged when utilizing CountVectorizer. These observations underscore the relationship between modeling techniques, stopwords, and the specific AA algorithms. The ensemble algorithms, such as XGBoost and RandomForest, exhibited a remarkable consistency in their performance across various configurations.

In alignment with the findings presented in the previous table, Table 5 presents the experimental results applied to the newsroom columns dataset. We observe similar results in this scenario, especially when using the feature set at the word level.

The language differences and style variations between literary works and newsroom columns profoundly influence AA methods. TF-IDF excelled at the char level in the English dataset, capturing relevant information from columns. Albanian's rich vocabulary promotes diverse word choice and styles, aiding authorship pattern

identification in word-level. However, standardized English challenges algorithms relying solely on word-level features.

Table 5. Results of classification algorithms, text modeling techniques, and their configurations, with the best scores highlighted in bold (CCAT10)

Features / Algorithms	***XGB***	***RF***	***LR***	***PAC***	***SVC***	***MNB***	***MLP***
character-level							
TF-IDF	*0.910*	***0.931***	*0.837*	*0.892*	*0.894*	*0.918*	*0.942*
CountVec	*0.788*	***0.813***	*0.812*	*0.517*	*0.801*	*0.809*	*0.696*
word-level							
TF-IDF	*0.917*	*0.953*	*0.961*	*0.977*	*0.975*	*0.966*	*0.971*
TF-IDF (SWR)	*0.921*	*0.944*	*0.957*	*0.976*	*0.973*	*0.966*	***0.982***
TF-IDF (CWR)	*0.892*	*0.927*	*0.939*	*0.975*	*0.966*	*0.964*	*0.980*
TF-IDF (SWR, CWR)	*0.894*	*0.937*	*0.946*	*0.953*	*0.955*	*0.946*	*0.962*
CountVec	*0.941*	*0.946*	*0.977*	*0.978*	***0.982***	*0.966*	*0.959*
CountVec (SWR)	*0.908*	*0.964*	*0.975*	*0.980*	*0.977*	*0.968*	*0.964*
CountVec (CWR)	*0.903*	*0.942*	*0.964*	*0.953*	*0.971*	*0.955*	*0.948*
CountVec (SWR, CWR)	*0.892*	*0.925*	*0.957*	*0.950*	*0.955*	*0.953*	*0.939*

SWR = Stop-Words Removal; **CWR** = Capitalized Words Removal.

We used the fastText model to conduct a comparative analysis. The experimental results are outlined in Table 6. The F1 scores obtained with fastText were slightly lower than those achieved with model-generated features in different configurations.

Table 6. Results of fastText across diverse settings, with the best scores highlighted in bold

Level	***char-level***	***word-level***			
Corpus	*fastText*	*fastText*	*SWR*	*CWR*	*SWR and CWR*
AAALitCorpus	***0.974***	*0.940*	*0.923*	*0.957*	*0.932*
CCAT10	*0.920*	***0.973***	***0.973***	*0.938*	*0.946*

Given that the SVC algorithm consistently achieved the highest F1 scores across various configurations, we utilized it as the primary algorithm for the subsequent experiments. Regarding text modeling, the best results were obtained when employing TFIDF with word n-grams. To assess the effectiveness of our approach, we present a comparative analysis in Table 7, which compares our method's results using the CCAT10 corpus with those of previous works in the field of AA.

Table 7. Comparison of our approach with previous works in the CCAT10 corpus

Reference	***Features***	***Method***	***Performance***
[6]	*lexical, syntactic, semantic, content-specific features, stop-words, n-grams*	*Logistic Regression*	*0.929*
[12]	*lexical and syntactic features*	*Bi-LSTM*	*0.924*
[13]	*character n-grams*	*under-sampling*	*0.794*
Our method	*word n-grams*	*Linear SVC*	*0.982*

Learning curves. We present learning curves (Figure 2 and Figure 3), illustrating the correlation between the training and validation test scores.

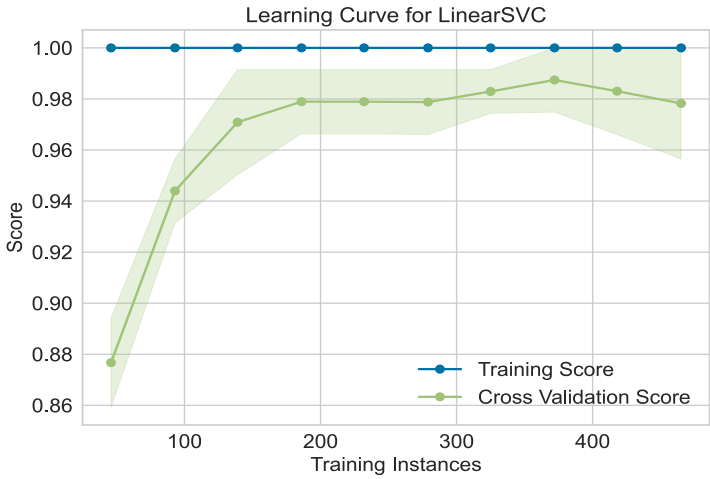

Figure 2. Learning curve for SVC on the TF-IDF model (AAALitCorpus)

These curves offer valuable insights into the performance of SVC as training data sizes vary.

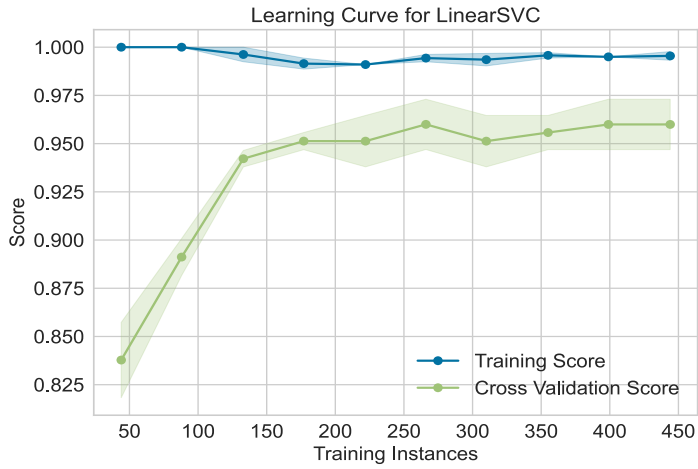

Figure 3. Learning curve for SVC on the TF-IDF model (CCAT10)

During the early training stages, the training score consistently outperformed the validation score, suggesting that the model demonstrated a strong ability to learn from the training data. As the training progresses, the validation score begins to show a gradual improvement. This phenomenon indicates the model's ability to generalize better to new data as it acquires additional knowledge from the training process.

Classification report. In our model's performance analysis, Figures 4 and 5 illustrate the classification reports for the SVC algorithm in the AAALitCorpus and CCAT10 datasets, respectively. This report provides a detailed overview of the algorithm's precision, recall, F1 score, and support for each class.

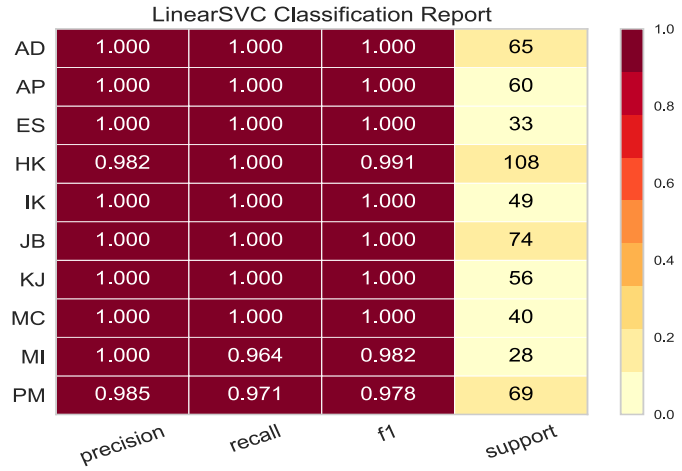

Figure 4. Classification report for model performance (AAALitCorpus).

The classification report visually represents the model's performance on a class-by-class basis, offering crucial insights for further model refinement.

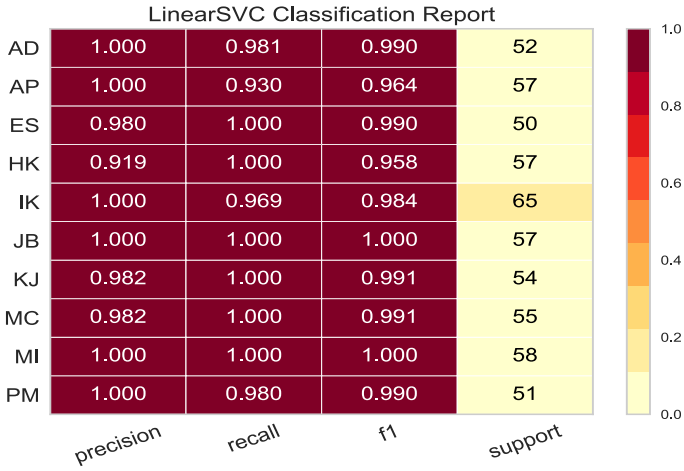

Figure 5. Classification report for model performance (CCAT10).

The classification report visually represents the model's performance on a class-by-class basis, offering crucial insights for further model refinement.

6. CONCLUSION

This study explores cross-linguistic authorship classification using corpora from two languages. We aimed to unravel hidden patterns in AA through ML-based algorithms. Our experiments compared features and linguistic contexts, with no significant differences observed. Seven classification algorithms are thoroughly evaluated, with SVC consistently outperforming using TF-IDF features at the word level. Notably, fastText demonstrated impressive performance close to ML-based algorithms. Our findings underscore the potential of algorithms in accurately attributing the authors behind written content, shedding light on the nuances of authorship classification in diverse textual sources. This research represents a novel effort in authorship classification, showcasing the depth achieved through extensive experimentation in Albanian and English. It's important to note that the methodology and findings of these experiments can also be extended to other languages. By offering insights into previously unexplored linguistic contexts, our study indicates the potential for advancing AA research on a more general dimension. While this study presents promising results, several avenues remain for further investigation. We intend to explore authorship attribution encompassing an expanded dataset, incorporating texts from diverse linguistic contexts and domains. Additionally, we aspire to explore additional features for authorship tasks, further enhancing our understanding of the characteristics that define authorial voices.

REFERENCES

[1] Romansky, R. Digital age and personal data protection. *International Journal on Information Technologies and Security*, vol.14, no.3, 2022, pp. 89-100.

[2] Hossain, M.R., Hoque, M.M., Dewan, M.A.A., Siddique, N., Islam, M.N. and Sarker, I.H. Authorship classification in a resource constraint language using convolutional neural networks. *IEEE Access*, vol.9, 2021, pp. 100319-100338. DOI: 10.1109/ACCESS.2021.3095967.

[3] Hriez, S. and Awajan, A. Authorship Identification for Arabic texts using logistic model tree classification. *Intelligent Computing: Proceedings of the 2020 Computing Conference*, 2021, pp. 656–666. DOI: 10.1007/978-3-030-52246-9_48.

[4] Paci, H., Kajo, E., Trandafili, E., Tafa, I., and Salillari, D. Author identification in Albanian language. *2011 14th International Conference on Network-Based Information Systems*, Tirana, Albania, September 2011, pp. 425–430. DOI: 10.1109/nbis.2011.71.

[5] Nazir, Z., Shahzad, K., Malik, M.K., Anwar, W., Bajwa, I.S. and Mehmood, K. Authorship Attribution for a Resource Poor Language—Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol.21, no.3, 2021, pp. 1–23. DOI: 10.1145/3487061.

[6] Wu, H., Zhang, Z. and Wu, Q. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*, vol.111, 2021, pp. 1–11. DOI: 10.1016/j.asoc.2021.107815.

[7] Ramezani, R. A language-independent authorship attribution approach for author identification of text documents. *Expert Systems with Applications*, vol.180, 2021, pp. 1–15. DOI: 10.1016/j.eswa.2021.115139.

[8] Misini, A., Kadriu, A. and Canhasi, E. Albanian Authorship Attribution Model. *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, Budva Montenegro, June 2023, pp. 1–5. DOI: 10.1109/MECO58584.2023.10155046.

[9] Alonso-Fernandez, F., Belvisi, N.M.S., Hernandez-Diaz, K., Muhammad, N. and Bigun, J. Writer identification using microblogging texts for social media forensics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol.3, no.3, 2021, pp. 405-426. DOI: 10.1109/tbiom.2021.3078073.

[10] Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I. and Demidov, P.G. A survey on stylometric text features. *2019 25th Conference of Open Innovations Association (FRUCT)*, Helsinki, Finland, November 2019, pp. 184-195. DOI: 10.23919/FRUCT48121.2019.8981504.

[11] Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, vol.60, no.3, 2009, pp. 538-556. DOI: 10.1002/asi.21001.

[12] Jafariakinabad, F. and Hua, K.A. A self-supervised representation learning of sentence structure for authorship attribution. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol.16, no.4, 2022, pp. 1-16. DOI: 10.1145/3491203.

[13] Stamatatos, E. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, vol.44, no.2, 2008, pp. 790-799. DOI: 10.1016/j.ipm.2007.05.012.

[14] Januzaj, Y., Luma, A., Aliu, A., Selimi, B. and Raufi, B. Web data scraping technique and preparation for comparison techniques between different documents. *International Journal on Information Technologies and Security*, vol.11, no.2, 2019, pp. 71-86.

[15] Ramnial, H., Panchoo, S. and Pudaruth, S. Authorship attribution using stylometry and machine learning techniques. *Intelligent Systems Technologies and Applications*, vol.384, 2016, DOI: 10.1007/978-3-319-23036-8_10.

[16] Hossain, A.S., Akter, N. and Islam, M.S. A stylometric approach for author attribution system using neural network and machine learning classifiers. *Proceedings of the International Conference on Computing Advancements*, Dhaka Bangladesh, January 2020, pp. 1-7, DOI: 10.1145/3377049.3377079.

[17] Kasabov, N. From multilayer perceptrons and neurofuzzy systems to deep learning machines: which method to use?-a survey. *International Journal on Information Technologies and Security*, vol.9, no.2, 2017, pp. 3-24.

[18] Misini, A., Kadriu, A. and Canhasi, E. A Survey on Authorship Analysis Tasks and Techniques. *SEEU Review*, vol.17, no.2, 2022, pp. 153-167, DOI: 10.2478/seeur-2022-0100.

[19] Abuhammad, Y., Addabe, Y., Ayyad, N. and Yahya, A. Authorship attribution of modern standard Arabic short texts. *The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research*, Sharjah United Arab Emirates, August 2021, pp. 1-6, DOI: 10.1145/3485557.3485563.

[20] Singh, P.K., Vivek, K.S. and Kodimala, S. Stylometric analysis of E-mail content for author identification. *IML '17: Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, Liverpool United Kingdom, October 2017, pp. 1-8, DOI: 10.1145/3109761.3109770.

[21] Romanov, A., Kurtukova, A., Shelupanov, A., Fedotova, A. and Goncharov, V. Authorship identification of a russian-language text using support vector machine and deep neural networks. *Future Internet*, vol.13, no.1, 2020, pp. 1-16, DOI: 10.3390/fi13010003.

[22] Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D. and Chanona-Hernández, L. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, vol.21, no.3, 2017, pp. 627-639, DOI: 10.1007/s00500-016-2446-x.

[23] Canhasi, E., Shijaku, R. and Berisha, E. Albanian fake news detection. *Transactions on Asian and Low-Resource Language Information Processing*, vol.21, no.5, 2022, pp. 1-24, DOI: 10.1145/3487288.

[24] Kadriu, A. and Abazi, L. A comparison of algorithms for text classification of Albanian news articles. *ENTRENOVA-ENTerprise REsearch InNOVAtion*, vol.3, no.1, 2017, pp. 62-68.

[25] Misini, A., Kadriu, A. and Canhasi, E. A3C: Albanian Authorship Attribution Corpus. In *International Scientific Conference on Business and Economics* (pp. 755-763). ISCBE 2023. Springer Proceedings in Business and Economics. Cham: Springer Nature Switzerland. DOI: 10.1007/978-3-031-42511-0_49.

Information about the authors:

Arta Misini is a teaching assistant at the Faculty of Computer Science, University of Prizren. She received a BSc in Computer Science in 2016 and an MSc in Computer Science and Technologies of Communication in 2018, both from the University of Prizren. She is currently a Ph.D. student in Computer Science at SEE University, North Macedonia.

Arbana Kadriu, Ph.D. in Computer Sciences from Ss. Cyril and Methodius University, Skopje (2008), specialized in NLP and information retrieval. As an full professor at the Faculty of Contemporary Sciences and Technologies, SEE University, North Macedonia, she covers various areas, including AI, ML, programming, software engineering, and e-learning. She has supervised numerous master's theses and authored over 50 research papers.

Ercan Canhasi, an associate professor at the University of Prizren's Faculty of Computer Science and CTO/CSO of Gjirafa, Inc. since 2014. He received his BSc and MSc degrees from Selcuk University. He completed his Ph.D. in Computer Science at the University of Ljubljana in 2013. Ercan has also authored several papers published in SCI-indexed journals and conferences.

Manuscript received on 17 December 2023