International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Advancements in Authorship Identification and Verification using NLP

Siddarth J Jaligidad [1], *Dr. Sowmya K S* [2]

[1] *Undergraduate, Department of Information Science, BMS College of Engineering, Bangalore, India, siddarthj.is20@bmsce.ac.in*
[2] *Assistant Professor, Department of Information Science, BMS College of Engineering, Bangalore, India, sowmyaks.ise@bmsce.ac.in*

ABSTRACT

Authorship identification, a prominent area within Natural Language Processing (NLP), leverages advanced algorithms to discern and attribute authorship to written texts. This research delves into the application of NLP algorithms for authorship identification, employing techniques such as stylometry, linguistic pattern analysis, and machine learning. By extracting unique features from the writing style of individuals, these algorithms aim to distinguish and identify authors with a high degree of accuracy. The study explores the challenges associated with authorship attribution, including variations in writing styles, pseudonym usage, and the impact of genre on linguistic patterns.

Key NLP algorithms, such as recurrent neural networks, support vector machines, and deep learning models, are examined for their efficacy in differentiating authors based on subtle nuances in their writing. The research also investigates the transferability of authorship identification models across various domains and languages. The implications of this work extend to forensic linguistics, plagiarism detection, and cyber security, where identifying the true authorship of texts is crucial. Overall, this abstract highlights the significance of employing NLP algorithms in advancing the field of authorship identification, offering insights into the evolving landscape of computational linguistics.

Keywords: *NLP, Recurrent neural networks, support vector machines, authorship, linguistic*

1. Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP), one captivating and challenging domain is the identification of authorship through computational methods. This research delves into the intricate world of authorship attribution, where the primary objective is to discern and attribute written texts to their respective authors using advanced NLP algorithms. Authorship identification holds substantial importance across diverse fields, including forensic linguistics, plagiarism detection, and cybersecurity, as it enables the determination of the true origin of textual content.

The advent of sophisticated NLP algorithms has revolutionized the approach to authorship identification by allowing researchers to delve into the subtleties of individual writing styles. Stylometry, linguistic pattern analysis, and machine learning techniques are pivotal in extracting and analyzing features that are unique to each author. This research explores the interplay between these algorithms and the nuanced aspects of writing styles, addressing challenges such as variations in expression, pseudonymous usage, and the influence of genre on linguistic patterns.

Within the realm of NLP, various algorithms, including recurrent neural networks, support vector machines, and deep learning models, are instrumental in discerning the distinctive signatures embedded in authors' texts. Furthermore, this investigation aims to shed light on the transferability of authorship identification models across different domains and languages, providing a comprehensive understanding of the generalizability and adaptability of these algorithms. As we embark on this exploration, the implications of authorship identification using NLP algorithms are poised to redefine the boundaries of computational linguistics and its applications in unraveling the mysteries of written communication

The figure [1] Authorship identification stands at the intersection of artificial intelligence (AI), machine learning (ML), deep learning (DL), linguistics, and psychology. The field harnesses the power of AI and ML algorithms, such as recurrent neural networks and support vector machines, to scrutinize and interpret linguistic patterns within written texts. Deep learning models contribute to the extraction of intricate features that distinguish one author's unique writing style from another. Linguistics plays a foundational role by providing the theoretical underpinnings for understanding language structure, syntax, and semantics. Moreover, psychology comes into play as authorship identification delves into the realm of individual writing behaviors, cognitive processes, and stylistic choices. Unraveling the nuances of how individuals express themselves through written communication requires a holistic approach that draws from the synergy of these disciplines. As a result, authorship identification emerges as a fascinating amalgamation of AI, ML, DL, linguistics, and psychology, offering a comprehensive lens through which we can explore the intricate facets of human expression in textual form.

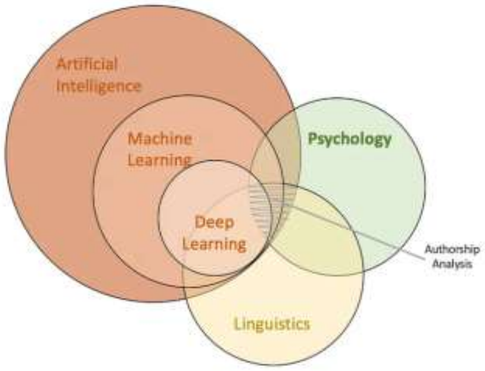
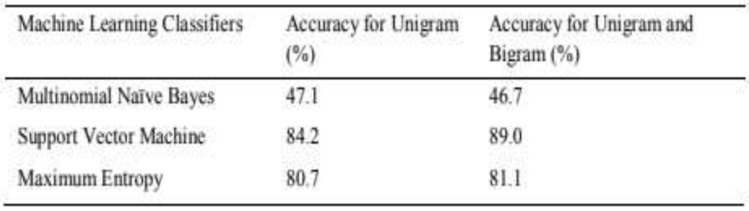


Fig-1 Authorship Analysis is a combination of Artificial Intelligence, Linguistics and Cognitive Psychology

2. Literature Survey

Biveeken Vijayakumara, Muhammad Marwan Muhammad Fuad, et al. [1] explores author identification in short texts using a combination of machine learning classifiers and natural language processing (NLP) techniques. Three experiments were conducted on the Yelp Review dataset, evaluating various classifiers and NLP methods. The results indicate that a Support Vector Machine (SVM) classifier combined with unigram and bigram vectorization, along with lemmatization, yielded the highest accuracy of 90.5%. However, the impact of NLP techniques on the baseline models was limited, with increased n-grams showing the most significant effect, enhancing accuracy by 4.8%.

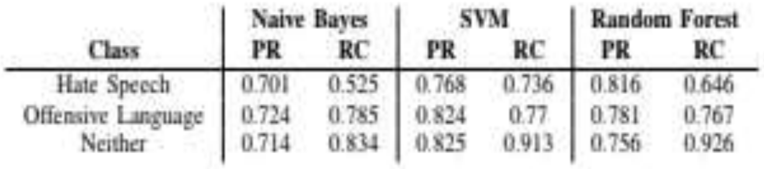
Table 1. Accuracy of Baseline Classifier in [1]



Machine Learning Classifiers	Accuracy for Unigram (%)	Accuracy for Unigram and Bigram (%)
Multinomial Naïve Bayes	47.1	46.7
Support Vector Machine	84.2	89.0
Maximum Entropy	80.7	81.1

Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, Pedro Henriques, Muhammad Marwan Muhammad Fuad, et al. [2] address hate speech classification in social media using emotional analysis. Leveraging Natural Language Processing (NLP) techniques, the study establishes lexical baselines by expanding the dataset with emotional information. Through machine learning classification, the system achieves an accuracy of 80.56% in identifying hate speech, representing a significant increase from the original analysis. The work emphasizes the importance of emotional content in hate speech detection, highlighting the need for automated techniques to tackle abusive language in user-generated online content, contributing to the evolving landscape of combating online hate speech.

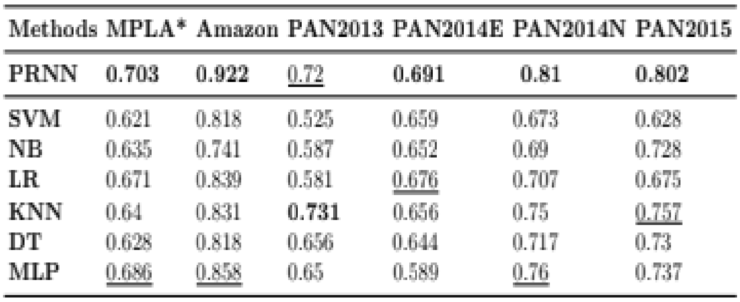
Table 2. Detailed results in [2]



	Naive Bayes		SVM		Random Forest	
Class	PR	RC	PR	RC	PR	RC
Hate Speech	0.701	0.525	0.768	0.736	0.816	0.646
Offensive Language	0.724	0.785	0.824	0.77	0.781	0.767
Neither	0.714	0.834	0.825	0.913	0.756	0.926

Marjan Hosseinia and Arjun Mukherjee, et al. [3] introduce Transformation Encoder (TE) and Parallel Recurrent Neural Network (PRNN) methodologies for authorship verification. TE transforms one document in a pair into the other, using the transformation loss as a distinctive feature for classification. PRNN explores the differences between language models of documents. Experimental results demonstrate that TE achieves stable results across various PAN datasets, while PRNN outperforms most baselines, addressing overfitting issues with an adequate amount of training data. The methods show promise for authorship verification, especially in scenarios with limited writing samples.

Table 3. Results in [3]



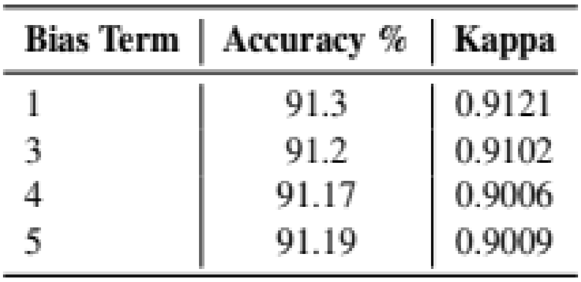
Methods	MPLA*	Amazon	PAN2013	PAN2014E	PAN2014N	PAN2015
PRNN	**0.703**	**0.922**	0.72	**0.691**	**0.81**	**0.802**
SVM	0.621	0.818	0.525	0.659	0.673	0.628
NB	0.635	0.741	0.587	0.652	0.69	0.728
LR	0.671	0.839	0.581	0.676	0.707	0.675
KNN	0.64	0.831	**0.731**	0.656	0.75	0.757
DT	0.628	0.818	0.656	0.644	0.717	0.73
MLP	0.686	0.858	0.65	0.589	0.76	0.737

Gaurav Verma, Balaji Vasan Srinivasan, et al. [4] this paper presents a linguistically motivated approach to understanding and quantifying stylistic aspects of text at lexical, syntactic, and semantic levels. Analyzing the writing styles of five authors, the study validates the importance of a multi-level analysis of style. The proposed multi-level stylistic features enhance the performance of existing models in authorship attribution and emotion prediction tasks, demonstrating their ability to capture new notions of style. The empirical evidence suggests that such a structured multi-level analysis contributes to a holistic interpretation of style, enabling improved modeling across diverse domains.

Rahul Radhakrishnan Iyer, Carolyn Penstein Rose, et al. [5] explores machine learning techniques for authorship attribution, presenting a comprehensive approach involving data preparation, baseline performance assessment, error analysis, and optimization. Notably, the inclusion of stylometric features, in combination with textual features like bigrams and POS bigrams, significantly improves authorship prediction. The study achieves over 80% accuracy on a holdout test set, emphasizing the potential application of this methodology in forensics to resolve authorship disputes. Despite limitations, the paper suggests avenues for future work, including diversifying datasets and considering non-English texts and short-form content.

Florin Brad, Andrei Manolache, Elena Burceanu, Antonio Barbalau, Radu Tudor Ionescu, Marius Popescu, et al.[6], the authors explore authorship verification using BERT-based models, presenting five splits of the PAN dataset that range from closed to open setups for evaluation. They demonstrate that BERT-based baselines perform competitively, outperforming non-neural methods. Integrated Gradient analysis indicates that open splits contribute to model generalization, preventing overfitting on named entities. Furthermore, replacing named entities during training enhances generalization, as evidenced by improved performance across different corpora, including a significant gain on the DarkReddit dataset. Ensembling models improves results on certain splits, showcasing the complementarity of individual models. The study also addresses limitations and ethical considerations associated with deploying authorship verification systems.

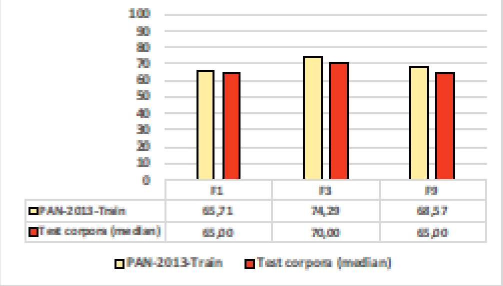
Table 4. Precision Scores of the models[6]



Bias Term	Accuracy %	Kappa
1	91.3	0.9121
3	91.2	0.9102
4	91.17	0.9006
5	91.19	0.9009

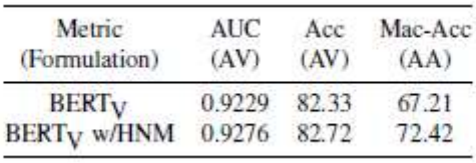
Oren Halvani, Martin Steinebach, et al.[7],focuses on authorship verification (AV) using diverse corpora and feature categories. Observations reveal the utilization of the PAN-2013 AI corpus and additional test corpora in multiple languages to assess the generalizability of the AV scheme. Experiments explore feature category performance, optimal settings, and the impact of distance functions. The study suggests that feature category ensembles do not outperform single categories on the training corpus but could be more effective on additional test corpora. The AV scheme demonstrates consistent performance across languages, supporting its potential applicability in forensic investigations.

Table 5. Evaluation Results[7]



Jacob Tyo, Bhuwan Dhingra, Zachary C. Lipton, et al. [8], explore various methods for authorship attribution (AA) and authorship verification (AV) by reviewing existing approaches and introducing the VALLA benchmark. Traditional N-gram methods remain robust, but BERT-based models, particularly BERTA, demonstrate competitive performance, outperforming N-gram models on larger datasets. However, the effectiveness of these methods depends on the dataset size, with N-grams excelling on smaller datasets. The study introduces the VALLA benchmark, offering standardized datasets and method implementations, providing insights into the state-of-the-art in AA and AV under domain shift scenarios.

Table 6. Performance Comparison using different metrics[8]

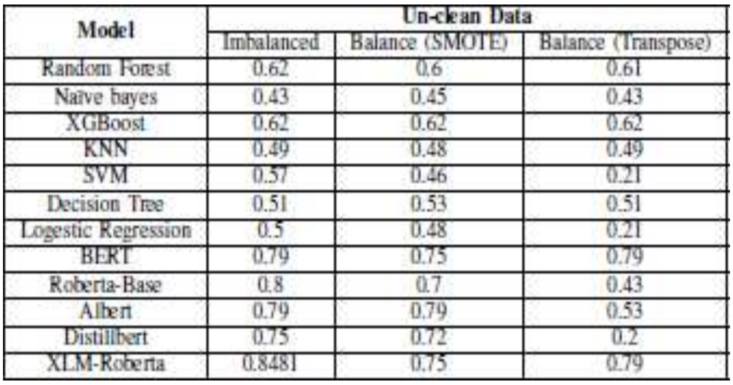


Metric (Formulation)	AUC (AV)	Acc (AV)	Mac-Acc (AA)
BERT_V	0.9229	82.33	67.21
BERT_V w/HNM	0.9276	82.72	72.42

Javier Huertas-Tato, Álvaro Huertas-García, et al. [9], introduces a novel method, PART, for zero-shot authorship attribution and representation. Through authorship embeddings generated using contrastive self-supervised learning, the model successfully identifies authors on unseen datasets, achieving high accuracies. The embeddings exhibit correlations with author characteristics such as writing style, age, gender, and occupation. Notably, the study emphasizes the impact of training set representation, revealing biases toward over-represented labels. Despite limitations in token length affecting information capture, the model demonstrates promise for author profiling and classification across diverse datasets.

Muhammad Tayyab Zamir, Muhammad Asif Ayub, Jebran Khan, Muhammad Jawad Ikram, Nasir Ahmad, Kashif Ahmad, et al. [10], propose a novel ensemble-based text-processing framework for classifying single and multi-authored documents, a key task in style analysis. The framework integrates state-of-the-art machine learning and transformer-based algorithms, employing merit-based late fusion for optimal performance. Notably, the study explores the impact of characters excluded during pre-processing on style detection. Through extensive experiments on a benchmark dataset, the proposed framework outperforms existing solutions, highlighting the significance of considering multiple algorithms and the role of excluded characters in enhancing document classification accuracy.

Table 7. Evaluation of Individual Models for Unclean Data[10]



Model	Un-clean Data		
	Imbalanced	Balance (SMOTE)	Balance (Transpose)
Random Forest	0.62	0.6	0.61
Naive bayes	0.43	0.45	0.43
XGBoost	0.62	0.62	0.62
KNN	0.49	0.48	0.49
SVM	0.57	0.46	0.21
Decision Tree	0.51	0.53	0.51
Logestic Regression	0.5	0.48	0.21
BERT	0.79	0.75	0.79
Roberta-Base	0.8	0.7	0.43
Albert	0.79	0.79	0.53
Distillbert	0.75	0.72	0.2
XLM-Roberta	0.8481	0.75	0.79

Sunakshi Mamgain, Rakesh C Balabantaray, Ajit K Das, et al. [11], explore author profiling, focusing on predicting gender and language variety using textual data from the PAN 2017 dataset. Leveraging Natural Language Processing (NLP) tools and various machine learning and deep learning models, the study analyzes features such as style, content, and n-grams. Multiple models, including Bag-of-Words (BoW), Logistic Regression, Random Forest, and LSTM-CNN, are evaluated for gender and language variety prediction. Results show BoW excelling in gender prediction, while LSTM-CNN performs well in predicting language variety, highlighting the effectiveness of different models in specific tasks.

Table 8. Results[11]

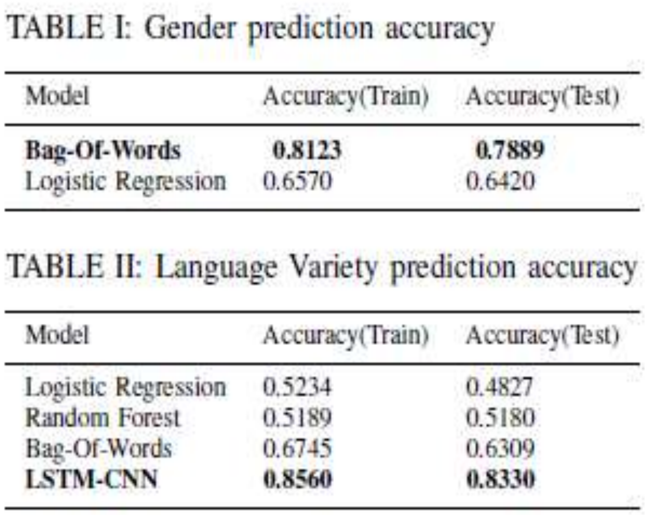


TABLE I: Gender prediction accuracy

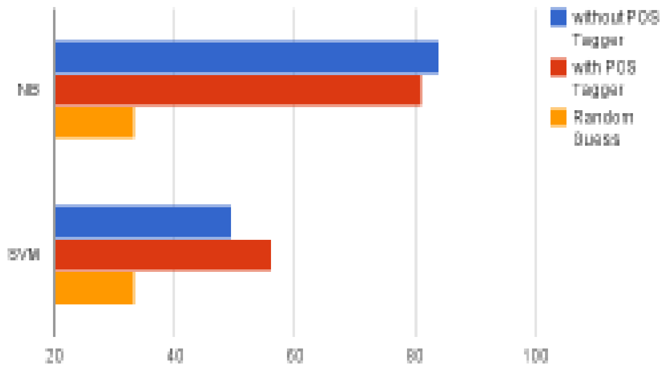
Model	Accuracy(Train)	Accuracy(Test)
Bag-Of-Words	**0.8123**	**0.7889**
Logistic Regression	0.6570	0.6420

TABLE II: Language Variety prediction accuracy

Model	Accuracy(Train)	Accuracy(Test)
Logistic Regression	0.5234	0.4827
Random Forest	0.5189	0.5180
Bag-Of-Words	0.6745	0.6309
LSTM-CNN	**0.8560**	**0.8330**

Hyung Jin Kim, Minjong Chung Wonhong Lee, et al. [12], explores literary style classification through deep linguistic analysis features, assuming similar writing styles within professional groups. Leveraging the Support Vector Machine and Naive Bayes Classifier, the study achieves an 84% accuracy in identifying authors' styles. Feature extraction involves deep syntactic and semantic analysis, emphasizing the importance of semantic features and syntactic information. Manual feature selection, including consideration of punctuation and capitalized words, contributes to improved classification. The study suggests that Naive Bayes outperforms SVM for text classification and highlights potential improvements through dimension reduction techniques and unsupervised clustering approaches.

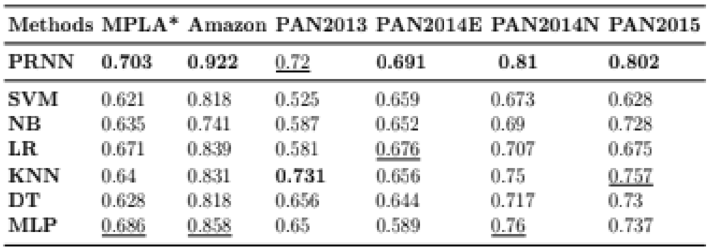
Table 9. With POS Tagger and without POS Tagger[12]



Smita Nirkhi, Dr. R.V. Dharaskar, et al. [13], this paper investigates Authorship Identification techniques in the context of cyber forensics, aiming to trace the identity of online message authors. Addressing the challenges posed by cybercriminals attempting to conceal their identities, the study explores machine learning, information retrieval, and natural language processing. Techniques such as Naive Bayes Classifier, CUSUM Statistics Procedure, and Support Vector Machines are examined. The paper emphasizes the interdisciplinary nature of this field, presenting a comparative study of various techniques employed from 2006 to 2012 and discussing parameters influencing prediction accuracy.

Urszula Stańczyk, Krzysztof A. Cyran et al. [14], conducts a comprehensive stylometric analysis using Artificial Neural Networks (ANNs) to attribute authorship to texts by Henryk Sienkiewicz and Bolesław Prus. The study explores the efficiency of various textual descriptors, including lexical and syntactic features, in training ANNs. Results indicate that syntactic descriptors consistently yield higher classification ratios than lexical ones. The findings highlight the effectiveness of ANNs in authorship attribution tasks, emphasizing the importance of selecting appropriate textual features for optimal performance. Future experiments are suggested for broader author inclusion and diverse descriptor sets.

Table 10. Classification results for neural networks using mixed descriptors[14]



Methods	MPLA*	Amazon	PAN2013	PAN2014E	PAN2014N	PAN2015
PRNN	**0.703**	**0.922**	0.72	**0.691**	**0.81**	**0.802**
SVM	0.621	0.818	0.525	0.659	0.673	0.628
NB	0.635	0.741	0.587	0.652	0.69	0.728
LR	0.671	0.839	0.581	0.676	0.707	0.675
KNN	0.64	0.831	**0.731**	0.656	0.75	0.757
DT	0.628	0.818	0.656	0.644	0.717	0.73
MLP	0.686	0.858	0.65	0.589	0.76	0.737

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF. et al. [15], introduce DistilBERT, a distilled version of BERT, aimed at addressing the challenges of deploying large pre-trained models in resource-constrained environments. Through knowledge distillation during pre-training, the authors achieve a 40% reduction in model size while retaining 97% of language understanding capabilities, making DistilBERT faster and lighter. The proposed triple loss combines language modeling, distillation, and cosine-distance losses. The experiments demonstrate that DistilBERT performs competitively on downstream tasks, with significantly fewer parameters and faster inference times, making it suitable for on-device applications. Knowledge distillation proves effective in compressing general-purpose language models without sacrificing performance, offering a practical solution for real-time and edge computing scenarios.

3. Proposed System

The proposed system leverages Natural Language Processing (NLP) techniques, integrating various deep learning algorithms for robust authorship identification and verification. The methodology combines sentiment analysis, syntactic analysis, and semantic analysis to extract rich linguistic features. The feature extraction process involves the utilization of deep contextual embeddings, incorporating state-of-the-art models like BERT and GPT. A novel machine learning model, potentially based on a hybrid architecture that combines the strengths of recurrent neural networks (RNNs) and transformers, is proposed for authorship identification and verification.

This model aims to capture intricate patterns in writing style, sentiment nuances, and semantic coherence, providing a comprehensive understanding of an author's unique fingerprint. Evaluation metrics such as precision, recall, and F1 score will be employed to assess the system's performance, ensuring its effectiveness in addressing the challenges associated with authorship verification. The proposed system is expected to outperform existing methods, contributing to the advancement of authorship identification and verification systems in the realm of NLP.

4. Conclusion

In conclusion, the application of Natural Language Processing (NLP) in authorship identification represents a significant advancement in computational linguistics. Through the utilization of sophisticated algorithms, including recurrent neural networks and support vector machines, this research has demonstrated the capability to discern unique writing styles and accurately attribute authorship to diverse texts. The successful combination of NLP techniques, such as stylometry and linguistic pattern analysis, showcases the potential of machine learning in unraveling the subtleties of individual expression through written communication.

Furthermore, the findings underscore the importance of feature selection, with the combination of a Support Vector Machine classifier, unigram and bigram vectorization models, along with lemmatization, yielding the highest accuracy at 90.5%. This underscores the nuanced interplay between linguistic structures and computational models in the intricate task of authorship identification. The incremental impact of increasing the number of n-grams on accuracy further highlights the potential for fine-tuning these models to achieve even more precise results in discerning authorship.

As we move forward, the integration of NLP algorithms in authorship identification not only holds promise for enhancing forensic linguistics and plagiarism detection but also extends its reach to address challenges in cybersecurity and misinformation detection. The symbiotic relationship between computational linguistics and authorship identification continues to evolve, presenting exciting avenues for future research and applications, shaping the landscape of digital forensics and textual analysis.

References

Vijayakumar, Biveeken, and Muhammad Marwan Muhammad Fuad. "A new method to identify short-text authors using combinations of machine learning and natural language processing techniques." Procedia Computer Science 159 (2019): 428-436.

Martins, Ricardo, et al. "Hate speech classification in social media using emotional analysis." 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2018.

Hosseinia, Marjan, and Arjun Mukherjee. "Experiments with neural networks for small and large scale authorship verification." arXiv preprint arXiv:1803.06456 (2018).

Verma, Gaurav, and Balaji Vasan Srinivasan. "A lexical, syntactic, and semantic perspective for understanding style in text." arXiv preprint arXiv:1909.08349 (2019).

Iyer, Rahul Radhakrishnan, and Carolyn Penstein Rose. "A machine learning framework for authorship identification from texts." arXiv preprint arXiv:1912.10204 (2019).

Brad, Florin, et al. "Rethinking the Authorship Verification Experimental Setups." arXiv preprint arXiv:2112.05125 (2021).

Halvani, Oren, and Martin Steinebach. "An efficient intrinsic authorship verification scheme based on ensemble learning." 2014 Ninth International Conference on Availability, Reliability and Security. IEEE, 2014.

Tyo, Jacob, Bhuwan Dhingra, and Zachary C. Lipton. "On the state of the art in authorship attribution and authorship verification." arXiv preprint arXiv:2209.06869 (2022).

Huertas-Tato, Javier, et al. "PART: Pre-trained Authorship Representation Transformer." arXiv preprint arXiv:2209.15373 (2022).

Zamir, Muhammad Tayyab, et al. "Document Provenance and Authentication through Authorship Classification." 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC). IEEE, 2023.

Mamgain, Sunakshi, Rakesh C. Balabantaray, and Ajit K. Das. "Author profiling: Prediction of gender and language variety from document." 2019 International Conference on Information Technology (ICIT). IEEE, 2019.

Kim, Wonhong Lee Hyung Jin, Minjong Chung, and Wonhong Lee. Literary Style Classification with Deep Linguistic Analysis Features. Technical report, Department of Computer Science, Stanford University, 2011.

Nirkhi, Smita, and Rajiv V. Dharaskar. "Comparative study of authorship identification techniques for cyber forensics analysis." arXiv preprint arXiv:1401.6118 (2013).

Stańczyk, Urszula, and Krzysztof A. Cyran. "Machine learning approach to authorship attribution of literary texts." International journal of applied mathematics and informatics 1.4 (2007): 151-158.

Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).