IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

Author Identification of IEEE Using ML

Prof. R. P. Kumawat[1], Prof. V. D. Vaidya[2], Sarode Kumar Ganesh[3], Varpe Gagan Nilesh[4], Jagtap Krushna Sanjay[5], Bari Dhiraj Bhaskar[6]

[1,2,3,4,5,6]Department of Cloud Computing and Big Data

Padmashri Dr. Vitthalrao Vikhe Patil Institute of Technology and Engineering (Polytechnic), Pravaranagar

Abstract: *Author identification, a captivating area of exploration within natural language processing, seeks to distinguish the individual writing styles of different authors. This project focuses specifically on English literature, a multifaceted and culturally significant literary tradition. The main aim is to develop a robust machine learning model capable of accurately assigning English texts to their respective authors by analyzing their unique writing patterns. Beginning with gathering and preprocessing a comprehensive dataset of English literary texts spanning various authors, genres, and historical periods, the project ensures data quality and consistency through text cleaning and tokenization. The critical phase involves selecting the most suitable machine learning algorithm for author identification. Techniques like Naive Bayes, Support Vector Machines (SVM), Random Forest, and transformer-based models are assessed for their effectiveness.*

Keywords: Author identification, SVM, machine learning, English literature, text analysis

I. INTRODUCTION

Author Identification of English Literature using Learning Techniques stands at the forefront of natural language processing and machine learning, captivating researchers and enthusiasts alike. This interdisciplinary field delves into the application of sophisticated algorithms and models to discern the probable author of a text, whether it's a timeless literary masterpiece, a historical document, or contemporary content, all without prior knowledge of the author's identity. Its ramifications extend across diverse domains including literary analysis, plagiarism detection, historical inquiries, and beyond.[1]

The crux of Author Identification lies in harnessing the distinctive writing styles, linguistic intricacies, and textual nuances that set one author apart from another. Employing learning techniques such as machine learning, deep learning, and natural language processing, practitioners aim to decode these subtle yet discernible characteristics. By meticulously analyzing and extracting features from texts – ranging from vocabulary choices to sentence structures, and even stylistic elements – these techniques facilitate the creation of predictive models capable of attributing authorship to a given text.[5]

In practice, machine learning models are honed on datasets comprising known texts authored by specific writers. Through this training process, models discern and quantify the unique patterns and idiosyncrasies associated with each author. Consequently, when presented with an anonymous text, these models can intelligently deduce the most likely author based on the text's resemblance to known authorial features. Thus, Author Identification using Learning Techniques not only enriches academic and literary research but also holds pragmatic significance in resolving authorship disputes, unraveling historical enigmas, and addressing contemporary challenges such as content verification and prevention of manipulation or ghostwriting.[3]

As the field of Author Identification continues to evolve, researchers and practitioners are incessantly refining their methodologies and experimenting with novel approaches. This evolution is driven by a relentless pursuit of higher accuracy and reliability in attributing authorship, prompting exploration into a diverse array of models and feature extraction techniques. With each iteration, the efficacy of these methods is enhanced, bringing us closer to unraveling the complexities of authorial signatures embedded within texts.[8]

Moreover, the advancements in natural language processing and machine learning technologies are propelling Author Identification to new heights of sophistication. With the advent of cutting-edge algorithms and computational resources,

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

researchers can delve deeper into the subtleties of language and style, uncovering previously unnoticed nuances that contribute to an author's distinct voice. This convergence of technology and scholarship opens up a plethora of possibilities for pushing the boundaries of what can be achieved in the realm of Author Identification.[2]

In essence, Author Identification using Learning Techniques represents a dynamic and multifaceted field with boundless opportunities for academic inquiry and practical application. As we continue to unravel the intricate tapestry of language and style, this discipline not only enriches our understanding of literary works but also empowers us to navigate the complexities of authorship in an increasingly interconnected world.

A. Motivation

Identifying the true author of a text is crucial for academic and professional settings to detect cases of plagiarism. Learning techniques can help in determining whether a particular piece of writing is the work of a specific author or if it has been copied from another source.Analyzing the writing style and characteristics of different au- thors is of great interest to literary scholars and researchers.

B. Objectives

To study existing literature related to Author Identification.
To design a machine learning model for author identification in English literature.
To process text using NLP and machine learning algorithms for author identification.
To evaluate performance of the proposed system.

II. LITERATURE REVIEW

Authorship Identification: Na¨ıve Bayes with XGBoost Approach, Dr. B.S. Daga, Jason Dsouza, Ryan Furtado, ManupendraTiwari,June 2021.

Description: When an individual writes, they subconsciously use a certain ar- ray of words or writing patterns and sentiments, and we could use this to deter- mine their writing style. The fundamental assumption of authorship identifica- tion is that each individual has a habit of subconsciously using certain words, patterns and emotions that make their writing style unique.This paper studies the need of Authorship Identification, and thereafter, proposes an architecture for it. It focuses on using N-gram with certain features and feeding them into a Na¨ıve Bayes classifier along with XGBoost for classification. The paper also tests the accuracy for various sub approaches like tfidf, word count and their combination with XGBoost to determine the best sub approach. Finally, a dataset is generated and the system is coded in Python using Anaconda3 and Jupyter Notebook along with WordCloud for display[5].

Author Identification Based On NLP, Noura Khalid Alhuqail, 2022.

Description: In this research, the study is performed with Bag of Words (BOW) and Latent Semantic Analysis (LSA) features. The “All the news” dataset on Kaggle is used for experimentation and to compare BOW and LSA for the best performance in the task of author identification. Support vector machine, random forest, Bidirectional Encoder Representations from Trans- formers (BERT), and logistic regression classification algorithms are used for author prediction. For first scope that have 20 authors, for each author 100 articles, the greatest accuracy is seen from logistic regression using bagof- words, followed by random forest, also using bag-of-words; in all algorithms, bag-of-words scored better than LSA. Ultimately, BERT model was applied in this research and achieved 70.33% accuracy performance. For second scope that increase the number of articles till 500 articles per author and decrees the number of authors till 10, the BOW achieves better performance results with the logistic regression algorithm at 93.86%. Moreover, the best accuracy per- formance is with LR at 94.9% when merged the feature together and it proved that it is better than applied BOW and LSA individual, with an improvement by almost 0.1% comparing with BOW only[3].

Author Identification with Machine Learning Algorithms Ibrahim, Feris¸tahDalkılıc¸,June 20, 2023

Description: In this study, we conducted an experiment for the identification of the author of a Turkish language text by using classical machine learning meth- ods including Support Vector Machines (SVM), Gaussian Naive Bayes (Gaus-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

sianNB), Multi Layer Perceptron (MLP), Logistic Regression (LR), Stochastic Gradient Descent (SGD) and ensemble learning methods including Extremely Randomized Trees (ExtraTrees), and eXtreme Gradient Boosting (XGBoost). The proposed method was applied on three different sizes of author groups including 10, 15 and 20 authors obtained from a new dataset of newspaper articles. Term frequency-inverse document frequency (TF-IDF) vectors were created by using 1-gram and 2-gram word tokens. Our results show that the most successful method is the SGD with a classification performance accuracy of 0.976% by using word unigrams and most successful method is the LR with a classification performance accuracy of 0.935% by using word bigrams[4].

A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," 2016 15th IEEE International Conference on Machine Learn- ing and Applications (ICMLA), 2016.

Description: In this paper, we apply basic classification models and explore GRU, LSTM and Bi-LSTM at the sentence and article levels to identify the authors of a given piece of text. We deal with the pre-processing and feature vectorization of texts from the Reuters 50 50 C-50 data-set. The features from these vectors are extracted and word embeddings using GloVe are created. The model is developed to deal with larger pieces of textual information and analyse semantic and metaphorical uses of words with the goal of improving authorship identification tasks[8].

III. DESIGN OF SYSTEM

The architecture of an author identification system for English literature involves a systematic process to analyze and discern the unique writing styles of various au- thors. Initially, a diverse dataset of English literature works is collected, encompass- ing different genres, time periods, and authors. Subsequently, the collected text data undergoes preprocessing to eliminate noise, such as punctuation and stop words, and is tokenized for further analysis[3]. The next step entails feature extraction, where relevant linguistic features are identified, including word frequency, sentence struc- ture, vocabulary richness, and potentially more advanced attributes like sentiment analysis[5]. A machine learning model, such as Support Vector Machines or Random Forests, is then trained on these extracted features to learn patterns associated with specific authors[4]. Cross-validation is employed to validate the model's performance and ensure its adaptability to unseen data. Fine-tuning and optimization are conducted to enhance the model's accuracy and address potential over fitting or under fitting. The model's effectiveness is evaluated using a separate test dataset. Upon successful training, the model is deployed into a functional system capable of taking input text, preprocessing it, extracting features, and predicting authorship. Optionally, a user interface may be developed for non-technical users, and a feed- back mechanism can be implemented for continuous improvement.

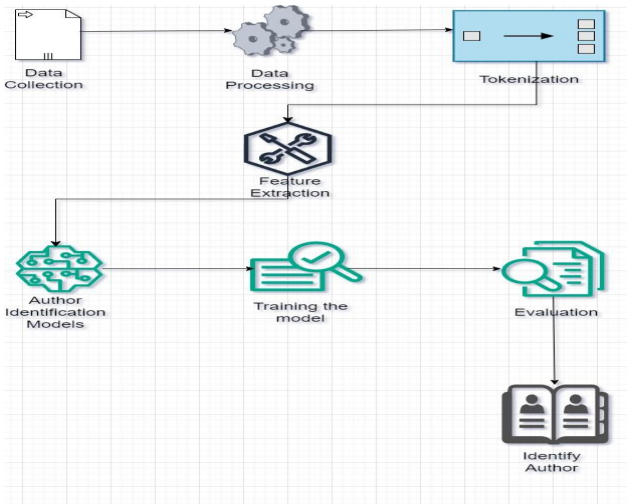


Fig. 3.1 System Architecture

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

A. Software Requirements Specification

Software Requirements:

- Python
- Flask
- Pycharm, Jupiter Notebook.
- Libraries: pandas, numpy, scikitlearn, matplotlib, etc.

Hardware Requirements:

- RAM 8 GB or Higher
- HDD 100 GB (Processor – Intel Core I4 or Higher)

B. Working

- Data Collection: We start by gathering a wide range of English texts, including novels, essays, articles, and more. These texts should cover various genres, time periods, and authors to ensure diversity.
- Preprocessing: Once we have our texts, we clean them up. This means getting rid of unnecessary elements like punctuation and common words (like "the" and "and") that don't give us much information. Then, we split the texts into smaller parts, like sentences or words, to make them easier to analyze.
- Feature Extraction: In this step, we look for important features in the texts that can help us identify authors. This might include things like how often certain words are used, the length and structure of sentences, and the complexity of vocabulary. We might even consider more advanced features like the overall sentiment or tone of the writing.
- Machine Learning Model Training: With our features identified, we use them to train a machine learning model. This model learns to recognize patterns in the data that are associated with specific authors. We can use various algorithms like Support Vector Machines or Random Forests for this task.
- Cross-Validation: To make sure our model is accurate and can work well with new data it hasn't seen before, we use a technique called cross-validation. This involves splitting our dataset into multiple parts, training the model on some parts, and then testing it on the remaining parts.
- Fine-Tuning and Optimization: Once we have a working model, we fine-tune and optimize it to improve its performance. This might involve adjusting parameters, tweaking the features we're using, or trying out different algorithms to see what works best.
- Evaluation: We evaluate the effectiveness of our model using a separate test dataset. This helps us determine how well our model can predict the authors of new texts.
- Deployment: Finally, once we're confident in our model's performance, we deploy it into a functional system. This system can take input text, preprocess it, extract features, and predict the most likely author. Optionally, we can create a user-friendly interface to make it easier for non-technical users to interact with the system.

C. Algorithms

Random Forest Algorithm:

Bootstrap Sampling:

Randomly select subsets of the training data with replacement. This creates multiple bootstrap samples, each potentially containing different instances and allowing some instances to appear multiple times while others not at all.

Feature Randomness:

At each split point of a decision tree, randomly select a subset of features from the available features.

This ensures that each tree in the forest is built on different subsets of features, reducing the correlation between trees and improving the diversity of the ensemble.

Decision Tree Construction:

Build multiple decision trees using the bootstrap samples and randomly selected features.

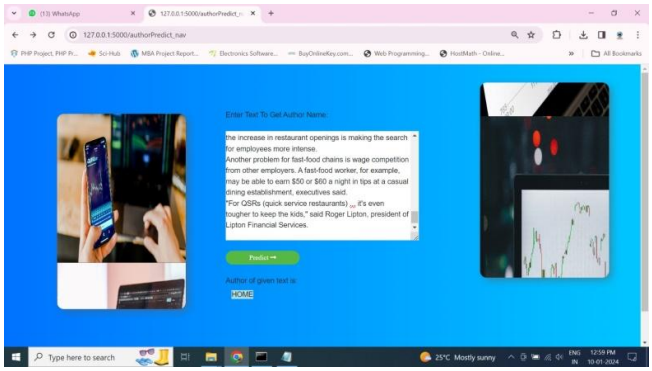
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

At each node of the tree, choose the best split among a random subset of features, based on a criterion such as Gini impurity for classification or mean squared error for regression.
Continue splitting the nodes recursively until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf node.

Voting or Averaging:
For classification tasks, each tree "votes" for the class label of a given instance, and the class with the most votes is chosen as the final prediction.
For regression tasks, the predictions of all trees are averaged to obtain the final output.

XGBoost (eXtreme Gradient Boosting):
Initialize Model:
Start with an initial prediction, often the mean value for regression or the log-odds for classification.
Compute Residuals:
Calculate the residuals (the difference between the predicted values and the actual values) for each instance in the training data.
Fit a Base Learner:
Train a weak learner (usually a decision tree) to predict the residuals. This tree is often shallow to avoid overfitting.
Update Predictions:
Add the predictions of the current weak learner to the previous predictions, adjusting the model towards the correct values.
Compute Pseudo-Residuals:
Calculate new residuals based on the difference between the updated predictions and the true target values.
Iterate:
Repeat steps 3-5 for a specified number of iterations (boosting rounds), each time fitting a new weak learner to the pseudo-residuals and updating the predictions.
Regularization:
Apply regularization techniques such as shrinkage (learning rate) and tree depth constraints to prevent overfitting and improve generalization performance.
Final Prediction:
The final prediction is obtained by summing up the predictions of all weak learners, typically weighted by a factor (learning rate) to control the contribution of each tree.

IV. RESULT & SUMMARY



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

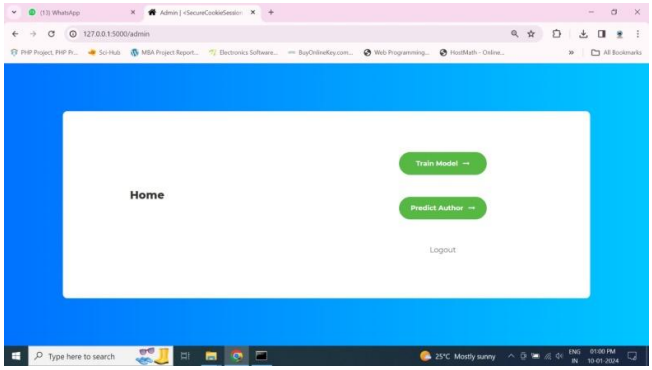


Fig. 4.2 Output

V. CONCLUSION

In conclusion, the development of an author identification system for English literature using learning techniques represents a significant advancement in computational linguistics. Through the integration of machine learning algorithms and natural language processing methods, the system offers a robust framework for extracting and analyzing textual features to accurately attribute authorship. As the system evolves, further enhancements and interdisciplinary collaborations hold the potential to refine its accuracy, usability, and ethical considerations, thereby enriching our understanding of writing styles and facilitating applications in various domains such as literary analysis, plagiarism detection, and historical research.

REFERENCES

[1]. Authorship Identification: Na¨ıve Bayes with XGBoost Approach, Dr. B.S. Daga, Jason Dsouza, Ryan Furtado, ManupendraTiwari,June 2021.
[2]. Author Identification Based on NLP, Noura Khalid Alhuqail, 2022.
[3]. Author Identification with Machine Learning Algorithms ˙Ibrahim Yu¨lu¨ce, Feris¸tahDalkılıc¸,June 20, 2023
A. M. Mohsen, N. M. El-Makky and N. Ghanem, ”Author Identification Us- ing Deep Learning,” 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.
[4]. Barlas, G., Stamatatos, E. (2020). Cross-Domain Authorship Attribution Us- ing Pre-trained Language Models. In: Maglogiannis, I., Iliadis, L., Pimenidis,
[5]. E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. 35.
[6]. Argamon, S., Saric, J., & Stein, S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 461-468.
[7]. Koppel, M., &Schler, J. (2004). Authorship verification as a one-class classification problem. Proceedings of the 21st International Conference on Machine Learning, 62.
[8]. Juola, P., &Baayen, H. (2005). A controlled-corpus experiment in authorship attribution by cross-entropy. Literary and Linguistic Computing, 20(1), 59-67.
[9]. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538-556.
[10]. Gupta, D., Shreevastava, A., &Saroha, R. (2013). Authorship attribution using machine learning algorithms. International Journal of Computer Applications, 79(6), 15-19.
[11]. Eder, M., &Rybicki, J. (2016). Stylometry with R: A suite of tools. Springer.
[12]. Hoover, D. L., & Jensen, K. (2016). The future of stylometry: A research agenda. Digital Scholarship in the Humanities, 31(1), 134-148.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, April 2025
Impact Factor: 7.67

[13]. Oliveira, R. H., & Gomes, A. T. (2017). On the effectiveness of preprocessing methods in authorship attribution. Expert Systems with Applications, 69, 88-103.
[14]. Ghosal, A., &Sarkar, R. (2018). Feature selection methods for authorship attribution: A comprehensive review. Artificial Intelligence Review, 49(2), 213-244.
[15]. Holmes, D. I., & Koppel, M. (2018). The authorship attribution of historical texts. Journal of Quantitative Linguistics, 25(2), 97-115.
[16]. Yang, M., Liu, Y., & Chen, X. (2019). A survey of machine learning methods for authorship attribution. Information Processing & Management, 56(6), 102088.
[17]. Rehurek, R., &Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45-50).
[18]. Gupta, S., Singh, S., &Goyal, D. (2016). A comparative analysis of authorship attribution techniques. Journal of Information Science, 42(3), 293-308.
[19]. Kestemont, M., Daelemans, W., & Van de Cruys, T. (2016). Authorship attribution and verification with many authors using transfer learning from pan. Literary and Linguistic Computing, 31(3), 443-457.
[20]. Evert, S., Proisl, T., &Jannidis, F. (2017). Do we need large amounts of data for authorship attribution? Digital Scholarship in the Humanities, 32(suppl_2), ii47-ii55.
[21]. Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification. In Proceedings of the IEEE International Conference on Semantic Computing (pp. 587-592).
[22]. Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.
[23]. Burrows, J. F. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267-287.
[24]. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538-556.
[25]. Kestemont, M., Daelemans, W., & Van de Cruys, T. (2014). Authorship attribution and verification with many authors using transfer learning from pan. Literary and Linguistic Computing, 31(3), 443-457