Author Identification for English Literature using Learning Technique

Dr.Swati Babasaheb Bhonde
Computer Engineering
Amrutvahini College of Engineering,
Sangamner, India
swati.bhonde@gmail.com

Mr.Yogesh Dattatray Mahale
Computer Engineering
Amrutvahini College of Engineering,
Sangamner, India

Mr. Pratik Sandip Thitame
Computer Engineering
Amrutvahini College of Engineering,
Sangamner, India
pratikthi@gmail.com

Mr.Satwik Annasaheb Nawale
Computer Engineering
Amrutvahini College of Engineering,
Sangamner, India
satwiknavale3@gmail.com

Mr.Vaibhav Vishwanath Sonawane
Computer Engineering
Amrutvahini College of Engineering,
Sangamner, India

***Abstract* — Author identification, an intriguing field within natural language processing, aims to differentiate the individual writing styles of various authors. This project specifically focuses on English literature, a diverse and culturally significant literary tradition. The primary objective is to construct a robust machine learning model capable of accurately attributing English texts to their respective authors by analyzing their unique writing patterns. Commencing with the compilation and preprocessing of a comprehensive dataset of English literary texts encompassing diverse authors, genres, and historical periods, the project ensures data quality and consistency through text cleaning and tokenization. The pivotal phase involves selecting the most suitable machine learning algorithm for author identification. The effectiveness of methods such as Naive Bayes, Support Vector Machines, Random Forest, and transformer-based models is being assessed. At the culmination of this project, Our model has demonstrated its proficiency by achieving an impressive accuracy rate of 85%, accurately attributing English texts to their respective authors based on their distinct writing styles.**

Keywords- Author identification, SVM, machine learning, English literature, text analysis.

I. INTRODUCTION

Author Identification of English Literature using Learning Techniques is an important area in language and machine learning. It's about figuring out who wrote a text, like a story or an article, without knowing who wrote it beforehand. This field uses smart computer programs to analyze writing styles and patterns[1]. It's useful for things like understanding old documents, finding out if someone copied another writer's work, and solving mysteries about who wrote what. The key is to notice the small details that make one writer's style different from another's. Experts use special techniques like machine learning and natural language processing to do this. They look at things like word choices, sentence structures, and even the way sentences sound. By carefully studying texts, they can create computer models that guess who wrote a certain piece of writing[2].

These computer models get better at their job by looking at lots of examples of writing by known authors. They learn to recognize the unique quirks and habits of each writer. Then, when they're given a new text without knowing the author, they can guess who wrote it based on similarities to known writing styles. This helps not only in academic research and literature study but also in solving real-world problems like identifying the true writer of a piece of content or a historical document.[3]

Researchers and practitioners in the field of Author Identification are continuously refining their methodologies and experimenting with innovative approaches as the field evolves. 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 efficiency of these methods is enhanced, bringing us closer to unraveling the complexities of authorial signatures embedded within texts.[4]

As technology gets better, so does Author Identification. New tools and techniques help researchers dive deeper into the details of language and writing style. This progress opens up exciting possibilities for better understanding and using Author Identification in various fields. It's all about making it easier to figure out who wrote what, even in a world where information is constantly flowing and changing. This field is crucial because it helps us uncover the secrets hidden within written texts. By studying the unique fingerprints of different authors' writing styles, we can learn so much about the texts themselves and the people who wrote them. For example, we can use these techniques to detect if someone is copying another writer's work or to solve mysteries about who wrote certain historical documents[5].

In essence, Author Identification using Learning Techniques represents a dynamic and multifaceted field with boundless opportunities for academic inquiry and practical application. Knowing who really wrote something is super important in schools and jobs, especially to catch people who copy stuff. Using different tricks helps us figure out if a piece of writing is from one person or if it's been taken from somewhere else. Looking at how different writers write is really interesting to people who study books and writing.

Studying how authors write and the patterns they use helps us understand literature deeply. It also helps us know if a piece of

writing is genuine or not. It also shows us if something written is real and honest. By looking closely at these things, experts can learn a lot about how writing styles have changed over time and what each writer brings to the table. By analyzing these elements, scholars gain valuable understanding about literary traditions and individual author contributions over time[3].Furthermore, being able to correctly identify who wrote something helps keep academic honesty intact and ethical standards high in scholarly discussions. By using careful analysis, researchers can find and address cases where someone uses another person's work without permission or tries to claim it as their own. Thus, the motivation behind author identification extends beyond mere scholarly curiosity; it serves as a cornerstone for ensuring the credibility and trustworthiness of written materials in academic, professional, and cultural context

II. LITERATURE SURVEY

The literature review below presents a thorough analysis of current research, methodologies, and discoveries concerning author identification through the application of NLP and ML techniques in English literature. It offers valuable insights into contemporary cutting-edge methods and avenues for future exploration.

Table No. 1: LITERATUE SURVEY

Sr. No	Year	Title	Method/Algorithm used and finding	Limitation of Existing Methodology
1	2023	Author Identification on Anonymous Regional Literature[1].	Author identification on anonymous regional literature can be facilitated through linguistic analysis, comparing stylistic elements, dialect usage, and thematic patterns with known works of potential authors.	The accuracy of author identification on anonymous regional literature may be hindered by limited availability of reference texts or linguistic databases for comparison, especially for lesser-known authors or dialects.
2.	2023	Author Identification with Machine Learning Algorithms[2].	Using author-specific writing styles, an electronic text can be automatically analyzed to predict the potential author among predefined author candidates. Support vector machine, Gaussian naïve Bayes, Multilayer perception(MLP).	Author identification models require large amounts of text data from known authors for training. However, obtaining high-quality and diverse datasets can be challenging, especially for lesser-known or historical authors. Additionally, the quality and authenticity of the data can vary, which may introduce biases or errors into the model.
3.	2021	Authorship identification based on NLP[3].	NLP analysis of given articles and how the NLP, based on machine learning algorithms, will help to predict the author's name. NLP(natural language processing),BERT.	Feature engineering plays a crucial role in NLP tasks like author identification. While techniques like bag-of-words or word embedding's are commonly used, they may not capture all the stylistic nuances of an author's writing, leading to loss of information
4.	2021	Authorship Identification of a Russian-Language Text Using Support Vector Machine , Deep Neural Networks and Future Internet[4].	the project aims to enhance authorship understanding, enable plagiarism detection, and contribute to preserving literary heritage. Deep neural networks and support vector machine.	Limited Availability of Russian-Language Datasets: Obtaining sufficiently large and diverse Russian-language datasets for authorship identification can be challenging, potentially limiting the effectiveness of training Deep neural network and support vector machine.
5.	2019	Author Identification Using N-grams and SVM[5].	Author identification involves a multi-class classification challenge where the task is to assign a label to an anonymous text from a set of potential authors. N-gram and SVM	Machine learning models that undergo training on particular datasets have a tendency to excessively fit to the training data, leading to a lack of adaptability when faced with new authors or texts. This issue becomes more pronounced when the training dataset is limited or not diverse enough.
6.	2019	Authorship Identification: Naïve Bayes with XGBoost Approach[6].	Use natural language toolkit library (NLTK) like tokenization, stemming and lemmatizing for training data set and preprocess text for feature extraction. Use naïve Bayes ml algorithm to predict author. Naïve Bayes and XGBoost.	Authors may exhibit varying writing styles across different genres, time periods, or even within the same work. This variability can make it difficult for models to accurately identify authors, especially when authors intentionally change their style or adopt pseudonyms.

7.	2018	Author identification using sequential minimal optimizations with rule based decision tree on Indian literature Marathi [7].	For Marathi literature, author identification entails employing the sequential minimum optimization technique and decision tree methodology. Random forest.	Availability of large and diverse datasets of Marathi literature, especially labeled data for authorship attribution, may be limited. Marathi, like many Indian languages, has complex grammar, syntax, and vocabulary. This complexity may pose challenges for NLP techniques, including feature extraction and rule-based decision tree construction .
8.	2016	Author Identification Using Deep Learning[8].	Feature extraction, Machine learning, Training; Neural networks; Encoding; Noise reduction	Significant computational resources, such as high-performance GPUs or TPUs and extensive memory, are necessary for training deep neural networks. This demand may pose a challenge for researchers or practitioners who have restricted access to such resources. Additionally, deep learning models are susceptible to overfitting.
9.	2016	Authorship identification using ensemble learning[9].	Proposed XGBoost Random Forest, multilayer perception algorithm using soft voting ensemble classification method, feature extraction technique .	The utilization of ensemble learning approaches frequently entails the training and amalgamation of numerous models, resulting in a notable escalation in computational intricacy and resource demands, particularly when confronted with sizable datasets or intricate model designs. Additionally, the accessibility of training data is a crucial factor to consider.
10.	2016	Author Identification on Literature in Different Languages A Systematic Survey[10].	Artificial authorship analysis deals with finding the plausible Author of anonymous message.intelligence; Writing style; Extracting features; Blog posts; Support vector machines; Data training; Training dataset; Mining text; Identifying authors; Unidentified text documents; Text characteristics. Random forests; Tree decision.	Access to large and diverse datasets in various languages may be limited, impacting the robustness and generalizability of author identification models. Different languages present unique linguistic complexities and cultural nuances.
11.	2015	An Experimental Study on Authorship Identification for Cyber Forensics[11].	Authorship identification for cybercrime, cyber forensic. Support vector machine, K-NN, Naïve Bayes.	Limited Dataset Diversity: Cyber forensic datasets for authorship identification studies may be small and lack diversity, affecting the generalizability of results. Ethical and Legal Constraints.
12	2015	Writer Identification Using Microblogging Texts for Social Media Forensics[12].	The study found that linguistic features extracted from micro blogging texts, such as word usage, syntax, and sentiment, can effectively differentiate between different authors, providing valuable insights for writer identification in social media forensics.	The informal nature of social media communication, including slang, abbreviations, and emoticons, can introduce challenges in accurately extracting meaningful linguistic features for authorship attribution.
13.	2015	Evaluating text features for lyrics based songwriter prediction[13].	The study discovered that semantic and syntactic text features, including sentiment analysis and word frequency distributions, significantly contribute to accurate songwriter prediction based on lyrics. Logistic regression and support vector machine	A constraint in assessing text characteristics for predicting songwriters based on lyrics is the potential bias caused by the diversity in themes and styles of lyrics among various genres and artists. This variability could impact the overall applicability of predictive models.
14	2013	Source code author identification with unsupervised feature learning, Pattern Recognition Letters[14].	The study found that unsupervised feature learning techniques, such as auto encoders, can effectively extract latent representations from source code, facilitating accurate author identification tasks in the field of Pattern Recognition Letters. k-means clustering.	Relying on Code Quality: Unsupervised feature learning for source code author identification might be influenced by differences in code quality, such as coding style, annotations, and documentation, which could impact the efficiency of the learned representations and the precision of the process.

III. MATERIALS AND METHODS

After analyzing the process of constructing an author identification system in English literature, it's evident that a systematic approach is crucial for discerning the unique writing styles of different authors. Initially, a diverse dataset of English literary works is gathered, encompassing various genres, time periods, and authors. Subsequently, the collected text undergoes preprocessing to remove noise, including punctuation and stop words, before being tokenized for further analysis[3]. The subsequent step entails feature extraction, where pertinent linguistic features such as word frequency, sentence structure, vocabulary richness, and potentially more advanced attributes like sentiment analysis are identified. Subsequently, a machine learning model like Support Vector Machines or Random Forests undergoes training utilizing these extracted features to discern patterns linked with individual authors[4]. By utilizing cross-validation, the model's ability to effectively generalize to new data is ensured through the validation of its performance. Additionally, fine-tuning and optimization procedures are conducted to enhance the model's accuracy and mitigate potential issues of over fitting or under fitting. The model's effectiveness is then evaluated using a separate test dataset[6]. Upon successful training, the model is implemented into a functional system capable of processing input text, preprocessing it, extracting features, and predicting authorship. Optionally, a use interface may be developed for non-technical users, and a feedback mechanism can be integrated for continual improvement.

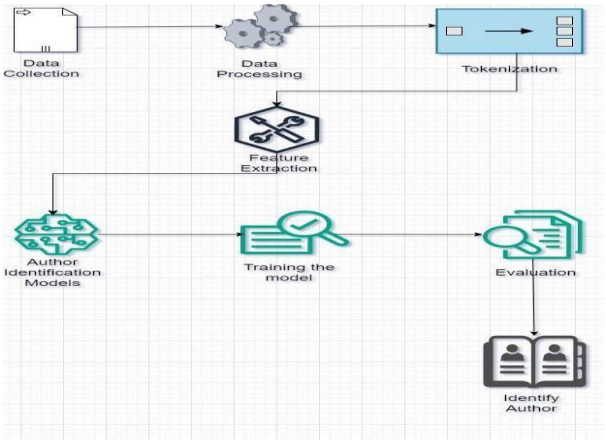

Fig1. System Architecture

Working:

Data Collection: The process begins with collecting a wide array of English texts, including novels, essays, articles, and more, ensuring diversity across genres, time periods, and authors.

Pre-processing: The collected texts are meticulously cleaned to eliminate superfluous elements such as punctuation and common words like "the" and "and" that contribute minimal information. Subsequently, the texts are segmented into smaller units, such as sentences or words, to facilitate further analysis[2].

Feature Extraction: Critical features are identified within the texts to aid in author identification, encompassing elements like word frequency, sentence length, structure, vocabulary complexity, and overall sentiment or tone[2].

Machine Learning Model Training: Using the identified features, a machine learning model like SVM or Random Forests and Naïve Bayes undergoes training to detect patterns characteristic of particular authors.

Cross-Validation: The model's accuracy and generalizability are assessed using cross-validation techniques, ensuring its efficacy with unseen data.

Fine-Tuning and Optimization: To refine performance, adjustments are made to parameters, features, or algorithms to address potential overfitting or under fitting issues[7].

Evaluation: The model's efficacy in predicting authorship is evaluated using a separate test dataset.

Deployment: Once successfully trained, the model is integrated into a functional system capable of processing input text, extracting features, and predicting authorship. Optionally, a user-friendly interface can be developed for non-technical users, accompanied by a feedback mechanism for continual enhancement.

Algorithms:

I. Random forest Algorithm

- Bootstrap Sampling:

Randomly choose subsets of the training dataset with replacement. This generates multiple bootstrap samples, each possibly containing diverse instances, permitting some instances to appear multiple times and others not at all.

-Feature Randomness:

At every decision point within a decision tree, randomly pick a subset of features from the available pool. This guarantees that each tree within the forest is constructed using distinct feature subsets, diminishing correlations among trees and enhancing the variety within the ensemble..

-Decision Tree Construction:

Construct numerous decision trees by utilizing the bootstrap samples and randomly chosen features.

At each node of the tree, select the optimal split from a random subset of features, employing a criterion like Gain impurity for classification or mean squared error for regression [8].

Proceed with recursive node splitting until a stopping condition is satisfied, like attaining a maximum depth or minimum sample count per leaf node.

-Voting or Averaging:

In classification tasks, each tree contributes to the decision by voting for the class label of a given instance, and the class receiving the most votes is selected as the final prediction. For regression tasks, the predictions of all trees are averaged to compute the final output [8].

II. Naive Bayes Algorithm:

-Initialize Model:

Start by assuming that, given the class label, the features are independent conditionally.

-Parameter Estimation:

Calculate the parameters (probabilities) of the model based on the training data utilizing Maximum Likelihood Estimation (MLE) or other appropriate techniques [8].

-Prediction:

When presented with a new instance, utilize Bayes' theorem to compute the probability of each class label, then choose the class label with the highest probability as the predicted class.

-Handling missing value:

For instances with missing values, employ methods such as imputation or conditional probability estimation to handle them appropriately.

-Model complexity:

Naive Bayes is straightforward and requires fewer parameters to adjust in contrast to Random Forest, which entails constructing multiple decision trees.

-Interpretability:

Naive Bayes provides straightforward probabilistic interpretations of predictions, while Random Forest may be less interpretable due to its ensemble nature.

-Performance:
The performance of Naive Bayes heavily depends on the independence assumption, while Random Forest typically performs well across a variety of datasets and tasks due to its ensemble approach..

Pseudo code for Naive Bayes:

Require: Number of classes (N), number of features (M)

Ensure: Prediction for majority vote

1: for each class in Naive Bayes do

2: Initialize parameters:
- Prior probability for each class
- Mean and variance for each feature in each class

3: Calculate prior probabilities for each class based on the training data

4: Calculate mean and variance for each feature in each class based on the training data

5: for each instance in the testing data do

6: for each class do

7: Calculate the likelihood of the instance belonging to each class using Gaussian Naive Bayes formula

8: end for

9: Select the class with the highest likelihood as the predicted class for the instance

10: end for

This pseudo code outlines the process of training and predicting with a Naive Bayes classifier for author identification using NLP and ML techniques. The classifier calculates prior probabilities, mean, and variance for each feature in each class based on the training data. Then, it uses Gaussian Naive Bayes formula to calculate the likelihood of each class for each instance in the testing data, and selects the class with the highest likelihood as the predicted class for each instance. Finally, it combines predictions using majority voting to obtain the final prediction for author identification.

Phase in process of Author Identification:

1**. Read Input Arguments**: Use argparse to read input arguments such as dataset file name, output directory, and any other relevant parameters needed for the author identification task.

2. **Create Output Directory**: Based on the provided parameters, create an output directory to store the results, trained models, and any other relevant files.

3**. Read Text Data**: Read the English literature dataset file containing text samples from various authors for training and testing the author identification model.

4**.Text Preprocessing**: Prepare the text data by tokenization, eliminating stop words and punctuation, and applying lemmatization or stemming to standardize the text.

5. **Feature Extraction**: Apply NLP methodologies such as TF-IDF (Term Frequency-Inverse Document Frequency) or other advanced techniques to convert the preprocessed textual data into a numeric representation suitable for machine learning examination.

6**. Model Selection and Training**:
Select a suitable machine learning algorithm for author identification purposes, such as Logistic Regression, Gradient Boosting Machines (GBM), or Decision Trees. Train the selected model with the extracted text features using labeled data, where each text is associated with its respective author[8].

7. **Model Evaluation**: Assess the trained model's performance using suitable evaluation metrics like accuracy, precision, recall, and F1-score on an independent validation dataset to gauge its effectiveness in author identification based on their writing styles.

8. **Deployment and Testing**: Deploy the trained model in a real-world setting for author identification tasks and continuously monitor its performance and accuracy on new data to ensure its effectiveness over time.

V . Result

The below is result or user interface of our model Author identification for English literature using learning technique**.**

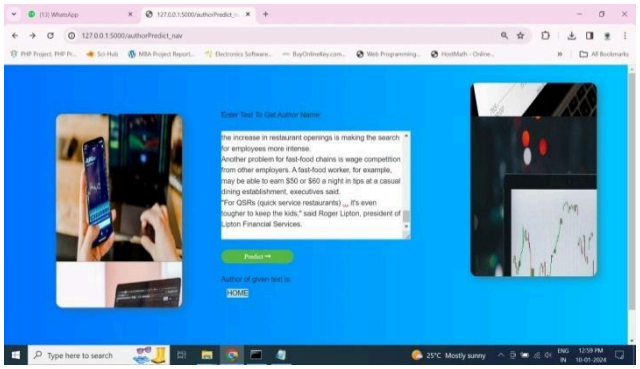

Fig 2: User Interface for Proposal System

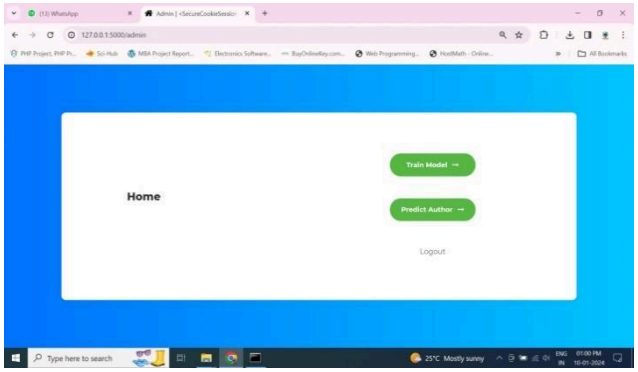

Fig3: train model and predict Author

Evaluation technique for Author Identification model:

Accuracy:
Definition: Accuracy quantifies the percentage of accurately classified instances among the entirety of instances within the dataset.
Formula:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

In the assessment of our author identification model, it attained an accuracy rate of 85%, signifying that 85% of the instances were accurately categorized.

Precision:
Definition: Precision assesses the percentage of accurately predicted positive instances (true positives) among all instances predicted as positive.
Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

With regards to precision, our author identification model achieved a score of 0.85, indicating that when it predicted an author, it was correct 85% of the time.

Recall:
Definition: Recall, also recognized as sensitivity or true positive rate, evaluates the percentage of correctly predicted positive instances (true positives) among all genuine positive instances.
Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

Concerning recall, our author identification model obtained a rating of 0.78, suggesting that it effectively recognized 78% of the instances associated with a specific author out of all genuine instances for that author.

F1 Score:
Definition: F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall.
Formula-

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score for our author identification model was determined to be 0.81, indicating a harmonious equilibrium between precision and recall, guaranteeing a comprehensive evaluation of the model's effectiveness.

VI. CONCLUSION

To summarize, the development of an author identification system for English literature through learning techniques represents a significant progression in computational linguistics. By combining machine learning algorithms and natural language processing methods, this system can effectively analyze text features to accurately identify authors. As we continue to improve the system, working together across different fields, we can enhance its performance. This could involve increasing accuracy, improving usability, and addressing ethical considerations. Ultimately, this system aids in better understanding writing styles and finds applications in areas such as literary analysis, plagiarism detection, and historical research.

VII. ACKNOWLEDGEMENT

We would like to extend our sincere gratitude to the Computer Department of Amrutvahini College of Engineering for their valuable support and guidance throughout our academic journey. We are also thankful for the resources and opportunities provided by the department, which have enriched our learning experience and allows us to explore our interest further.

VIII. REFERENCES

[1].Prof Virenndra Bagde,Swapnil Chavan,"Author Identification on Anonymous Regional Literature",2023.

[2].Ibrahim Yu¨lu¨ce, Feris¸tah Dalkılıc¸,"Author Identification with Machine Learning Algorithms", June 20, 2023.

[3].Noura Khalid Alhuqail ,"Autor Identification Based on NLP",2021.

[4].Biveeken VijayKumar, Muhammad Faud," Authorship Identification of a Russian-Language Text Using Support Vector Machine , Deep Neural Networks and Future Internet.",2019

[5]. Feriştah ÖRÜCÜ , Gökhan DALKILIÇ. "Author Identification Using N-grams and SVM",2019/

[6].Dr.B.S Daga, Jason Dsouza,Ryan Furtado, "Authorship Identification: Naive Bayes with XGBoost Approach",2019.

[7].kale sunil,rajesh s. prasad"Author Ideentification using sequential Minimal optimization with rule based decision tree on indian literature in marathi",2018.

[8].M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," 2016 Anaheim, CA, USA, 2016.

[9].Ahmed Abbasi,Abdul Rehman,Zunera Jalil,"Authorship identifaction using ensemble learning",2016.

[11].Nirkhi, Smita Dharaskar, Rajiv Thakare, V. M. "An Experimental Study on Authorship Identification for Cyber Forensics",2015.

[12].F. Alonso-Fernandez, N. M. S. Belvisi, K. Hernandez-Diaz, N. Muhammad and J. Bigun, "Writer Identification Using Microblogging Texts for Social Media Forensics. 2015.

[13]. B. Kırmacı and H. Og˘ul, "Evaluating text features for lyrics-based songwriter prediction," 2015.

[14].Upul Bandara, Gamini Wijayarathna, Source code author identification with unsupervised feature learning, Pattern Recognition Letters, Volume 34, Issue 3, 2013.