Multilingual Author Identification: Integrating Machine Learning and Deep Neural Networks forEnglish and Arabic Text

Ahmed Anwar, Arabi E.Keshk, Eman M.Mohamed

*Department of Computer Science, Faculty of Computers
and InformationMenofia University
Shebin El Kom, Egypt

Abstract
The article describes a new method for author identification across English and Arabic texts using machine learning and deep neural networks, based on a comprehensive framework that employs five algorithms—Random Forest, Logis- tic Regression, k-nearest Neighbors, Support Vector Machines, and Naive Bayes, alongside deep learning approaches. Using the TF-IDF method for feature extraction, the authors analyze two datasets: the Victorian Era Authorship Attribution and a dataset for Arabic author identification. The results show that Support Vector Machines and Logistic Regression demonstrate strong per- formance in authorship attribution, effectively capturing nuancedwriting styles with accuracy rates reaching 95%. The authors illustrate the proposed method through detailed comparative analyses and highlight its applicability in forensic linguistics, pla-giarism detection, and literary analysis. The method significantly improves accuracy and computational efficiency in authorship attribution tasks. The new methods effectiveness evaluation is confirmed by the higher precision, recall, and F1-scores obtained compared to traditional methods. New research results develop the field of text mining and can be used for enhancing security measures and understanding linguistic patterns across diverselanguages. This studys novelty and scientific contribution lie in its cross-lingual approach and integration of multiple advanced algorithms, providing robust tools for text analysis in multilingualsettings.

Keywords
Author Attribution, Forensic Linguistics,Cross- Linguistic Analysis, Machine Learning

1. Introduction

Identifying authors from their texts is a critical challenge in digital forensics, academic integrity, and literary analysis. As digital content proliferates, the ability to ascertain authorship with precision becomes imperative, especially in languages as linguistically and culturally diverse as English and Arabic. Traditional authorship attribution methods often struggle with the complexity and subtleties of multilingual texts, resultingin suboptimal accuracy and efficiency.

The advent of machine learning (ML) and deep learn- ing (DL) technologies offers transformative potential for addressing these challenges. These advanced computationalapproaches can adapt to and learn from the nuanced pat- terns of language use, significantly enhancing the accuracyof author identification. This paper introduces an innovative framework that leverages a combination of machine learn- ing algorithms—namely, Random Forest, Logistic Regression, k-nearest Neighbors, Support Vector Machines, and Naive Bayes—alongside deep neural network models, tailored specif- ically to improve author identification across English and Arabic texts.

The primary objectives of this research are as follows:

1) To develop a robust framework that integrates both machine learning and deep learning algorithms to im- prove author identification across English and Arabic texts.
2) To evaluate the performance of various algorithms, including Random Forest, Logistic Regression, k- nearest Neighbors, Support Vector Machines, and
3) To demonstrate the effectiveness of these algorithms through rigorous testing against benchmark datasets, aiming for high accuracy, precision, recall, and F1-score
4) To explore the practical applications of these ad- vanced computational techniques in enhancing lit- erary analysis, forensic linguistics, and plagiarismdetection.

Problem Statement : .

Our goal is to develop a machine learning or deep learning model that can accurately attribute authorship to texts across different languages, particularly focusing on English and Ara- bic. Given a set of documents, the objective is to identify the most likely author or a set of probable authors for each doc- ument. The outcome could be a system capable of attributing authorship to unknown

Manuscript received May 5, 2025
Manuscript revised May 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.5.6

texts or a better understanding of the challenges in cross-lingual author identification.

Design and implement machine learning or deep learn-ing models capable of learning the unique writing styles of different authors. Models could include traditional classifiers (e.g., SVM, Random Forests , Naive Bayes ,) or more complexarchitectures such as recurrent neural networks (RNNs), con- volutional neural networks (CNNs). for better context under- standing.Also Assess the model's performance using appropri-ate evaluation metrics such as accuracy, precision, recall, and F1-score. Perform cross-validation and potentially employ ad-ditional techniques like ensemble methods or transfer learning to enhance performance.

The paper makes several significant contributions to the field of authorship attribution across English and Arabic texts using machine learning and deep neural networks:

- **Comprehensive Framework:** Developed an efficient framework integrating five advanced machine learning and deep neural network algorithms (Random For- est, Logistic Regression, k-nearest Neighbors, SupportVector Machines, and Naive Bayes) for author identification.
- **Cross-Linguistic Analysis:** Addressed the challenges of authorship attribution in both English and Arabic, showcasing the ability to handle multilingual datasets effectively.
- **Enhanced Accuracy:** Demonstrated high accuracyrates, with certain models reaching up to 95% accu- racy in identifying authors, significantly outperform- ing traditional methods.
- **Algorithm Comparison:** Provided a detailed com- parative analysis of different algorithms, highlighting the strengths of Support Vector Machines and LogisticRegression in capturing nuanced writing styles.
- **Application Potential:** Illustrated the practical appli- cations of the proposed methods in forensic linguistics, plagiarism detection, and literary analysis, thereby enhancing security and academic research.
- **Contributions to Text Mining:** Advanced the field of text mining by developing and implementing robustand the hyper-parameters tuning. The experimental results are shown in section IV. Section V concludes the paper and provides some future works

2. Related Work

In the realm of Arabic opinion mining (AOM), numerous techniques for author identification have been explored. The primary challenge addressed is the large-scale identification of authors within extensive text collections such as emails and speeches, with the aim of pinpointing the authors of specific documents or confirming their authorship. Utilizing a range of datasets, including text categorization collections from Reuters and listserv discussion groups, researchers have employed di- verse methodologies such as stylometric analysis and machine learning techniques, including Bayesian multinomial logistic regression. Results demonstrate varying success, with errorrates between 0.02 to 0.08 depending on the combination of target and decoy authors. Through experimental validation, the efficacy of different representations for author identification was assessed, revealing that some machine learning models, particularly Support Vector Machines and Logistic Regression, offer strong performance by effectively capturing nuanced writing styles. The study significantly contributes to the fields of forensic linguistics and text mining by enhancing the accuracy and computational efficiency of authorship attributiontasks across multilingual texts. [1]

In addition to stylistic and probabilistic methods, advancedNatural Language Processing (NLP) techniques such as stem- ming, lemmatization, and stop words removal have been em- ployed to refine the process of author identification in Arabic texts. These methods aim to reduce words to their root forms, thus simplifying the textual data for further analysis [2].

Furthermore, the integration of Machine Learning classi- fiers enhances the precision of authorship attribution. Tech- niques like Support Vector Machines (SVM), Multinomial Na¨ıve Bayes (MNB), and Maximum Entropy (ME) classifiershave been utilized. SVMs are preferred for their computationalefficiency and ability to handle multi-class classifications witha linear kernel. MNB focuses on the frequency of word occurrences, whereas ME leverages the Principle of Maximum Entropy to consider dependencies among features, which is crucial for handling the complexity of Arabic texts.

The combined use of NLP techniques with machine learn- ing models has shown promising results, particularly the SVMclassifiers enhanced with unigram and bigram models andlemmatization, achieving an accuracy of 90.5%. This approach not only proves effective in author identification but also opens avenues for further research in author verification and background identification. The study highlights the limitedimpact of most NLP techniques except for lemmatization, which notably increases the accuracy of SVM classifiers by 1.5% [2].

In exploring the robustness of author identification meth- ods, two specific scenarios were examined using stylometric and content features to detect authorship deception [3]. The first scenario involved identifying blog authors, particularly those attempting to conceal their identities or impersonate others. The study achieved an impressive accuracy of 89% in identifying deceptive authors, with stylometric features alone reaching 88% accuracy, and content features alone at 83%. This method significantly outperformed other machine learning approaches, underscoring the effectiveness of combining stylo-metric and content features for detecting deceptive authorship

However, the challenges highlighted include the variedcomplexity of datasets. The blog dataset presented a more diverse and complex challenge in identifying distinct writing styles, whereas the imitation attack dataset, with its clearer distinctions, allowed for more straightforward identification.

The paper addresses the significant challenge faced bylaw enforcement agencies in tracing identities within the vast online landscape, particularly in cybercrime investigations[4]. The primary objectives are to develop a comprehensive framework for analyzing online messages to ascertain the most plausible author of an anonymous email and to craftdata mining models that classify and cluster emails based on undiscovered relationships.

The methodologies employed encompass statistical anal- ysis techniques, social network analysis, author identification via frequent pattern mining, and advanced classification and clustering strategies. The research utilizes the extensive Enron corpus, containing 619,446 messages from 158 users, to im- plement and test the proposed models. This dataset facilitates arich analysis of communication patterns and social interactions among users.

The results demonstrate that the framework is effective not only in identifying plausible authors of anonymous emails but also in classifying these emails into various categories or clustering them based on hidden relationships. Notably, the study finds that machine learning methods, which leverage a broad set of features, tend to outperform traditional statistical methods that often rely on more rigid mathematical models.

However, the paper does highlight certain limitations, particularly its focus solely on statistical and social network analysis modules without a thorough evaluation of the entire framework. It also notes a specific challenge in analyzingtexts shorter than 250 words, which presents a fundamental limitation in message-level analysis within email datasets.

The research by [5] tackles the complex challenge of multi-class text classification of Khmer news articles, aiming to effectively categorize them into multiple predefined categoriesusing machine learning techniques. This task is particularly challenging due to the unique characteristics of the Khmer language and the diversity of news content. The study's objectives include utilizing ensemble learning methods to im- prove classification accuracy, optimizing hyperparameters for machine learning classifiers, and evaluating the effectiveness of these methods in the Khmer language context. Methodologies employed range from data collection from various online news portals to sophisticated text preprocessing and the application of ensemble learning methods across multiple machine learn- ing models. The research utilized a dataset of Khmer news articles across nine major categories, achieving an averageclassification accuracy of 81% and a peak F1-score of 85.3%. However, the study notes potential limitations in its approach, particularly the assumption that each article fits only one category, which might not accurately represent the complexity of news content and could affect the generalizability of the results across other datasets and languages.

The study by [5] tackles the intricate challenge of multi- class text classification of Khmer news articles, aiming to categorize them effectively into multiple predefined categoriesusing machine learning techniques. This task is particularly challenging due to the unique characteristics of the Khmer language and the diversity of the news content. The primary objectives of the study are to utilize ensemble learning methodsin machine learning classifiers, optimize hyperparameters to enhance classification accuracy, and assess the effectiveness of these methods for news categorization in the Khmer language. Methodologically, the research involves data collection from various online news portals, applying advanced text prepro- cessing techniques, and using ensemble learning with a varietyof machine learning models. The dataset comprises Khmernews articles categorized into nine major categories, achieving an average accuracy of 81% and a peak F1-score of 85.3%. De- spite these achievements, the study acknowledges limitations such as the assumption that each article fits only one category, which could potentially underestimate the performance of the proposed approach and misrepresent the articles' contents. The paper also suggests the need for further exploration of the dataset limitations and the generalizability of the results to other datasets and languages.

The literature survey by [6] addresses the vulnerabilities and potential malicious uses of machine-based automatic text generation in adversarial machine learning, focusing on the risks such as politically manipulated texts that could influence public opinions. The main objectives of the survey are to sum- marize recent literature in this field, provide an introduction and comprehensive resource for researchers, and outline future research directions and potential improvements. Methodolog- ically, the survey employs a top-down ontological approachto identify key themes, trends, and challenges, reviewing and analyzing recent studies to determine the state of the art and future directions. The datasets discussed include those usedin Generative Adversarial Networks (GANs) applications like MS COCO Image Caption and EMNLP2017 WMT, addressing challenges in creating and evaluating text generation mod- els. The survey highlights ongoing growth in text generation applications and associated adversarial attacks, noting trends such as the use of pre-trained models and development of universal text perturbations. It calls for further research in areas like GAN text generation, development of evaluation metrics, and adversarial training strategies. A notable limitation of the survey is its focus on literature from the Scopus database, which may not encompass all relevant research in the field.

The survey conducted by [7] explores the complexities of author profiling on social networks, addressing the challengeof analyzing unstructured and anonymous textual data to identify demographic and psychological aspects of authors. The objectives of the survey are multifaceted, including pro- viding an overview of author profiling processes, proposing a taxonomy of features for profiling, presenting main classifi- cation techniques, and analyzing the most relevant literatureto equip researchers with effective tools for this domain. Methodologically, the survey involves an extensive literature review using databases such as Google Scholar and IEEEXplore, analyzing 650 relevant articles out of an initial 1020to address specific research questions about author profiling. The datasets used include the Enron corpus and social media data, with the survey presenting a detailed analysis of various studies, highlighting the effectiveness of different profiling methods and discussing the challenges like the need for better distinction between fake and authentic profiles. Results indi- cate varying successes in profiling accuracy, effectiveness of feature extraction methods, and the importance of developing multilingual datasets to overcome monolingual limitations. The survey identifies a need for broader research inclusionand addresses the underrepresentation of non-English studies, emphasizing the importance of enhancing dataset diversity to improve the general applicability of author profiling tools in social media contexts.

The research by [8] focuses on developing a system for author identification and plagiarism prevention in Devanagari script literature, particularly Marathi texts. The project aimsto enhance the accuracy of authorship determination using machine learning techniques. The objectives include imple- menting machine learning algorithms such as SVM, Na¨ıve Bayes, and K-NN to identify authors accurately and refine these algorithms by removing stop-words to improve the sys- tem's precision. The methodologies involved separating data into training and testing sets, removing stop-words for better feature extraction, and employing these algorithms to predict authorship based on the stylistic nuances of the texts.

While the paper does not provide specific numerical results, it discusses the comparative effectiveness of each algorithm in identifying authors, suggesting an iterative approach to optimizing accuracy. However, the study faces limitations suchas insufficient detail on the feature extraction process and the lack of detailed numerical results which might affect the assessment of the system's effectiveness. Additionally, thefocus on Marathi literature using the Devanagari script raises questions about the generalizability of the proposed system to other languages or scripts. The paper concludes with a call for further research to refine these techniques and enhance the general applicability of the system across various linguistic contexts.

The study by [9] addresses the challenge of accurately iden-tifying authors and genres in Turkish news texts by developing and applying various machine and deep learning algorithms. The primary objectives were to conduct a modeling study using algorithms like Multinomial Na¨ıve Bayes (MNB), Random Forest (RF), Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM), create large-scale, multi-class datasets for this purpose, and evaluate the performance of these models. Thirteen datasets were created and labeled according to genre and author, facilitating extensive testing through 10- fold cross-validation. The results were promising, with the best author identification models using RF achieving accuracies up to 85.17%, and LSTM performing best for genre identificationwith accuracies as high as 96.73%. However, the study notes challenges such as potential biases in dataset collection and calls for more detailed comparisons of model strengths and weaknesses. Additionally, the generalizability of these findingsto other languages or genres remains undiscussed, and theavailability of the code on platforms like Kaggle is mentioned without providing access details, which could improve the study's transparency and reproducibility.

The study by [10] explores the effectiveness of shallow parsing and machine learning in authorship attribution, fo-cusing on syntax-based and token-based features to predict authors of texts. Using a corpus of newspaper articles from the Belgian newspaper De Standaard, the research investigates the potential of differentiating authors in the context of national current affairs articles. The methodology involves extracting various syntactic, token-based, and lexical features, applying machine learning algorithms like TiMBL and WEKA to de- velop implicit profiles that characterize an author's style. The results demonstrated a promising 72.6% accuracy in authorship prediction, comparing favorably with other studies in the field and suggesting that the combination of syntax-based features can be as effective, if not more, than traditional lexical and token-level features. However, the study also notes potential issues such as the modification of texts by editors, which could impact the authenticity of authorial styles in the corpus. Additionally, the paper calls for further research on the integra- tion of automatically extracted features with machine learning and the application of these methods to various text types, highlighting the need for broader validation of the techniques used.

The study by [11] tackles the complex task of manuscript writer identification, particularly challenging due to the vari- ability in handwriting styles and document degradation over time. This paper critiques the limitations of traditional letter or word recognition methods and explores deep learning'spotential, albeit its dependency on substantial labeled data. The methodology encompasses a three-stage system: text line detection using object detectors, classification via deep neural networks, and writer identification through a weighted ma- jority vote mechanism employing transfer learning and data augmentation via a handwriting synthesis tool. Tested on the Avila Bible dataset, comprising 749 pages with annotations foreight writers, the system achieved a peak accuracy of 96.48% and an F1 score of 96.56%, outperforming traditional and some transfer learning methods but not those utilizing global featureswith larger datasets. Future improvements could include en- hancing text line detection accuracy, exploring different meta- architectures for classification, and adding a reject option to boost precision. The study also acknowledges the constraints posed by small training datasets and the adaptation challenges to manuscripts of varying styles and languages.

The paper by [12] addresses the challenges of authorship attribution in social media, specifically focusing on Twitter messages, through the lens of stylometry and machine learning. The objective is to analyze tweets using stylometric tech-niques—specifically, extracting frequencies of various featureslike function words, lexical traits, and syntactic patterns—to ascertain authorship. The methodology involves using recur- rent neural networks (RNNs) and committee machines, which aggregate outputs from multiple RNNs to enhance the accuracyof authorship determination. The dataset comprises 17,000tweets from 34 different authors, with the study achieving an author attribution accuracy ranging between 70% and 80%. The research highlights a transition from traditional statistical methods to more robust machine learning approaches in han- dling the complexity and brevity of tweets. Despite successes, the paper identifies significant challenges, such as the short and informal nature of tweets which complicates traditional stylometric analysis, and the scarcity of standardized data for validating these methods. This study underscores the evolving nature of authorship attribution in adapting to the constraints of modern communication platforms like Twitter.

The study by [13] investigates authorship attribution for news articles using the Reuter-50-50 dataset, aiming to eval- uate the efficacy of various algorithms and new stylometric features, including the use of POS tags. The methodology employs multiple algorithms such as LibLINEAR SVM, SMO, Logistic Regression, Na¨ıve Bayes, and J48 Decision Tree, with LibLINEAR SVM selected based on its superior performance and compatibility with the dataset characteristics. Evaluation through 10-fold cross-validation on the development set revealsthat linear models like SVM excel, demonstrating that the dataset is linearly separable, whereas Na¨ıve Bayes underper- forms due to its inappropriate assumption of feature indepen- dence. Despite achieving high accuracy and kappa statistics, the study acknowledges limitations such as its sole focus on English texts, restricted evaluation metrics, and challenges in generalizing findings to shorter texts. Future directions include broadening the evaluation scheme and exploring the applicability to diverse text formats and languages.

The study by [14] focuses on developing methods for author identification in emails, a crucial task in the realmof cybercrime investigations. The paper sets out to enhance authorship attribution techniques for emails, leveraging fre- quent pattern mining and a variety of stylometric features, with the FP-Growth algorithm central to the methodology.Objectives include creating a model to identify the most plausible email authors, extending the use of frequent pattern mining, and automating the analysis of email ensembles for forensic investigation. Despite the comprehensive methodology involving data collection, feature extraction, and model gener- ation, the paper does not specify the datasets used, nor does it provide empirical results or detailed measures for evaluating the model's effectiveness. The study discusses potential appli- cations and the adaptability of the model to various cybercrime scenarios but acknowledges limitations like the absence ofempirical findings, the need for a broader evaluation scheme, and the challenge of short email texts. Future work could aim to validate the proposed model with real-world data, compare it against existing methods, and explore its ethical and privacy implications in forensic settings.

The research by [15] addresses the challenge of online anonymity and its misuse, proposing a variable length char- acter n-gram framework combined with a genetic algorithm for effective writeprint identification. This approach is aimed at enhancing social accountability by identifying the authors of online messages. The methodology includes collecting on- line messages, employing variable length n-grams for feature extraction, and using genetic algorithms for optimal feature selection. The approach is validated on a dataset of Amazon customer reviews, involving 20 participants, to assess the effectiveness of the proposed system in distinguishing betweenauthors based on their writing styles. The results indicate significant performance improvements; the genetic algorithm-based feature selection method outperformed the baseline and traditional methods like Information Gain, reducing feature dimensionality considerably and improving accuracy by up to 15% compared to no featureselection. The variable length n-gram method proved superior to fixed length in scenarios with fewer than 6,000 featuresbut showed diminishing returns at higher dimensions. Despite these promising results, the study faces limitations such as the small dataset size, lack of broader comparisons with other state-of-the-art methods, and potential generalizability issues due to the specific nature of the dataset used. Future research could expand on these findings by incorporating largerand more diverse datasets, comparing more varied methods, and using additional evaluation metrics to provide a more comprehensive validation of the approach.

Algorithm 1 Text Processing and Machine Learning Pipeline

```
procedure DATA COLLECTION
    Collect English and Arabic text files from specified directory.
    Process each file, including '.properties' files and corresponding text files.
end procedure
end procedure
procedure DATA PREPROCESSING
    Convert text data to lowercase.
    Remove non-alphabetic characters.
    Adjust 'author' column for compatibility with the classifier.
end procedure
end procedure
procedure FEATURE ENGINEERING
    Employ TF-IDF vectorization for feature representation.
    Perform feature scaling using MinMaxScaler.
end procedure
end procedure
procedure LABEL ENCODING
    Apply label encoding to convert author labels to numerical format.
end procedure
end procedure
procedure DATA SPLITTING
    Split data into training, validation, and testing sets.
    train_test_split with proportions: 70% train, 15% validation, 15% test.
end procedure
end procedure
procedure OPTIMIZATION PARAMETERS
    Explore parameters using GridSearchCV for:
        Multinomial Naive Bayes.
        Logistic Regression.
    Incorporate Adam and SGD optimizers in the training loop for DL models.
end procedure
end procedure
procedure CLASSIFICATION
    Utilize multiple ML classifiers: Naive Bayes, Logistic Regression, K-NN, Random Forest, SVM.
    Train and evaluate each ML model for accuracy.
    Utilize a Multi-layer Perceptron (MLP) for deep learning.
    Train and evaluate MLP using both Adam and SGD optimizers.
end procedure
end procedure
procedure PREDICTION AND EVALUATION
    Test models and generate:
        Confusion Matrix.
        Classification Reports.
    Calculate Evaluation Metrics (Accuracy, Recall, Precision, F1-score).
end procedure
end procedure
procedure VISUALIZATION
    Plot relationship between number of estimators and testing accuracy for Random Forest.
    Plot training and validation curves for optimization algorithms.
end procedure
end procedure
```

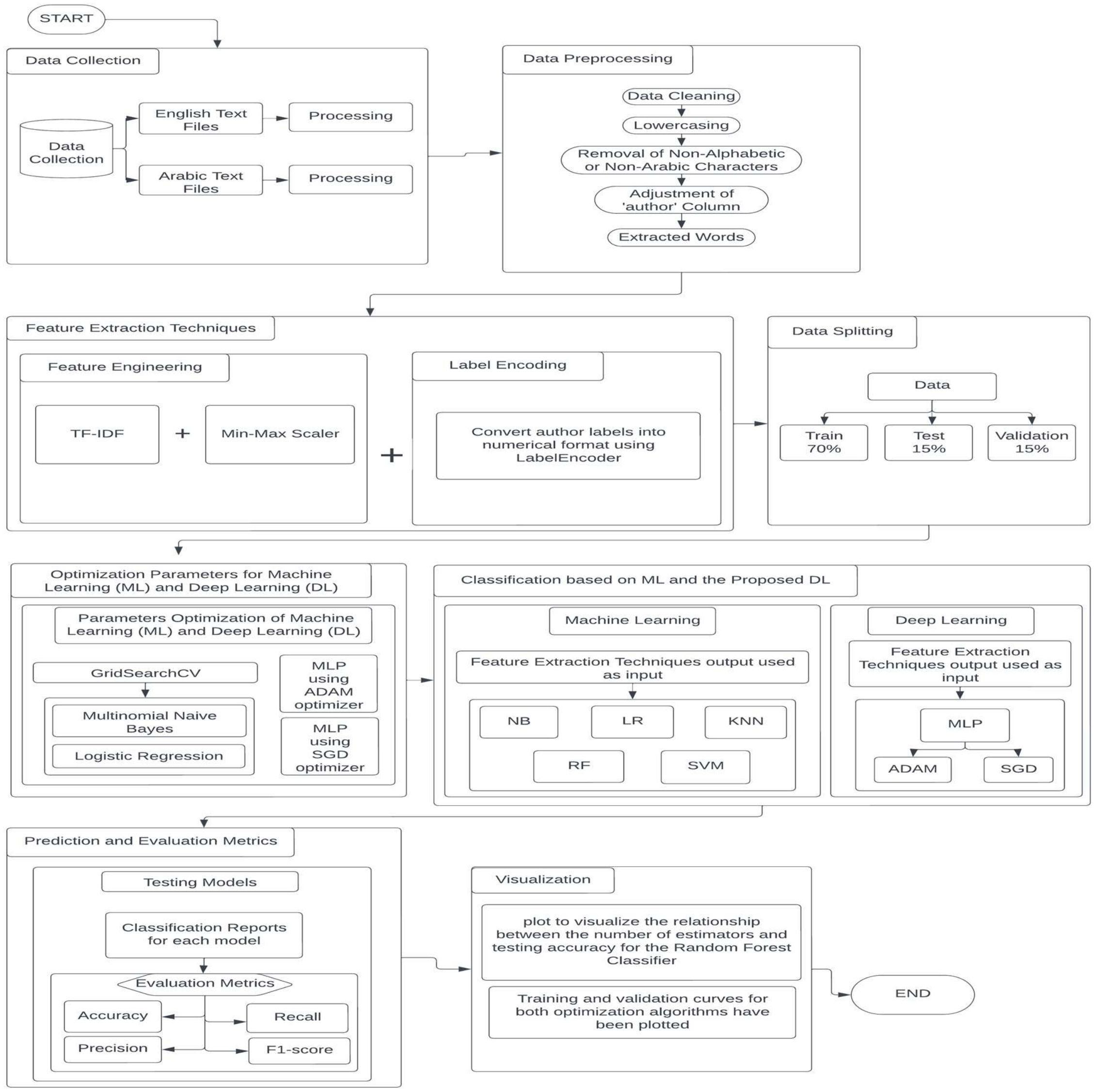
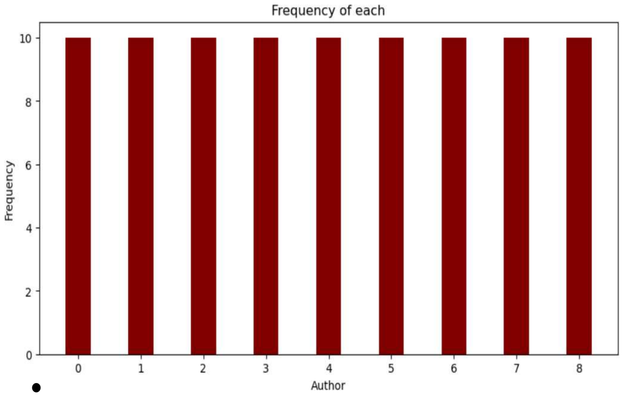

START
Data Collection
Data Collection
English Text Files
Processing
Arabic Text Files
Processing
Data Preprocessing
Data Cleaning
Lowercasing
Removal of Non-Alphabetic or Non-Arabic Characters
Adjustment of 'author' Column
Extracted Words
Feature Extraction Techniques
Feature Engineering
TF-IDF
+
Min-Max Scaler
+
Label Encoding
Convert author labels into numerical format using LabelEncoder
Data Splitting
Data
Train 70%
Test 15%
Validation 15%
Optimization Parameters for Machine Learning (ML) and Deep Learning (DL)
Parameters Optimization of Machine Learning (ML) and Deep Learning (DL)
GridSearchCV
Multinomial Naive Bayes
Logistic Regression
MLP using ADAM optimizer
MLP using SGD optimizer
Classification based on ML and the Proposed DL
Machine Learning
Feature Extraction Techniques output used as input
NB
LR
KNN
RF
SVM
Deep Learning
Feature Extraction Techniques output used as input
MLP
ADAM
SGD
Prediction and Evaluation Metrics
Testing Models
Classification Reports for each model
Evaluation Metrics
Accuracy
Recall
Precision
F1-score
Visualization
plot to visualize the relationship between the number of estimators and testing accuracy for the Random Forest Classifier
Training and validation curves for both optimization algorithms have been plotted
END

Fig. 1. The Proposed System of Author Identification (Arabic and English).



•

Fig. 2. Frequency of Each Author in Arabic Dataset

3. Methodology

Figure 1 illustrates the proposed framework for English and Arabic author attribution, which contains mainly ten significant steps:

B. Data collection

The data collection process for this research aimed at constructing a robust dataset for solving the author attribution problem. The objective was to curate a diverse and repre-sentative collection of texts to enable effective training and evaluation of the authorship attribution model.

Victorian Era Authorship Attribution Dataset: This Dataset has 93600 instances, with 1000 wordsin each instance. Authors were restricted to those whowrote in English and had a substantial body of work, with a minimum of 5 books available in the database.Authors considered for inclusion were required tohave been active during the 19th century, aligning with the Victorian Era. They enhanced the dataset uniformity and reduce noise, the top 10,000 words occurring across the entire corpus of texts from the45 authors were selected. Any words outside this set were removed, preserving the structural integrity of the sentences. The books were then segmented into text fragments, each comprising 1000 words.

A dataset for Arabic author identification : This dataset has 9 authors and 10 articles for each author. So it totals 90 instance, it is a small dataset. This datais stored as text file so in order to begin processing them we have to write this dataset into a CSV file.

C. Data Preprocessing

Data preprocessing phase plays a pivotal role in shaping the effectiveness of the authorship attribution model by refiningraw textual data into a format suitable for analysis. This sectionoutlines the key steps taken to clean, standardize, and prepare the dataset for training and evaluation. Many preprocessing techniques were used before training. The techniques that wereused are Data Cleaning, Removing stop words, Tokenization.

Phase 1: Data Cleaning: In this phase, we removed all special character like ($, #, $, , etc.), removed all digits also. We kept only the characters, Arabic char- acters in the Arabic dataset, and English characters in the English dataset.

Phase 2: Removing Stop Words Stop words are common words that frequently occur in a languagebut often carry little semantic meaning. Examples include articles (e.g., ”the,” ”a,” ”an”), conjunctions (e.g., ”and,” ”but,” ”or”), and prepositions (e.g., ”in,” ”on,” ”at”). Stop word removal directs the model’s attention towards content words—those carrying morespecific and meaningful information.

Phase 3: Tokenization In this Phase, the collected datasets are tokenized by splitting each tweets into tokens which are basically a piece of text formed in a smaller units.

D. Applying Feature Extraction techniques for ML and Pro-posed DL Models

In this step, we implemented Feature Extraction techniquesfor Standard ML models, DL algorithms as shown in the following subsections:

1) *Feature Extraction Method :* In this step, the Frequency- Inverted Document Frequency (TF-IDF) Feature ExtractionTechnique. It is a numerical statistic that’s used in natural language processing and information retrieval to assess a word’s significance in a document in relation to a corpusof texts. Information retrieval, document grouping, and text analysis are among its frequent uses. We ran multiple trials using different maximum number of features. After the trials, 1000 maximum features as a limit was used for machine learning and deeplearning was trained without limit The weight

$$IDF\,(i,j) = \log \frac{\text{Total number of document in the datasets}}{\text{Number of document which include } i \text{ term}} \qquad (2)$$

$$W(i,j) = TF\,(i,j) \times \mathrm{IDF}(i,j) \qquad (3)$$

Where $TF\,(i,j)$ is the frequency of term i in review j,$IDF\,(i,j)$ is the frequency of feature with respect to all reviews. Finally, the weight of feature i in review j , $W\,(i,j)$is calculated by Eq.(3)

E. Data splitting

In this step, the data was split into training, testing and validation sets (75% for training, 15% for validation, 15% for testing).

E. Hyperparameters Optimization Methods

A crucial step in machine learning is hyperparameter tuning, which is choosing a model's ideal values. Hyperparameters are those parameters that are predetermined by the practitioner rather than being learned during training. These values play a major role in a model's performance.

Therefore, in order to get the best performance out of a model, it is imperative that these parameters be tuned. By adjusting these parameters, a model's accuracy, generalizationcapacity, and reduction of overfitting can all be enhanced. It guarantees that the model is most appropriate for resolving thecurrent issue in this way.

1) Hyperparameters Optimization Methods for standardML techniques :

Using Grid search we used this algorithm because it is a straightforward and thorough search method that looks for the best combination by trying every possible combination. Although this approach may incur significant computational costs, it can ensure theoptimal hyperparameters within a specified range.

In Naive Bayes : We used grid search to get the optimum values for 1.alpha which is a smoothing parameter that is added to the term frequencies to avoid the problem of zero probabilities. In the case of text classification using a Multinomial Naive Bayesmodel, it is used to handle words that do not appear incertain documents. A small alpha value (like 0.0001 that we used) means stronger smoothing. 2. Fit Prior: This is a boolean parameter. If fit-prior is set to True, the class prior probabilities will be adjusted based on the training data. If set to False, no adjustment will be made, and all classes will be presumed to have a prior probability proportional to the number of samples in them. we used false which indicates a uniform prior. for example By tuning the hyperparameters we reached 79 % accuracy instead of 76 % in the English dataset.In the Arabic dataset accuracy reached 85.7 %.

In KNN : We used grid search to get the optim-ium values for 1.n neighbors: Which is the numberof neighbors to consider when making predictions. In your case, n neighbors=5 means that the modelwill consider the labels of the 5 nearest neighbors 2. weights: This parameter determines how the contributions of neighbors are weighted. In your case, weights='distance' means that closer neighbors will have more influence than farther neighbors, and the weight is inversely proportional to the distance. 3.metric: This specifies the distance metric used to mea-sure the distance between instances. In our case, metric='euclidean' indicates that the Euclidean distanceis used. By tuning the hyperparameters we reached 59.5 % accuracy instead of 54.5 % in the English dataset and when we made the same optimization in the Arabic dataset we got 64%.

In Random Forest : We used grid search to get the optimum values for 1.max depth: This parameter controls the maximum depth of each tree in the forest.A higher max depth allows the trees to grow deeper and capture more complex patterns. In your case, max depth=70. 2. n estimators: This parameter deter- mines the number of trees in the forest. Increasing the number of trees can lead to better generalization. In your case, n estimators=400.By tuning the hyperpa- rameters we reached 63.3 % accuracy instead of 61.5% in the English dataset and when we made the same optimization in the Arabic dataset we got 92 % insteadof 78.5 %.

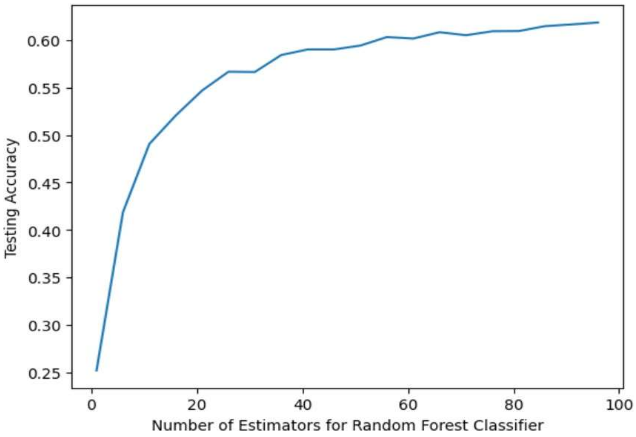


Fig. 3. Increasing Estimators vs testing accuracy

The relationship between the number of estimators (trees) in a Random Forest Classifier and testing accuracy is often illustrated by an increase in accuracyas the number of estimators increases. This behavior is known as the ”ensemble effect,” and it’s a key char- acteristic of ensemble learning methods like Random Forests.

As you increase the number of estimators in a Random Forest, the model becomes more robust to noise and outliers. Each tree in the forest is trained on a random subset of the data and makes decisions independently and become more general.

In Logistic Regression the solver is optimized using different numerical optimization algorithms, and the solver parameter specifies which algorithm to use. In our case, solver=’saga’.’saga’: Stochastic Average Gradient Descent. It’s a variant of the gradient descent optimization algorithm that incorporates variance reduction techniques. It is particularly suitable for large datasets and supports both L1 (Lasso) and L2 (Ridge) regularization.By changing the solver We reached 91.4% accuracy inthe English dataset and 71 % instead of 64 % in the Arabic dataset.

In SVM 1. C: This parameter controls the trade-off between having a smooth decision boundary and classifying the training points correctly. A smaller value of C makes the decision boundary smoother, while a larger C aims to classify all training points correctly. 2. gamma acts as a regularization parameterin the SVM algorithm. It controls the shape of the decision boundary and can significantly impact the performance of the SVM on different datasets. 3. kernel : which is the type of kernel function we will use. we used Radial basis function (RBF) kernel and C=1 which make an impressive result in the English dataset by accuracy 92 %. In addition to trying to scalethe x test , x train features using Standard •Scaler before training a Support Vector Machine (SVM) model can lead to higher accuracy (92.4 %) because SVM is sensitive to the scale of features. Features withlarger scales might dominate those with smaller scales during the training process, leading to a biased model.Scaling ensures that all features contribute equally to the decision-making •process , and can lead to faster convergence.

2) Hyperparameters Optimization Methods for the pro- posed DL models :

In this phase, the neural network wasoptimized by running multiple trials using different numberof layers, different optimizer, and different learning rates. We tried a one layered model, two layers model, and a three layers model, the one layer model was underfitting while the two layers model was best and the three layers model was overfitting the data. In figure 4 and 5, learning curves are shown, using different optimizers (Adam, SGD with Nestrov).Until we came with the best results with 2 layers network, Adam optimizer and using a learning rate of 0.0001.

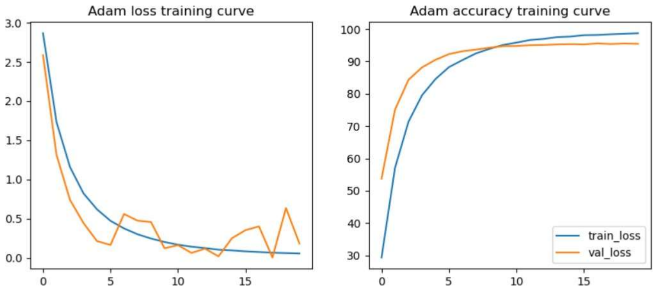


Fig. 4. Adam Optimizer Learning Curves

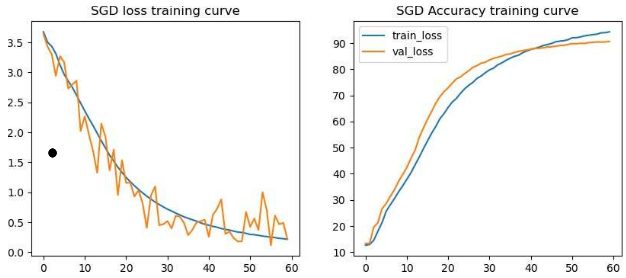


Fig. 5. SGD Optimizer Learning Curves

F. Classification based on ML models

In this step, six regular ML algorithms have been used including (DT), (RF), (KNN), (LR), (SVM) , and (NB) to classify Arabic opinion Tweets datasets.

Random Forest (RF) A classifier that uses multiple decision trees on different subsets of a given dataset and averages the results to increase the predicted accuracy of that dataset. The random forest gathersthe results from each decision tree and bases its expectation of the result on the majority votes of the predictions, as opposed to relying solely on one decision tree.

Naive Bayes (NB) The Naive Bayes algorithm is a probabilistic machine learning algorithm commonly used for classification tasks, especially in natural language processing (NLP) and text classification. Despite its simplicity, Naive Bayes often performs well and is particularly effective for large datasets.

$$P(A|B) = P(B|A).P(A)/P(B) \quad (4)$$

In the context of Naive Bayes: 1. P(A—B) is theprobability of the author being A given the text features B. 2. P(B—A) is the probability of the text features B given the author is A. 3. P(A) is the prior probability of author A. 4. P(B) is the probability of observing the text features B.

3. Classification Decision: To classify a new text into a particular author, the algorithm calculates the posterior probability for each author and selects the author with the highest probability.

Logistic Regression A logistic function, also calleda sigmoid function, is a supervised machine learning algorithm, used to produce a probability value between0 and 1 based on inputs that are independent variables. As an illustration, there are two classes: Class 0 and Class 1. An input is classified as Class 1 or Class 0if the logistic function value for it is greater than the threshold value of 0.5. Since it is a continuation of linear regression and is primarily applied to classifi- cation problems, it is known as regression.The output of logistic regression predicts the likelihood that an instance will belong to a specific class or not, whereasthe output of linear regression is a continuous value that can be anything. There are three categories for logistic regression: ordinal, multinomial, and binary. Both in theory and in practice, they are different. The two possible outcomes in binary regression are yes andno. When there are three or more values, multinomial logistic regression is employed which we used here inour problem.

K-nearest-neighbors (KNN) The k-Nearest Neigh- bors (KNN) algorithm is a simple yet effective ma- chine learning algorithm used for both classification and regression tasks. It belongs to the family of instance-based learning, where predictions are made based on the similarity of new data points to existing data points in the training set. KNN is particularly useful when dealing with smaller datasets and prob- lems where instances of similar classes tend to cluster together.

$$d_{\text{euclidean}} = \sqrt{\sum_{i=1}^{k}(x_{2i} - x_{1i})^2} \quad (6)$$

$$d_{\text{manhattan}} = \sum_{i=1}^{k}|x_{2i} - x_{1i}| \quad (7)$$

G. Association Rule Mining Algorithms :

Association rule mining algorithms like Apriori and FP-Growth are not suitable for text classification problems like author attribution. Both algorithms usually applied to transactional datasets, where there are multiple transactions and each transaction lists the items that apply to it. When applied to text classification, where the objective is to predict authors based on individual documents, several challenges appear that show that this type of algorithms isn't suitable for

Individual Document Treatment: Apriori and FP-Growth are designed for transactional datasets, assum-ing associations between items within transactions. In text classification, each document is treated as a separate entity, and the algorithms would struggle to understand meaningful relationships between words across different documents.

Unsupervised Nature: Both Apriori and FP-Growth are unsupervised learning algorithms focused on dis- covering patterns or associations within the data. Author attribution, however, is a supervised learning problem where the goal is to predict authors basedon labeled training data. These algorithms do not usually handle the supervised learning aspect and lack the capability to utilize labeled data for making predictions.

Sparse and High-Dimensional Data: Text data, es- pecially after feature extraction using techniques like TF-IDF, often results in high-dimensional and sparse feature matrices. Apriori and FP-Growth may not efficiently handle such high-dimensional data, and their performance might be compromised. Traditional supervised learning algorithms like Naive Bayes, Lo- gistic Regression, and Support Vector Machines are better equipped for handling sparse text data.

Interpret-ability and Relevance: Apriori generates association rules, and FP-Growth focuses on frequent itemsets, which may not provide direct insight fortext classification tasks. The relationships between individual words or features might not be as relevant or interpretable in the context of predicting authors in text data.

In summary, both Apriori and FP-Growth are better suited for mining associations in transactional datasets, where items co- occur in sets of transactions. However, for text classification problems like author attribution, traditional supervised learningapproaches, such as Naive Bayes, Logistic Regression, SupportVector Machines, or neural networks, are more appropriate. These algorithms are designed to handle the unique character- istics of text data and are well-suited for the goal of predicting labels based on feature representations of the text.

H. The proposed DL models :

The proposed deep neural network architecture is a regularDeep neural network with different hidden layers sizes for eachdataset:

Arabic Dataset: The neural network architecture comprises three layers: a 1024-unit input layer, fol- lowed by a 64-unit hidden layer, and a final output layer with the number of units corresponding to the classes in the dataset. Each hidden layer is equipped with a ReLU activation function to introduce non- linearity. Dropout regularization with a probability of 0.5 is applied between the layers to prevent overfitting and enhance generalization.

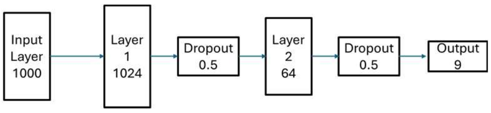


Fig. 6. Arabic Dataset NN Architecture

2) Victorian Era Dataset: The architecture consists of three layers: an expansive 1024-unit input layer, a compressed 64-unit hidden layer, and a final output layer with the number of units corresponding to the classes in the dataset. Each hidden layer employs the ReLU activation function, ◦ facilitating non-linearity in the model's representations. To enhance generaliza- tion and combat overfitting, dropout regularization with a dropout probability of 0.5 is applied between the layers.

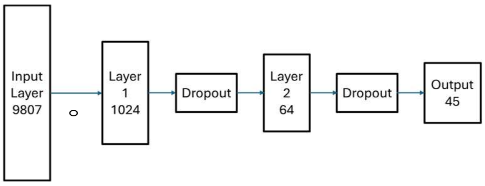


Fig. 7. English Dataset NN Architecture

These models are trained using the Adam optimizer with a learning rate of 0.0001 and weight decay set to 1e-5. Using 32 as a batch size. The training process spans 30 epochs, during which the model aims to minimize the cross-entropy loss. Comprehensive metrics, including accuracy, are monitored to assess the model's performance on both the training and validation sets. But in the Arabic dataset the hyperparameters was similar to the English dataset except the batch size wasthe size of all training dataset and it was trained on 40 epochs.

I. *Performance metrics*

Four standard performance metrics; Accuracy (ACC), Pre- cision (PREC), Recall (REC), and F1-score (F1) are used to evaluate the performance of the proposed models. They are calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

4. Experimental Results and Discussion

To examine the efficiency of Machine Learning (ML methods) and Deep Learning models (DL) over Arabic and English text Mining to predict the author, we have assessed two datasets: Victorian Era Authorship Attribution, A dataset for Arabic author identification dataset using strategy of learning including Holdout validation (70% in training and 30% intesting).

A. Case Study I

Random Forest

Results: Achieved accuracy of 0.615 for the English dataset and 0.666 for the Arabic dataset, indicating good performance for the model.
Rationale for Usage: The decision to use Random Forest was based on its efficiencywith categorical data, its ability to address overfitting, and its superior performance even with large datasets and missing data. Further- more, Random Forest aggregates multiple de- cision trees to enhance model accuracy throughmajority voting.

Naive Bayes

Results: For the English text dataset, accuracywas initially 0.761, later adjusted to 0.794. Forthe Arabic dataset, the accuracy was around 0.857.
Rationale for Usage: Naive Bayes was chosen for its efficiency in text classification tasks due to its computational simplicity and its robust performance across large volumes ofmultilingual data.

Support Vector Machine (SVM)

Results: SVM achieved an accuracy of ap-proximately 0.924 on the English dataset and 0.643 on the Arabic dataset.
Rationale for Usage: SVM is favored in scenarios with high-dimensional feature spaces due to its capability in handling large feature sets effectively, its resilience against overfit- ting, and its utilization of the kernel trick to manage non-linear data.

Logistic Regression

Results: On the English author dataset, accu- racy reached 91.4% using the saga Solver. For the Arabic dataset, accuracy started at 64% andimproved to 71.4% after solver adjustments.

K-Nearest Neighbors (KNN)

Results: KNN achieved 59.5% accuracy in the English dataset and 64.3% in the Arabic dataset.
Rationale for Usage: The method was tested despite its sensitivity to the curse of dimen- sionality and its dependency on the discrimi- native power of the neighbors, which can be a challenge in author attribution tasks.

Tuning of the Algorithm: Adjusting the number of neighbors and experimenting with distance-weighted approaches improved KNN's performance, highlighting the impor- tance of parameter tuning in achieving optimal results.

B. Deep Neural Network Results

Results

Using a deep neural network on Arabic and English dataset, we had 90 documents in the Arabic dataset and 53678 documents in the English dataset. As anticipated, the deep learning is performing much better than most of the machine learning models. It performed on the English dataset with validation accuracy of 95.8% and testing accuracy of 95% as shown in figure 8 and 95% as a final precision, f1 score and recall. A confusion matrix was plotted in figure 10. In the other hand it performed on the Arabic dataset with validation accuracy of 92.9% and with 100% as a testing accuracy as shown in figure 9 and 100% precision, 100% recall and 100% f1-score. A confusion matrix was plotted in figure 11.

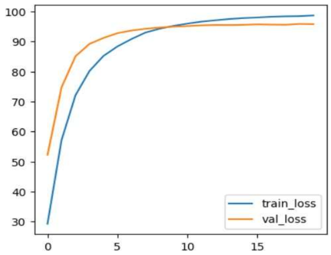


Fig. 8. Accuracy on English Dataset

Fig. 10. Confusion Matrix for English Dataset

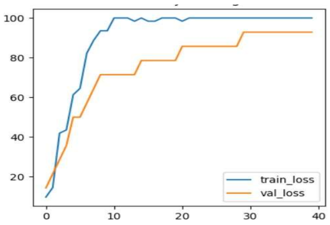


Fig. 9. Accuracy on Arabic Dataset

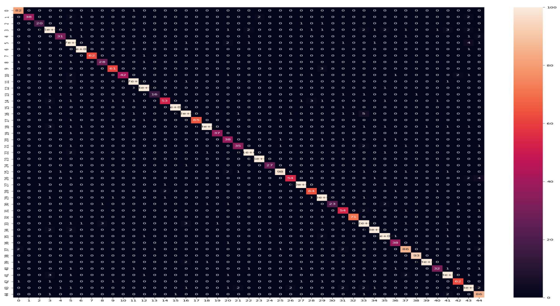


Fig. 10. Confusion Matrix for English Dataset

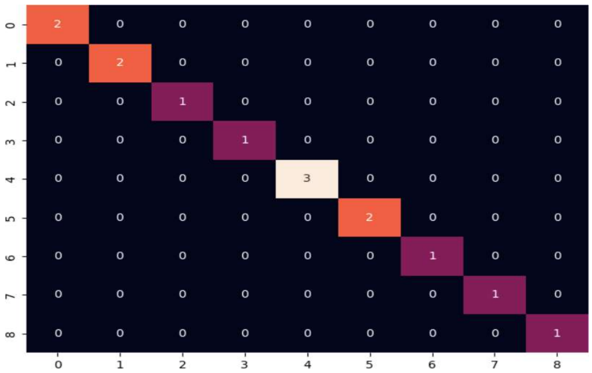


Fig. 11. Confusion Matrix for Arabic Dataset

Why we use this algorithm? Neural networks was used due to their ability to extract complex patterns and features from text data. This means they can analyze and extract patterns from the writing style of an author, such as sentence structure, vocabulary, and grammatical nuances.

To sum up, the researches in author identification is important in many domains, especially when it comes to languages that aren't as similar as English and Arabic. Criminal investigations are aided by the capacity to identify the writers of anonymous communications, especially when threatening emails or ransom notes are involved. Determining

TABLE I. Performance Metrics for English Dataset afteroptimization

2*Algorithm	Metrics			
(lr)2-5	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Random Forest	0.63	0.49	0.85	0.43
Naive Bayes	0.793	0.77	0.77	0.79
Support Vector Machine	0.924	0.85	0.87	0.84
Logistic Regression	0.914	0.90	0.93	0.88
K-nearestNeighbor	0.595	0.51	0.67	0.48
Neural Networks	0.95	0.93	0.94	0.93

The objective of creating a deep learning or machine learning model for cross-lingual author identification, withan emphasis on Arabic and English, is in line with recent technical developments. The discipline has undergone a revolution because of machine learning, particularly deep learning as it is well-suited for applications like author identification because it concentrates on multi-layered neural networks, which enable it to handle complex data with proficiency.

We went through different algorithms which proved their efficiency for solving our problem using ML models like SVM , Random Forest and logistic regression got very impressive accuracy between 65 to 92 % on both Arabic and English datasets but deep neural networks proved to be much better using fully connected layers and dropout got us95 % accuracy on English dataset and 100 & on the

Arabic dataset which is very beneficial and away from sophisticated architectures like CNN and RNN techniques. Assessing and enhancing the model's efficacy will require performance evaluation utilizing metrics like accuracy, precision, recall,and F1-score in addition to approaches like ensemble methods and cross-validation.

This comprehensive approach to author identification, which incorporates cutting-edge machine learning techniques, computational methodologies, and linguistic analysis, reflects the interdisciplinary character of the undertaking.

5. Conclusion and Future Results.

The exploration of machine learning and deep neural networks for author identification across English and Arabic texts has demonstrated significant advancements in text mining capabilities. The implementation of five distinct algo- rithms—Random Forest, Logistic Regression, k-nearest Neigh-bors, Support Vector Machines, and Naive Bayes—augmented with deep learning techniques, has effectively addressed the complex challenges of multilingual authorship attribution.

The comparative analysis of these methods revealed that Support Vector Machines and Logistic Regression notably excelled in identifying nuanced writing styles, thereby confirm- ing their robustness and suitability for complex text analysis tasks. The application of these algorithms on two specific datasets, the Victorian Era Authorship Attribution and an Arabic author identification dataset, highlighted their efficacy, with a remarkable accuracy of up to 95%.

This research underscores the potential of advanced com- putational techniques to enhance the precision and efficiency of author identification processes. Such improvements are cru- cial for applications spanning forensic linguistics, plagiarism detection, and broader literary studies, where accurate author attribution is paramount.

The success of these techniques in handling the linguistic and stylistic complexities of English and Arabic texts pavesthe way for future research to explore further enhancements and to apply these methodologies across even more diverse languages and textual formats.

Ultimately, this study contributes significantly to the field of text mining, offering a robust framework for effectively tackling the challenges of authorship attribution in a multilin- gual context. It also sets a precedent for future technological advancements in the security and academic sectors, where the accurate identification of authorship is increasingly vital.

References

[1] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and
L. Ye, "Author identification on the large scale," in *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*, 2005.

[2] B. Vijayakumar and M. M. M. Fuad, "A new method to identify short-text authors using combinations of machine learning and natural language processing techniques," *Procedia Computer Science*, vol. 159,pp. 428–436, 2019, Elsevier.

[3] L. Pearl and M. Steyvers, "Detecting authorship deception: A supervisedmachine learning approach using author writeprints," *Literary and lin- guistic computing*, vol. 27, no. 2, pp. 183–196, 2012, Oxford UniversityPress.

[4] S. M. Nirkhi, R. V. Dharaskar, and V. M. Thakre, "Analysis of online messages for identity tracing in cybercrime investigation," in *Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, pp. 300–305, 2012, IEEE.

[5] R. Phann, C. Soomlek, P. Seresangtakul, et al., "Multi-Class Text Classification on Khmer News Using Ensemble Method in Machine Learning Algorithms," *Acta Informatica Pragensia*, vol. 12, no. 2, 2023,Prague University of Economics and Business.

[6] I. Alsmadi, N. Aljaafari, M. Nazzal, S. Alhamed, A. H. Sawalmeh, C. P.Vizcarra, A. Khreishah, M. Anan, A. Algosaibi, M. A. Al-Naeem, et al.,"Adversarial machine learning in text processing: a literature survey," *IEEE Access*, vol. 10, pp. 17043–17077, 2022, IEEE.

[7] S. Ouni, F. Fkih, and M. N. Omri, "A survey of machine learning-basedauthor profiling from texts analysis in social networks," *Multimedia Tools and Applications*, pp. 1–34, 2023, Springer.

[8] S. Amidwar, S. Baxi, K. Rao, and S. Kale, "Text Analysis for Author Identification Using Machine Learning," *Journal of Emerging Technolo-gies and Innovative Research*, vol. 4, no. 6, pp. 138–141, 2017.

[9] P. TÜFEKCİ and M. Bektaş, "Author and genre identification of Turkish news texts using deep learning algorithms," *Sa̅dhana̅*, vol. 47, no. 4, p. 194, 2022, Springer.

[10] K. Luyckx and W. Daelemans, "Shallow text analysis and machine learning for authorship attribtion," *LOT Occasional Series*, vol. 4, pp. 149–160, 2005, LOT, Netherlands Graduate School of Linguistics.

[11] N. D. Cilia, C. De Stefano, F. Fontanella, C. Marrocco, M. Molinara, and A. S. Di Freca, "An end-to-end deep learning system for medieval writer identification," *Pattern Recognition Letters*, vol. 129, pp. 137– 143, 2020, Elsevier.

[12] N. M. Demir, "Authorship Authentication of Short Messages from So- cial Networks Using Recurrent Artificial Neural Networks," *Southeast Europe Journal of Soft Computing*, vol. 7, no. 2, 2018.

[13] R. R. Iyer and C. P. Rose, "A machine learning framework for authorship identification from texts," *arXiv preprint arXiv:1912.10204*,2019.