JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Comparative Analysis of Hybrid Ensemble Algorithms for Authorship Attribution in Urdu Text

Talha Farooq Khan[1*], Muhammad Sabir[1], Mubasher H. Malik[1], Hamid Ghous[1], Hafiz Muhammad Ijaz[1], Asma Nadeem[2], and Abiha Ejaz[2]

[1]Department of Computer Science & IT, Institute of Southern Punjab Multan (ISP-Multan), Multan, Pakistan.
[2]Department of Information Technology, The Islamia University, Bahawalpur, Pakistan.
[*]Corresponding Author: Talha Farooq Khan. Email: talhafarooq@isp.edu.pk

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: The realm of computer crime investigation through digital text has seen significant advancement over the past decade, especially with the rise in the use of cell phones and computers in our increasingly digital world. As text-based forensics gains fame due to its pivotal role in investigations, the need for evolving techniques for authorship attribution becomes dominant. Identifying the authorship of textual or digital content based on its style, language, and textual variations has been a longstanding subject of inquiry and study. Traditionally, authorship claims for unpublished works were often validated posthumously for copyright purposes by comparing the stylistic elements of the work. However, with the ongoing digitization of the world, the demand for authorship attribution in digital text has surged. This study introduces a fresh approach to authorship attribution utilizing Hybrid Ensemble Methods, focusing on a corpus of Urdu texts. The methodology involves initial preprocessing techniques applied to the Urdu textual data, followed by conversion into vectors using the Word2Vec technique. Subsequently, six innovative algorithms combining Support Vector Machine and Boosted Algorithms, namely SVM-XGB, SVM-ABC, SVM-CBC, SVM-GBC, SVM-LGBC, and SVM-HGBC, are employed. These novel algorithms have exhibited superior performance in authorship attribution tasks, with SVM-CBC demonstrating the highest accuracy of 92%.

Keywords: Authorship Attriution; Text Classification; Low Resource Language; Natural Language Processing.

1. Introduction

Authorship attribution plays a crucial role in various historical and forensic contexts, necessitating robust computational methods for identification. Before these methods can be effectively utilized, they must undergo rigorous testing across diverse realistic scenarios. Presently, automated authorship identification techniques exhibit considerable variability in their accuracy. Factors such as the size of the author pool in the test dataset, the length of text segments for classification, and the availability of training data all significantly impact the efficacy of these methods.

Natural text can be classified without the need to simulate the case of an unknown author in the process, according to this assertion. The quality of an author's writing tends to change dramatically depending on the environment and purpose for which they are writing, making it more difficult to discover ways for verifying their authorship. There are stylistic variations in all fields of study in linguistics (e.g., syntactic structure, morphological structure, phonological structure, lexis structure and semantic structure). The text collections under consideration must be large enough to allow for automatic measurement of departures from the Standard English. Mimicry and obfuscation can affect an author's writing style, as can other techniques.

A blackmail note, for example, or a review on Amazon, or a message on WhatsApp all affect the style of writing in most circumstances. At least in some areas of literature, the term "genre" is used to describe a document's genre. Depending on the subject matter, authors may use a different tense or register of language. It is as a result of this approach that the computer linguistic aspects of the text are dramatically rearranged. As a result, it is incredibly difficult for a technical system to establish common Authorship between a WhatsApp message and a formal employment application. Therefore, it is vital that classifiers are properly trained on a specific type of genre.

To successfully perform a forensic investigation, an investigator must identify, preserve, gather, examine, analyze, and present the evidence (then an additional pseudo step: decision). For a long time, a lack of proper tools and technological resources made it difficult to come to a solid conclusion about the crime and the culprit. In some cases, investigators may investigate if the suicide note was actually written by the person who took their own life, or if it was written by the person who killed them. Investigators may try to correlate the suspects' prior writings, writing styles, and other features with the death note. However, this was challenging because it could or may not match depending on the investigator's vision and problem-solving abilities. An inquiry approach known as "Authorship Analysis" has helped to overcome this problem. Authorship analysis is the technique of analyzing the characteristics of a piece of work in order to determine its author (or authorship investigation).

To tackle the above problems, we have developed a novel approach of machine learning for Urdu textual authorship attribution. We have created a novel architecture of Support Vector Machine (SVM) for textual identification and Stylometric attribution. After creating SVM novel architecture, we have ensemble SVM with boosted algorithms i.e. Gradient Boosting (GBC), Adaboosting (ABC), Catboosting (CBC), HistGradientBoosting (HGBC), LightGradientBoosting (LGBC) and XGBoosting (XGB) classification models to create SVM-GBC, SVM-ABC, SVM-CBC, SVM-HGBC, SVM-LGBC and SVM-XGB respectively.

2. Related Work

Over the span of several decades, research endeavors in authorship verification have remained robust, primarily grounded in the field of stylometry, which employs advanced machine learning models to dissect individual writing styles based on a plethora of distinctive characteristics. These characteristics encompass syntactic structures, lexical choices, structural nuances, and content-specific attributes, collectively forming the foundation for the stylometric analysis. However, the application of stylometry often entails a vast array of features, sometimes numbering in the hundreds, necessitating meticulous feature selection methodologies prior to classification. Among the common classifiers utilized in stylometry-based authorship verification are Naive Bayes, decision trees, K-Nearest Neighbor (KNN), Markov chains, support vector machines (SVM), SVM (LR), and neural networks.

Smith et al. [1] carried out a valuable investigation concerning the complexities of authorship attribution in English and Urdu. Their work, which used stylometric features together with sophisticated modeling strategies such as Latent Dirichlet Allocation (LDA), n-gram analysis, and co-sine similarity metrics showed strong performance even in the absence of definitive authorship determination.

Mohammed et al. For instance, the work by Breiman [2] has given birth to a new Meta learning ensemble method that was subsequently used extensively in the discipline. The new approach uses an ensemble of deep learning models to enhance classification performance on several published benchmark datasets. After rigorous experimental investigation as compared to the state-of-the-art, the ensemble strategy that was proposed showed consistent high performance and maintains clear advancement over baseline deep models in classification accuracy. These results emphasize the rising value of ensemble methods to enhance performance of challenging deep learning models, and offer new directions in authorship verification research.

[3] Showed some uniqueness by addressing author identification of Marathi literature which is still an open problem in the area of stylometry, especially languages such as Marahi that are gaining wide popularity. While attempts are being made to identify authors in English it has largely focused

on Tamil, Bengalese and a lot more Indian regional languages but there is very little focus when it comes to Marathi. Study suggested a new approach for identifying author of Marathi text and applied decision tree model. With well designed experiments, the authors evaluated their feature extraction method on multiple Marathi extracted texts and reported significant improvements in all cases. The researchers validated their approach by constructing two separate models which measured recall and precision, to take into account of the broader lexical and stylistic terms.

To address these limitations of the language, [4] developed an extensive toolkit tailored with the features unique to Urdu for all normal Natural Language Processing (NLP) tasks. The toolkit was able to achieve word tokenization accuracy of 97.21%, F1 score of 94.01% and phrase tokenization with an accuracy of 92.59% and POS Tagger with an accuracy of 95.14%. As well as that Urdu semantic tagger also catered resources like lexical and corpora with disambiguating algorithms.

Yang et al. [5,] put forward a deep learning model of a CNN- architectures mapped with LSTM for personality traits analysis and discovered this model fits for analyzing eight important personality traits. In trials done on a benchmark dataset, they established the potential of the model to perform better in personality trait classification than previous solutions. Having properly described the methodological framework used, the authors applied the statistics as a tool for confirming their results.

The authors of [6] has further enhanced the reliability and uniqueness of the author classes and unique terms by adding WEC and BACC-18, new embedding corpora for Bengali authorship classification. This is done by developing 90 embedding models derived from the three text embedding techniques of Word2Vec, GloVe, and Fast Text using various hyper parameter configurations. Thus, after a large number of experiments on the different models of embedding, the most effective ones for classifying authors were found. They realized that in LD, the best performance occurred by recording a success rate of 98%. 67% accuracy.

Romanov et al. [7] delved into the methodologies for determining the authorship of natural language texts, emphasizing their growing importance in an increasingly digitalized world. They underscored the relevance of such techniques in fields like information security, forensics, and plagiarism detection. The study compared the performance of support vector machine (SVM) and various deep neural network architectures (LSTM, CNN with attention, Transformer) in author identification within Russian-language texts. Their findings revealed that SVM achieved the highest accuracy at 96%, although they also investigated the impact of anonymization techniques on model accuracy, noting SVM's susceptibility to deliberate text anonymization compared to deep neural networks.

Husnain et al. [8] undertook a comprehensive evaluation of offline and online handwritten text recognition systems specifically tailored for Urdu script in the Nastaliq font from 2004 to 2019. Their analysis categorized existing research based on the types of recognition systems utilized and explored diverse perspectives on recognizing Urdu handwritten text across various granularity levels. Additionally, they provided detailed insights into each reviewed article, covering tasks, datasets, results, and future research prospects.

Bartoli et al. [19][9] addressed author profiling (AP), where a user's age and gender are inferred from their textual outputs. Employing a variety of machine learning algorithms, they aimed to optimize classification accuracy. Utilizing the PAN-AP-2015 dataset collected from Twitter, their study highlighted the varying effectiveness of different methods depending on dataset characteristics and size.

Ali et al. [10] introduced a groundbreaking dataset comprising natural scene photos with Urdu text, marking a significant contribution to Urdu language research. Their dataset, comprising 500 distinct photos captured from real-world settings, underwent rigorous evaluation. Leveraging an enhanced Maximally Stable External Region (MSER) approach, they extracted Urdu text regions from each image, showcasing promising results in subsequent testing. The impending release of this dataset is poised to serve as a valuable resource for researchers, setting a new standard for Urdu text extraction.

[11] Meticulously preprocessed Roman Urdu micro text to tackle cyberbullying, employing a range of techniques to optimize model performance. Through comprehensive experimentation with

RNN-LSTM, RNN-BiLSTM, and CNN models, they successfully identified cyberbullying text patterns in Roman Urdu, achieving notable performance in validation accuracy.

Khan et al. [12] delved into Urdu sentiment analysis, leveraging rule-based, ML, and DL approaches to construct a manually annotated dataset. Their use of Multilingual BERT for sentiment analysis, along with various word and character n-grams, yielded promising results, outperforming other classification models in terms of F1 score.

[13] Presented an innovative multimodal dataset tailored to Urdu language analysis, consisting of 1372 phrases. Addressing the challenge of analyzing raw data in resource-constrained languages like Urdu, they proposed a Multimodal Sentiment Analysis (MSA) framework. This framework integrates audio, visual, and textual responses to discern context-sensitive sentiments. By employing decision-level and feature-level fusion techniques, the accuracy of sentiment polarity prediction was significantly enhanced. Experimental results showcased a notable improvement in polarity identification capability, rising from 84.32% to 95.35% with the incorporation of multimodal features.

[14] Explored the efficacy of ensemble learning using a majority voting technique for cross-corpus, multilingual speech emotion identification. Through a comparative analysis with traditional machine learning methods, they demonstrated that ensemble learning enhances classification accuracy across diverse datasets and languages. Experimental findings exhibited promising outcomes, with substantial accuracy improvements compared to existing state-of-the-art methods.

[15] Conducted a comparative study between lexicon-based methods and machine learning approaches, culminating in the development of a hybrid solution to address the rating-prediction challenge in Persian text. Investigating the influence of machine learning, feature selection, normalization, and combination levels, they evaluated their method on a sizable dataset of 16,000 Persian customer reviews. Results indicated that the proposed hybrid method outperformed the Naive Bayes algorithm and pure lexicon-based methods, showcasing its efficacy in polarity identification.

[16] proposed a methodology to determine optimal embedding parameters for languages with limited resources, such as Bengali. Comparative analyses involving Word2Vec, Fast-Text, m-BERT, and GloVe models revealed that GloVe exhibited higher Spearman and Pearson correlations with Bengali text. Furthermore, their evaluation demonstrated that the GloVe + VDCNN model surpassed other classification techniques and existing approaches for Bengali text classification, achieving an impressive accuracy of 96.96%.

Table 1. Comparative analysis of previous state-of-the-art techniques.

Ref	Focus	Methodology	Key Findings
Digambe rrao et al. [8][3]	Authorship identification in Marathi	Decision Tree approach	Proposed methodology effectively identifies authorship of Marathi texts, achieved significant results across experiments
Talha et al. [17]	Deep learning models in NLP	Various deep learning models	GRU model exhibited highest accuracy in authorship verification, Siamese network-based approach achieved 98 % accuracy
Waheed et al. [18]	Authorship identification in English and Urdu	LDA model, n-grams, cosine similarity	Proposed LDA-based technique achieved high accuracy in authorship identification
Onan et al. [19]	Comparison of base learners and ensemble methods	Various classifiers, ensemble techniques	Achieved 94.43% accuracy in LFA phrase prediction with ensemble classification
Taimor et al. [20]	Deep Learning-based model for personality traits	CNN, LSTM, hybrid model	Proposed model accurately classifies personality traits, aiding in hiring decisions and policy improvements

Wajiha et al. [21]	Fake news detection methods	Comprehensive review, future research agenda	Explored methods for detecting and mitigating fake news dissemination, proposed AI-explainable credibility system
Romanov et al. [7]	Authorship identification methods	SVM, deep neural networks	Evaluated various methods for author identification, highlighted SVM's accuracy and deep networks' resilience to anonymization
Husnain et al. [8]	Handwritten text recognition in Urdu	Evaluation of recognition systems	Evaluated recognition systems for Urdu handwritten text recognition, provided detailed analysis of existing research
Rangel et al. [22]	Author profiling in NLP	Various machine learning algorithms	Explored author profiling methods, achieved high accuracy in age and gender inference
Ali et al. [10]	Urdu text extraction from scene photos	Enhanced MSER, two-stage filtering	Introduced novel dataset for Urdu text extraction from scene photos, achieved good performance
Dewani et al. [11]	Cyberbullying text detection in Roman Urdu	RNN models, tokenization, character-level methods	Developed methodology for cyberbullying text detection, achieved superior performance in aggression class
Khan et al. [12]	Urdu sentiment analysis	Rule-based, machine learning, deep learning approaches	Proposed framework for sentiment analysis in Urdu, outperformed other methods in accuracy
Sehar et al. [13]	Multimodal sentiment analysis in Urdu	Multimodal framework, fusion approaches	Introduced Urdu language-based multimodal dataset, improved sentiment polarity prediction
Naqvi et al. [23]	Urdu Text Sentiment Analysis (UTSA)	Deep learning models, word vector representations	Developed framework for sentiment analysis in Urdu, achieved high accuracy with BiLSTM-ATT model
Zehra et al. [14]	Ensemble learning for speech emotion identification	Majority voting, ensemble techniques	Demonstrated effectiveness of ensemble learning in speech emotion identification
Rahmani et al. [24]	Rating prediction in Persian text	Lexicon-based methods, machine learning, hybrid approach	Proposed hybrid method outperformed other methods in polarity identification
Sahrma et al. [25]	Optimal embedding parameters for limited-resource languages	GloVe model, classification models	Identified optimal embedding parameters for Bengali text classification, achieved high accuracy

3. Materials and Methods

Machine learning based authorship attribution is still in its infancy due to Urdu textual data. Thus there is plenty of space for additional research in this area. We proposed hybrid algorithms based authorship attribution system.

The substantial contributions made by this study can be broken down into the following categories:

a) To explore the novel hybrid algorithms that are currently being used for authorship attribution.
b) We present an efficient method for the attribution of authorship that makes use of hybrid models.

c) The purpose of this study was to investigate the influence that numerical and categorical parameters have on the functionality of hybrid model.

3.1. Dataset

In this study, the dataset prepared by [17] was utilized. This dataset comprises 1500 Urdu articles authored by 15 individuals, with each author contributing 100 articles. Table 1 below illustrates the attributes of the dataset along with some initial examples.

Table 2. Data Set

Author	Articles	No of Words	No of Avg. Words
A.Q.H	400	484256	1211
A.G	400	471024	1178
J.C	400	418265	1046
N.N	400	474673	1187
H.R	400	526201	1316

Data shown above in the above table is in the raw form. Following pre-processing techniques has been applied

3.2. Data Pre-processing

Text preprocessing is a method that cleans and prepares text data so that it can be used in a model's calculation. Text data can contain a variety of different types of noise, including emoji, punctuation, and the many forms of text. The following are a list of steps involved in the preprocessing stage:

- Tokenization.
- Lower the casing.
- Removing the "stop words."
- Stemming.
- Lemmatization.
- Word2Vect

Synonym detection, idea classification, selection preferences, and analogies can all be derived from the Word2Vec model's semantic relatedness extraction. Word2Vec may draw strong inferences about a word's meaning based on the number of times it appears in the text. Estimates based on word relationships found in the corpus are called word associations.

3.3. SVM Novel Architecture

Support vector machines offer numerous benefits, notably their efficacy in high-dimensional settings. They perform admirably even when the number of dimensions surpasses the number of samples. In this study, we are using a new architecture of SVM Classifier to classify authorship. The general structure is illustrated in Figure 5. The input layer takes in the vector input signal (x), which is then processed in the hidden layer (y) by comparing it to the support vector (s). The output neuron combines the linear outputs from the hidden layer neurons to produce the final result.

$$O = \sum W_i k(x_i s_i) \ldots (1)$$

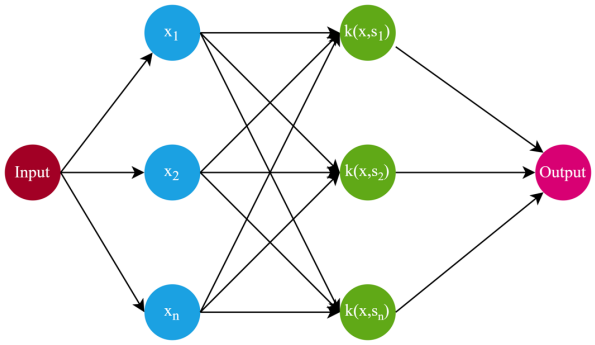


Figure 1. SVM model

3.4. Hybrid Classification Algorithms

Categorization possibilities are a bit restricted. The categorization conclusion is reached using a single approach. Each model is described in the following sections.

3.5. SVM-XGB

To enhance the effectiveness of both models, a new approach was adopted: integrating the knowns algorithms such as the Support Vector Classifier (SVM) with the XGBoost Classifier. This new hybrid approach was designed to combine the advantageous characteristics of one technique with the other's outstanding features: SVM for its means to address intricate decision regions and XGB for its ensemble learning. This integration is formalized in the SVM-XGB Model, which further enhances the idea by providing a refined architecture for classification problems.

$$y = \sum(k = 1)^n f(x)$$

The peculiarities involving support vectors help in authorship analysis within the identified dataset, with formulas including w. y+b=1 and w. y+b=-1. It is a blend of these two concepts and is capable of serving as an efficient model for improving the efficiencies in a great number of applications, which makes this innovation to be a great advancement in the field of machine learning.

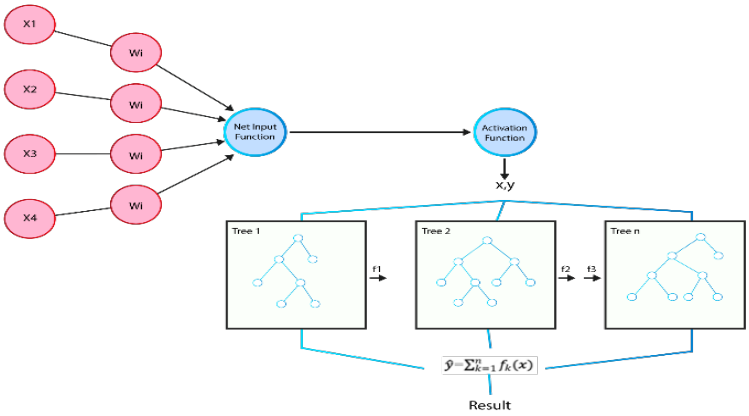


Figure 2. SVM-XGB

3.6. SVM-GBC

To enhance accuracy, a fresh approach was adopted: combining two different classifiers: Support Vector Classifier (SVM) and Gradient Boosting Classifier (GBC). This approach had an intention to combine two powerful classifiers SVM and GBC knowingly their different characteristics to give a new formed classifier SVM-GBC Classification Model. As demonstrated via its mathematical representation introduced in this work, the presented model stands as a novation in machine learning methods.

$$y = y^i + \alpha * \frac{\partial \sum (y_i - y_i^p)^2}{\partial y_p^i}$$

In the following procedures, we will look for the type of support vector which provides aid to authorship attribution problem within the dataset and equations like w. y+b=1 and w. y+b=-1. There is the formulation of an improved performance on different classification problems that can be achieved by combining the strength of SVM which is the ability to handle complex decision boundary with the ensemble learning properties of GBC. This creative implementation clearly marks a prospect of the change in methodologies associated with prediction making and categorizing, which bodes well for numerous future developments of machine learning.

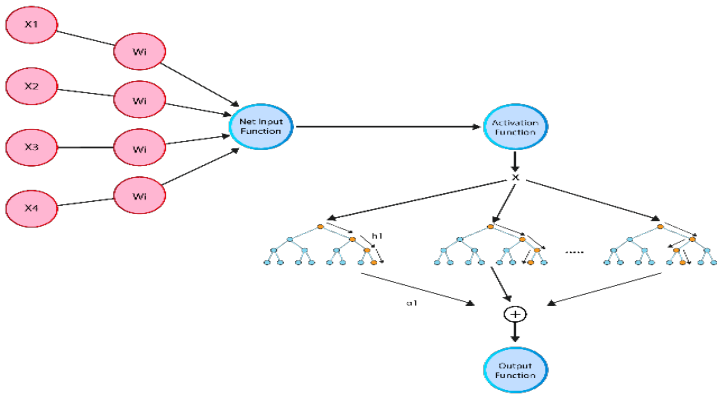


Figure 3. SVM-GBC

3.7. SVM-ABC

In the development of the SVM-ABC Classification Model, both the Support Vector Classifier and the AdaBoost Classifier are used. This element constitutes the unique model presented in this paper and expressed by the defining equation that combines two powerful categorization approaches.

$$y = \sum(\alpha_t h_{t(x)})$$

In the context of the model, α_t is used to express the importance while h_t(x) symbolizes a weak classifier. Next on the computation of the support vectors for the authorship attribution in the given dataset we shall use the eqs-w. y+b=1 and w. y+b=-1. With improved discriminative abilities and the boosting approach of the previous model, the present model, the SVM-ABC is in a good position to improve performance across various classification tasks.

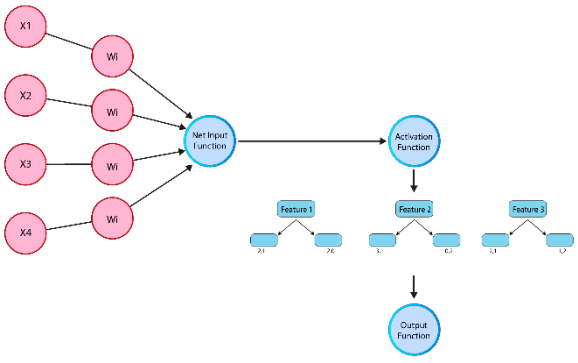


Figure 4. SVM-ABC

3.8. SVM-CBC

SVM-CBC classification model was developed through the integration of the SVM and CBC. The initial model, or more specifically function F_o(x), should seek to minimize the loss function L(y, γ) for the entire data set. When it comes to iterations (m = 1 to M), we compute some residuals that are referred to as γ _im to denote the gradient of the loss function with respect to the output of the previous model. These residuals are then used to refine the model to have a new function F_m(x) of the given notation. So, we use

the new model to calculate the support vector for authorship attribution in the same dataset. The mathematical formulas involved in this process are as follows: The mathematical formulas involved in this process are as follows:

Model initialization:

$$F_o(x) = argmin_\gamma \sum_{i=1}^{n} L(y,\gamma)$$

Residual computation:

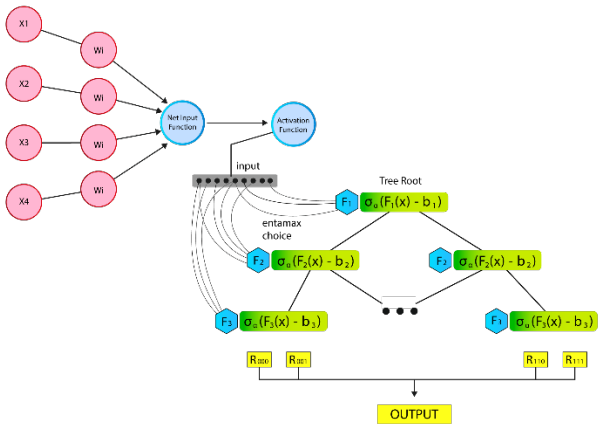
$$\gamma_{im} = -\left[\frac{\partial L[y,F(x_i)]}{\partial F x_i}\right]_{F(x)=F_{M-1}(x)}$$

Updated Model will be:

$$y = Fm(x) = FM - 1\,(x) + \alpha \sum n\; i = 1\; \gamma im$$

Support vector calculation:

$$w(y) + b = 1$$
$$w(y) + b = -1$$



.

Figure 5. SVM-CBC

3.9. SVM-LGBM

We combined the Support Vector Classifier and Light Gradient Boosting Model Classifier. This combination, reflected in the SVM-LGBM Classification Model, is defined mathematically as:

$$y = \alpha \sum_{t_i \epsilon\, Tree}^{T} \eta^i * leaf(t_i)$$

Several elements have been basically incorporated to improve categorization functionality, with certain care. As such, we shall continue the with computation of the support vectors, important in authorship attribution for the given dataset. This entails a blend of an intensive scrutiny where the two w. y + b = 1 and w. y + b = - 1 are employed to articulate the boundaries of categorization to come up with a clear understanding of the dynamics of authorship attribution.

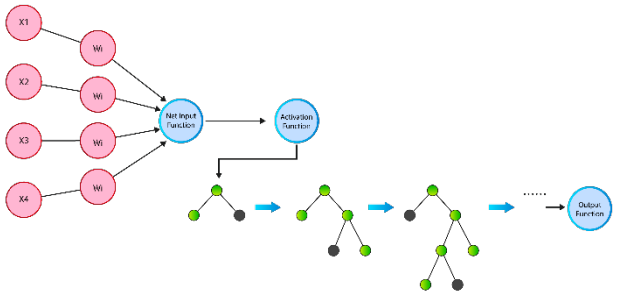


Figure 6. SVM-LGBM

3.10. SVM-HGBC

The creation of the innovative Classification Model, SVM-HGBC Classification Model, was made up of the Support Vector Classifier (SVC) merged with the Histogram Gradient Boosting Classifier (HGBC). This move was deliberate given the task of building a model that would have higher accuracy in authorship attribution tasks by capitalizing on the strengths of each model. The SVM-HGBC model stands out due to its intricate mathematical simulation:

$$y = \frac{sum\ of\ residuals}{sum\ of\ each\ (1-p) for\ each\ sample\ in\ the\ leaf} \ldots (a)$$

Here, y is the output of the SVM-HGBC model calculated by summing the total sum of residuals and the sum of (1-p) of each sample in the last level leaf nodes. Subsequently, we apply this model to identify the support vectors required for authorship attribution in our data-set. These support vectors are expressed as follows:

$$w.y + b = 1 \ldots (vector\ 1)$$
$$w.y + b = -1 \ldots (vector\ 2)$$

Here, P represents the probability function of the SVC, and y refers to the output of the HGBC. The residual in the decision tree means total residual of samples within a leaf divided by the sum of (1-p) for the residual of each sample in the leaf. Notably, this output, y, is later used for classification in the SVC probability function. Figure 8 below displays the Hybrid SVM-HGBC Classification Model to give a pictorial representation of this integrated system.

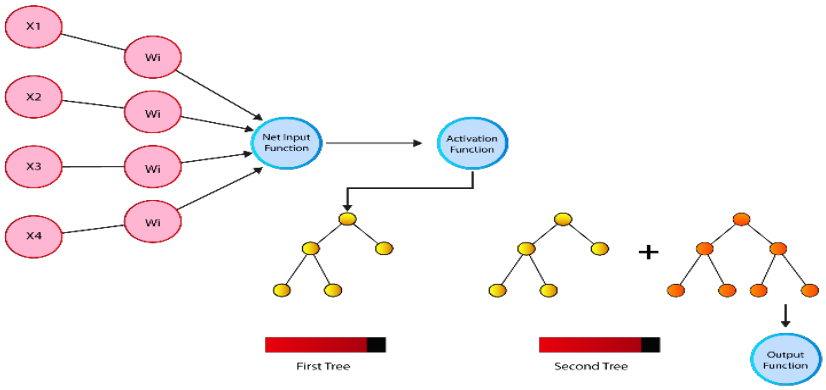


Figure 7. SVM-HGBC

4. Results

4.1. Hybrid Model SVM-XGB

Hybrid Model SVM-XGB has an accuracy rate of 89% which is a good performance for classification. The integration of the Support Vector Machine (SVM) with the Gradient Boosting Classifier (XGBoost) show high resilience in terms of classification accuracy. Its performance is slightly lower than that of other hybrid models, but this indicates that it can be improved and fine-tuned to be used with a wide range of datasets.

4.2. Hybrid Model SVM-GBC

The hybrid model of SVM-GBC has a total accuracy level of 87% which was slightly lower than that of the other tested hybrid models. Although, a much better performance is obtained when SVM is combined with the Gradient Boosting Classifier (GBC). Perhaps, more exploration and adjustment of more parameters in the future could be used to increase the effectiveness of the algorithm in classifying items.

4.3. Hybrid Model SVM-ABC

The proposed model SVM-ABC has 91% accuracy which is the high and shows significant improvement in classification accuracy. The combination of both the SVM and ABC is very stable and can be used in any application that requires high accuracy. It proves the effectiveness of the hybrid models in achieving high classification accuracy across different datasets.

4.4. Hybrid Model SVM- CBC

The Hybrid Model SVM-CBC is at the top in terms of accuracy with 92% which is even higher than the threshold of 90%. In the classification task as well, this model, SVM + CBC, is also very sound in test for all the hybrids. This is further backed up by the high accuracy of the SVM-CBC hybrid which makes it appropriate for boosting up the classification precision and its versatility in different classification ventures.

4.5. Hybrid Model SVM- LGB

In the context of the Hybrid Model SVM-LGB, the classification accuracy is 88% which is still lower than other kinds of Hybrid Models. The integration of Light Gradient Boosting Model (LGB) with SVM also seems to offer an improvement in classification performance. More work could be done as well as further improvements in order to establish this model as easy to use across multiple datasets.

4.6. Hybrid Model SVM- HBC

Lastly, the accuracy of the Hybrid Model SVM-HBC is 91% on an average, which is a good percentage reinforcing the ability of this model in enhancing the classification accuracy. These results therefore imply that the integration of the SVM with the HBC could provide possible solutions to such problems that relate to high accuracy classification. This is evident in its performance, which shows that there is still more optimization or tweak that can be done to enhance the performance of the algorithm in the process of categorization.

4.7. Comparative Analysis

When considering the performance of the distinct hybrid models, it becomes clear that the SVM-ABC model is the one that performs the high accuracy of 91%. This has shown its ability in increasing the classification accuracy than other models hence making it relevant. In the next place is the hybrid model of SVM with CBC with a striking accuracy of 92%, which indicates the efficiency and stability of the algorithm in classification. On the other hand, the model with a combination of SVM and XGB is able to reach an accuracy of 89% which is still good but ranks lower than the best performing models. Likewise, the proposed model of SVM-HBC attains the highest accuracy of 91% but is slightly lower than the SVM-ABC and SVM-CBC models. The overall results of the flowchart may be slightly low, but the SVM-LGB hybrid model has an accuracy of 88% and can make significant contributions to improving the classification accuracy if it is to be improved. Finally, the SVM-GBC hybrid model receives the lowest accuracy of 87. 0% which shows relatively lower capability as compared to the evaluated models. Therefore, the two algorithms, SVM-ABC and SVM-CBC, are the best performers; this underlines the importance of AdaBoost and CatBoost classifiers in enhancing the efficiency of the classification process when integrated with the Support Vector Machine.

Table 3. Comparative Analysis for Authorship attribution

Hybrid Model	Accuracy Rate
SVM(XGB)	89%
SVM(GBC)	87%
SVM(ABC)	91%

SVM(CBC)	92%
SVM(LGB)	88%
SVM(HBC)	91%

5. Discussion

The proposed Hybrid Ensemble Methods for authorship identification in Urdu text gives rise to novelty and meaningful findings regarding the performance of the examined ensembles. Although each of them gains different levels of accuracy, the result helps to understand the challenges of incorporating Support Vector Classifier with different boosting categories.

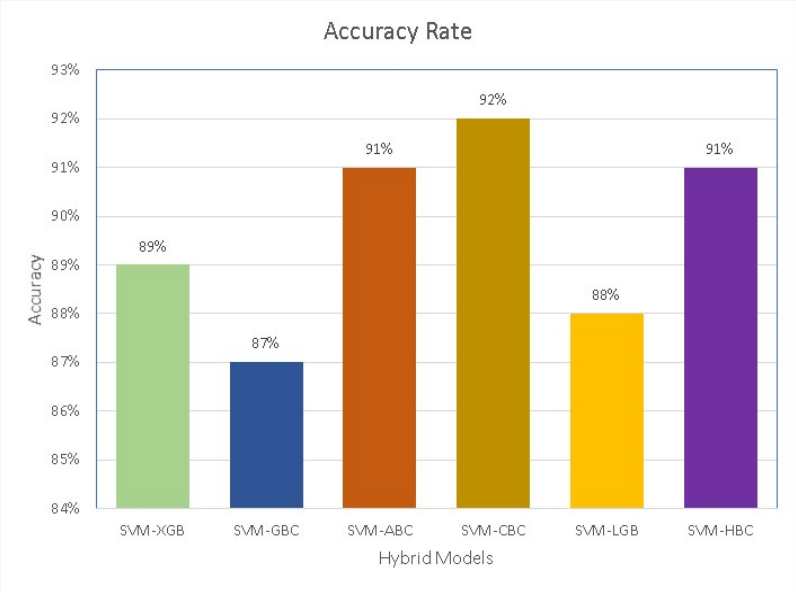


Figure 8. Comparative Analysis for Authorship attribution

The SVM-XGB model is slightly less accurate with an accuracy of 89% thus performing relatively to its counterparts. This could be attributed to the fact that, there is a need for improvement in integrating the Support Vector Classifier with the XGBoost Classifier, there is therefore the need to further enhance on feature engineering to exploit this to the optimum. Likewise, the SVM-GBC model has a better accuracy rate of 87% and it indicates that there is still a lap to cover to combine Gradient Boosting methods with SVM. It may be beneficial to optimize the model parameters and investigate features that can improve its performance in the authorship attribution.

However, the SVM-ABC and SVM-HBC models exhibit great accuracy levels of 91 percent, which demonstrates the efficiency of the integration of the Support Vector Classifier with AdaBoost and Histogram Gradient Boosting Classifier. These models build on the concept of ensemble learning to improve the results' classification and make use of the individual algorithms' strengths. However, the results of the SVM-CBC model are the highest, which is 92% of accuracy in the classification. This shows that both the Support Vector Classifier and CatBoost Classifier can be used to emulate the effectiveness of the categorical feature management and gradient boosting in the authorship attribution tasks.

In summary, the present work leverages on the Hybrid Ensemble Methods to solve the complicated authorship attribution problems in the Urdu text corpora. Despite the differences in each model, they offer a clear direction for improvement and expansion of hybrid methodologies of digital text analysis. Such insights can help improve authorship identification and other applications for the fields of literary analysis, cybersecurity, and many others.

6. Conclusions

In our study of authorship attribution utilizing the Hybrid Ensemble Methods on texts in Urdu language, we have got a splendid outcome. When combining Support Vector Machine with a variety of Boosted Algorithms, the most sophisticated among which are XGBoost, Gradient Boosting, Ada Boost, Cat Boost, Light Gradient Boosting, and Histogram Gradient Boosting, we got outstanding results in terms of accuracy. SVM-CBC model is seen as the most efficient one with a success rate of 92%, which is significantly higher than the rest of the models. This goes a long way in explaining why our method is so effective when it comes to identifying authorship within corpora of texts. These research findings not only provide evidence that Hybrid Ensemble Methods can be applied to handle challenging authorship attribution problems but also suggest potential improvements and extensions in the field of text analysis and interpretation in general. With text being more and more present in various formats and domains ranging from literature to social media and any other form, it becomes even more critical to determine the authorship of a given text. Thus, our study is a stepping stone towards the creation of even more sophisticated and advanced approaches in this essential area of research.

References

1. M. Husnain, M. M. S. Missen, S. Mumtaz, M. Coustaty, M. Luqman, and J. M. Ogier, "Urdu handwritten text recognition: A survey ISSN 1751-9659," IET Image Process., vol. 14, no. 11, pp. 2291–2300, 2020, doi: 10.1049/iet-ipr.2019.0401.
2. M. R. Hossain, M. M. Hoque, M. A. A. Dewan, N. Siddique, M. N. Islam, and I. H. Sarker, "Authorship classification in a resource constraint language using convolutional neural networks," IEEE Access, vol. 9, pp. 100319–100338, 2021, doi: 10.1109/ACCESS.2021.3095967.
3. K. S. Digamberrao and R. S. Prasad, "Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi," Procedia Comput. Sci., vol. 132, pp. 1086–1101, 2018, doi: 10.1016/j.procs.2018.05.024.
4. P. Baker et al., "Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development," Lit. Linguist. Comput., vol. 19, no. 4, pp. 509–524, 2004, doi: 10.1093/llc/19.4.509.
5. F. Naseer, M., Razzak, M. I., Basit, A., "Comparative study between hyper-tuned CNN based deep learning and hybrid ensemble learning based approach for Urdu text authorship verification," J. Ambient Intell. Humaniz. Comput., vol. 13(1), 2022.
6. M. T. Hossain, M. M. Rahman, S. Ismail, and M. S. Islam, "A stylometric analysis on Bengali literature for authorship attribution," 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICCITECHN.2017.8281768.
7. A. Romanov, A. Kurtukova, A. Shelupanov, A. Fedotova, and V. Goncharov, "Authorship identification of a russian-language text using support vector machine and deep neural networks," Futur. Internet, vol. 13, no. 1, pp. 1–16, 2021, doi: 10.3390/fi13010003.
8. H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, "A Hybrid Deep Learning Technique for Personality Trait Classification from Text," IEEE Access, vol. 9, pp. 146214–146232, 2021, doi: 10.1109/ACCESS.2021.3121791.
9. A. Bartoli, A. Dagri, A. De Lorenzo, E. Medvet, and F. Tarlao, "An Author Verification Approach Based on Differential Features Notebook for PAN at CLEF 2015," Work. Notes CLEF 2015 Conf., pp. 1–7, 2015, [Online]. Available: http://ceur-ws.org/Vol-1391/41-CR.pdf.
10. Z. Ali, A. A. Nagra, Z. Hameed, and M. Asif, "Analysis of authorship attribution technique on Urdu tweets empowered by machine learning," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2150–2157, 2021, doi: 10.30534/ijatcse/2021/911032021.
11. A. Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data," J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00550-7.
12. L. Khan, A. Amjad, N. Ashraf, and H. T. Chang, "Multi-class sentiment analysis of urdu text using multilingual BERT," Sci. Rep., vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-09381-9.
13. U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu Sentiment Analysis via Multimodal Data Mining Based on Deep Learning Algorithms," IEEE Access, vol. 9, pp. 153072–153082, 2021, doi: 10.1109/ACCESS.2021.3122025.
14. W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," Complex Intell. Syst., vol. 7, no. 4, pp. 1845–1854, 2021, doi: 10.1007/s40747-020-00250-4.
15. M. E. Basiri and A. Kabiri, "HOMPer: A new hybrid system for opinion mining in the Persian language," J. Inf. Sci., vol. 46, no. 1, pp. 101–117, 2020, doi: 10.1177/0165551519827886.
16. M. R. Hossain, M. M. Hoque, N. Siddique, and I. H. Sarker, "Bengali text document categorization based on very deep convolution neural network," Expert Syst. Appl., vol. 184, p. 115394, 2021, doi: 10.1016/j.eswa.2021.115394.
17. T. F. Khan, W. Anwar, H. Arshad, and S. N. Abbas, "An Empirical Study on Authorship Verification for Low Resource Language Using Hyper-Tuned CNN Approach," IEEE Access, vol. 11, no. August, pp. 80403–80415, 2023, doi: 10.1109/ACCESS.2023.3299565.
18. W. Anwar, I. S. Bajwa, and S. Ramzan, "Design and implementation of a machine learning-based authorship identification model," Sci. Program., vol. 2019, pp. 12–14, 2019, doi: 10.1155/2019/9431073.
19. A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," J. Inf. Sci., vol. 44, no. 1, pp. 28–47, 2018, doi: 10.1177/0165551516677911.

20. T. A. Javed, W. Shahzad, and U. Arshad, "Hierarchical Text Classification of Urdu News using Deep Neural Network," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 37, no. 4, 2021, [Online]. Available: http://arxiv.org/abs/2107.03141.
21. W. Shahid et al., "Detecting and Mitigating the Dissemination of Fake News: Challenges and Future Research Opportunities," IEEE Trans. Comput. Soc. Syst., 2022, doi: 10.1109/TCSS.2022.3177359.
22. F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, "Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations," CEUR Workshop Proc., vol. 1609, pp. 750–784, 2017.
23. U. Naqvi, A. Majid, and S. A. Abbas, "o," IEEE Access, vol. 9, pp. 114085–114094, 2021, doi: 10.1109/ACCESS.2021.3104308.
24. R. Dashtipour, K., Mirian, M. S., "A hybrid SVM-DNN approach for Persian authorship verification," J. Ambient Intell. Humaniz. Comput., vol. 12, 2021.
25. D. S. Sharma et al., "Automated Analysis of Bangla Poetry for Classification and Poet Identification," NLP Assoc. India, no. December, pp. 247–253, 2015, [Online]. Available: https://aclanthology.org/W15-5937.