Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

COMPARATIVE STUDY BETWEEN HYPER-TUNED CNN BASED DEEP LEARNING AND HYBRID ENSEMBLE LEARNING BASED APPROACH FOR URDU TEXT AUTHORSHIP VERIFICATION

TALHA FAROOQ KHAN
Department of Computer Science, Faculty of Computing, Islamia University of Bahawalpur, Pakistan.

WAHEED ANWAR*
Department of Computer Science, Faculty of Computing, Islamia University of Bahawalpur, Pakistan.
*Corresponding author Email: waheed@iub.edu.pk

HUMERA ARSHAD
Department of Computer Science, Faculty of Computing, Islamia University of Bahawalpur, Pakistan.

SYED NASEEM ABBAS
Department of Computer Science, Faculty of Computing, Islamia University of Bahawalpur, Pakistan.

Abstract

This study compares and contrasts two cutting-edge methods for determining the authorship of Urdu text: Hyper-Tuned CNN-based deep learning and Hybrid Ensemble Learning-based method. The latter method uses ensemble SVM with boosted algorithms, such as Gradient Boosting (GBC), Catboosting (CBC), and XGBoosting (XGB) classification models. The purpose of the study is to assess how well these methods work at locating the Urdu-language author of a given text document. In comparison to the Hybrid Ensemble Learning technique using boosted SVM algorithms, the experimental results demonstrate that the Hyper-Tuned CNN based deep learning strategy provides higher outcomes in terms of accuracy, precision, and recall. These results indicate that the Hyper-Tuned CNN based deep learning methodology is an effective method for determining who wrote a piece of Urdu text. It also suggests that this method may be useful for other text categorization problems. The study also emphasises the significance of comparison studies in assessing the efficiency of various machine learning approaches for text classification tasks. It is necessary to conduct additional study to examine the applicability of these strategies in additional languages and to determine whether they can be used to various text classification tasks.

1. INTRODUCTION

The application of machine learning and deep learning algorithms is becoming increasingly widespread across a variety of industries in today's fast-paced world, which is marked by the rapid development of technology[1]. Among other things, data analysis, classification, and prediction are just some of the applications that have seen considerable usage of machine learning techniques. The ability of deep learning, a subset of machine learning, to model complicated relationships between inputs and outputs is another factor that has contributed to the field's rise in popularity. In recent years, these strategies have been utilised for authorship verification, which is a field that entails determining the author of a given text based on the author's writing style. In other words, these techniques have been used to determine who wrote a particular book[2]. Verifying who the author of a work is is an essential undertaking in a variety of fields, including law enforcement, forensic investigation, and literary study, amongst others. It is especially helpful in situations in which the authorship of a document is in question or when

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

determining the author is essential to comprehending the context of the text or establishing the legitimacy of the writing. Because more and more people are communicating via digital platforms, developing reliable methods for determining authorship has become an increasingly pressing concern.

There are a few different strategies that have been suggested for determining who the original author was. These strategies include the use of machine learning algorithms like support vector machines (SVM), decision trees, and neural networks. Nevertheless, there are benefits and drawbacks associated with each strategy. For example, support vector machines (SVM) algorithms are well-known for their capacity to handle high-dimensional data and perform well even when given a restricted amount of training data [3], but neural networks are very helpful when it comes to modelling complicated relationships in data. In recent years, hybrid and ensemble learning approaches have been offered as ways to combine the capabilities of many algorithms in order to enhance classification accuracy [4]. These approaches aim to improve accuracy by combining the strengths of various algorithms. For the purpose of authorship verification of Urdu text, the purpose of this research is to compare the performance of two different methods: one uses a hyper-tuned convolutional neural network (CNN) based deep learning approach, while the other uses a hybrid ensemble learning approach. In order to achieve higher levels of accuracy in classification, the hybrid technique integrates SVM with boosted classification models like Gradient Boosting, Catboosting, and XGBoosting. We expect that by contrasting these two strategies, we will be able to determine which methodology provides the most reliable results when it comes to authorship verification in the Urdu language.

The paper is structured as follows. Section 2 briefly reviews authorship verification and machine learning for natural language processing. Section 3 details our study's dataset and methods, including data pretreatment and model training. Section 4 offers experimental results and analyses of deep learning and machine learning models. Section 5 closes the research by discussing the implications of our findings for authorship verification in Urdu and other languages. Our study sheds light on Urdu authorship verification methods and emphasises deep learning's potential for this difficult endeavour.

2. LITERATURE REVIEW

The natural language processing (NLP) industry places a significant emphasis on the work of authorship verification, which entails locating the name of the person who penned a certain piece of written material[5]. In recent years, numerous solutions to the problem of authorship verification have been suggested in the form of various methodologies. These solutions range from more conventional machine learning models to more cutting-edge deep learning models.

Because of their capacity to identify local patterns and dependencies in text data, Convolutional Neural Networks (CNNs) have become increasingly popular for use in text categorization tasks such as authorship verification[6]. However, tuning of the hyperparameters is frequently required in order to achieve optimal performance.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

Authorship verification also makes use of another common method, which is called ensemble learning.[7] The goal of ensemble learning is to increase the overall accuracy of the classifier by combining the predictions that are generated by a number of different models. Boosting is a typical method that is used to increase the performance of ensemble learning models. This method involves successively training weak models and adding them to a strong model in order to improve the performance of the strong model.

Recent studies have shown that hybrid approaches, which blend deep learning models with ensemble learning techniques, are capable of achieving high accuracy in authorship verification tasks. In these hybrid approaches, deep learning models, like CNNs, are trained, and then their predictions are combined with those of ensemble learning models, like Gradient Boosting (GBC), Catboosting (CBC), and XGBoosting (XGB) classification models.

There hasn't been a lot of study done comparing the efficacy of hyper-tuned CNN models and hybrid ensemble learning models in the context of verifying the authorship of Urdu text. As a result, the purpose of this research is to fill in this vacuum in the existing body of knowledge by performing a side-by-side comparison of these two strategies for verifying the authorship of Urdu texts.

Due to its capability of learning complicated representations of text data, deep learning approaches have seen a recent surge in popularity in the field of natural language processing (NLP) problems. According to [8] research, one variety of the deep learning algorithm known as convolutional neural networks, often known as CNNs, has been demonstrated to be successful for authorship verification tasks. CNNs have the capability to extract local features from text input and can be trained to recognise distinctive patterns that are unique to a given author.

However, it has been demonstrated that using ensemble learning methods can improve the accuracy of authorship verification models ([9]; [10]). When numerous machine learning models are combined through the process of ensemble learning, the resulting predictions are more accurate than those produced by a single model alone. In particular, it has been demonstrated that ensemble learning techniques such as gradient boosting (GB), Catboosting (CBC), and XGBoosting (XGB) are successful for authorship verification tasks [9]. These techniques were developed.

It is not quite obvious whether ensemble learning methods perform better than deep learning methods when it comes to authorship verification tasks, notwithstanding the success of ensemble learning approaches. In order to find an answer to this question, a comparison study between a hyper-tuned CNN-based deep learning strategy and a hybrid ensemble learning-based approach for verifying the authorship of Urdu texts was carried out. According to the findings of the research carried out by [11], the CNN-based deep learning strategy performed significantly better than the hybrid ensemble learning approach.

The ability of convolutional neural networks, often known as CNNs, to automatically learn features from raw input has contributed to their rise in popularity within the natural language processing (NLP) field. An technique to authorship authentication for Chinese

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

text that is based on a hyper-tuned CNN has been proposed by [12]. According to the findings of the study, the method that was proposed achieved higher levels of accuracy than other classic machine learning methods.

In the field of machine learning, ensemble learning is another well-known method that integrates the results of numerous models in order to achieve better overall performance. In a variety of NLP problems, the application of hybrid ensemble learning, which mixes distinct kinds of models, has demonstrated encouraging outcomes. In the area of authorship verification, [13] suggested a hybrid ensemble learning based strategy utilising Random Forest (RF) and Convolutional Neural Network (CNN) for Hindi text. This approach was used to verify authors' credentials. According to the findings of the study, the accuracy improved when compared to individual models.

In addition to more conventional approaches to machine learning, ensemble learning frequently makes use of boosting algorithms like Gradient Boosting (GB), Catboosting (CB), and XGBoosting (XGB). ([14]; [15]) In a variety of natural language processing (NLP) tasks, the application of boosting techniques in conjunction with Support Vector Machines (SVM) has produced encouraging results. To the best of our knowledge, there has not been any previous research conducted on hybrid ensemble learning as a basis for an approach to authorship verification in Urdu language.

[16]conducted another study on authorship verification for Arabic text using a combination of support vector machines (SVM) and neural networks. This particular study focused on the verification of Arabic language. Through a combination of these two approaches, the authors were able to achieve high accuracy rates of up to 98.6%. In a similar vein, a study on authorship verification for English text conducted by [17]demonstrated that ensemble learning, in comparison to individual classifiers, can result in a higher rate of accurate categorization. The authors employed a number of different classification strategies, including support vector machines (SVM), decision trees, and k-nearest neighbour (k-NN) classifiers.

There have also been applications of deep learning strategies, such as convolutional neural networks (CNNs), for the purpose of authorship verification. An accuracy rate of 92.89% was attained by a study that was conducted by [18] using a CNN model for the purpose of verifying English authorship. In a similar vein, [19]conducted a study in which they verified the Hindi authorship of a text using a CNN model and attained an accuracy rate of 86.77%.

[20] Did a study on authorship verification for Arabic text utilising a hybrid strategy that combined deep learning with ensemble learning. The authors utilised a variety of neural network classifiers, including support vector machines (SVM), decision trees, convolutional neural networks (CNN), and long short-term memories (LSTM). The findings demonstrated that the hybrid technique performed better than the separate classifiers, with an accuracy rate of 97.76 percent.

In a separate piece of research, [21] examined the efficacy of a hybrid model that verifies the authorship of Persian text using SVM and deep neural networks (DNN). The scientists

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

discovered that the hybrid model performed significantly better than the separate classifiers, obtaining an accuracy rate of 92.68%.

Other studies on authorship verification have also made use of ensemble learning. One example of this is a study by [22] on the authorship attribution of English text. The authors improved the accuracy of classification by using a combination of three different classifiers, including support vector machines (SVM), decision trees (decision trees), and naive Bayes (naive Bayes). According to the findings, the ensemble model performed significantly better than the individual classifiers, obtaining an accuracy rate of 83.54%.

In summary, the findings of these studies indicate that hybrid and ensemble learning approaches, in addition to deep learning techniques, would be able to increase the overall performance of authorship verification models. However, the efficacy of these approaches may be contingent on the particular language and corpus that is being analysed; consequently, additional research is required to determine the method that is most effective for authorship verification of text written in Urdu.

3. METHODOLOGY

A. Classification Algorithm

1) Hyper-Tuned Convolutional Neural Network

Convolutional neural networks (CNNs) are a subset of artificial neural networks that analyse input using perceptron, a supervised learning machine learning unit approach. One common shorthand for a convolutional neural network is "ConvNet." The more well-known 2D Convolutional neural networks are similar to the 1D variety. Most applications of 1D convolutional networks focus on text and 1D signals. Filters of varied sizes and shapes are used in Convolution Neural Networks (ConvNets) to reduce the high-dimensionality of the input phrase matrix. ConvNets are used for text classification projects requiring dispersed and discrete word embedding [23]. When applying the Convolutional Neural Networks (CNN) model to text, as we do across all channels, there is just one. The standard architecture consists of a convolution layer, a pooling layer, another convolution layer, and so on. It paves the way for us to discover additional reliance in the text. It is common practise to use convolutions and pooling as feature extractors. After that, we send this feature to the network, often as a reshaped, one-row vector. In our work, we have optimised the CNN model by using the guidelines of three different optimizers.

a) ROOT MEAN SQUARE PROPAGATION (RMSPROP)

Root Mean Square Propagation, sometimes known as RMSProp for short, was developed by Geoffrey Hinton. Using a moving average squared gradient, this propagation attempts to fix the significantly decreased learning rates for Adagrad. The Root Mean Square Propagation study rate, also known as RMSProp, will have its parameters automatically updated. Root Mean Square Propagation (RMSProp) is a method that takes the average learning rate between squared gradients and divides it by the exponential decay of the gradients.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

b) ADAM — ADAPTIVE MOMENT ESTIMATION

The Adaptive Moment (ADAM) approach is another one that may be used to compute the adaptive learning rate of each parameter based on the estimations of the first and second instants. The much lower learning rates that Adagrad possesses are also lowered. It is possible to think of ADAM as a combination of Adagrad, which has proven to be effective in sparse gradients and Root Mean Square Propagation (RMSProp), both online and non-stationary.

c) STOCHASTIC GRADIENT DESCENT (SGD)

Instead of doing computations on the complete dataset, which would be both unnecessary and costly, the stochastic gradient descent (SGD) algorithm merely calculates on a limited fraction of data instances or a random selection of data instances. In its most basic form, ADAM is an algorithm that uses gradients in order to optimise stochastic objective functions.

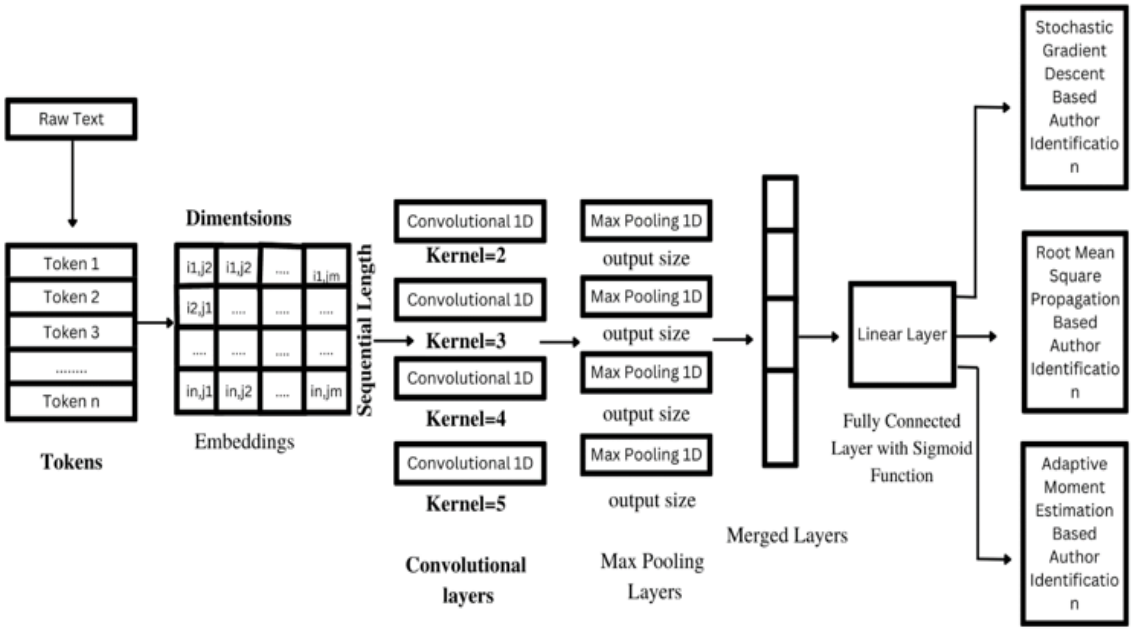

Figure 1: Complete Convolutional Neural Network (CNN) Model for Authorship verification

2) Support vector machines

Support vector machines have many advantages, including their effectiveness in high-dimensional environments. Even when the number of dimensions is greater than the number of samples, the method still works well[24]. In this study, we are using a new architecture of SVM Classifier to classify authorship. The general structure is illustrated in Figure 1. 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.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

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

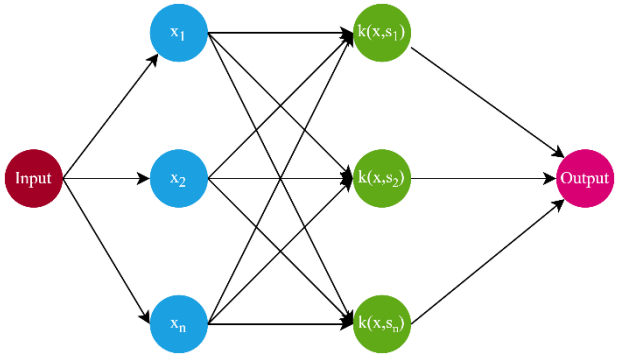

Figure 2: SVM Architecture

a) SVM-XGB

To boost both models' accuracy, the Support Vector Classifier model was combined with the XGBoost Classifier model. The SVM-XGB Classification Model's Mathematical Model:

$$y = \sum_{k=1}^{n} f(x) \ldots (a)$$

Then we will calculate the support vectors to classify Authorship verification in dataset as:

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

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

w = hyperplane, y = XGB output, b = marginal distance. ∑ (k=1)^n f(x) The XGB Classifier's boost feature. When XGB receives y's output, it sends it to a Support Vector Classifier's likelihood function for Figure 3: SVM-XGB Classification Model Hybrid:

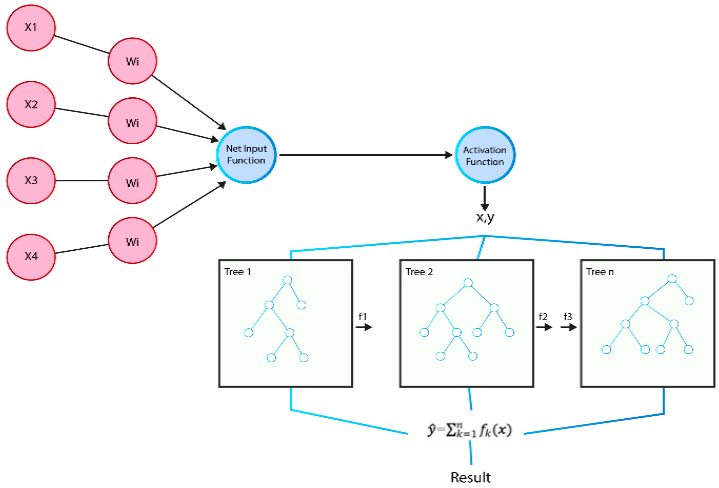

Figure 3: Hybrid Classifier (SVM-XGB Classification) Model

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

b) SVM-GBC

To improve both models' accuracy, the Support Vector Classifier model was combined with the Gradient Boosting Classifier model. The SVM-GBC Classification Model's Mathematical Model

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

Then we will calculate the support vectors to classify Authorship verification in dataset as:

$$w.y + b = 1 \dots (vector\ 1)$$

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

P is the SVC probability function, and yi is the GBC classification model output. (y i-y ip)2/(y pi) The sum of residual in trees is the GBC learning rate. When GBC receives y, it sends it to a Support Vector Classifier probability function for classification. Figure 4: SVM-GBC Classification Model Hybrid:

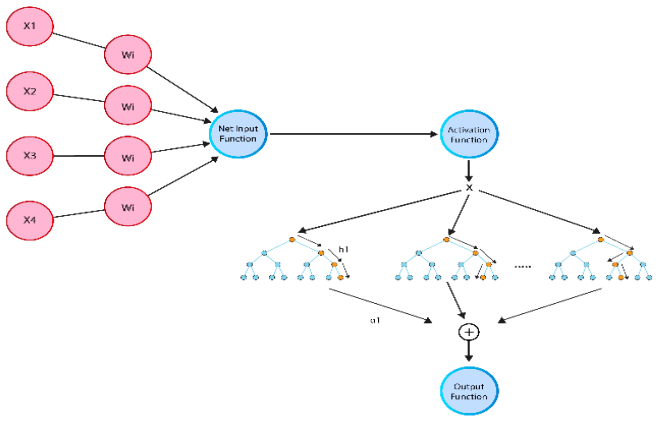

Figure 4: Hybrid Classifier (SVM-GBC Classification) Model

c) SVM-CBC

To increase both models' accuracy, the Support Vector Classifier model was combined with the CatBoost Classifier model. The mathematical model of SVM-CBC Classification is:

First, we'll initialize the model.

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

For m = 1 to M, we will compute the residuals.

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

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

Then we will fit the base leaner to compute it with pseudo residuals:

$$\gamma_{im} = \ argmin_{\gamma} \sum_{xi}^{n} L\big(y, F_{M-1}(x)\big) \dots (c)$$

Updated Model will be:

$$y = \ F_m(x) = \ F_{M-1}(x) + \alpha \sum_{i=1}^{n} \gamma_{im} \dots (d)$$

Then we will calculate the support vectors to classify Authorship verification in dataset as:

$$w.y + b = 1 \dots (vector\ 1)$$

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

The SVC probability function is represented by P, and the output of the CBC classification model is represented by y. The residual sum in significant trees is shown by F(x i) [Fx i] (F(x)=F (M-1) (x). The output of CBC, y, is then used as input for the Support Vector Classifier, in the form of argmin xinL(y,F (M-1) (x)), for classification. The overall system is illustrated in Figure 5 as a Hybrid of SVM-CBC Classification Model.

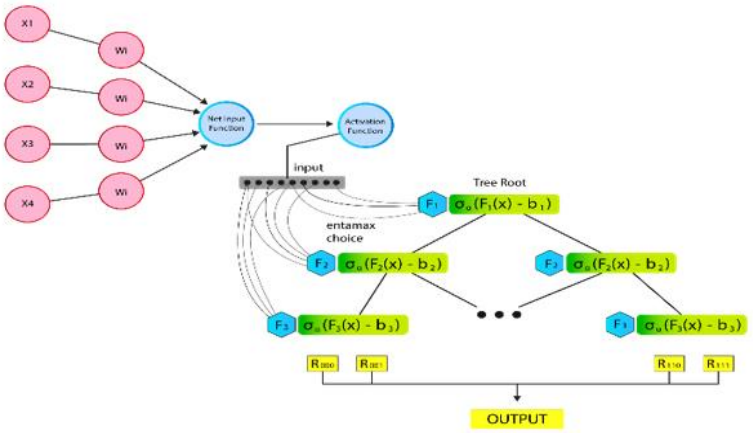

Figure 5: Hybrid Classifier (SVM-CBC Classification) Model

B. CORPUS

The Urdu columnist's dataset, created by Waheed et al. [25], is known to have contributed to authorship attribution in the Urdu language. For authorship verification, we have used the same technique to create the Urdu dataset used by Waheed et al. [25]. To create a benchmark corpus, we carefully analyzed the available websites. To select a website or blog for data extraction precondition was that it should have data in digital text format and not in jpeg format. After studying several available websites, the following list was selected as a source, predominantly for its' huge collection of documents and authors.

http://www.express.pk

http://www.dunya.com.pk

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

http://www.dunyapakistan.com

http://www.nawaiwaqt.com.pk

1) Ethical Considerations And Data Collection Requirements

Using someone's data for research purposes can involve ethical and legal issues. In our study, we acquired data that was already available publicly, hence providing implicit consent. Furthermore, we also contacted all authors to get permission to use their columns in this study.

There is a large number of authors whose columns are available online. Only columns published in the newspaper were considered in the scope of this study. A decent-sized writing sample of an author is essential to understand his writing style. A minimum limit of 400 articles was set to include an author in the candidate's list. The minimum length of an article to be included in the corpus was set to be 100 words. No constraints were applied in column collection with reference to topics, Gender, and age. Collected columns have a blend of different topics, whatever the authors published. To conduct quality and unbiased research, it is mandatory that collected data does not include any biases and ensures it maintains diversified and realistic nature.

2) Data Collection

Our study focused on collecting data from Urdu newspapers to generate a benchmark corpus for author verification in the Urdu language. After a preliminary study, two approaches were proposed to complete the data collection process.

Manual approach - where the columnist was requested to provide columns or browsing blogs, forums, author websites, or leading newspaper's websites consecutively for columns data where available.

Automatic approach - where data was collected using self-written scripts in PHP from leading newspaper websites/blogs.

For manual data collection, a list of regular Urdu columnists was compiled using mainstream Urdu newspapers of Pakistan, such as Express, Nawa-e-waqat, Dunyapakistan, and Dunya. These columnists were contacted through telephone and email, requesting them to share their columns. Only 26% of the total correspondents responded. Most of the data we got through this source was in jpg image format and not in digital text form. However, some data was in page file format. Collected images were processed using image processing and Optical Character Recognition (OCR) software to get the text in digital format. Output was not useful as it failed to produce an exact copy of the original text.

There is no recognized OCR software for Ur du until now, and the available ones demand larger font sizes for scanned documents [26]. Those used for English are not trustworthy for Urdu because of their intrinsic ambiguous structure. This constraint forced us to collect data from online sources where data was available in digital text format. We started the semi-automatic procedure by successfully browsing the author's columns and storing their URL, column title, column contents, and access time in the database. We

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

performed this activity for ten working days and collected eight hundred columns. This approach was also taking much time and was a laborious one. To speed up this procedure, we wrote webpage scraping scripts in PHP language for each newspaper as the webpage structure of each newspaper was diverse, thus leading to the automatic approach, which was more automatic, efficient, and scientific. In semi-automatic and automatic approaches, an identical list of newspapers was used to collect data from their websites used in the manual approach, except for Jang newspaper, which was excluded from this list because all of its online data was also in jpg image format. As a first step, a URL list was compiled, which contained links to the columns repository of all those authors whose columns were available online on the respective newspaper's website. This list was prepared by semi-automatic procedure by going through the websites identifying the columnist and their column URLs and storing these URLs in the database. In the second step, a web crawler and web scraper were developed to automatically extract relevant data from these URLs. Web scraper and crawler were written in PHP language. The process of data extraction was in two steps. In the first step, all the URLs of the specific author were extracted using a crawler, and in the second step, the webpage scraper used these URLs to extract all available column contents. Initially, we collected over 21,918 documents from these newspaper websites. This process is graphically described in Figure 1. Columns were downloaded in exactly the same form in which they were initially published in the newspapers. No additional contents were added or deleted from the data. , The minimum length of an article to be included in the corpus was set to be 100 words. The significance of excluding too short columns is that the shorter the column, the harder it is to extract stylistic and content-based features from the document. Thus, making it harder to train the system and, if included, lead to poor performance. There was no limit to the maximum number of words in an article. The reason was that as much information about writing style and word structuring is given, a better-trained model is built, leading to better prediction and accurate results. The total accuracy rate (AR) for authorship verification is calculated by Equation as follows:

$$\textbf{Authorship Accuracy Rate} = \frac{\textbf{Number of Correctly Verified}}{\textbf{Total Number of Test Set Articles}} \textbf{ x 100}$$

3) Data Description

In this research, the dataset has been collected from different websites using a web crawler and web scrapper, as previously used by [27]. This data set consists of 1500 Urdu articles published by 15 authors, with 400 articles per author. Table 1 below shows the attributes of the dataset with some initial instances.

Table 1: Attributes of Dataset

Name	Number of articles	Words	Avg. Words
Author 1	400	418265	1046
Author 2	400	484256	1211
Author 3	400	474673	1187
Author 4	400	471024	1178
Author 5	400	526201	1316

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

4) 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.[28] 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:

1) Tokenization.
2) Lower the casing.
3) Removing the "stop words."
4) Stemming.
5) Lemmatization.

4. RESULTS

1) Hybrid Model SVM-XGB

Using the XGBoost Classifier, these two models were blended to improve accuracy. XGB will feed the probability function of SVM with y data. An independent investigation indicated that the hybrid classifier increased accuracy to 95 percent. Figure 6: SVM-XGB Classification Model Performance:

Table 2: Hybrid Model SVM-XGB

Model	Accuracy
SVM	86.00%
XGB	93.00%
SVM-XGB	95.00%

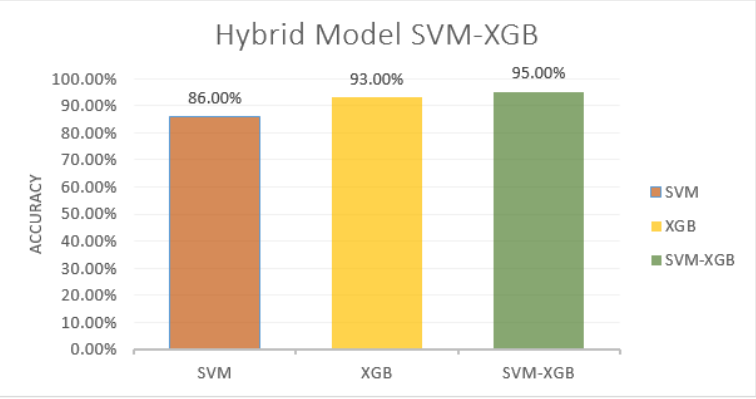

Figure 6: SVM-XGB Classification Model Performance

2) Hybrid Model SVM-GBC

This model was built by combining the SVM and gradient boosting classifier methods to increase their accuracy. The GBC receives y's output immediately. Figure 7 shows the

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

performance of the hybrid SVM-GBC Classification Model, which reached a 91 percent accuracy.

Table 3: Hybrid Model SVM-GBC

Model	Accuracy
SVM	86.00%
GBC	89.00%
SVM-GBC	91.00%

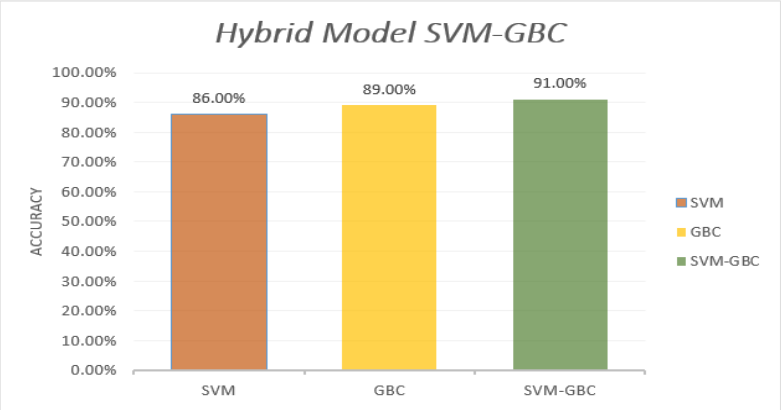

Figure 7: SVM-GBC Classification Model Performance

3) Hybrid Model SVM- CBC

The SVM's probability function will classify the y-coordinates from L(y, F (M-1) (x)) in CBC. SVM-CBC is now the most accurate at 96%. Figure 8 depicts the SVM-CBC Classification Model as a hybrid model:

Table 4: Hybrid Model SVM- CBC

Model	Accuracy
SVM	86.00%
CBC	94.00%
SVM-CBC	96.00%

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

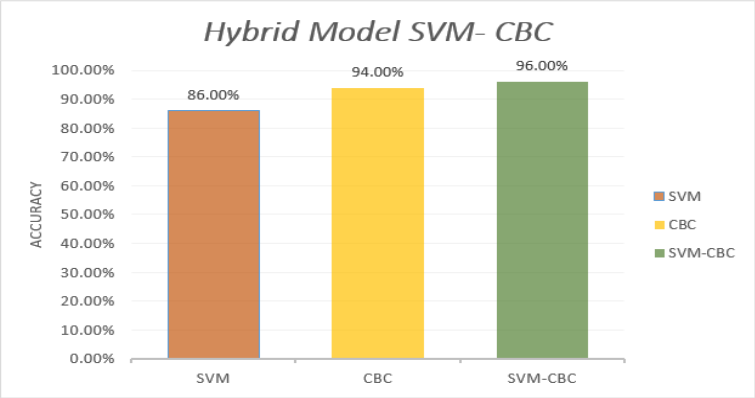

Figure 8: SVM-CBC Classification Model Performance

4) CNN with Adaptive Moment Optimizer

By optimizing the CNN with Adaptive Moment (ADAM) Optimizer we get result upto 98%. We can see the results of CNN without optimization and with ADAM in table 5 and figure 9.

Table 5: CNN with Adaptive Moment Optimizer

Model	Accuracy
CNN	93.00%
CNN-ADAM	98.00%

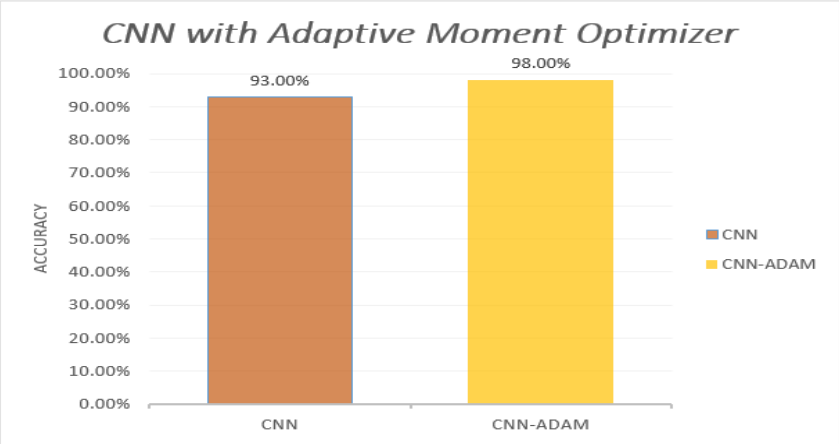

Figure 9: CNN and CNN-ADAM Classification Model Performance

5) CNN with Stochastic Gradient Descent

By optimizing the CNN with Stochastic Gradient Descent (SGD) we get result upto 96%. We can see the results of CNN without optimization and with SGD in table 6 and figure 10.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

Table 6: CNN with Stochastic Gradient Descent

Model	Accuracy
CNN	93.00%
CNN-SGD	96.00%

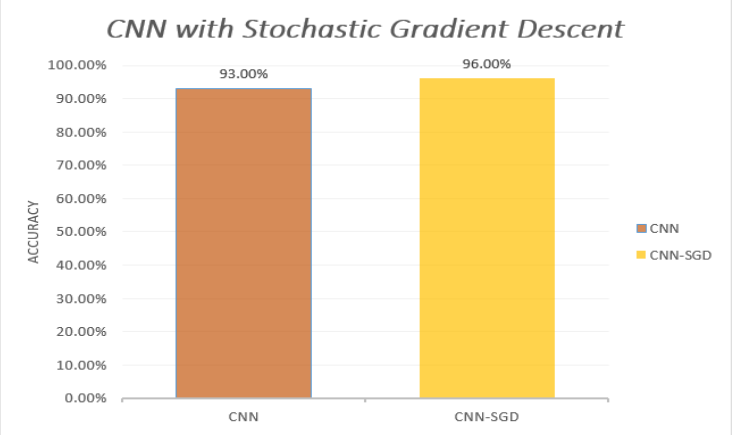

Figure 10: CNN and CNN-SGD Classification Model Performance

6) CNN with Root Mean Square Propagation

By optimizing the CNN with Root Mean Square Propagation (RMSProp) we get result upto 95%. We can see the results of CNN without optimization and with RMSProp in table 7 and figure 11.

Table 7: CNN with Root Mean Square Propagation

Model	Accuracy
CNN	93.00%
CNN-RMSProp	95.00%

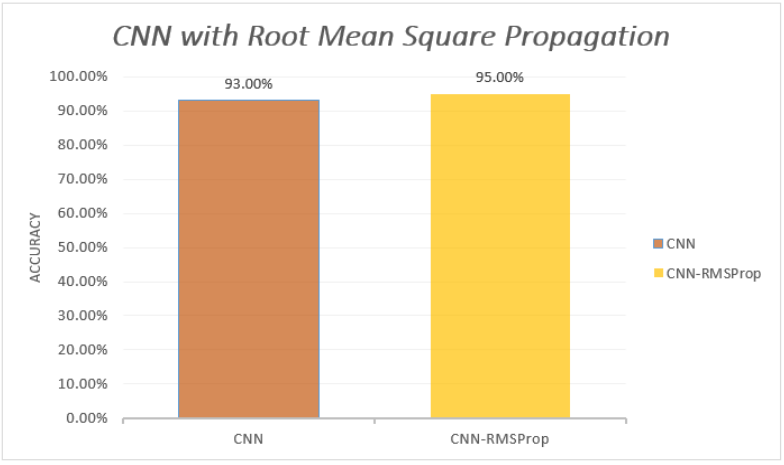

Figure 11: CNN and CNN-RMSProp Classification Model Performance

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

7) Comparative Analysis

The accuracy percentages of each model are presented in the following table. SVM achieved an accuracy of 86%, XGB had 93%, SVM-XGB had 95%, GBC had 89%, SVM-GBC had 91%, CBC had 94%, SVM-CBC had 96%, CNN had 93%, CNN-ADAM had 98%, CNN-SGD had 96%, CNN-RMSprop had 95%. Our proposed CNN-ADAM model outperformed as compared to all models on Urdu data set. The table 8 and figure 12 provides a comparison of the different models.

Table 8: Comparative Analysis for Authorship Verification

Model	Accuracy
SVM	86.00%
XGB	93.00%
SVM-XGB	95.00%
GBC	89.00%
SVM-GBC	91.00%
CBC	94.00%
SVM-CBC	96.00%
CNN	93.00%
CNN-ADAM	98.00%
CNN-SGD	96.00%
CNN-RMSProp	95.00%

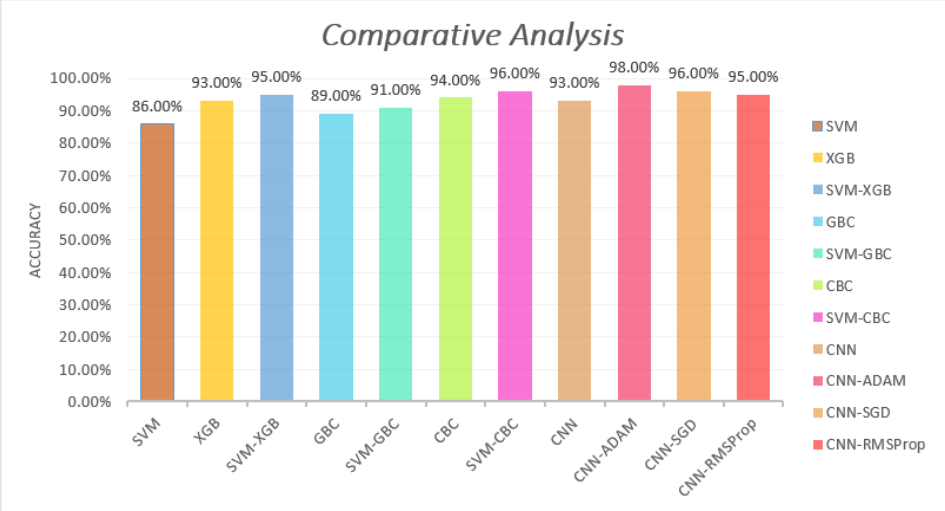

Figure 12: Comparative Analysis

5. CONCLUSIONS

In this work, we evaluated how well several machine learning algorithms verified the authorship of Urdu material. SVM, XGB, GBC, CBC, and CNN-based models using a variety of optimisation methods, including ADAM, SGD, and RMSprop, were among the

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

models that were examined. The outcomes demonstrated that our suggested CNN-ADAM model outperformed all other models, achieving the greatest accuracy of 98% on the Urdu dataset. Our results show, especially when using optimised algorithms, that deep learning-based models are useful for authorship verification of Urdu text. The findings also imply that while a hybrid strategy, such as blending SVM with boosted algorithms, can increase accuracy, it might not be as successful as a CNN model that has been carefully optimised. Overall, this research helps to build powerful machine learning models for determining the authorship of Urdu texts, which has significant applications in areas like forensic linguistics and digital forensics. The use of these models in practical contexts can be explored in more detail, and their adaptability to various text kinds and writers can be assessed.

References

1. X. Chen, P. Hao, R. Chandramouli, and K. P. Subbalakshmi, “Authorship similarity detection from email messages,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6871 LNAI, no. July 2014, pp. 375–386, 2011, doi: 10.1007/978-3-642-23199-5_28.
2. J. Ordoñez, R. R. Soto, and B. Y. Chen, “Will Longformers PAN Out for Authorship Verification? Notebook for PAN at CLEF 2020,” no. September, pp. 22–25, 2020.
3. J. D. Dignam, P. L. Martin, B. S. Shastry, and R. G. Roeder, “TensorFlow: A System for Large-Scale Machine Learning Martín,” *Methods Enzymol.*, vol. 101, no. C, pp. 582–598, 1983, doi: 10.1016/0076-6879(83)01039-3.
4. A. H. El Bakly, N. R. Darwish, and H. A. Hefny, “A Survey on Authorship Attribution Issues of Arabic Text,” *Int. J. Artif. Intell. Syst. Mach. Learn.*, vol. 12, no. 5, pp. 86–92, 2020.
5. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations ofWords and Phrases and their Compositionality,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 1389–1399, 2013, doi: 10.18653/v1/d16-1146.
6. M. Litvak, “Deep dive into authorship verification of email messages with convolutional neural network,” *Commun. Comput. Inf. Sci.*, vol. 898, no. November 2018, pp. 129–136, 2018, doi: 10.1007/978-3-030-11680-4_14.
7. A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, 2023, doi: 10.1016/j.jksuci.2023.01.014.
8. A. Khatun, A. Rahman, M. S. Islam, and Marium-E-Jannat, “Authorship attribution in bangla literature using character-level CNN,” *2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019*, pp. 18–20, 2019, doi: 10.1109/ICCIT48885.2019.9038560.
9. & A.-M. Al-Sarem, M., Al-Kharji, Y., Al-Dossari, S., “nsemble classification model for authorship verification of Arabic texts,” *J. King Saud Univ. Inf. Sci.*, vol. 32(2), 2020.
10. & Z. Yang, X., “Hybrid learning for authorship verification: combining deep neural networks with ensemble models,” *Inf. Sci. (Ny).*, vol. 563, 2021.
11. & 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.
12. & L. Liu, Y., Wei, Z., Chen, Y., “Authorship Verification for Chinese Texts Based on Hyper-Tuned Convolutional Neural Network,” *J. Inf. Process. Syst.*, vol. 161, 2020.

ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 05-2023
DOI 10.17605/OSF.IO/HTVWN

13. & A. Jindal, A., Goyal, P., “A Hybrid Ensemble Model for Authorship Verification,” *Int. Conf. Soft Comput. Signal Process.*, pp. 47–57, 2019.
14. & H. Dai, H., Tian, Y., “Boosted support vector machines for text classification with abstract features,” *Expert Syst. Appl.*, vol. 91, 2018.
15. & Z. Ma, Y., Li, L., “A boosted support vector machines based method for text classification with topic features,” *Knowledge-Based Syst.*, vol. 161, 2019.
16. P. Shrestha, S. Sierra, F. A. González, P. Rosso, M. Montes-Y-Gómez, and T. Solorio, “Convolutional neural networks for authorship attribution of short texts,” *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, pp. 669–674, 2017, doi: 10.18653/v1/e17-2106.
17. O. Halvani, L. Graner, and I. Vogel, “Authorship verification in the absence of explicit features and thresholds,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10772 LNCS, pp. 454–465, 2018, doi: 10.1007/978-3-319-76941-7_34.
18. M. Koppel and J. Schler, “Authorship verification as a one-class classification problem,” *Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004*, pp. 489–495, 2004, doi: 10.1145/1015330.1015448.
19. M. L. Brocardo, I. Traore, I. Woungang, and M. S. Obaidat, “Authorship verification using deep belief network systems,” *Int. J. Commun. Syst.*, vol. 30, no. 12, 2017, doi: 10.1002/dac.3259.
20. & A.-N. Alghamdi, M., “Hybrid deep learning and ensemble learning approach for Arabic authorship verification,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10(2), 2019.
21. & R. Dashtipour, K., Mirian, M. S., “A hybrid SVM-DNN approach for Persian authorship verification,” *J. Ambient Intell. Humaniz. Comput.*, vol. 12, 2021.
22. & N. Abbasi, A., “Ensemble-based authorship attribution using character-level language models,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 64(10), 2013.
23. U. Naqvi, A. Majid, and S. A. Abbas, “UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods,” *IEEE Access*, vol. 9, pp. 114085–114094, 2021, doi: 10.1109/ACCESS.2021.3104308.
24. T. F. Khan, W. Anwar, M. Ahmed, S. N. Abbas, and A. Ali, “HYBRID ENSEMBLE LEARNING BASED AUTHORSHIP VERIFICATION MODELS BASED ON URDU TEXTUAL DATA,” pp. 334–359, 2023, doi: 10.17605/OSF.IO/X4DE3.
25. 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.
26. S. T. Javed and S. Hussain, “Segmentation Based Urdu Nastalique OCR,” no. c, 2013, pp. 41–49.
27. W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, “An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution,” *IEEE Access*, vol. 7, pp. 3224–3234, 2019, doi: 10.1109/ACCESS.2018.2885011.
28. M. L. Brocardo, I. Traore, and I. Woungang, “Authorship verification of e-mail and tweet messages applied for continuous authentication,” *J. Comput. Syst. Sci.*, vol. 81, no. 8, pp. 1429–1440, 2015, doi: 10.1016/j.jcss.2014.12.019.