

Report on Bengali Literature Author Identification Using Pre- trained Models

by Ch Zakauddin Md Ruslan

Submission date: 22-Feb-2023 12:02AM (UTC+0530)


Submission ID: 2019700543

File name: sc_paper_5.pdf (268.76K)

Word count: 3217

Character count: 16354

Bengali Literature Author Identification Using Pre-trained Models

 Yumna Islam
Department of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
180204046@aust.edu

Ch Zakauddin Md Ruslan
Department of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
180204058@aust.edu

Salehin Bin Iqbal
Department of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
180204062@aust.edu

 Farhana Azad
Department of CSE
Ahsanullah University of
Science and Technology
Dhaka, Bangladesh
180204068@aust.edu

Abstract—There are millions of people around the world speaking in Bangla language but not much work has been done on Bangla language because it is not a language that has been digitized effectively, and the reason is the lack of annotated computer-readable datasets and inadequate resources. This work focuses on author identification from bengali literature that can be used for various purposes like identifying plagiarism or anonymous authors. Although few works are done on Bengali language there are some previous works on bengali literature author identification. This work tries to improve the previous works (using pre-trained models) by achieving a better accuracy.

Index Terms—Author Identification, BERT, Transformer models

I. INTRODUCTION

People around the world speak a different languages. There are around 7,100 languages all around the world. Among them, Bangla or Bengali is the 7th largest language with over 265 million people talking in this language all around the world [1]. But there are very few works done on Bangla Language in general as it is not a very well-developed language in terms of digitization. It is due to the scarcity of annotated computer-readable datasets and the limited support for resource building. Authorship attribution is an arising area of research in the field of Natural Language Processing (NLP) [2]. The process of identifying the author of a specific text from a group of authors is known as authorship attribution. It enables us to identify the most likely author of a given text. Authorship identification can be used for purposes like identifying ghost writers, anonymous authors and catching plagiarism [2].

Every author has a unique writing style and human readers can distinguish at some extent. The importance of author identification basically relies on its wide applications. It offers a creative way to introduce readers to writers, such as anonymous writers or authors with striking stylistic similarities to their favorite writers. So working on authorship attribution on Bengali literature is very fruitful. There are some works on this author identification, so this work is an attempt to be a better version of the previous works by achieving better accuracy.

In order to achieve a better accuracy in identifying authors on bengali literature we have used different pre-trained models (e.g. BERT) and showed a comparative study. The corpus that we have used for this paper was highly imbalanced. So, the data has been stratified and each class has same amount of data for more accurate result.

II. RELATED WORKS

Das et al. (2011) [2] worked on a comprehensive study on 'Author Identification in Bengali Literary works'. It can identify the author from a piece of literary work in Bengali. They have used a self-made dataset in this work and have used Stylometric Feature or uni-gram bi-gram classification to classify the data from the dataset. KNN and Naive Bayes Classifier are used to Classify the data. The model that has an almost around 91.67% accuracy rate.

Hossain et al. (2021) [3] worked on 'Authorship Classification in a Resource Constraint Language Using Convolutional Neural Networks'. Authorship classification is done by CNN which consists of four modules. BAAD16, LD, and BACC-18 dataset have been used for this purpose. Four classification models (CNN, LSTM, SVM, SGD) and three embedding techniques (Word2Vec, GloVe, and FastText) have been used to construct embedding models. The result of this system (accuracy) is 93.45% for BACC-18, 95.02% for BAAD-16, and 98.67% for LD.

Ontika et al. (2020) [4] proposed a study on 'Author Identification from Song Lyrics'. In this paper, the author is identified from a given song lyrics. The authorship of a song is determined using verbal information as well as six machine learning techniques: Ridge Regression, Perception, Passive-Aggressive, Support Vector Machine, Stochastic Gradient Descent, and Naive Bayes. The result of this system is 80.70% using Ridge Regression, 79.10% using Perception, 83.50% using Passive Aggressive, 77.60% using SVM, 84.50% using SGD, and 86.70% using Naive Bayes.

Khatun et al (2019) [1] did Authorship Attribution in Bangla literature using Character-level CNN. Here they used the BAAD-16 dataset consisting of 1,122,875 samples. Here they used the BAAD-16 dataset consisting of 1,122,875 samples. They used Character level CNN here because it can sufficiently

replace words for classifications. The CNN model is almost 81.3 to 98 percent accurate for the given dataset.

Dhar et al. (2020) [5] worked on 'Author Identification from Literary Articles with Visual Features: A Case Study with Bangla Documents' in which he proposed a CNN based author identification from articles. 93.58% accuracy was achieved by working with 1200 articles and 50 authors by using image based features.

III. BACKGROUND STUDY

We have used BERT based models which are pre-trained for bengali language in our work like namely Bangla-BERT Base [6], BanglaBERT generator [7], BanglaBERT [8] and SahajBERT [9].

A. BERT model

The abbreviation of the model BERT [8] is Bidirectional Encoder Representations from Transformers is a transformer based model which is developed by Google widely used in NLP (natural language processing). We are using BanglaBERT of CSEBUETNLP available on the HuggingFace website. The model is based on the BERTBase model. We have made changes in the model to better suit our requirements. We have made the weight decay 0.01. We have changed the learning rate to $2e-5$. The maximum length of a sentence in our dataset is 1482 but we have shortened it by truncating it to 512, as BERT cannot take a sentence length of more than 512. We have split our data 9:1 as training and testing data. And from the training data we have split in 9:1 as training data and validation data.

B. Bangla-BERT-Base

Bangla Bert base [6] is a pre-trained language model of bengali language. Masking language modelling is used to build Bangla-bert base. Masking language modelling is a bert based model which takes a sentence as input and masks 15 percent of the words and runs the model with masked word to predict the context of the word. Here we are using Bangla-BERT base available on HuggingFace website.

C. Bangla-BERT generator

BanglaBERT base pretrained generator model [7]. This model is an ELECTRA generator model which has been pre-trained with masked language modeling (MLM) works for large amount of corpora.

D. BanglaBERT

This [8] is a pre-trained ELECTRA discriminator model with the Replaced Token Detection (RTD) objective. Many tasks related to NLP like Sentiment classification, Named Entity Recognition, Natural Language Inference etc can be done by finetuning this model to achieve state-of-the-art results.

E. SahajBERT

Pre-trained BERT model on bengali language by using masked language modeling and sentence order prediction. This SahajBERT [9] model is composed of tokenizer and ALBERT architecture. Sequence classification, token classification, making decisions and question answering are done by this model.

IV. DATASET

The dataset from paper [10] is used which is BAAD16 an Authorship Attribution corpus for Bengali literature. The data was acquired and analyzed by the study's authors. It was created by extracting content from an online Bangla e-library using a custom web crawler and includes literary works by various well-known Bangla writers. It contains the books, stories, series, and other works of 16 authors. Every sample document contains 750 words. An imbalanced corpus which accurately reflects actual circumstances, where not all authors will have a large number of sample texts.

Our Dataset is a multiclass dataset. There are many authors in Bengali literature. So our dataset has multiple classes depending on the number of authors. Our whole dataset contains works of 16 authors so there are 16 classes. There are a total of 17,966 number of samples in the dataset with 590,660 unique words and the word count is 13,474,500.

Description of the dataset is given below in the table [10]

Author Name	Number of Samples	Word Count	Unique Word
zahir rayhan	185	138k	20k
nazrul	223	167k	33k
manik band-hopaddhay	469	351k	44k
nihar ronjon gupta	476	357k	43k
bongkim	562	421k	62k
tarashonkor	775	581k	84k
shottojit roy	849	636k	67k
shordindu	888	666k	84k
toslima nasrin	931	698k	76k
shirshendu	1048	786k	69k
zafar iqbal	1100	825k	53k
robindronath	1259	944k	89k
shorotchanra	1312	984k	78k
shomresh	1408	1056k	69k
shunil gongopad-dhay	1963	1472k	109k
humayun ahmed	4518	3388k	161k

V. METHODOLOGY

At first we have done some preprocessing on our corpus and then we have run several bert models in order to identify

author. Our work can be divided into three stages which are as follows:

1. Preparing Dataset
2. Execution of different Bert models
3. Performance Metrics

A. Preparing Dataset

The BAAD'16 corpus was employed in our study. From the original corpus's four features, we extracted two features (label, text). As this corpus is extremely imbalanced, we have only used 6 of the 16 classes in this corpus for the majority of the Bert models. After that, we normalized our labels and took 1100 data points for each class. Then, we divided our corpus into training, testing, and validation sets, with 80% of the data being in the training set, 10% being in the validation set, and 10% being in the testing set. The Bert tokenizer is used for tokenization. Normalization and truncation are carried out during the tokenization process.

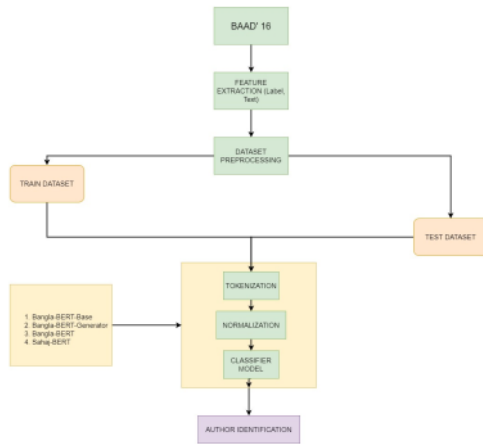


Fig. 1. proposed methodology

B. Execution of Different Bert Models

We have used several Bert models for our work. 3 different Bert models are used here in order to show a comparative study between the models. At first, we have worked with BanglaBERT. This model is executed twice: one execution with 16 classes and another execution with 6 classes. For 16 classes, we have worked with a batch size of 16 and a learning rate of $2e-5$ which executed for 5 epochs, and for 6 classes, we have worked with a learning rate of $2e-5$, batch size of 16 which executed for 10 epochs. The second model that we have used is Bangla-BERT-Generator with a batch size of 16 and a learning rate of $2e-5$ where 15 epochs have been executed for 6 classes. Furthermore, for our third Bert model which is BanglaBERT-Base, here we have taken batch size 16 and learning rate $2e-5$ which executed for 10 epochs. Then, lastly, we have worked with Sahaj-BERT with batch size 6, learning rate $5e-6$ which executed for 3 epochs.

C. performance Metrics

An easiest way to evaluate classifier's performance is using accuracy metric. Each data point's true values and predicted values are compared in this case, and a match counts as one accurate prediction. The accuracy is then calculated as the ratio of accurate forecasts to all other incorrect ones.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

Precision is the ratio of true positive which is correctly classified positive samples to the total number of positively classified samples (either true positive or false positive). Precision is calculated as shown below:

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

The recall is basically the ratio of total number of positive samples and total number of positive samples that were predicted positive or correctly. The recall measures the model's ability to detect positive samples. Higher recall means more positive samples are found. Recall can be calculated as follows:

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

F1 takes into consideration both Precision and Recall. It is calculated as follows:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

F1-score is basically the harmonic mean of precision and recall and provides a balance between them.

VI. EXPERIMENTAL RESULT

For BanglaBERT model with 16 classes, we have achieved 89% accuracy where overall precision score is 0.8994, recall score is 0.9803, and f1-score is 0.8933. Classwise precision, recall, and f1-score are shown below:

Author Name	Precision Score	Recall Score	F1 Score
humayun ahmed	0.98	0.98	0.98
shunil gongopad-dhay	0.97	0.95	0.96
shomresh	0.84	0.86	0.85
shorotchandra	0.95	0.95	0.95
robindronath	0.88	0.88	0.88
MZI	0.92	0.95	0.94
shirshendu	0.81	0.82	0.84
toslima nasrin	0.91	0.84	0.87
shordindu	0.79	0.91	0.85
shottojit roy	0.86	0.95	0.91
tarashonkor	0.78	0.78	0.78
bongkim	0.83	0.79	0.81
nihar ronjon gupta	0.80	0.69	0.74
manik band-hopaddhay	0.57	0.74	0.65
nazrul	0.78	0.32	0.45
zahir rayhan	0.29	0.11	0.16

Epoch wise training and validation loss are shown in the following graph for BanglaBERT model with 16 classes.



Fig. 2. Epoch vs loss graph for BanglaBERT for 16 classes

For the rest of the approaches we have taken the data of the first 6 class with 1100 data.

For Bangla-BERT-Base model we have achieved 99% accuracy. The total Precision score of the model was 0.9896. The recall score was 0.9894 and f1 score was 0.9894.

Author Name	Precision Score	Recall Score	F1 Score
humayun ahmed	1.00	1.00	1.00
shunil gongopad-dhay	0.98	1.00	0.99
shomresh	0.99	0.98	0.99
shorotchandra	0.99	1.00	1.00
robindronath	0.97	0.97	0.97
MZI	1.00	0.98	0.99

Epoch wise training and validation loss are shown in the following graph for Bangla-BERT-Base model with 6 classes.

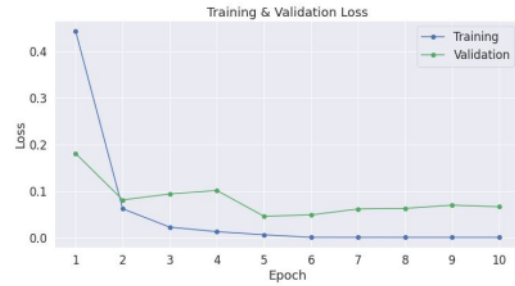


Fig. 3. Epoch vs loss graph for Bangla-BERT-Base

BanglaBERT-Generator model gives 97% accuracy where the precision score is 0.9700, recall score is 0.9696 and f1-score is 0.9696.

Author Name	Precision Score	Recall Score	F1 Score
humayun ahmed	0.97	0.93	0.95
shunil gongopad-dhay	0.96	0.98	0.97
shomresh	0.91	0.94	0.92
shorotchandra	1.00	1.00	1.00
robindronath	1.00	0.97	0.99
MZI	0.98	1.00	0.99

Epoch wise training and validation loss are shown in the following graph for BanglaBERT-Generator model with 6 classes.

For the BanglaBERT model we have achieved 98% accuracy. The overall precision score is 0.98057, recall score is 0.9803 and f1-score is 0.9803.

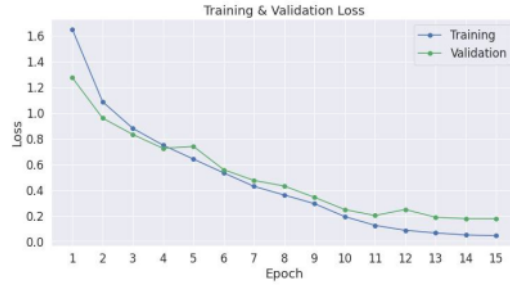


Fig. 4. Epoch vs loss graph for BanglaBERT-Generator

Author Name	Precision Score	Recall Score	F1 Score
humayun ahmed	0.97	0.95	0.96
shunil gongopad-dhay	0.97	1.00	0.99
shomresh	0.95	0.95	0.95
shorotchandra	1.00	1.00	1.00
robindronath	0.99	0.98	0.99
MZI	0.99	1.00	1.00

Epoch wise training and validation loss are shown in the following graph for BanglaBERT model with 6 classes.

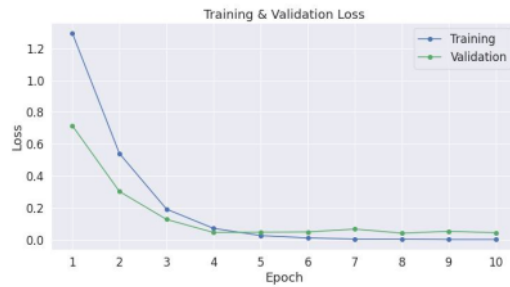


Fig. 5. Epoch vs loss graph for BanglaBERT for 6 classes

For the Sahajbert model we have achieved 98% accuracy. The overall precision score is 0.9838, recall score is 0.9833 and f1-score is 0.9833.

Author Name	Precision Score	Recall Score	F1 Score
humayun ahmed	0.99	0.98	0.99
shunil gongopad-dhay	0.96	1.00	0.98
shomresh	0.99	0.98	0.99
shorotchandra	0.96	1.00	0.98
robindronath	1.00	0.95	0.97
MZI	0.99	0.99	0.99

Epoch wise training and validation loss are shown in the following graph for SahajBERT model with 6 classes.

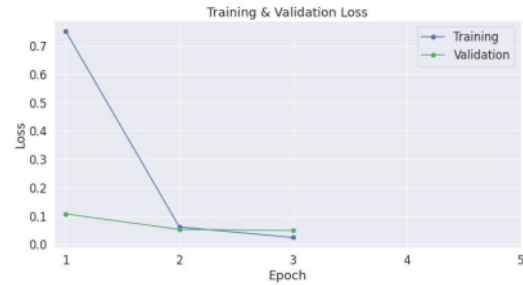


Fig. 6. Epoch vs loss graph for SahajBERT

VII. RESULT ANALYSIS

A comparative study has been shown using different Bert models. We have achieved highest accuracy for Bangla-BERT-Base model (99%) and lowest accuracy (89%) for BanglaBERT model (16 classes). We have achieved low accuracy for 16 classes as the corpus is highly imbalanced. However, while working with 6 classes we achieved high accuracy. SahajBERT and BanglaBERT model (with 6 classes) give similar (98%) accuracy. As Bangla-BERT-Base has comparatively more parameters than other models we have achieved highest accuracy for this model. Having more parameters enable this model to learn better while working with this specific corpus. In addition to that we have achieved better accuracy than the previous work that has been done [1] as we have used transformer based models. The previous paper [1] achieved highest accuracy for fastText (98%) model where in our study we have achieved highest accuracy for Bangla-BERT-Base model (99%).

VIII. CONCLUSION

Identification of authors is a very crucial task to do in this era. There are very few works that have been done on this topic. Many new deep learning models were used on this system to detect the identity of the authors of given works of literature. BERT, BANGLABERT-BASE, CSEBUET-NLP, BANGLABERT, SAHAJBERT, and various classifier models were used here to identify the authors using the BAAD 16 dataset. After pre-processing the dataset, the model was trained and tested based on the significant attributes.

REFERENCES

- [1] A. Khatun, A. Rahman, M. S. Islam *et al.*, "Authorship attribution in bangla literature using character-level cnn," in *2019 22nd International conference on computer and information technology (ICCIT)*. IEEE, 2019, pp. 1–5.
- [2] S. Das and P. Mitra, "Author identification in bengali literary works," in *Pattern Recognition and Machine Intelligence: 4th International Conference, PReMI 2011, Moscow, Russia, June 27-July 1, 2011. Proceedings* 4. Springer, 2011, pp. 220–226.
- [3] 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.
- [4] N. N. Ontika, M. F. Kabir, A. Islam, E. Ahmed, and M. N. Huda, "A computational approach to author identification from bengali song lyrics," in *Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2018*. Springer, 2020, pp. 359–369.
- [5] A. Dhar, H. Mukherjee, S. Sen, M. O. Sk, A. Biswas, T. Gonçalves, and K. Roy, "Author identification from literary articles with visual features: A case study with bangla documents," *Future Internet*, vol. 14, no. 10, p. 272, Sep 2022. [Online]. Available: <http://dx.doi.org/10.3390/fi14100272>
- [6] S. Sarker, "Banglabert: Bengali mask language model for bengali language understading," 2020. [Online]. Available: <https://github.com/sagorbrur/bangla-bert>
- [7] A. Bhattacharjee, T. Hasan, K. Mubasshir, M. S. Islam, W. A. Uddin, A. Iqbal, M. S. Rahman, and R. Shahriyar, "Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla," in *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*, 2022. [Online]. Available: <https://arxiv.org/abs/2101.00204>
- [8] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: <https://aclanthology.org/2022.findings-naacl.98>
- [9] "neuropark/sahajbert." [Online]. Available: <https://huggingface.co/neuropark/sahajBERT>
- [10] A. Khatun, A. Rahman, and M. S. Islam, "Baad16: Bangla authorship attribution dataset," <https://data.mendeley.com/datasets/6d9jrkgtvv/4>, 2019.

Report on Bengali Literature Author Identification Using Pre-trained Models

ORIGINALITY REPORT

22%
SIMILARITY INDEX

20%
INTERNET SOURCES

17%
PUBLICATIONS

9%
STUDENT PAPERS

PRIMARY SOURCES

1 huggingface.co 4%
Internet Source

2 arxiv.org 3%
Internet Source

3 rifatshahriyar.github.io 2%
Internet Source

4 Ibrahim Al Azhar, Sohel Ahmed, Md. Saiful Islam, Aisha Khatun. "Identifying Author in Bengali Literature by Bi-LSTM with Attention Mechanism", 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021 1%
Publication

5 aclanthology.org 1%
Internet Source

6 www.mdpi.com 1%
Internet Source

7 ojs.academypublisher.com 1%
Internet Source

8	downloads.hindawi.com Internet Source	1 %
9	"Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2022 Publication	1 %
10	Submitted to Cranfield University Student Paper	1 %
11	analyticsindiamag.com Internet Source	1 %
12	escholarship.org Internet Source	1 %
13	Yusuke Mori, Youichiro Miyake. "Ethical Issues in Automatic Dialogue Generation for Non- Player Characters in Digital Games", 2022 IEEE International Conference on Big Data (Big Data), 2022 Publication	1 %
14	curve.carleton.ca Internet Source	1 %
15	Omar Sharif, Mohammed Moshiul Hoque. "Tackling Cyber-Aggression: Identification and Fine-Grained Categorization of Aggressive Texts on Social Media using Weighted Ensemble of Transformers", Neurocomputing, 2021 Publication	1 %

16	www.kdnuggets.com Internet Source	1 %
17	Aisha Khatun, Anisur Rahman, Md Saiful Islam, Hemayet Ahmed Chowdhury, Ayesha Tasnim. "Authorship Attribution in Bangla Literature (AABL) via Transfer Learning using ULMFiT", ACM Transactions on Asian and Low-Resource Language Information Processing, 2022 Publication	<1 %
18	Submitted to Eastern Mediterranean University Student Paper	<1 %
19	www.e-informatyka.pl Internet Source	<1 %
20	Atish Kumar Dipongkor, Md. Saiful Islam, Humayun Kayesh, Md Shafaeat Hossain, Adnan Anwar, Khandaker Abir Rahman, Imran Razzak. "DAAB: Deep Authorship Attribution in Bengali", 2021 International Joint Conference on Neural Networks (IJCNN), 2021 Publication	<1 %
21	Submitted to University of Durham Student Paper	<1 %
22	Submitted to Guru Jambheshwar University of Science & Technology Student Paper	<1 %

23

videomeansbusiness.com

Internet Source

<1 %

24

Arefin Niam, Avijit Das, Summit Haque.
"Chapter 46 Performance Analysis
and Implementation of Pre-trained Model
Using Transfer Learning on Bangla Document
Clustering", Springer Science and Business
Media LLC, 2022

Publication

<1 %

25

lib.buet.ac.bd:8080

Internet Source

<1 %

26

pure.ulster.ac.uk

Internet Source

<1 %

27

Ergün Yücesoy. "Speaker age and gender
classification using GMM supervector and
NAP channel compensation method", Journal
of Ambient Intelligence and Humanized
Computing, 2020

Publication

<1 %

28

Fred Nwanganga, Mike Chapple. "Evaluating
Performance", Wiley, 2020

Publication

<1 %

29

Md. Rajib Hossain, Mohammed Moshiul
Hoque, M. Ali Akber Dewan, Nazmul Siddique,
Md. Nazmul Islam, Iqbal H. Sarker.
"Authorship Classification in a Resource

<1 %

Constraint Language Using Convolutional Neural Networks", IEEE Access, 2021

Publication

30

Ankita Dhar, Himadri Mukherjee, Shibaprasad Sen, Md Obaidullah Sk, Amitabha Biswas, Teresa Gonçalves, Kaushik Roy. "Author Identification from Literary Articles with Visual Features: A Case Study with Bangla Documents", Future Internet, 2022

<1 %

Publication

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off