iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.51223

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Author Identification on Anonymous Regional Literature

Prof. Virendra Bagade[1], Swapnil Chavan[2], Suyash Joshi[3], Koushik Batgiri[4], Mehul Patil[5]
[1, 2, 3, 4,5]*Computer Engineering Department Pune Institute of Computer Technology, Pune*

Abstract: ***In order to identify the author of a given text, researchers have used many different methods. One such method is forensic linguistics which examines stylistic and content-based aspects of a text in an effort to determine who wrote it. Another area where author identification is often useful is bot detection; this involves identifying automated accounts on social media or other websites. Finally, marketing research can also be utilized for identifying the authors behind advertisements. Despite significant progress having been made with English texts, there remains much work to be done when it comes to regional variations in language use. In this study, we aim to build a machine learning model that will be able to identify the probable author of the provided anonymous text using various lexical and syntactic variables of the literature.***
Keywords: Author Identification, Classification, Text Processing, Feature Extraction, Multi-label text classification

I. INTRODUCTION

Establishing connections between authors and their literary work is the primary goal of the scientific field of authorship identification. Nirkhi S. and Dharaskar R. state that as persons differ in their word choices, sentence structure methods and punctuation usage, academics in this subject believe that writers unintentionally establish their distinctive writing styles [1]. Early linguists discovered that, like biological fingerprints, most people have distinctive stylistic distinguishing characteristics and peculiarities of their own. Although these specific traits differ from person to person, according to Holmes D, they typically hold true across the writings of the same individual despite their diverse writing styles [2]. This concept has served as the foundation for linguistic scholars' work on a variety of analysis features and methodologies, which has led to notable results in the authorship identification study field. Finding the author of a particular piece of text from a given list of candidate writers (suspects) is known as authorship identification. From the perspective of machine learning, it may be understood as a multiclass single-label text classification task where the author represents a class (label) of a given text. Topic-based classification makes use of semantic features, whereas author-based classification makes use of stylometric features. The features that focus on the patterns appearing in the text for the same author are known as Stylometric features.

In the earlier research on this subject, classification models were solely built using lexical characteristics and a Sequential Minimal Optimizer with Rule Based Decision Tree.

The model can be enhanced by use of transfer learning on newer transformer-based models which are known for their better accuracy on NLP tasks including text classification, language translation, language modelling, sentiment analysis, and text summarization.

II. MOTIVATION

Author identification on anonymous regional literature is an interesting and a topic full of curiosity to explore. With this technique we can help identify or receive recognition of an author who has been cloaked in mystery for decades or centuries. The topic is really very important to expand the research and knowledge gates of computer science to other literature which have been confined only to English literatures. The discovery of the author of a specific literature helps us in gaining insights into personal experiences, beliefs, philosophy, etc. With authorship identification we can explore in detail themes, style, grammar, sentence construction of an author. Authorship identification is an interesting intellectual challenge. It requires critical thinking, creative problem solving, analysis of minute details sculpted all together and arrive at a conclusive identification.

III. RELATED WORK

In this section, we review earlier authorship analysis research with a focus on stylometric traits that were applied to the text representation. In the literature, a few characteristics have been put out to characterize a particular author's writing style.

The most popular elements that have been utilized to represent the text in earlier research are mentioned below.

A. Lexical Features

Lexical characteristics refer to the idea of reading a text as a collection of tokens put together in sentences. Character-based and word-based lexical features could be used to categorize these characteristics.

B. Syntactic Features

The use of syntactic data to identify the authors' unintentional sentence-level syntactic patterns, such as typos, part-of-speech, sentence structure, function word frequency.[5]

C. Character Features

The use of syntactic data to identify the authors' unintentional sentence-level syntactic patterns, such as typos, part-of-speech, sentence structure, function word frequency.[5]

D. Semantic Features

According to Pandian et al., semantic characteristics encompass a collection of structures that exaggerate a word's significance [8]. For managing semantic investigations, Natural Language Processing (NLP) methods are outdated and insufficient. The implementation of semantic characteristics, including semantic dependencies and Systemic Functional Linguistics (SFL), synonyms, which explains intentional words in combination with POS features, has therefore been quite limited. Using 4 the right tools, it is possible to extract writing style traits from many scripts, including lexical, character, syntactic, and semantic traits.

IV. PROPOSED IMPLEMENTATION

The problem in this study falls in the category of classification problem, so to solve this problem we will make use of the machine learning classification techniques. We will make use of some neural network based classifiers and compare their results in order to find out which model serves the purpose in a better way.

A. Classification Algorithm

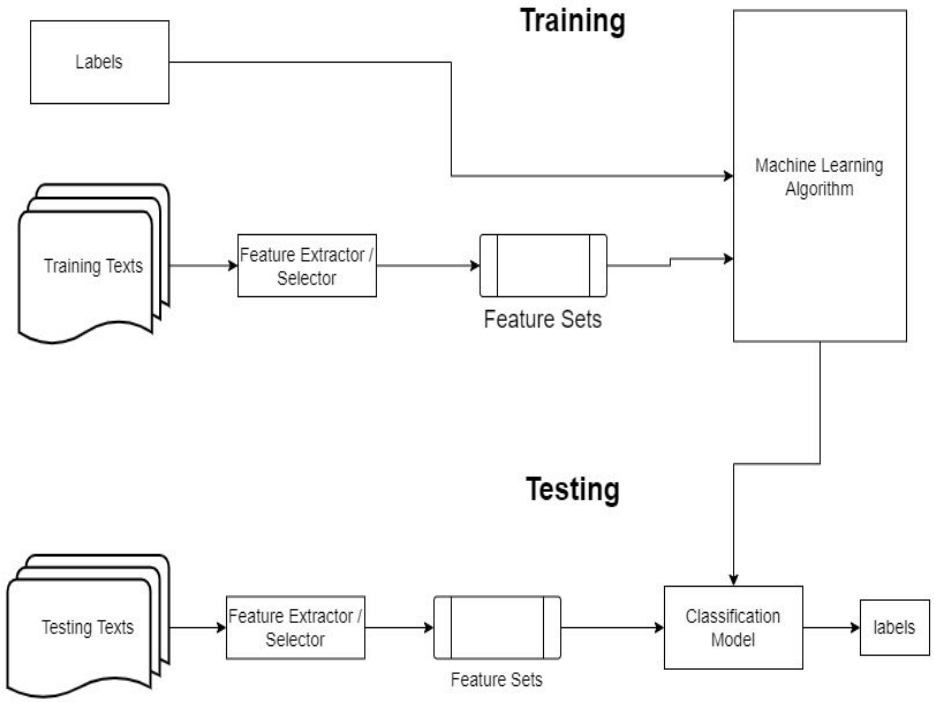


Fig : A classification block diagram

B. Proposed Workflow

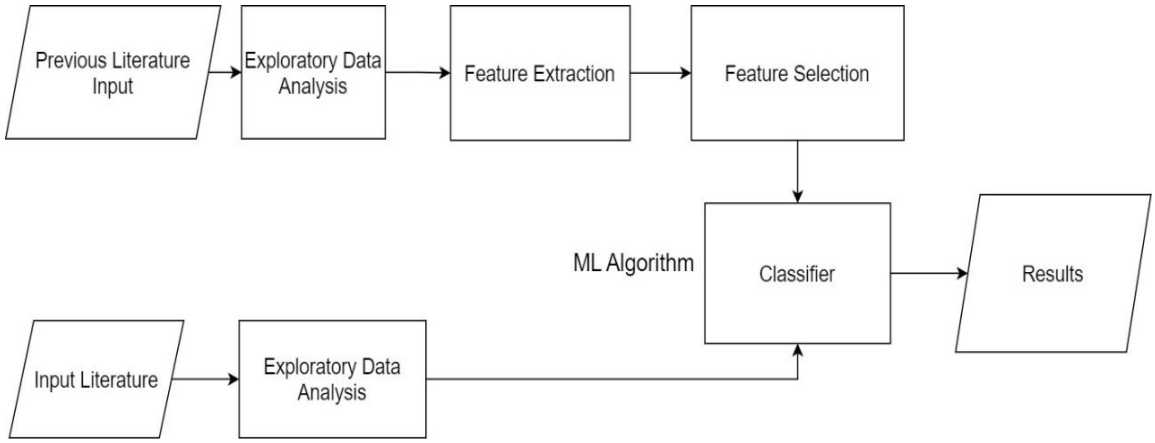


Fig : Proposed workflow for implementation

V. METHODOLOGY

In this study, we intend to create a classification model that will take the previous literature text of the probable authors as its input and train itself on the input data to be able to classify the test data. Dataset used for this study is csv file having two cloumns namely content and authors. We made this dataset by extracting the text from epub files of the books and labelling it with author's name. We used 6 books written by 4 different authors.

1) *Model Building:* Transformer models have revolutionized the field of Natural Language Processing (NLP) with their exceptional ability to learn long-term dependencies and capture contextual information in text data. It was introduced in 2017 by Vaswani et al. [9] and has since been used in a variety of NLP applications, including language translation, language modeling, sentiment analysis, and text summarization. The Transformer architecture relies on attention mechanisms that allow it to consider all words in a sequence simultaneously, which makes it more efficient than previous models, such as recurrent neural networks (RNNs), while also achieving state-of-the-art performance. The pre-trained Transformer models, such as BERT, GPT-2, and XLNet, have shown remarkable performance on various benchmark.

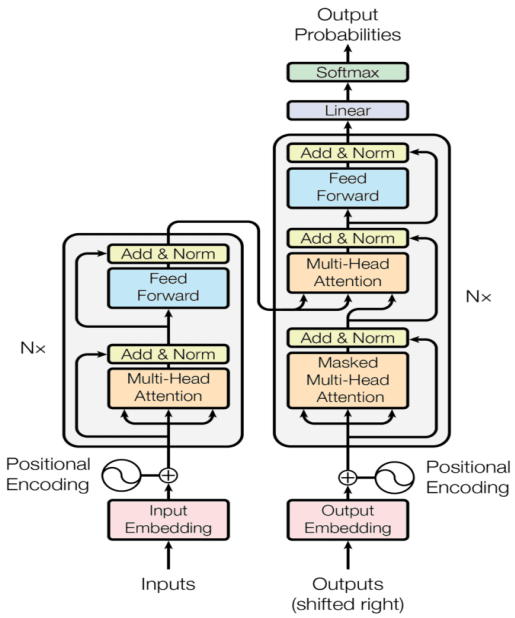


Fig : The Transformer-Model Architecture [9]

BERT (Bidirectional Encoder Representations from Transformers), a pre-trained Transformer-based model, has shown remarkable performance on various NLP tasks, including text classification [10]. BERT is 7 particularly well-suited for author identification because it can learn complex linguistic patterns and relationships from large amounts of unlabelled text data. By pre-training on massive amounts of text data, BERT can learn a deep understanding of language that can be fine-tuned to specific NLP tasks such as author identification. BERT is multilingual and has been trained on 104 different languages with large corpus of texts. Marathi is one of the 104 languages which makes it perfect for task of Author identification on Marathi literature, steps involved are as follows:

a) Pre-processing data for BERT: In BERT, Longer sequences are disproportionately expensive because attention is quadratic to the sequence length. Hence, BERT will not consume more than 512 tokens. Any input greater than 512 tokens is truncated. Author identification dataset contains large lengths of text, pre-processing of input data is required to restrict sequence length. Splitting the input string into separate strings results in loss of information regarding context between two individual sub-strings. In order to avoid this, "Sliding Window" mechanism is used which splits the text into multiple strings such that the contents of sub-string overlap.

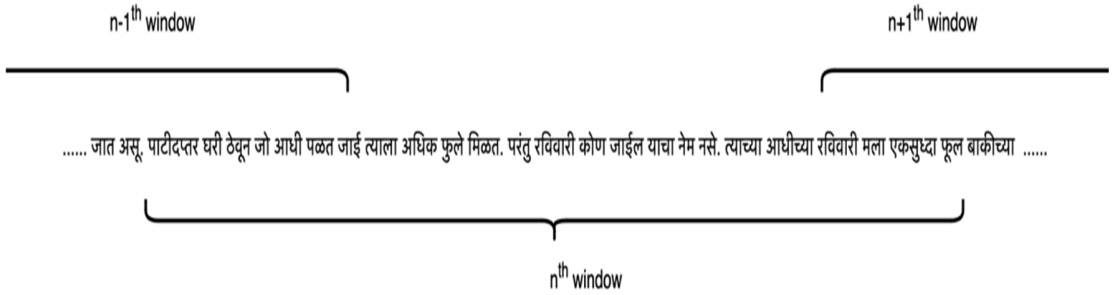


Fig : Sliding Window Technique for splitting large input text

In our implementation, we overlap 20% of part of previous window to next window, allowing BERT to learn from long texts without significant loss of information.

b) *BERT Pre-processing:* Convert the pre-processed text into BERT input format, which is tokenized and padded to a fixed length.

c) *BERT Fine-Tuning:* Fine-tune the pre-trained BERT model on the training set by optimizing the model's parameters for the author identification task. The BERT model is typically trained using a supervised learning approach where the input text is paired with the author label. Conversion of author labels from categorical to numeric datatype is performed.

d) *Model Evaluation:* Evaluation of performance of the BERT model on the testing dataset using metrics such as accuracy, precision, recall, and F1-score. Comparison of these scores on train and test dataset gives information about overfitting or underfitting.

e) *Model Tuning:* If the model's performance is not satisfactory, tuning the hyperparameters, such as learning rate, batch size, and number of epochs is performed which is followed by retraining the model.

VI. CONCLUSIONS

Transformer based models are more accurate than traditional analytical techniques and Neural Networks. Due to their deep understanding on NLP and multilingual nature they are the State of the Art (SOTA) models for text classification tasks. Following are the performances of various models on task of author identification on Marathi literature which involves classification of literature text between three well known Marathi authors namely V. S. Khandekar, Shivaji Sawant and Ranjeet Desai.

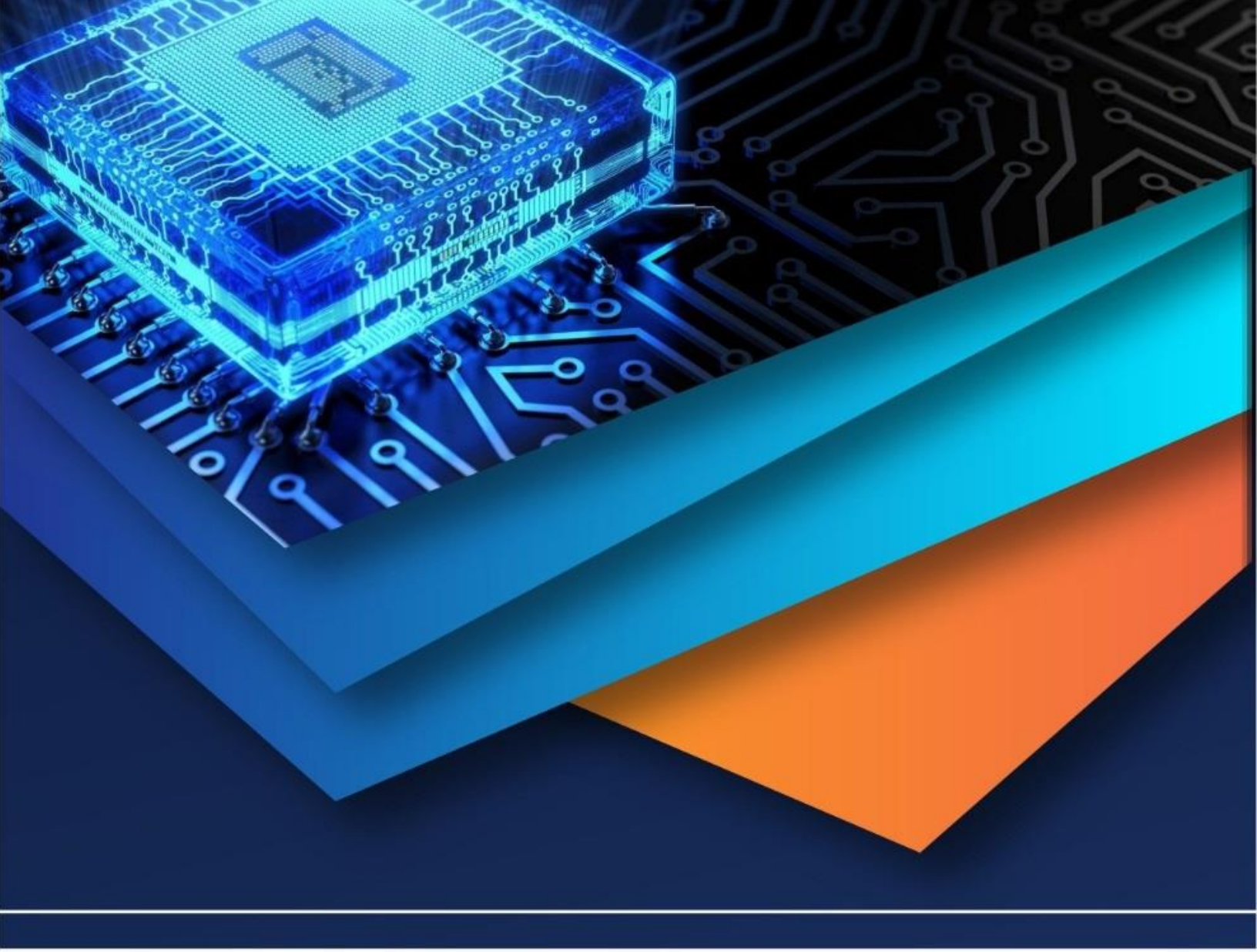
Table: The performance of different models on Author Identification task in Marathi Literature

Models	Train Accuracy (%)	Test Accuracy (%)	Validation Accuracy (%)
Base-BERT	100	99.4	90
RoBERTa	99.85	99.21	86
AlBERT	100	98.41	90
MuRIL	100	99.28	90

In conclusion, this research paper explored the feasibility of using state-of-the-art text classifier transformer models like Bert for author identification in Marathi literature. The findings indicate that these models can achieve almost 100% train accuracy and approximately 99% train accuracy, which suggests that they can effectively distinguish between authors in Marathi literature.
Furthermore, the models achieved an impressive 90% validation accuracy, indicating their ability to generalize well to new data. These results demonstrate the potential of text classifier transformer models for author identification tasks in Marathi literature and provide a foundation for further research in this area.

REFERENCES

[1] Nirkhi, Smita & Dharaskar, Rajiv & Thakare, V. M.. (2015). An Experimental Study on Authorship Identification for Cyber Forensics. 4. 0-417
[2] Holmes, David & Kardos, Judit. (2003). Who Was the Author? An Introduction to Stylometry. Chance. 16. 10.1080/09332480.2003.10554842.
[3] A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," 2016 15th IEEE International 9 Conference on Machine Learning and Applications (ICMLA), 2016, pp. 898-903, doi: 10.1109/ICMLA.2016.0161..
[4] Romanov, Aleksandr & Kurtukova, Anna & Shelupanov, Alexander & Fedotova, Anastasia & Goncharov, Valery. (2020). "Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. Future Internet." 13. 3. 10.3390/fi13010003.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[5] Rexha, A., Kröll, M., Ziak, H. et al. "Authorship identification of documents with high content similarity." Scientometrics 115, 223– 237 (2018). https://doi.org/10.1007/s11192-018-26M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
[6] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pages 77– 86. Springer, 2006. "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
[7] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writingstyle features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3):378–393, 2006.
[8] Dr. A. Pandian, Paritosh Maurya, Nitin Jaiswal . Author Identification of Hindi Poetry. International Journal of Scientific and Technology Research Volume 9, Issue 03, March 2020
[9] A. Vaswani et al., 'Attention is all you need', arXiv [cs.CL], 12-Jun-2017
[10] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, 'BERTweet: A pre-trained language model for English Tweets', in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. arXiv, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY