




Machine Learning based Plagiarism Detection for Marathi Language

Ramesh R Naik ^{1*}, Sunil Gataum ^{2*}, Maheshkumar Landge³, Rajkumar Jagdale⁴, C Namrata Mahender ⁵

^{1,2}Dept of CSE, Nirma University, Sarkhej, Ahmedabad. Gujarat. India

³Dept of MCA, School of business, Dr. Vishwanath karad MIT, WPU, Pune.

⁴Dept of Computer Science, Vishwakarma University, Kondhwa pune.

⁵ Dept of CS and IT, Dr. Babasaheb Ambedkar, Marathwada University Aurangabad, Maharashtra, India

Author Email: maheshkumar.landge@mitwpu.edu.in, rajkumar.jagdale@vupune.ac.in, cnamrata.csit@bamu.ac.in

* **Corresponding Author:** rameshram.naik@nirmauni.ac.in, sunil.gautam@nirmauni.ac.in.

ARTICLE INFO

ABSTRACT

Received: 20 Dec 2024

Revised: 25 Jan 2025

Accepted: 14 Feb 2025

In today's modern environment, when data are easily accessible, plagiarism is the most pervasive problem. Hence, a system for identifying and controlling it is crucial. In a variety of languages, there are numerous approaches that may be used for the purpose, but they are insufficient for literature that is based in the Marathi language. Plagiarism detection is a critical aspect of maintaining academic integrity and ensuring the originality of content in various languages. The detection of plagiarism in languages with relatively less computational research, such as Marathi, presents unique challenges due to its complex linguistic structure, syntax, and morphology. This paper explores a machine learning-based approach for efficient plagiarism detection specifically tailored for the Marathi language. We introduced a machine learning-based plagiarism detection method in this research study. We utilised the learning techniques of naive bayes, svm and artificial neural networks. SVM research have shown an average accuracy of 90%, while Naive Bayes studies have shown an average accuracy of 71%. Studies employing a Neural Network for Marathi Language Plagiarism Detection reported an average accuracy of 95%. The results demonstrate that the proposed approach can effectively detect plagiarism in Marathi texts, offering a promising tool for researchers, educators, and content creators to uphold content authenticity and originality.

Keywords: Plagiarism detection, Support vector machine, neural network, Machine learning, Marathi language

INTRODUCTION

The Finding the copied text using a reliable source or technology is known as plagiarism detection. Plagiarism is considered to occur when the content of two or more files is too identical to one another beyond a particular threshold. The process entails a number of processes, collecting data in a specified format, calculating related terms, counting the occurrences of a single word in both files, and finally revealing a similarity score. Several approaches are being employed nowadays to evaluate and comprehend the behaviour of documents that are similar to those that are used to expand the organisations. The rapid growth of digital content and the ease with which information can be accessed and reproduced have led to an increasing concern regarding plagiarism, particularly in academic and professional settings. Plagiarism detection has become essential in ensuring the integrity and originality of written material. While much of the existing research on plagiarism detection has focused on languages such as English, there is limited exploration of the issue in regional languages like Marathi, despite its rich literary heritage and widespread use in India. The challenges of detecting plagiarism in Marathi are compounded by its unique linguistic characteristics, including complex morphology, syntax, and a vast variety of regional variations. Traditional plagiarism detection methods, such as fingerprinting and string-matching algorithms, often struggle with these challenges, especially when dealing with languages that lack large-scale annotated datasets and comprehensive linguistic resources. To address these limitations, recent advancements in machine learning and natural language processing offer promising solutions.

These techniques enable the development of more sophisticated models capable of understanding and processing the nuances of Marathi language structure. This paper proposes a machine learning-based approach for detecting plagiarism in Marathi text. The focus is on utilizing various NLP techniques, such as text pre-processing, feature extraction, and the application of machine learning algorithms, to build an effective plagiarism detection system. The goal is to create a system that can identify instances of plagiarism, whether they involve direct copying or more subtle paraphrasing, across a wide range of Marathi text sources. By addressing the unique challenges posed by the Marathi language, this research contributes to the development of more robust and language-agnostic plagiarism detection tools, ensuring content authenticity in diverse linguistic contexts.

LITERATURE REVIEW

Explaining the act of converting one sentence into another by using different words or rearranging the words of a sentence is known as paraphrasing or rephrasing. In Natural Language Processing, the detection of paraphrasing is regarded as a challenging job. This study uses the Recurrent Neural Network algorithm model to detect plagiarism that takes the form of paraphrase. Since it is frequently impossible to determine the precise context of brief content, paraphrasing detection is a challenging operation [1].

This study's goal is to suggest a single method to identify plagiarism. The study makes use of 25 novels by different writers, and It uses the Most Common Terms usage trends to calculate the results. [2]

A novel approach to identify machine learning and natural language processing to detect plagiarism across languages is proposed by the study [3]. The three main steps of this system's operation are text input, translation detection, internet search, and report production. Most electronic-based input documents can be used using this method. With the primary goal of the study being the detection of plagiarism in source codes, the study proposes a programmatic statement or identifier order independent plagiarism detector. The author compares their point of view to a plagiarism simulator. The system in this work makes advantage of sequence

Alignment as well as other Syntax tree components. [4]

The paper suggests a model that makes use of Deep Learning features to detect plagiarism in Arabic writings. A technique to be employed in this article is the word2vec model, which recognises semantic similarities between Arabic words. Word2vec is a straightforward deep learning technique that accurately represents words as features of vectors. To determine how similar the vectors are, it makes use of the idea of cosine similarity. [5].

Text classification

Text classification is used in numerous real-world applications and plays a significant part in data mining. [6] Three stages make up the modified Lingo algorithm. Cluster label generation, the initial phase, aids in determining the cluster's label Cluster formation, the final stage, is when labelled clusters are noticed. Lingo is outperformed by the modified Lingo algorithm2.1.1 Plagiarism detection using support vector machine From both original and suspect texts, the statistical characteristics of sentences were retrieved and categorized using SVM. When two sentences are marked as plagiarism, that signifies the original sentence has been copied; otherwise, the suspect sentence has been stolen. In 1998, Vapnik proposed support vector machines, which employ statistical learning methods. These algorithms find an appropriate hyperplane to accurately categorise the data by maximising the distance between it and the training samples. Since it is hard to classify training data using a liner classifier when there is noise, main samples are non-linearly translated onto a higher space. Data will be linearly categorized in the new, larger space by a kernel function utilising the right hyperplane without increasing computing complexity. In fact, the kernel function finds similarities between vectors in a bigger space by using the similarity between data in the original space. Polynomial functions, RBF functions, hyperbolic tangents, and other suitable functions can be chosen as the kernel function. [13] Following Table1 shows literature review on classification techniques used for different languages

Table1.literature review on Techniques used for different languages.

Refere nce	Techniques used	Different languages
[7]	LSTM, BERT, ULMFiT and CNN.	Marathi
[8]	5 Superior Arabic TC Deep-Model.	Arabic
[9]	CNN, RNN,	Arabic
[10]	Logistic regression, SVM and Naive Bayes,	Telugu
[11]	SVM, RF, KNN,	Chinese
[12]	LSTM, LASER, BERT	Hindi

Naive Bayes Model for Detecting Plagiarism

Source code author identification can be done using naive Bayes classifiers, which are good for pattern recognition. Based on the Bayes theory, this classifier. When employing the Bayes theorem, S with few classes or outcomes is dependent on a number of traits denoted by t_1 , t_2 , and so on. [14]

Neural Network

As a result of a significant number of neurons being connected to one another, learning is the result of communication between several neurons. The extremely interconnected nature of human neurons makes learning appear possible. Although the neural network does not perfectly replicate the biological neural architecture, it does approach the biological neural network in some ways. Actually, it is a model for information processing that draws inspiration from how the biological nervous system handles information [15].

Two kinds of neural networks exist

i) Feed forward NN

This kind of NN is made up of layers of processing units, each of which uses connection strength also known as weight to forward information to the subsequent layer. There is no allowable reverse propagation of the input

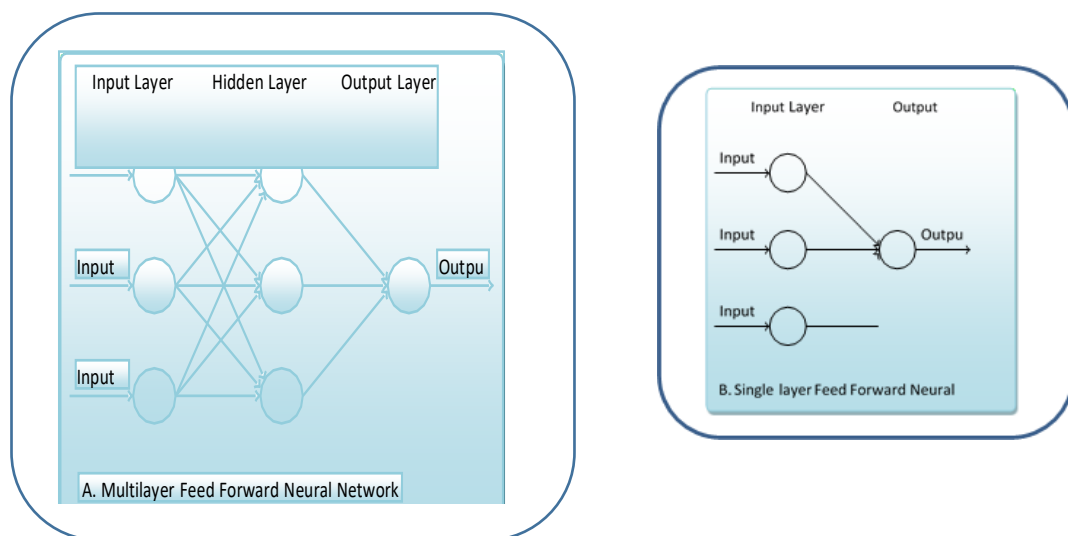


Figure1. A and B Feed forward NN

ii) Feed backward RNN

It allow the output of one node to be supplied to both other nodes and the same node simultaneously. In figure 2, the architecture of this type of network is depicted neural networks with Backpropagation

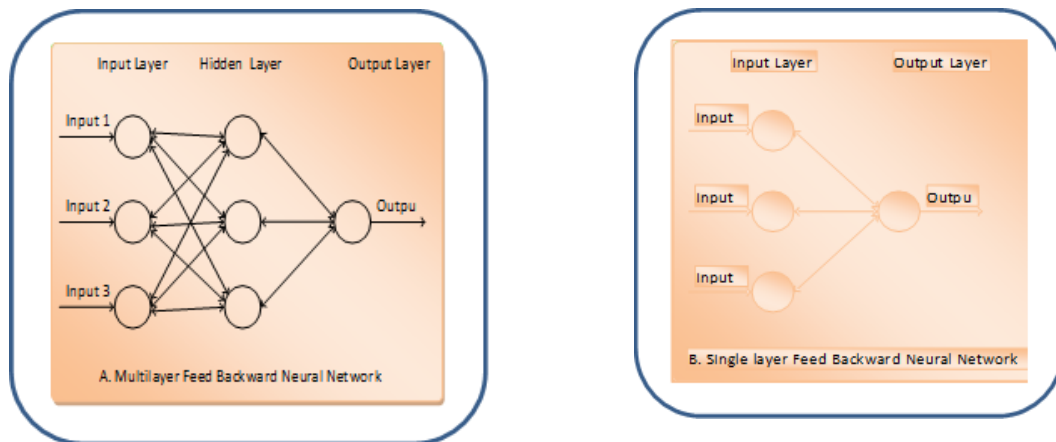


Figure 2. Feed Backward Neural Network

The Backpropagation algorithm is well known for neural network training. It minimizes error using the gradient descent method. The discrepancy between the intended and actual results is called an error. The Backpropagation algorithm comprises four phases, which are as follows: Initialize the network, Initialize the weights and biases, Feed Forward, Backpropagation of Errors, and Update the weights and biases. These steps are described below

Weights and biases are initialized

There are a set of weights that must be maintained for each neuron. There are two weights: one for the bias and one for each input link. These biases and weights are selected at random from very modest values. This is how the network is being set up.

Feed Forward

At this phase, input signals are sent to each layer up until the output layer, and output of the Network is determined. We refer to this as forward propagation.

Backpropagation of errors

At this phase, the difference between the actual and desired outputs is calculated, and the difference which is actually the error is transmitted back to the layer below. By weighting the mistake according to its weight in the layer before it and the gradient of the associated activation function, the error is back-propagated.

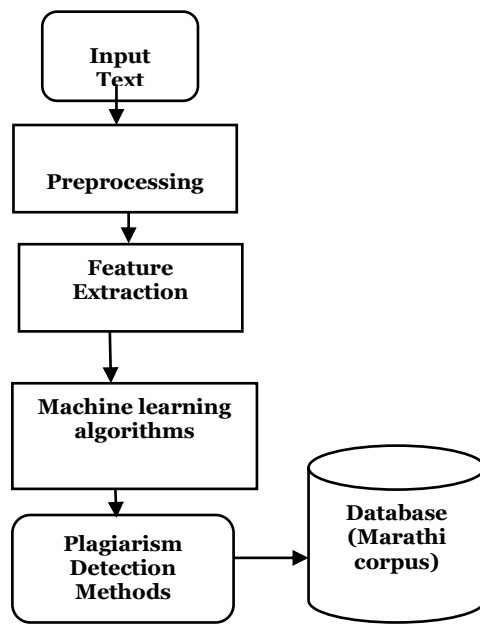
Update weight and biases

Lastly, the weights and biases related to each unit are updated using the activation and the factor discovered in the previous stage. Updates are made until the threshold is reached or an error occurs

METHODOLOGY

Proposed System

The following figure 3. Shows the detail outline of the complete proposed process



Database

Database files were utilised as the input for this stage. The system's corpora-development phase is likewise a crucial one. The following factors were taken into account when constructing the corpus. 300 Marathi text files from Ajanta Prakashan publications' Marathi research papers were used. 250 of these files are used for training, while 50 text files are used for testing. Thus, several fonts were employed to write in Marathi in the publication. Bringing it together in one format for further processing was the first task. Thus, they were all converted to the UTF-8 Unicode standard.

Table 2. Statistics of corpus for Ajanta Prakashan research papers

Sr.no.	Training files	Testing files
1	250	300

Preprocessing method covers punctuation removal, tokenization and Stopword removal.

Punctuation Removal Punctuation is a set of marks that regulates and clarifies the meaning of different text is called punctuation. = ""!()-[]{};:'"\<>./?@#\$\$%^&*~""

अन्न हे पूर्णब्रह्म आहे अन्न हे रुचकर करण्यासाठी त्यात आपले मनशांती असणे गरजेचे आहे जेव्हा आपण जेवणात अन्न ग्रहण करतो तेव्हा आपण व्यवस्थित चावणे हे गरजेचे असते.जेवणात वरण,भात,चपाती,वा भाकरी, उसळ,कोशिंबीर असणे हे स्थूलपनाला रामराम ठोकणे आहे पण हे अन्न आपण नित्यनेमाने करणे आवश्यक आहे.

अन्न हे पूर्णब्रह्म आहे अन्न हे रुचकर करण्यासाठी त्यात आपले मनशांती असणे गरजेचे आहे जेव्हा आपण जेवणात अन्न ग्रहण करतो तेव्हा आपण व्यवस्थित चावणे हे गरजेचे असते.जेवणात वरणभातचपातीवा भाकरी उसळकोशिंबीर असणे हे स्थूलपनाला रामराम ठोकणे आहे पण हे अन्न आपण नित्यनेमाने करणे आवश्यक आहे

Figure 5. Original sample file

Figure 6. The output of the original sample file after punctuation removal

Feature extraction

Feature extraction is the process of obtaining a fresh set of features from the pool of features obtained during the feature selection step. Lexical characteristics and vocabulary richness features were our two main areas of focus. These characteristics include Hapax legomenon and Hapax dislegemena, additionally to characteristics like Average sentence length by character, Average sentence length by word, Average word frequency class, and Average sentence length. A word called hapex legomena only appears once in a context either in the corpus of a language as a whole or in a particular document. A Greek expression that meaning "something that is told once and only" is Hapax legomenon.

Table4: Sample result for extracted feature

mai n files	avg word freq class	Avg_se n length by word	HL	HD	Avg_S entenc e_len
1	4.49	11.81	848.9 6	0.17	9.94
2	4.50	11.16	848.2 8	0.17	8.80
3	3.25	15.71	679.57	0.15	10.43
4	4.28	14.27	834.7 5	0.16	8.58
5	4.05	17.62	819.7 8	0.16	16.32
6	4.11	12.09	811.16	0.17	9.98
7	4.46	16.10	870.3 0	0.18	10.23
8	4.58	13.01	818.7 8	0.16	8.70
9	4.71	11.55	860.5 2	0.18	7.62
10	3.68	10.64	801.2 6	0.16	8.97

Results and discussion

After completing all the necessary preliminary work for classifying, we used three methods to build the model. 1. SVM 2.Naive Bayes 3.Neural Network

Table5. Accuracy using Support vector machine

Training data	Testing data	Accuracy
250	300	90.00

Table6. Accuracy using Naive Bayes

Training data	Testing data	Accuracy
250	300	71.00

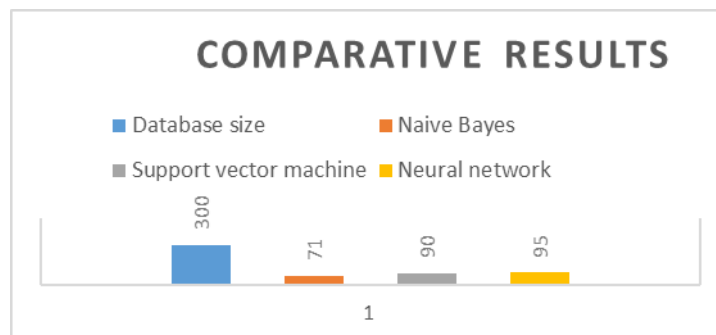
Table7. The backpropagation's Results

Used Database	Used algorithms	Number of Epochs	error occurred
300 Ajanta Prakashan papers	Backpropagation	30	21.333
300 Ajanta Prakashan papers	Backpropagation	200	8.99

When comparing the papers line by line and by paragraph, backpropagation's mean accuracy was found 95 percent.

Table8.Comparative Results

Sr.no	Database size	NB	SVM	NN
1	300	71.00	90.00	95.00



Graph1. Comparative results

Above graph1 shows Comparative results in that SVM gives average accuracy of 90%, while Naive Bayes gives average accuracy of 71%. And Neural Network reported an average accuracy of 95%. For Marathi Language Plagiarism Detection.

CONCLUSION

Plagiarism is the act of stealing information from another person's work without giving due credit. Hindi and other regional languages are used for plagiarism detection. Yet, there hasn't been much work done in Marathi. We are mainly focused Machine learning-based plagiarism detection for the Marathi language. We employed tokenization, stop word removal, and feature extraction. SVM yields an average accuracy of 90%, whereas Naive Bayes delivers an average accuracy of 71%. The average accuracy of Neural Network, was 95%. For the detection of plagiarism in Marathi. All researchers and students will benefit from using this method. This research is important because it uses state-of-the-art machine learning techniques to address a pressing issue plagiarism in an underserved language Marathi. By assisting organizations, teachers, and content producers in properly handling the problem of plagiarism in Marathi and possibly other languages as well it has the potential to significantly advance both the

technological and social spheres of language processing.

REFERENCES

- [1] Hunt, E. et al. Machine learning models for paraphrase identification and its applications on plagiarism detection. (2019) IEEE International Conference on Big Knowledge (ICBK) (pp. 97-104). IEEE.
- [2] AlSallal, et al. An integrated machine learning approach for extrinsic plagiarism detection. In 2016 9th International Conference on Developments in eSystems Engineering (DeSE) (pp. 203-208). IEEE.
- [3] Anguita, et al. Automatic cross-language plagiarism detection. In 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (pp. 173-176). IEEE.
- [4] Kikuchi, et al. (2014, June). A source code plagiarism detecting method using alignment with abstract syntax tree elements. In 15th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 1-6). IEEE.
- [5] Suleiman et al., Deep Learning Based Technique for Plagiarism Detection in Arabic Texts. 2017.
- [6] Shraddha et al., Text categorization of Marathi documents using modified LINGO. In 2017 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE.
- [7] Kulkarni, et al. (2022). Experimental evaluation of deep learning models for Marathi text classification. In Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021 (pp. 605-613). Springer Singapore.
- [8] Alhawarat, et al. (2020). A superior Arabic text categorization deep model (SATCDM). IEEE Access, 8, 24653-24661.
- [9] Elnagar, et al. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.
- [10] Sudha, D. N. (2021). Semi Supervised Multi Text Classifications for Telugu Documents. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(12), 644-648.
- [11] Yan, L., & Lu, C. (2019). Multi-label Chinese comments categorization: comparison of multi-label learning algorithms. Journal of New Media, 1(2), 51.
- [12] Shahin, et al. (2022). Machine learning approach for autonomous detection and classification of COVID-19 virus. Computers and Electrical Engineering, 101, 108055.
- [13] Joachims, T. (2005, June). Text categorization with support vector machines: Learning with many relevant features. In Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings (pp. 137-142). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [14] Kolter, et al. (2004, August). Learning to detect malicious executables in the wild. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 470-478).
- [15] Sivanandam, et al. (2006). Introduction to neural networks using Matlab 6.0. Tata McGraw-Hill Education.