IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

Plagiarism Detection with Paraphrase Recognizer Using Deep Learning

Yogesh Wadekar[1], Tushar Shendge[2], Manali Dhokale[3], Vaishnavi Ohol[4], Prof. Sagar Dhanake[5]
Students, Computer Engineering[1, 2, 3,4]
Assistant Professor, Computer Engineering[5]
D Y Patil Institute of Engineering and Technology, Ambi, Pune, Maharashtra, India

Abstract: *Plagiarism is a progressively widespread and growing issue within the educational field. Many plagiarism techniques square measure utilized by fraudsters, starting from a straightforward word replacement, phrase structure modification, to additional advanced techniques involving many varieties of transformation. Primarily human-based plagiarism detection is troublesome, not much accurate, and time-consuming method. In this paper, we tend to propose a plagiarism detection framework supported by 3 deep learning models: Doc2vec, Siamese Long Short-term Memory (SLSTM), and Convolutional Neural Network. Our system uses 3 layers: Preprocessing Layer together with word embedding, Learning Layers, and Detection Layer. To judge our system, we tend to dispense a study on plagiarism detection tools from the educational field and build a comparison supported a group of options. Compared to alternative works, our approach performs an honest accuracy of 97.26% and might notice differing kinds of plagiarism, permits to specify another dataset, and supports to check the document from an internet search.* *Plagiarism is a progressively widespread and growing issue within the educational field. Many plagiarism techniques square measure utilized by fraudsters, starting from a straightforward word replacement, phrase structure modification, to additional advanced techniques involving many varieties of transformation. Primarily human-based plagiarism detection is troublesome, not much accurate, and time-consuming method. In this paper, we tend to propose a plagiarism detection framework supported by 3 deep learning models: Doc2vec, Siamese Long Short-term Memory (SLSTM), and Convolutional Neural Network. Our system uses 3 layers: Preprocessing Layer together with word embedding, Learning Layers, and Detection Layer. To judge our system, we tend to dispense a study on plagiarism detection tools from the educational field and build a comparison supported a group of options. Compared to alternative works, our approach performs an honest accuracy of 97.26% and might notice differing kinds of plagiarism, permits to specify another dataset, and supports to check the document from an internet search.*

Keywords: Plagiarism detection, Plagiarism detection tools, Deep learning, Doc2vec, Stacked Long Short-Term Memory (SLSTM), Convolutional Neural Network (CNN), Siamese neural network

I. INTRODUCTION

"The Plagiarism can be conceptualized as the theft of others efforts or words without citing the right reference and therefore while not giving the correct credit to the correct person and original author [9]". [1] Depending on the depth of transformation performed on the original text, plagiarism can be classified into different categories as Copy paste plagiarism [11], Paraphrasing [12], Use of false references [13], Plagiarism with translation [14], and Plagiarism of ideas [15]. Plagiarism is applied in various areas which include literature, music, software, scientific articles, newspapers, advertisements, websites, etc. As the use of the internet increases plagiarism becomes a big challenge in schools, institutions, and universities to maintain academic integrity. Web search engines become the common point of view to retrieve and find needed information. Hence, evaluating search engine quality [18] is a hot topic that attracts many researchers' attention [18]. People commonly use web search engines to find what they want. However, as search engines become a very efficient and effective tool [17], plagiarists can grab and redistribute text contents without much difficulty [17]. TensorFlow is an open-source python library which is used for diverse applications of deep learning programming tasks [16]. The deep learning model can be trained using high-level Keras API [16].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

1.1 Motivation

Increase in plagiarism in an every sector is the main motivation for this study. It is observed that, people copy paste others work and pretend it as there own. By this way, people get attention and rewards for that work which is hard work of another person. Like this way the right person is kept apart and the thief gets the credit. To avoid this condition and help people to secure there own work from such interactions plagiarism concept comes into picture.

II. DESCRIPTION OF THE PROBLEM

2.1 Problem Definition

Design and develop the plagiarism detection system which can detect different types of plagiarisms and the fraud submissions which might be copied from others work using deep learning and machine learning algorithms.

In 2020 El Mostafa Hambi and his team came with a Research Paper entitled ”A New Online Plagiarism Detection System based on Deep Learning” [1] at IJACSA. In this paper, Deep Learning based model is proposed which identifies the plagiarism using Doc2vec, Long Short-Term Memory (LSTM) and CNN algorithms and gives respective plagiarism percentage.

Considering the above-mentioned problems in mind we decided to design a system that will help to identify the uniqueness of the document uploaded.

2.2 Project Idea

We are developing a plagiarism detection website which can detect the percent similarity of the contents from your uploaded document or paragraph comparing with the contents from various websites.

2.3 Goals and Objectives

To protect the originality of the work and detect the similarity of your work with the content available on the internet.

III. LITERATURE REVIEW

“A New Online Plagiarism Detection System based on Deep Learning [1]” paper proposed an online plagiarism detection system which is based on Doc2vec technique for word embedding in coordination with SLSTM and CNN deep learning algorithms. “Code Plagiarism Detection Method Based on Code Similarity and Student Behavior Characteristics [2]” paper proposed the concept of code similarity concentration. We learn to focus on the similarity between two documents from this paper. “Machine Learning Models for Paraphrase Identification and its Applications on Plagiarism Detection [3]” paper identify if a sentence is a paraphrase of another one. Paraphrasing [5] is what where you copy someone’s exact words and put them in quotation marks. From this paper we studied LSTM and RNN algorithms. Plagiarism detection is done using string matching algorithm, k-gram and karp Rabin algorithm. This algorithms are used to detect the originality of students work based on similarity concentration.

“Plagiarism Detection Using Machine Learning-Based Paraphrase Recognizer [4]” paper studied a support vector machine based paraphrase recognition system. SVM is one of the algorithm in machine learning, which works by extracting lexical, syntactic, and semantic features from input text has been used. In “Extending Web Search for Online Plagiarism Detection [17]”, Yi Ting Liu and team developed an online plagiarism detection system to reduce misapplication of search engines. They extracted and verify the suspicious documents through the collaboration of plagiarism detection system and search engines.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

IV. SYSTEM DESIGN AND FLOW

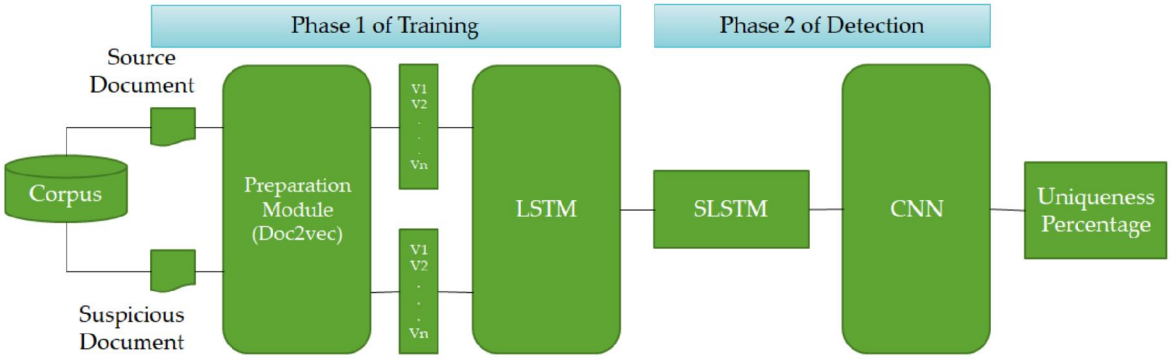

Fig.:- System Architecture

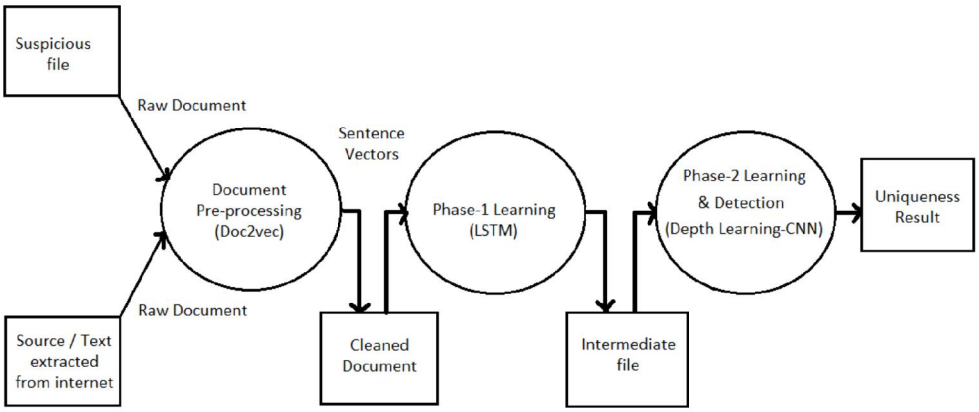

Fig. Flow Diagram

Mathematical Model:

More precisely we assume that the document contains N paragraphs.

For example the first paragraph contains S sentences, so we launch S internet searches.

Now we have S x N results.

Assume that each result will offer us P paragraphs which are considered as suspected initials. So, the first paragraph of the analysis document is compared with N x S x P paragraph.

V. PROJECT IMPLEMENTATION

Algorithm Details:

Doc2vec

Doc2vec is a deep learning technique that is used to represent words as features of vectors with high precision [10]. In this method, a text is considered as bag of words where there is no more order, and with each word we associate a weight which makes it possible to measure its importance in the text. A text is transformed into a vector in a large space where each coordinate corresponds to the degree of importance of a given word in the text. This new illustration contains a serious a part of syntactical in addition as linguistics rules of the text information. A lot of larger units like "phrases, sentences and documents" ought to be represented as a vector. The paragraph vector learning approach is based on word vector learning methods. The inspiration is that the vector words are asked to contribute to a prediction task regarding consecutive word at intervals the sentence. [6]

LSTM

LSTM learns the long-term dependencies of the text. After taking the word embedding as input it captures most data from the text. The primary step in our LSTM is to return to a decision what data we're attending to throw away from the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

cell state. Sequent step is to come to a choice what new data we're attending to store at intervals the cell state. This has 2 parts. First, a sigmoid layer stated because the "input gate layer" decides that values we'll update. [7]
Next, a tanh layer creates a vector of recent candidate values that might be added to the state. Within the next step, we'll mix these 2 to form Associate in nursing update to the state. Finally, we need to make your mind up what we're attending to provide as an output. A common LSTM unit consists of a cell, input gate, output gate and a forget gate. The cell remembers values over discretionary time intervals and also the 3 gates regulate the flow of knowledge into and out of the cell.

CNN

CNN is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle) that uses a variation of multilayer perceptions designed to require minimal preprocessing. These are inspired by animal visual cortex. CNNs are usually utilized in computer vision; but, they need recently been applied to numerous NIP tasks sort of a text classification. It reduces the no. of features in dataset by creating the new and reduced dataset gives us information contained in original set of features. It is performed by convolutional layer and sub sampling layer. And the classification is performed by dense layer and soft max layer.[8]
It reduces the no. of features in dataset by creating the new and reduced dataset gives us information contained in original set of features.

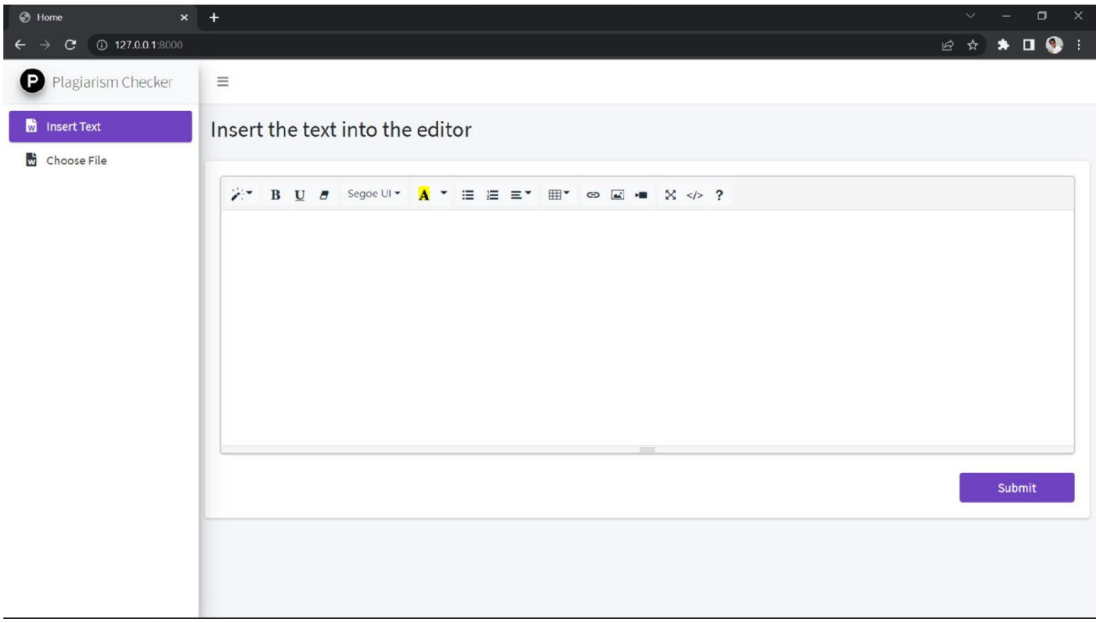

Fig.: Home Page

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

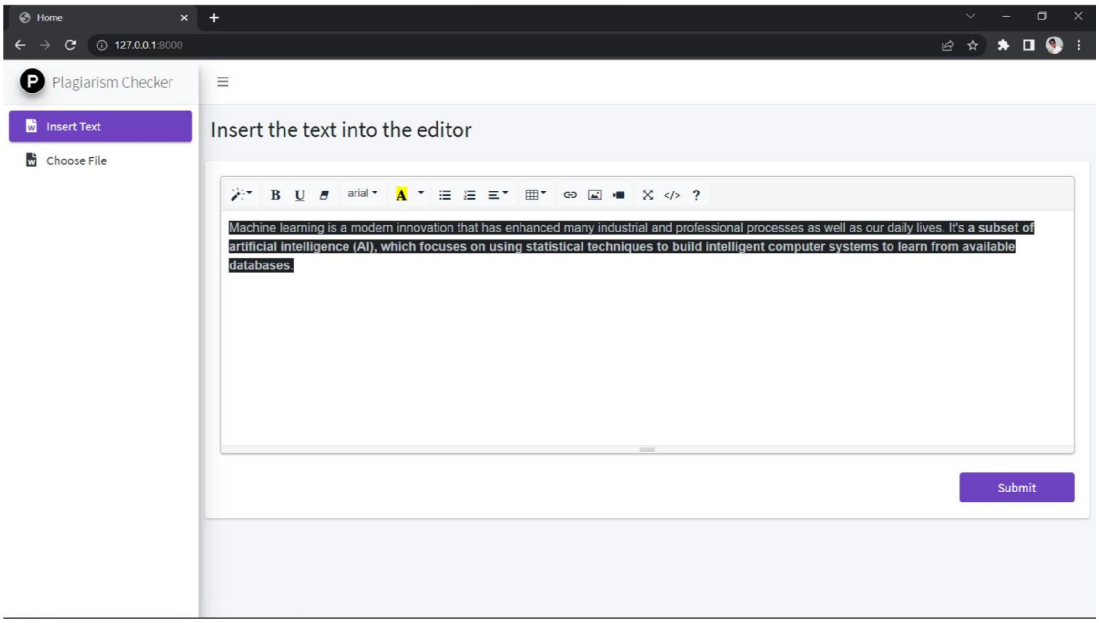

Fig. Paragraph writing

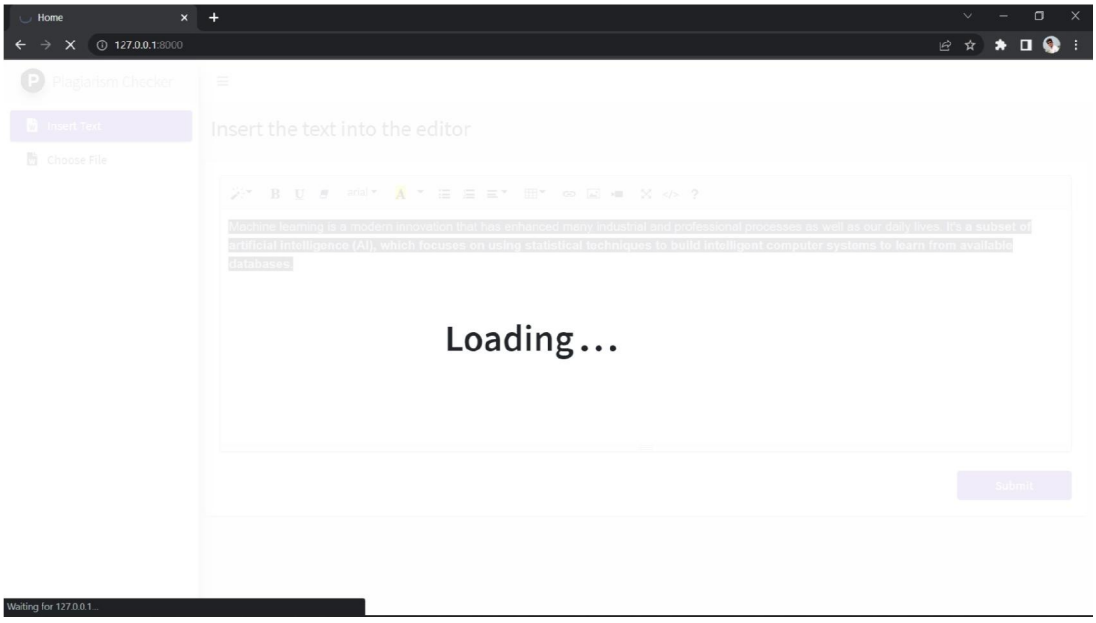

Fig. Checking

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

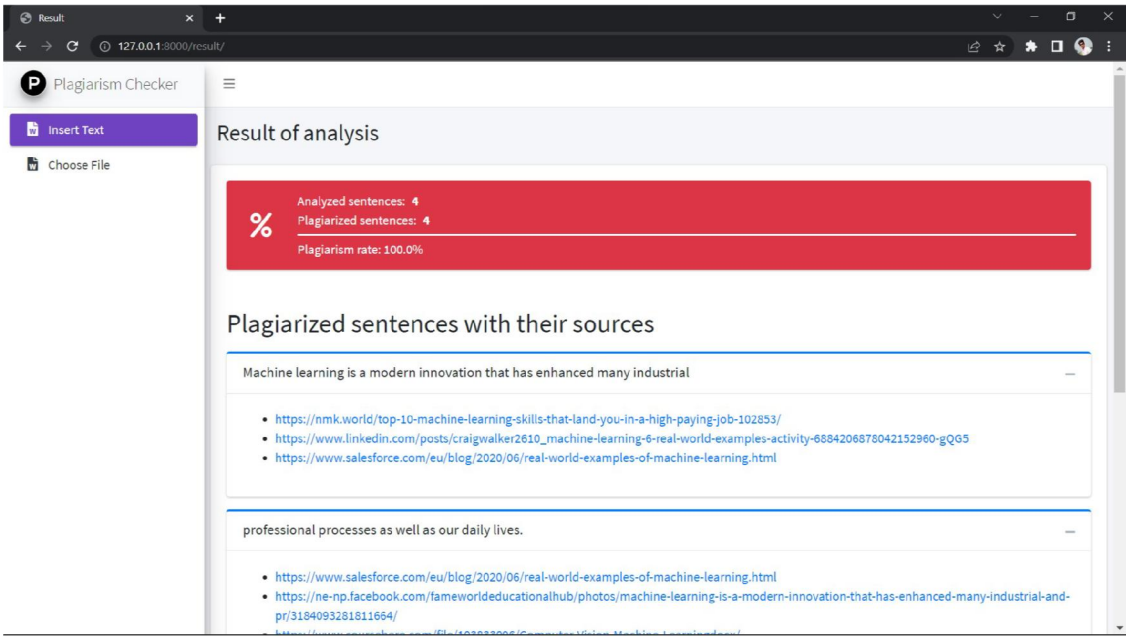

Fig. Paragraph Plagiarism Result 1

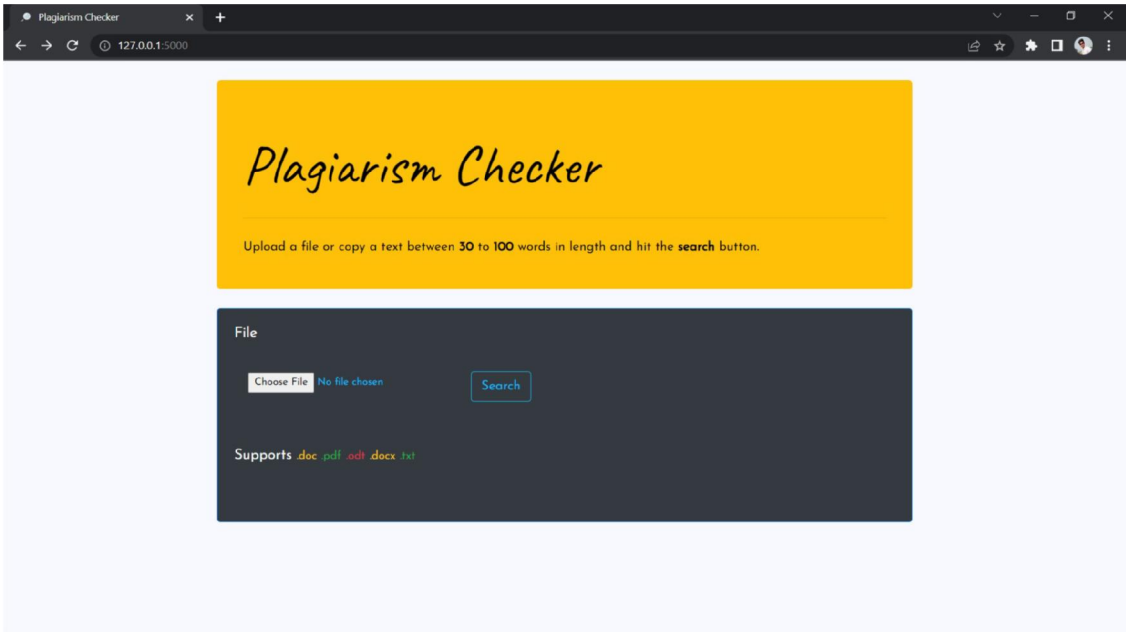

Fig. Home Page

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

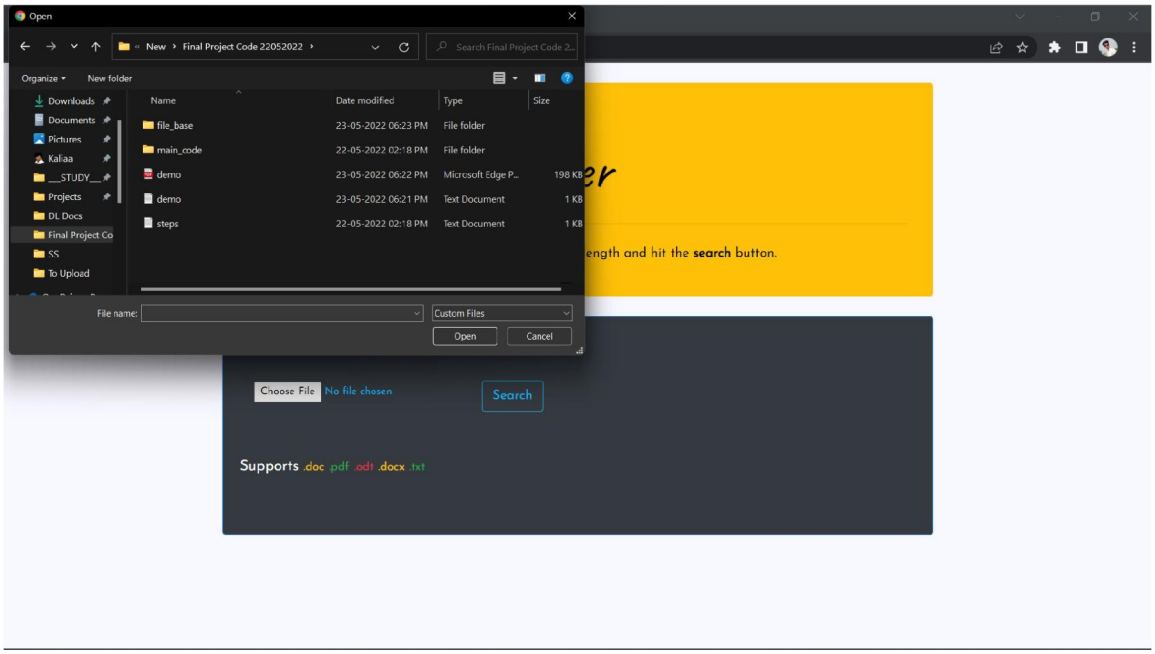

Fig. File Upload

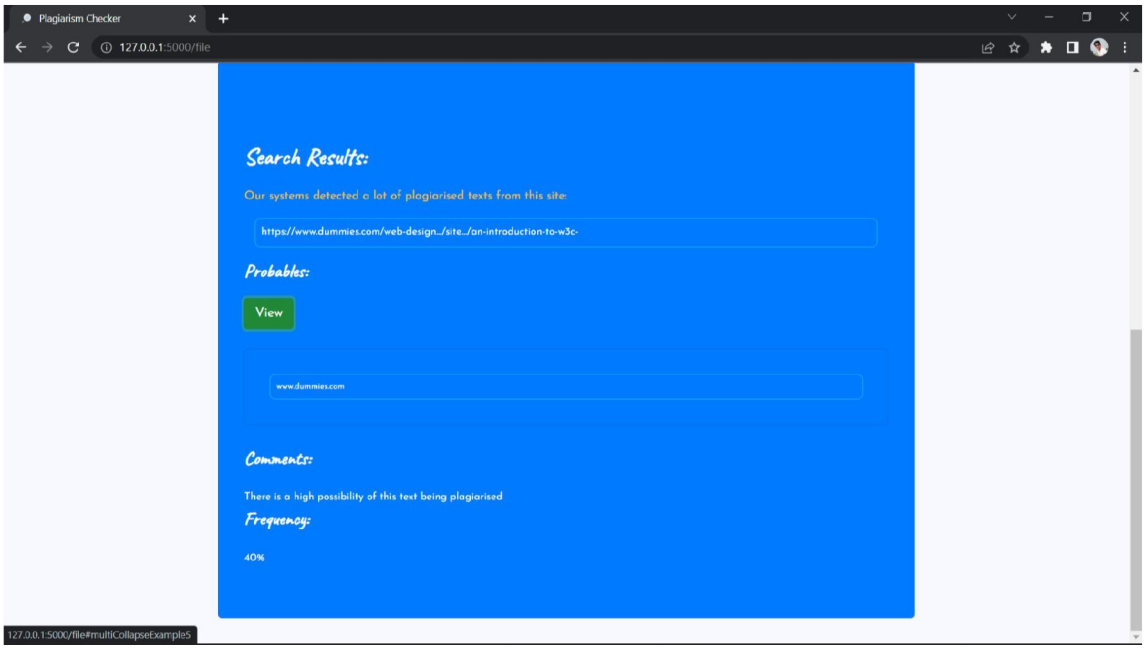

Fig. Document Plagiarism Report

VI. ADVANTAGES AND DISADVANTAGES

Advantages:-

- The advantage of a plagiarism detector is that it is free of cost and you can easily use it.
- Even if it is a free service, still it is the most trustworthy service that you can use to check plagiarism.
- Through this application, you can also check stuff relevant to other software.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

- It not only checks plagiarism but also provides you editing services for any mistake or blunder so that you will have perfect writing.
- It also checks grammatical errors, detects spelling mistakes.
- It tells you the right way for placing the references.

Limitations:-

- Some disadvantages such as the plagiarism tool can accurately detect plagiarism when the text is comprised of seven or more than seven words but it cannot work as well with the text of smaller words.
- And because of this, we cannot say that it is an infallible application.
- However, it plays a good role when working with barefaced cases of plagiarism.
- Although, it cannot completely find the mistakes it can help up to some extent.

Applications:-

- Writer:-
  - Help writers to write their own content uniquely which will not be subjected to copyright after publication.
  - Help them to grow.
- Teacher
  - To design courses (assignments and assessments) to encourage honest work.
  - Helps in the assessment evaluation process for assigning marks based on the uniqueness of the answer.
- Researchers
  - Helps researchers to publish unique research papers.
  - Avoids unnecessary research.
- Government
  - Helps in the evaluation process of competitive exams like essay writing based on uniqueness.
  - It helps to proceed with the further process.
- Students
  - Helps students to design their work reports and assignments.
  - Guide students with a clear understanding of plagiarism and cheating and how to collaborate.

VII. CONCLUSION

In this paper, we proposed a new system for the detection of plagiarism which is based on the deep learning strategies. Its interest is the extraction of characteristics without losing the sense of the document by using doc2vec word embedding technique. The planned system has the ability to detect not solely that there's plagiarism however additionally the chances of the existence of every form of plagiarism. We presented the various services offered by our system, either at the level of the personalized learning phase or the various ways of detecting plagiarism offered. When compared to the other tools studied, As for our views, we will improve the various interfaces of the application to make it more accessible to the general public and improve the response time due to the learning time. It would also be fascinating to compare the performance of various approaches in a quantitative way.

In future, We want to read some more algorithm of plagiarism detection which further optimize result and execution speed. Based on the already developed software's and tools, We want to understand there implementation and how that tools and algorithms can increase the efficiency of our project for plagiarism detection.

ACKNOWLEDGEMENTS

The completion of our project brings with it a sense of satisfaction, but it is never complete without those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is a great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

REFERENCES

[1] El Mostafa Hambi , Faouzia Benabbou, ”A New Online Plagiarism Detection System based on Deep Learning”, (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 11, No. 9, 2020.
[2] Qiubo Huang, Xuezhi Song, Xuezhi Song, ”Code Plagiarism Detection Method Based on Code Similarity and Student Behavior Characteristics”, IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2020
[3] Ethan Hunt, Ritvik Janamsetty and team, ”Machine Learning Models for Paraphrase Identification and its Applications on Plagiarism Detection”, IEEE International Conference on Big Knowledge (ICBK) 2019
[4] Yahia Jazyah, ”Open Learning, the Issue of Plagiarism - Efficient Algorithm”, International Journal of Computers, Volume 3, 2018
[5] Chitra and Anupriya Rajkumar, ”Plagiarism Detection Using Machine Learning- Based Paraphrase Recognizer”, J. Intell. Syst. 2016; 25(3): 351–359
[6] Kim, Do-Guk Ko, Bonggyun. (2019). Investment Universe Construction Based on the Theme Keyword Search. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2939414.
[7] Janardhanan, Deepak Barrett, Enda. (2018). CPU Work load forecasting of Machines in Data Centers using LSTM Recurrent Neural Networks and ARIMA Models. 10.23919/ICITST.2017.8356346.
[8] Miralles, Luis Rosso, Dafne Jim´enez, Fernando Garc´ıa, Jos´e. (2017). A methodology based on Deep Learning for advert value calculation in CPM, CPC and CPA networks. Soft Computing. 21. 1-15. 10.1007/s00500-016-2468-4.
[9] Risquez, A., Dwyer, M. O.; Ledwith, A. (2011). ”Thou shalt not plagiarize’: from self-reported views to recognition and avoidance of plagiarism”. Assessment Evaluation in Higher Education, vol. 2, no. 1, p. 34-43. http://doi.org/10.1080/02602938.2011.596926. 3 Ruip´erez, G.; Garc´ıa-Cabrero, J.C. (2016). Plagiarism and Academic Integrity in Germany. Comunicar, vol. 24, no. 48, p. 9-17. http://doi.org/10.3916/C48-2016-01.
[10] Suleiman, D., Awajan, A., Al-Madi, N. (2017). ”Deep Learning Based Technique for Plagiarism Detection in Arabic Texts”. 2017 International Conference on New Trends in Computing Sciences (ICTCS). doi:10.1109/ictcs.2017.42
[11] Liu, Y.-T., Zhang, H.-R., Chen, T.-W., Teng, W.-G. (2007). ”Extending Web Search for Online Plagiarism Detection”. 2007 IEEE International Conference on Information Reuse and Integration. doi:10.1109/iri.2007.4296615
[12] Sousa-silva, r. -”detecting trans lingual plagiarism and the backlash against translation plagiarists language and law” / linguagem e direito, vol. 1(1), 2014, p. 70-94.