International Journal of Research in Engineering, Science and Management
Volume 4, Issue 4, April 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Plagiarism Detector Using Machine Learning

Hiten Chavan[1], Mohd. Taufik[2], Rutuja Kadave[3], Nikita Chandra[4*]
[1,2,3,4]*Department of Information Technology, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India*

***Abstract*: Plagiarism is the act of stealing someone's idea or work and representing it as one's own. Plagiarism has been identified as a violation of moral rights in various countries. Today in the world of evolving technology and ever-growing usage of the Internet, the unacceptable act of plagiarism has been increasing on a large scale. It is often observed in many educational areas such as research papers, blogs, articles, assignments, etc. This paper majorly focuses on the plagiarism that is frequently found in schools and colleges. Many students can be found to have copied assignments from their classmates. A system can be developed for the convenience of teachers that could check the amount of plagiarism in students' assignments. This system could be mentioned as an improvement from the old manual way as it eliminates the tedious work with increased speed and efficiency.**

***Keywords*: Plagiarism, Detector, Machine Learning, Cosine similarity, TF-IDF, Plagiarism checker.**

1. Introduction

Plagiarism detection is the process of spotting the plagiarised content via a trustable source or system. The similarity of content beyond a certain limit between two or more files is not acceptable and hence, recognized as plagiarism. The task requires many steps such as accepting the input in a particular format, computing the resembling words and counting the occurrences of a single word in both the files and finally disclose a similarity score. Now-a-days, different kinds of strategies are being implemented to analyze and understand the similarity behavior in documents as like in used in growth of the business [7].

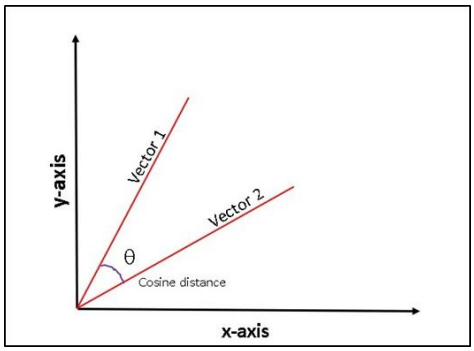

Fig. 1. Illustration of the cosine of vectors (dot product)

The system proposed is a machine learning-based model. It uses module incorporated features in the sci-kit-learn. The Tf-Idf Vectorizer converts the text into vector form thereupon the dot product or the cosθ (θ being the angle between the two vectors) is computed. This result gives us the resemblance of data between both the text files, or in this case, vectors.

With the outbreak of the COVID-19 pandemic, the whole education system has been proven to be dependent on technology through online lectures, assignments, and examinations. Through this theory, an easier spotting of plagiarism in student's assignments and online examinations could be done.

2. Literature Review

Paraphrasing or rephrasing is the conversion of a sentence into another with alternate use of words or changing the sequence of words in a sentence. The recognition of paraphrase in Natural Language Processing (NLP) is considered a rigorous task. This study aims to identify plagiarism in the form of paraphrasing through the application of the Recurrent Neural Network (RNN) algorithm model. Paraphrasing detection is a difficult process as it is not always possible to get the correct context of short-length content [1].

The objective of this study is to propose a unified technique to detect plagiarism. It makes use of four well-known models namely, Bag of Words (BOW), Latent Semantic Analysis (LSA), Support Vector Machine (SVM), and Stylometry. The study uses 25 books of various authors and computes the results using the usage patterns of the Most Common Words (MCW). [2]

The study [3] suggests a new way to recognize cross-language plagiarism using machine learning and natural language methods. The modus operandi for this system involves three major steps, namely, textual input, translation detection, internet search, and report generation. The approach applies to most of the electronic-based input documents.

Detection of plagiarism in source codes, being the core objective of the study, the study proposes a plagiarism detector that is not influenced by changing the identifier or program statement order. It compares the perspective with that of a sim plagiarism detector. The study uses Sequence Alignment and various Syntax tree elements in the system. [4]

The study proposes a model to spot plagiarism in Arabic texts using Deep Learning features. It puts forward an approach to use the word2vec model which detects the semantic similarity between Arabic words. Word2vec is a simple deep learning method used to portray words as features of vectors with great

*Corresponding author: nikitachandra2@gmail.com

accuracy. It uses the concept of cosine similarity to check the similarity between the vectors. [5]

Xinhao Wang et al. [6] have proposed a plagiarism detection model for non-native English speakers. The model differentiates between plagiarised and non-plagiarised content, which is present in two forms, namely, text-based and speech-based data. The system is designed for detecting plagiarism in speaking proficiency voluntary speech is extracted using an automated speech scoring system. In the recent years spam became as a big problem of Internet and electronic communication. So for overcoming these problems some techniques are developed to fight with them [8].

3. Proposed System

In the era of Internet revolution, the abrupt act of plagiarism has been highest than ever. Hence the world is in dire need of up plagiarism detectors. Most of such systems in the market ask for personal details of the user. The aim of our group through this project is to provide the user, especially teachers and educational institutions with a freely available, easy to use plagiarism detector.

We propose an idea to build a plagiarism detector using built-in machine learning libraries. We focus on sci-kit library which contains various useful and efficient machine learning tools. Basic techniques like Vectorization and cosine similarity together could be used in building an efficient plagiarism detection system.

4. Implementation

Sci-kit-learn is a built-in library that is used for machine learning tools. It contains tools for machine learning and statistical modeling. This library has been used in the proposed system for feature extraction from the text. The Tf-idf vectorizer is used for word embedding, i.e., conversion of textual data into an array of numbers.

This converted form of textual data into the form of a vector is now utilized to detect the similarity between two text files. Cosine similarity computes the cosine of the angle between the two vector forms of text files. This computation results in a score that ranges from 0-1, hence providing us the information about the extent of similarity between the two input files.

The implementation approach involves four crucial steps that includes,

1) Input File

The file is supposed to be the input for the plagiarism detection system. It should be in text format (.txt extension).

2) Vectorization of text

Sci-kit built-in features make sure that the words obtained from the textual input get converted into a vector format.

3) Compute similarity

The resemblance of two text files is computed using the basic concept of Cosine Similarity. The similarity between two text files depicted in the form of vectors is computed using the dot product of both the vectors, i.e., cosθ (θ being the angle between the two vectors).

4) Similarity Score

A similarity score is generated that signifies the amount of similarity detected between the two text files. The score is on a scale of 0-1(positive values of cosθ ranges from 0 to 1).

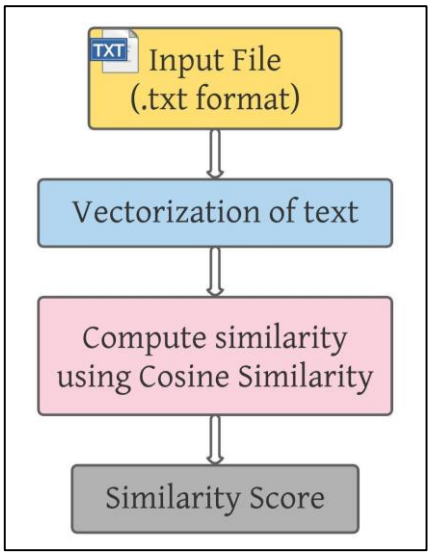

Fig. 2. System flowchart

Algorithm:

a) The TF-IDF technique has been used in the mentioned system. TF-IDF stands for Term Frequency- Inverse Term Frequency. This algorithm emphasises on the frequency of a recurring word and its importance in the given context of input.
   - Term Frequency = (count of the term) / (total word count in the document)
   - Inverse Document Frequency = log (number of docs) / (docs containing keyword)

TF-IDF formula,

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

b) Cosine similarity makes use of the vectors as an input and fetches the cosine of those vectors. This algorithm takes the data vector and calculates the cosine of the two vectors using the angle between them. It provides the output in 0-1 format, signifying the similarity score.

5. Result

The system has been presented using an easy-to-use User Interface (UI). It accepts the input text file (.txt extension) using the “Choose Files” button. The chosen 2 files get uploaded to the database. These uploaded files are then checked for plagiarism.

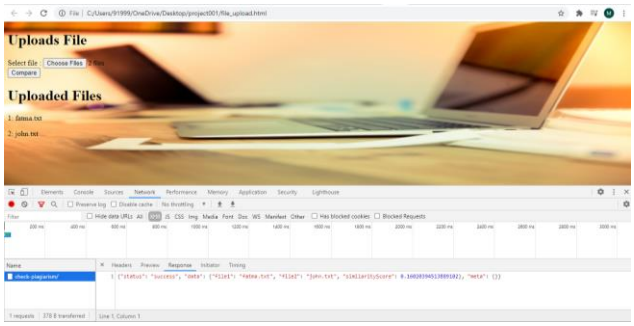

Fig. 3. Plagiarism detector: System output

The system displays the output as a “Similarity Score” which shows the extent of similarity between the two text files. The score ranges between 0 to 1.

6. Conclusion

In this paper, a plagiarism detector has been implemented using machine learning features like word2vec and cosine similarity. The system works efficiently and detects the extent of plagiarism between the given text files. The incorporation of a user interface makes it easier for a layman to utilize the service of the system. This system can easily be used in institutions like schools and colleges to detect plagiarism in students' assignments. The system can also be used for the evaluation of students' examination mark sheets to detect the expected technical words in the answer sheet. Operation of the system does not require any complex directions or training. It is a time-efficient, easy to use, and effective plagiarism detection system.

References

[1] E. Hunt et al., "Machine Learning Models for Paraphrase Identification and its Applications on Plagiarism Detection," 2019 IEEE International Conference on Big Knowledge (ICBK), 2019, pp. 97-104.

[2] M. AlSallal, R. Iqbal, S. Amin, A. James and V. Palade, "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), 2016, pp. 203-208.

[3] A. Anguita, A. Beghelli and W. Creixell, "Automatic cross-language plagiarism detection," 2011 7th International Conference on Natural Language Processing and Knowledge Engineering, 2011, pp. 173-176.

[4] H. Kikuchi, T. Goto, M. Wakatsuki and T. Nishino, "A source code plagiarism detecting method using alignment with abstract syntax tree elements," 15th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014, pp. 1-6.

[5] Suleiman, Dima & Awajan, Arafat & Al-Madi, Nailah. (2017). Deep Learning Based Technique for Plagiarism Detection in Arabic Texts.

[6] Wang, Xinhao & Evanini, Keelan & Mulholland, Matthew & Qian, Yao & Bruno, James. (2019). Application of an Automatic Plagiarism Detection System in a Large-scale Assessment of English Speaking Proficiency. 435-443.

[7] Aditya Ambre, Praful Gaikwad, Kaustubh Pawar, Vijaykumar Patil."Web and Android Application for Comparison of E-Commerce Products", International Journal of Advanced Engineering, Management and Science, vol. 5, no. 4, pp. 266-268, 2019.

[8] Bhawana S. Dakhare and Ujwala V. Gaikwad. Article: Spam Detection and Filtering using Different Methods. IJCA Proceedings on National Conference "MEDHA 2012" MEDHA (1):1-5, September 2012.