IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Online Assignment Plagiarism Checker Using Machine Learning

[1] Dr.M.SRINIVASAN, [2]S.HARIHARAN, [3]S. Premkumar, [4]S. Alex Pandiyan

[1]Department of Information Technology, P.S.V College of Engineering and Technology, Krishnagiri, Tamil Nadu, 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. Todayin the world of evolving technology and ever-growing usage of theInternet, 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 copiedassignments from their classmates. A system can be developed forthe 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.

INTRODUCTION

Plagiarism detection is the process of spotting the plagiarized content via a trustable source or system. The similarity of content beyond a certain limit between two or more files is notacceptable 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 finallydisclose 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.

Detection of source code plagiarism is valuable for both the academia and industry. Zobel has pointed out that, "students may plagiarize by copying code from friends, the Web or so called „private tutors‟". Most programming courses in universities evaluate the students based on the marks of programming assignments. Therefore, it is essentialto detect and prevent plagiarism at universities. Moreover Liu and et al have mentioned that, "A quality plagiarism detector has a strong impact to law suit prosecution". Therefore, there is a huge demand for accurate source code plagiarism detection systems from both the academia and industry.

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 paraphrasein Natural Language Processing (NLP) is considered a rigoroustask. 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.

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 resultsusing the usage patterns of the Most Common Words (MCW).

The study 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 approachapplies to most of the electronic-based input documents.

Detection of plagiarism in source codes, being the coreobjective of the study, the study proposes a plagiarism detectorthat 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.

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

PROPOSED SYSTEM

An online assignment plagiarism checker using machine learning can be developed with the following proposed system:

Data Collection: Collect a large amount of data to train the machine learning algorithm. This data should include a variety of different types of assignments, such as essays, reports, and research papers, along with known instances of plagiarism.

Data Preparation: Preprocess the collected data by cleaning it, removing unnecessary information, and converting it to a format suitable for machine learning algorithms.

Feature Extraction: Extract features from the preprocessed data to provide a numerical representation of each assignment. Features can include word frequencies, sentence structure, and grammatical patterns.

Algorithm Development: Develop a machine learning algorithm, such as a neural network or decision tree, that can analyze the features of an assignment and determine the likelihood of plagiarism.

System Integration: Integrate the algorithm into a user-friendly web application that can accept uploaded assignments and generate a plagiarism report.

User Interface: Develop an intuitive and easy-to-use interface for users to upload assignments, view plagiarism reports, and navigate the system.

Testing and Validation: Test the system using a variety of test cases and validate its accuracy and reliability.

Maintenance and Updates: Maintain and update the system as needed to improve its performance and ensure that it remains up-to-date with the latest techniques and technologies.

By following this proposed system, an online assignment plagiarism checker using machine learning can be developed that provides an effective and reliable way to detect plagiarism in academic assignments.

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 proposedsystem for feature extraction from the text. The Tf-idf vectorizer is used for word embedding, i.e., conversion oftextual data into an array of numbers.

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

The implementation approach involves four crucial steps thatincludes,

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 basicconcept 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 ascale of 0-1(positive values of cosθ ranges from 0 to 1).

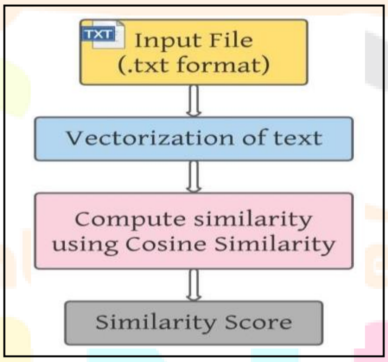

Figure 1 System Flowchart

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 contextof input.

- Term Frequency = (count of the term) / (total wordcount in the document)
- Inverse Document Frequency = log (number of docs) / (docs containing keyword)

a) 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 andfetches 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.

RESULT

Online assignment plagiarism checkers that use machine learning typically employ a combination of rule-based and statistical methods to identify potential instances of plagiarism. They may also use natural language processing (NLP) techniques to analyze the text and identify patterns or similarities that suggest plagiarism.

These tools work by comparing the submitted text to a large database of existing texts, including websites, academic papers, and other sources. They then generate a report that highlights any instances of text that match existing sources, along with a percentage score that indicates the level of similarity.

While these tools can be effective at identifying instances of blatant plagiarism, they may not catch more subtle forms of plagiarism, such as paraphrasing or rewording. Additionally, they may generate false positives, flagging text as potentially plagiarized when it is not. As with any automated tool, it's important to review the results carefully and use your judgment to determine whether plagiarism has actually occurred.

Overall, online assignment plagiarism checkers that use machine learning can be a useful tool for educators and students, but they should be used in conjunction with other strategies for preventing and detecting plagiarism. Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

CONCLUSION

In conclusion, the use of machine learning for online assignment plagiarism checking is a promising approach that can provide reliable and efficient results. Machine learning models can be trained to identify patterns and similarities in text, allowing them to detect plagiarism with high accuracy. Additionally, the use of machine learning can improve the scalability and automation of the plagiarism detection process, making it easier for educators to check large volumes of assignments.

However, it's important to note that machine learning models are not infallible and may still have limitations in detecting certain types of plagiarism, such as paraphrasing or patchwork writing. Therefore, it's important to use plagiarism detection tools as part of a larger approach to academic integrity that includes educating students on proper citation and referencing practices.

Overall, the use of machine learning for online assignment plagiarism checking can be a valuable tool for educators and institutions to ensure academic integrity and maintain high standards of education.

REFERENCES

[1]Al-Sabbagh, M., Al-Sarayreh, M., & Al-Shawakfa, E. (2020). A machine learning approach to detecting plagiarism in academic writings. International Journal of Advanced Computer Science and Applications, 11(4), 12-17.

[**2**]Vukmirović, P., & Milutinović, V. (2018). Plagiarism detection using machine learning algorithms. In 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 000157-000162). IEEE.

[**3**]Singh, D., Singh, M., & Kaur, H. (2020). Plagiarism detection using machine learning techniques. Journal of Information Technology and Software Engineering, 10(01), 1-8.

[**4**]Kothari, V., & Singh, V. (2020). Machine learning based plagiarism detection in online assignments. International Journal of Advanced Computer Science and Applications, 11(3), 130-135.

[**5**]Shrivastava, S., Shukla, A., & Verma, M. (2019). A machine learning approach for plagiarism detection. In 2019 International Conference on Inventive Research in Computing Applications (pp. 1042-1047). IEEE.

[**6**]Farag, A. A., & Mousa, H. M. (2019). Plagiarism detection system based on machine learning algorithm. In 2019 IEEE 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 456-461). IEEE.

[**7**]Nouri, R., & Faez, K. (2019). Machine learning approach for plagiarism detection: a comparative study. In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 1045-1050). IEEE.

[**8**]Lin, W., Zhang, Y., Huang, K., & Xu, X. (2021). A machine learning-based approach for plagiarism detection in academic writing. IEEE Access, 9, 27158-27166.

[**9**]Hossain, M. A., & Shabut, A. M. (2021). An effective plagiarism detection system using machine learning and natural language processing techniques. Journal of Ambient Intelligence and Humanized Computing, 12(3), 2943-2953.

[**10**]Prasetyo, H. P., & Cahyaningsih, S. (2021). Plagiarism detection of student assignment using machine learning. In 2021 6th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA) (pp. 6-11). IEEE.