# Detection of reuse texts for research and educational documents

Roberto Valdez, Renata de Luna, Eduardo Acosta

**1. Abstract**

The detection of plagiarism and reuse of text is of the utmost importance at present, since falling into this is committing a crime. For this reason, various techniques and methods have been developed to detect it, in order to publicize and differentiate those documents that really must be accredited for the work done on their documents, as well as differences in the types of reuse text present in each of the documents which are suspected of not being a genuine document. In this paper, we will develop and describe the implementation of software developed to achieve both plagiarism and text reuse detection.

To achieve this, it was necessary to make the difference between plagiarism and text reuse, finding identical texts to the original text or only passages used without having been cited. Knowing this, we decided that the developed software would be capable of both recognizing plagiarism and text reuse.

For this, we implement the pre-processing of the texts, including the separation and cleaning through various methods, to later carry out a comparison of the suspicious text. Once said comparison is made, we will obtain as a result a list with the percentage of similarity and the occurrence between both texts, with which we can determine whether or not the suspicious text is text reuse.

**2. Introduction**

Today, copyright laws protect many different types of works such as paintings, photographs, illustrations, musical compositions, sound recordings, computer programs, books, poems, blog posts, movies, architectural works, works of theater and much more. This is why copyright protection is of paramount importance.

Plagiarism detection or detection of reuse text, is the process of locating instances of plagiarism or copyright infringement within a work or document. However, when credit is given to the work or quotes are used, it can become a reference, which affects its detection in the texts.

Now it should be noted that copying is not the same as a copy, since copying does not imply the theft or reuse of information, which can generate a totally different idea based on the information reviewed, making it known that this does not necessarily carry criminal or intentional charges for using that information.

That is why it is also known as content similarity detection or text reuse. This situation escalates to various fields such as software development. There are code similarity detection tools that allow you to determine how similar the source code of different developers is. Developing computational tools that cover all types is very complex and time-demanding, which could mainly be concentrated on different levels of text reuse:

- Level 1: unchanged copy of the document.
- Level 2: Copy of text from different documents united into one.
- Level 3: Generate a paraphrase, verb change, tense change, person change, generative change, data change of one or more documents in one.

In the development of this document, we will implement a tool to be able to identify text reuse, which will be based on finding if the analyzed text has a high percentage of similarity with the original text, thus identifying if the suspicious text incurs text reuse .
With this paper , we generate the following hypothesis:

*"The developed software will be able to detect 90% of the text reused from an original source in another plain text generated with the help of different techniques such as lemmatization, stemming, n-grams, cosine similarity and distance between paragraphs, which will initially be tested for the best methods with texts extracted from the resources of the digital library of the Tecnológico de Monterrey."*

**3. Problem Statement**

In this paper, a text reuse detection tool will be designed and developed, which uses various techniques, Quantitative methods. data cleaning, classification and root word recognition, to produce an effective and efficient solution that can be projected as a text reuse detection service. To determine the above, we will compare the proposed solution with a reading of texts and documents within the database carried out with methodologies and processes that will be seen later.

In solving the problem, techniques seen in this document should be applied, such as the longest common subsequence, generation of root words within the sentences present between each document, comparison of n-grams, among others. Initially, we will use lemmatization and stemming, n-gram generation, and cosine similarity to be able to make comparisons between genuine and suspect texts.

In the review of related literature, competence in the use of bibliographic databases, available in the library, must be demonstrated to identify the most outstanding methods in relation to the problem.

**4. Justification**

In this new era of technology in which we find ourselves, where most of the world is immersive, with artificial intelligence, information in a large number of different categories, social networks, among other thousands of tools and data present on the internet, they have played a very important role for the new way of life that we are living today, as well as the new methods of searching and using the information present on the internet with a minimum investment of time when carrying out the activity.

Before this, students and professional researchers invest a large amount of time in collecting, validating data and generating a valid or correct enough document to be part of their work, however, with these new technologies at hand, the time of searching and generation of documents is faster than ever in recent years, going through an artificial intelligence or tool that supports and generates the search for the information that is needed, validating it during the process and generating a new document based on the discoveries of what the tool does for you.

The latter can benefit and harm researchers and students, as well as the institutions that represent them, since, as these tools can generate good documents for their work, they can also use text or exact copies of documents, which can cause problems of legitimacy on a document that is not properly referenced or, in the upper case, doing a job that is affected by being plagiarized from one or more documents that were collected by the tool.

The purpose of this research is to be able to implement a software tool which can detect the reuse of texts with various data cleaning methods and denote which of these was better based on the precision of the similarities presented by the distance cosine, which this tool can increase the chances of detecting the reuse of texts and that it releases an optimal result of three genuine documents that is presented in the suspect document by means of percentages given by the same distances of both types of documents.

## 5. Methodologies of Related Works

In order to achieve the objectives of this paper, we investigate and search for information regarding the problem statement that is being addressed in this paper were carried out, which leads us to the following documents that we take to identify methods and processes that other authors addressed the same problem with bases, points of view and methods different from this paper. These external papers are the following:

[1] In the first paper, named *Plagiarism detection: Methodological approaches*, the author explains the differences between plagiarism and copyright infringement, as we do in our work. The author also explains the latest research done in computer-based plagiarism detection, the examination of frameworks for plagiarism, and the importance of interpreting the data correctly in detailed step by step analysis for a live case of plagiarism between translators.

[2] The next paper, titled *Chatting and cheating: Ensuring academic integrity in the era of ChatGPT*, the author explains that artificial intelligence is a great tool for students but also a big challenge to avoid cheating. The author proposes different ways in which universities can prevent and detect cheating, such as implementing different methods, policies and procedures that can address these concerns by taking a proactive and ethical approach to use this tools .

[3] The third paper we studied is titled *The impact of text pre-processing to determine the similarity in students assignments* and tries to identify a copy of a student's assignment and the original one. As we did in our work, the authors used different pre-processing methods, such as lemmatization, replacement of synonyms and suppression of stop words, which the combinations of pre-processing techniques and methods for determining the similarity of students assignments that they used are the most suitable, if we want to detect similarity as exactly as possible and for particular techniques to find out the extent in detection of categorized types of plagiarism.

[4] On the other hand, we studied a paper titled *Some students plagiarism tricks, and tips for effective check*. In this paper, the authors explain how students try to attempt to cheat with different tricks, some plagiarism detectors, and how institutions can prevent this. As in our work, this paper tries to give solutions to identify plagiarism, but it gives us a lot of ideas in which the software developed can be vulnerable to some student's tricks.

[5] The next text studied is titled *Combating unethical publications with plagiarism detection services.* This article explains that the plagiarism detection softwares that are free do not have access to all the bibliographical sources, so this has as consequence that plagiarism cannot always be detected. Mainly as a consequence of the availability and completeness of the literature bases to which new queries are compared. Most of the commercial software has been designed for detection of plagiarism in high school and college papers. However, there is at least a fee-based service which is designed to target the needs of the biomedical publication industry. Information on these various services, examples of the type of operability and output, and things that need to be considered by

publishers, editors, and reviewers before selecting and using these services is provided.

[6] The text titled *Taxonomy of academic plagiarism methods* explains what plagiarism is, the terms related and the subdomain of plagiarism in texts. This text explains the different types of plagiarism, and clears the idea of plagiarism and text similarity and suggests the new classification of academic plagiarism, describes sorts and methods of plagiarism, types and categories, approaches and phases of plagiarism detection, the classification of methods and algorithms for plagiarism detection. The title of the article explicitly targets the academic community, but it is sufficiently general and interdisciplinary, so it can be useful for many other professionals like software developers, linguists and librarians.

[7] In the paper titled *A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm*, the authors explore a new technique to detect plagiarism based on attention mechanism-based long short-term memory (LSTM). The usage of the BERT model can detect different linguistic characteristics. The authors of this text use, as we did, AI to give a solution to plagiarism detection, training different models classifying the words into different classes. These AI models are very useful and effective for detecting plagiarism.

[8] The next paper titled *Model of Lexico-Semantic Bonds between Texts for Creating Their Similarity Metrics and Developing Statistical Clustering Algorithms.* The authors of this experiment explained how they created and established different metrics to determine the similarity between texts, and when to establish when it is plagiarism or not. This paper is a great start to establish metrics that help the software to determine when to classify a suspicious text as plagiarism or text reutilization or not.

[9] The next studied text is titled *Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches.* The authors of this document explained how they did to detect plagiarism using vectorization of texts, cosine similarity, and other techniques. The authors addressed the problem in a similar way as we did: using text vectorization and cosine similarity. They concluded that these two methods are very effective to detect plagiarism and text reutilization. The approach studied is an excellent way to pre-process the studied text and, in an accurate way, the software that is going to be developed can calculate the cosine similarity yo determine if the text is a plagiarism of another text.

[10] The last paper titled *Text Mining for Plagiarism Detection: Multivariate Pattern Detection for Recognition of Text Similarities* explains how text similarity is detected using text mining, analyzing the texts. Using text mining can be an excellent approach to analyze the texts in an efficient and accurate way.

## 6. Our Methodology

For the development of the solution to the planted problem, cleaning, classification, ordering and decision-making methods will be used, which can be of great help for the development of the solution. Some of the methodologies that are intended to be used within the investigation are divided into the following sections:

- Specification of the preprocessing stage
    - Reading Text from the documents: The preprocessing stage is where we take care of reading the suspicious and genuine texts included in the dataset from a local folder stored as a dataset.
    - Stemming: In addition to opening the files, we convert all text to lowercase letters and tokenize words to separate them by whitespace. We carry out the stemming on the text, with the help of the LancasterStemmer being the most aggressive method to find the root words and automatically cleaning the texts with previous functions that support obtaining the root words.
    - Lemmatization: On the other hand, in addition to Stemming, also in this solution, we generate the tokenization of the words by means of the spaces and we carry out the search for root words, but unlike the previous method, the lemmatization method is carried out by rules, which we create to be able to find the root only of the words that are verbs, nouns and adjectives that were previously identified. With lemmatization, morphological affixes are eliminated, thus identifying the roots of each of the words is easier. Once the texts are lemmatized, we save the root words per sentence and the root words per paragraph in a list. In addition, we generate a list that contains a single root, without repetitions.
    - Text Cleaning of symbolism and using compilers: Subsequently, we clean the texts, in order not to increase the percentage of similarity between two texts due to elements that are found

recurrently in all or most of the texts. In our case, we remove punctuations and ignore both expressions attached to an "@" symbol and links beginning with http or https.

- Preparation stage specification
    - Preprocessed texts: At this stage, both genuine and suspect texts should already be open and preprocessed with the steps performed in the previous stage.
    - N-grams: N-grams are generated in order to analyze the keywords within a context, going through the sentences and counting the frequency in which they appear within a text. Linear sequences of various sizes are generated, which in our case we are using unigrams, bigrams and trigrams, to identify keywords within the context in which they are found. Sentences are separated into groups of words, in the case of unigrams the groups consist of only one word, in bigrams they are groups of two words and finally in the case of trigrams they are groups of three words. We generate the n-grams of both the genuine texts and the suspicious texts.

- Specification of the decision-making stage
    - Preprocessed texts: For this stage, we use the texts from the previous stages, both the original texts, lemmatized and stemming texts, and the texts divided with n-grams. In the case of preprocessed texts, we only keep a list with the words, so as not to work on tuples.
    - Cosine similarity: To calculate the similarity between both texts, we calculate the cosine similarity between two vectors belonging to a genuine text and a suspect text. We use a suspect text and the dataset of the genuine texts so that the suspect text can be compared with each of the genuine texts belonging to the dataset, obtaining as a result a matrix with numbers from 0 to 1, the highest being that with the one with the greatest similarity, which allows us to determine if the text incurs in text reuse.
    - Results: After calculating the cosine similarity, we obtain for each suspect text a matrix containing N cosine similarity results, corresponding to the N number of genuine texts included in the dataset.

## 7. Expected Contributions

The purpose of this paper is to develop a text reuse detection software tool that can identify the used text or texts from genuine texts based on a single suspect text and demonstrate the percentage of similarity between of both types of text and highlighted in order to make an optimal prediction for the detection of reuse of texts in one or more documents.

## 8. Conclusions

In the presented document we carry out an experiment on the detection of text reuse using different computational tools, helping us with the experimentations, discoveries and conclusions analyzed from other related work.

To achieve text reuse detection, we performed the process in three stages, starting with the preprocessing consisting in cleaning up texts that removed existing symbols and expressions from all or most of the files. We also carry out with this process the lemmatization of the files, obtaining the root words. On the other hand, on the original texts, we carry out stemming, which performs both the cleaning of the texts and the obtaining of root words in a deeper way. In our second stage, we carry out the generation of n-grams. Useful for finding keywords within an n-word context, traversing the entire list of words.

Finally, with the results generated in the previous stages, we perform the similarity comparison by means of the cosine similarity of a suspicious text with all the genuine texts contained in the data set, in order to obtain the text with which it has the greatest similarity. and falls into text reuse.

We determined the appropriate metrics to classify a text as reuse of text or not. With this process, we were able to successfully and effectively detect the reuse of text in a high percentage when using the Stemming and Lemmatization method thanks to its realistic and specific approaches to each document compared to the genuine data present in the project compared to the original ones.

The approach of bigrams and trigrams threw results of false positives in most of the documents analyzed with the genuine documents, since, having a high proportion of repetition within the list created for each document, which leads us to highlight the approach and precision of the Stemming and Lemmatization method, thanks to the rules that were carried out during the cleaning process of each of the documents, we analyze the text by cosine similarity once on suspicious documents that was judged and analyzed against the genuine documents.

9. **References**


[1] Guillén-Nieto, V. (2022). Plagiarism detection: Methodological approaches. Language as evidence: Doing forensic linguistics (pp. 321-372) doi:10.1007/978-3-030-84330-4_10 Retrieved from www.scopus.com

[2] Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, doi:10.1080/14703297.2023.2190148

[3] Chudá, D., Chlpek, J., & Kumor, A. (2011). The impact of text pre-processing to determine the similarity in students assignments. Paper presented at the ACM International Conference Proceeding Series, , 578 417-422. doi:10.1145/2023607.2023677 Retrieved from www.scopus.com

[4] Elkhatat AM,Elsaid K, Almeer S.(2021).Some Students Plagiarism Tricks and Tips for Effective Check.Paper presented at the Web of Science Researcher and ORCID.Retrieved from: https://www.webofscience.com/wos/woscc/full-record/WOS:000677748300001

[5] Garmer, HR.(2011).Combating Unethical Publications with Plagiarism Detection Services.Paper presented at the Web of Science Researcher and ORCID.Retrieved from:https://www.webofscience.com/wos/woscc/full-record/WOS:000286412500016

[6] Vrbanec T,Mestrovic.(2021).Taxonomy of Academic Plagiarism Methods. Paper presented at the Web of Science Researcher and ORCID.Retrieved from: https://www.webofscience.com/wos/woscc/full-record/WOS:000655021100017

[7] Moravvej SV, Mousavirad SJ, Oliva D, Mohammadi F. A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm. 2023. Accessed May 20, 2023. https://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2305.02374&lang=es&site=eds-live&scope=site

[8] Demidova L, Zhukov D, Andrianova E, Kalinin V. Model of Lexico-Semantic Bonds between Texts for Creating Their Similarity Metrics and Developing Statistical Clustering Algorithm. Algorithms. 2023;16(4):198. doi:10.3390/a16040198

[9] Ayob Ali, Alaa Taqa. Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches. مجلة التربية والعلم. 25-8:(1)31;2022. doi:10.33899/edusj.2021.131895.1192

[10] Xylogiannopoulos K, Karampelas P, Alhajj R. Text Mining for Plagiarism Detection: Multivariate Pattern Detection for Recognition of Text Similarities. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Advances in Social Networks Analysis and Mining (ASONAM), 2018 IEEE/ACM International Conference on. August 2018:938-945. doi:10.1109/ASONAM.2018.8508265