EasyChair Preprint

№ 12504

Semantic Exploration of Textual Analogies for Advanced Plagiarism Detection

Elyah Frisco Andriantsialo, Volatiana Marielle Ratianantitra and Thomas Mahatody

March 15, 2024

Semantic Exploration of Textual Analogies for Advanced Plagiarism Detection

Elyah Frisco Andriantsialo
GLoRe, Madagascar
elyahfrisco7@gmail.com

Volatiana Marielle Ratianantitra
GLoRe, Madagascar
volatianamarielle@yahoo.fr

Thomas Mahatody
GLoRe, Madagascar
tsmahatody@gmail.com

Abstract

This study explores textual analogies in French within the context of plagiarism detection, adopting a semantic approach. By combining traditional methods with advanced models such as BERT and GPT, the paper proposes a hybrid model to enhance detection efficiency. Comparative evaluation highlights the model's ability to detect subtle similarities and paraphrases. The approach represents a significant advancement in accurate plagiarism detection by leveraging deep contextual understanding and the reformulation capabilities of integrated models.

1 Introduction

Plagiarism detection, constantly evolving, remains a crucial challenge in the field of digital content management. The emergence of new copying methods and circumvention of traditional systems necessitates the exploration of more advanced and adaptive solutions. In this perspective, our research positions itself at the forefront of innovation by focusing on Semantic Exploration of Textual Analogies.

The current landscape, marked by the sophistication of plagiarism practices, underscores the urgency of adopting more complex and sophisticated approaches. Our research capitalizes on the latest advancements in natural language processing (NLP), thus laying the groundwork for a more robust plagiarism detection system.

2 Basic Theory

2.1 Plagiarism

Plagiarism is a term with moral and aesthetic connotations, used in literature to describe the act of incorporating, in an undisclosed and more or less faithful manner, textual elements from another author. This term is not commonly used in legal contexts, where one would rather refer to infringement and violation of copyright law (Vandendorpe, 1992).

A document is considered plagiarized when it is produced by applying a series of transformations to an original document. The plagiarized document should retain the same function as the original but may take on a different form. There are several types of plagiarism, including copy-paste, paraphrasing, the use of false references, and plagiarism of ideas. (Mostafa, 2016)

2.2 Natural Language Processing

Natural Language Processing (NLP) is a multidisciplinary field involving linguistics, computer science, and artificial intelligence (AI) with the aim of creating NLP tools capable of automatically processing linguistic data for various applications.

. Some of the most well-known applications include automatic translation, information extraction, text summarization, spell checking, automatic generation, voice synthesis, speech recognition, and the detection of specific topics (sentiment analysis, etc.) (Ratianantitra, 2023).

One outcome of the progress in NLP is GPT (Generative Pre-trained), a language model employing deep learning to generate text resembling human speech. In simpler terms, it's a computational system created to produce sequences of words, code, or other data from an input source known as the prompt. GPT finds applications in various tasks like machine translation, where it predicts word sequences statistically. The model is trained on an unlabelled dataset comprising texts from sources like Wikipedia, available mostly in English but also in other languages. This computational approach serves diverse purposes, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and more (Floridi, Luciano, Massimo Chiriatti, 2020).

3 Literature review

The literature review emphasizes the importance of detecting plagiarism by examining similarities between documents. Different approaches, can be explored. Table 1 summarizes these plagiarism detection methods, ranging from simple algorithms to advanced approaches.

Method	Description
Fingerprinting	Represents the document in the form of fingerprints (n-grams) and utilizes algorithms such as "Rabin-Karp" for plagiarism detection.
String Matching	Compares documents word by word using algorithms such as "Brute Force"
Bag of Words	Utilizes a vector space model with vectors representing documents, calculating cosine similarity to measure the similarity between texts.
Citation Analysis	Analyzes citations within texts to detect similar patterns in citation sequences, adapted for academic and scientific texts
Stylometry	Utilizes statistical methods to quantify and analyze the writing style of an author based on features such as word frequencies.
Rule-Based Algorithms	Simple to implement and quick, but limited to predefined rules and less suitable for different languages and sentence structures.
Neural networks	Achieves high performance on complex texts, detects paraphrases and similarities, but requires a high level of implementation complexity and massive amounts of data for training.
Bidirectional Encoder Representations from Transformers (BERT)	Utilizes a pre-trained model on a large corpus of text, provides a deep understanding of the text but requires high computational power and is not designed for text generation.

Table 1: Overview of Plagiarism Detection Methods

When comparing documents to detect plagiarism, the search for similarities is crucial. Word-for-word comparison, while effective in identifying "copy and paste" instances, becomes insufficient in the face of sophisticated paraphrasing and rephrasing. The work of Barron-Cedeño et al. (2013) highlights the challenges posed by these practices. Detecting paraphrases and rephrases requires distinct approaches, although they are semantically related (Harris, 1957; Martin, 1976; Duclaye, 2003).

To overcome these challenges, alternative methods can explored:

- Stylometric Approaches, this method employs statistical techniques to analyze various aspects of writing style, focusing on features such as word frequencies, sentence lengths, punctuation usage, and syntactic structures. By quantifying these features, the method aims to capture unique patterns and characteristics specific to each author's writing style.

- Neural networks, this method achieves high performance in detecting plagiarism, especially in identifying paraphrases and subtle similarities within complex texts. It utilizes advanced techniques such as deep learning models, which have demonstrated superior capabilities in capturing intricate patterns and nuances in language.

- BERT is a pre-trained natural language processing (NLP) model developed by Google. It uses Transformer architecture and is trained on large unlabeled text corpora. BERT is designed to understand the context of words in a sentence by looking at both preceding and succeeding words, allowing it to capture nuances of meaning and context.

Methods for detecting paraphrastic rephrasing are common, using alignment methods (Callison-Burch et al., 2008) or more advanced techniques (Shen et al., 2006). The work of Fenoglio et al. (2007) emphasizes fundamental elementary transformations, while Mel'cuk's Sense-Text theory (1967) is often adopted.

These advancements in NLP and models like BERT contribute not only to the efficiency of plagiarism detection but also to a more nuanced understanding of language use.

4 Methodology

Our approach is to combine the traditional method, which is Direct Textual Comparison, with Natural Language Processing techniques such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) to provide a more robust and accurate approach.

We will proceed as follows:

Step Name	Description
Phase 1 Text Preprocessing: Tokenization	Using text cleaning techniques to eliminate irrelevant elements such as whitespaces, punctuation, etc. Tokenizing the texts to prepare them for processing by models.
Phase 2 Direct Textual Comparison	Using traditional methods such as string comparison to detect direct copy-pasting.
Phase 3 Semantic Analysis with BERT	Converting texts into embeddings (vector representations) using BERT and comparing the embeddings to evaluate semantic similarity between texts. If the embeddings are very similar, it could indicate paraphrasing or plagiarism
Phase 4 Generation and Comparison with GPT	Using GPT to rephrase one of the texts and comparing the rephrased text with the other. If GPT generates a text very similar to the other text, it could indicate plagiarism.texts
Phase 5 Stylometric Analysis	Using stylometric analysis techniques to compare the writing style of both texts. If the styles are very similar, it could indicate plagiarism, especially if the content is also similar.
Phase 6 Evaluation and Decision	Combining results from all methods to make a final decision on plagiarism. For example, if direct textual comparison, semantic analysis with BERT, and stylometric analysis all indicate plagiarism, you can be reasonably certain that the text is plagiarized.

Table 2: Different phases of designing the stages of the multi-level combination method.

By combining these models, we can leverage multiple architectures and achieve better results, obtaining superior performance compared to each individual model. However, this requires careful planning and implementation.

5 Evaluation

We assess the effectiveness of our plagiarism detection approach by applying various methods. The tests encompass diverse datasets containing authentic texts and examples of plagiarism with varying levels of complexity.

Evaluation Metrics: Precision, recall, and F-measure are employed for a balanced assessment of the model.

Test Dataset: Various texts representing different styles are utilized, including simulated cases of plagiarism to test the model's sensitivity.

Comparison with Other Methods: Our performance is compared to traditional methods and others.

Let's take a look at two text extracts::

- Text A: " Les avancées technologiques ont révolutionné notre quotidien."
- Text B: " Les progrès technologiques ont bouleversé la vie quotidienne."

For testing, we used two sentences in French, as it is the most widely used language for publishing articles or writing theses in Africa. However, it can also be used with various languages such as English and Spanish, as BERT and GPT already support multiple languages.

Method	Precision	Limitation
Textual Comparison	0.85	Limited to copy-paste cases, less effective on longer texts.
Semantic Analysis with BERT	0.92	High computational costs, requires large amounts of training data.
Generation and Comparison with GPT	0.88	It can generate text, unlike BERT, but it's not primarily designed for plagiarism detection and requires finesse in hyperparameter tuning
Stylometric Analysis	0.80	May be sensitive to intentional stylistic variations.

Table 3: Results of the approaches used

To obtain the result, we followed a rigorous evaluation procedure using diverse datasets, including authentic texts and plagiarism examples of varying levels of complexity. Each method was evaluated based on its precision performance, taking into account its specific advantages and limitations.

The textual comparison method was applied to authentic texts and copy-paste cases, evaluating accuracy and identifying limitations on lengthy texts. Semantic analysis with BERT converted texts into embeddings, measuring semantic similarity with paraphrase examples while assessing computational costs.

Generation and comparison with GPT involved rephrasing a text and adjusting hyperparameters, evaluating accuracy and detecting creative similarities. Stylometric analysis assessed writing style with tests sensitive to stylistic variations, measuring accuracy.

The overall process encompassed a comparison of the performance of each method, identifying and analyzing the limitations of each approach for a comprehensive evaluation.

Following the evaluation of these results, it was observed that the plagiarism detection approach combined with the natural language processing methods Bert and GPT reflects effectiveness in several key aspects: improved accuracy, detection of complex plagiarism patterns, scalability, and generalization.

6 Conclusion

Our innovative semantic approach, integrating BERT and GPT, has demonstrated increased effectiveness in detecting various forms of plagiarism, including subtle paraphrasing. Despite challenges related to complexity and computational costs, significant benefits, such as accurately detecting paraphrased content, make our model a promising solution for meeting the requirements of plagiarism detection in diverse digital contexts. Moreover, our approach is designed to be easily implementable, utilizing programming languages compatible with OpenAI's library and BERT. Ongoing research is necessary to optimize the model and explore emerging domains, underscoring our commitment to evolving plagiarism detection tools and preserving the integrity of digital content.

7 References

Vandendorpe, C. (1992). Le plagiat.

Hambi El Mostafa, Faozia Benabbou, El Habib Ben LahMar. (2016, June). Comparaison Des Techniques De Détection Du Plagiat Académique.

Ratianantitra, V. M. (2023, December). A State of the art review on Natural Language Processing applied to the Malagasy Language. In International Conference on Artificial Intelligence and its Applications (pp. 1-5).

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics, 39(4), 917-947.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. Language, 33(3), 283-340.

Fenoglio, I., Lebrave, J. L., & Ganascia, J. G. (2007). EDITE MEDITE: un logiciel de comparaison de versions.
En ligne: http://www. item. ens. fr/index. php.

Martin, R. (1976). Inférence, antonymie et paraphrase: éléments pour une théorie sémantique.

Duclaye, F. (2003). Apprentissage automatique de relations d'équivalence sémantique à partir du Web (Doctoral dissertation, Télécom ParisTech).

Callison-Burch, C., Cohn, T., & Lapata, M. (2008, August). Parametric: An automatic evaluation metric for paraphrasing. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) (pp. 97-104).

Shen, S., Radev, D., Patel, A., & Erkan, G. (2006, July). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (pp. 747-754).