Exploring the Performance of Proximity Measurements for Paper Plagiarism Detection

Haeyun Choi
AIGS

Chaebin Lee
AIGS

Youngjun Sim
AIGS

Abstract

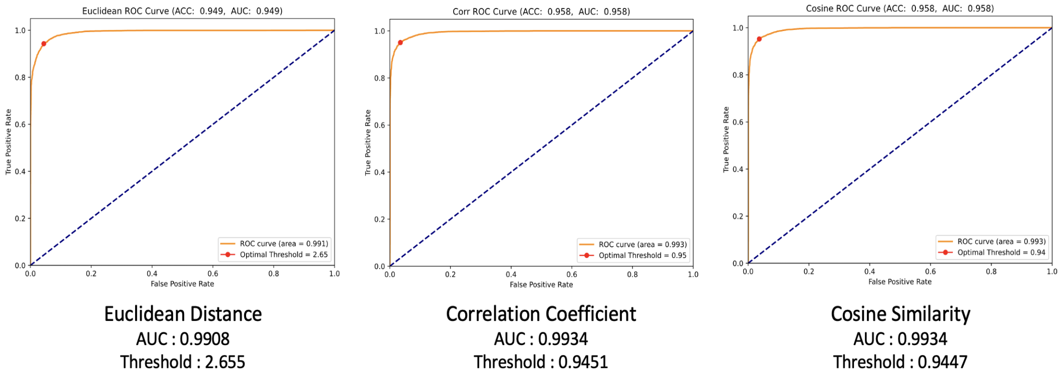
This project examines the effectiveness of various proximity measurement methods (Correlation, Euclidean distance, and Cosine similarity) in detecting paper plagiarism. Utilizing a dataset of 300 abstracts collected from three different fields, we simulate plagiarism through text augmentation techniques. By employing diverse text augmentation methods, including Easy Data Augmentation (EDA) and Parrot paraphraser, we generated a substantial dataset comprising 12,600 pairs of plagiarism and non-plagiarism instances. The performance of three proximity measurement methods in detecting plagiarism was analyzed by comparing them with the trained BERT classification model based on this dataset. The experimental results showed that all three metrics approached an accuracy of 95%, and a failure case analysis was conducted to identify underlying issues.

1 Introduction

Plagiarism in academic papers is a persistent issue that occurs regardless of intent, posing significant challenges in the field of academic integrity. Recognizing this, we explore plagiarism detection using three proximity measurement methods discussed in our course: correlation, Euclidean distance, and cosine similarity. Our investigation focuses on plagiarism in abstracts, assuming that abstracts represent papers. To address the lack of data for original and plagiarized pairs, we create our dataset, employing Easy Data Augmentation (EDA) and Parrot paraphrase techniques.

EDA [3] is a simple yet effective way to augment textual data [3]. It generates new text based on simple operations: synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD). These strategies result in new text that is relatively simple and could be structurally similar. On the other hand, Parrot Paraphrase is a paraphrasing tool designed to rephrase text while retaining the original meaning [1]. It uses advanced machine learning techniques to generate contextually similar yet structurally diverse paraphrases. This approach aids in preparing a dataset that challenges the detection capabilities of our chosen proximity measurement methods, mimicking more intricate forms of potential academic plagiarism.

Our project offers the following novelty and contributions: 1) We constructed a unique dataset comprising 12,600 pairs of original and plagiarism to investigate the effectiveness of proximity measurement methods in identifying plagiarized content. 2) We constructed a deep learning model for plagiarism detection utilizing pre-trained BERT and compared results from BERT and proximity methods to confirm the effectiveness of the proximity measurement. Through experiments, we substantiated the performance of our proximity measurement method, offering empirical evidence of its utility in plagiarism detection. 3) We uncovered underlying issues through a detailed failure case analysis, offering valuable insights into the limitations and potential improvements of current plagiarism detection methodologies.



Euclidean Distance
AUC : 0.9908
Threshold : 2.655

Correlation Coefficient
AUC : 0.9934
Threshold : 0.9451

Cosine Similarity
AUC : 0.9934
Threshold : 0.9447

Figure 1: ROC curves of three proximity measurement methods.

2 Method

We outlines the methodology used in this project to evaluate the effectiveness of different proximity measurement methods for plagiarism detection in this section. The methodology is divided into three main parts: Dataset preparation, Proximity Measurement Methods, and Plagiarism Detection using BERT.

Dataset To create a comprehensive dataset for this study, we first collected 300 abstracts from academic papers across three distinct fields: Dialogue System, Voice Conversion, and Neural Radiance Fields. Each field contributed 100 abstracts sourced from Google Scholar. We applied six different text augmentation techniques to each abstract: the four Easy Data Augmentation (EDA) strategies (Synonym Replacement (SR), Random Insertion (RI), Random Deletion (RD), Random Swap (RS)), a combination of all four EDA strategies (MIX), and the use of the Parrot API for advanced paraphrasing. Including the original abstracts, this resulted in seven versions per abstract. For the creation of plagiarism and non-plagiarism pairs, we formed 21 unique combinations (${}_7C_2$) from these seven versions for each abstract, leading to 21 plagiarism pairs. An equal number of non-plagiarism pairs were generated by sampling 21 different abstracts from the dataset, ensuring no augmented data from the same field was included. This approach yielded a total of 12,600 pairs.

Proximity Measurements In this project, we computed the proximity between pairs of abstracts embedded by a pre-trained BERT model [2] [1]. A pair of abstracts was classified as plagiarism if the proximity score exceeded a certain threshold. However, defining a precise threshold for plagiarism detection is challenging and subjective. To address this, thresholds for the three proximity measurement methods—correlation, Euclidean distance, and cosine similarity—were determined based on their performance in the training set. Specifically, we drew Receiver Operating Characteristic (ROC) curves for each method and selected the threshold that maximized the Area Under the Curve (AUC). Visual representations of these ROC curves for each metric can be seen in Figure 1.

Plagiarism Detection BERT Beyond the proximity measurement methods, we also trained distinct natural language classification models based on BERT to assess the effectiveness of each proximity measurement method and understand their respective tendencies in plagiarism detection. BERT, known for its groundbreaking results in various NLP tasks, was utilized to classify abstract pairs as plagiarism or non-plagiarism. This approach provided a comprehensive analysis of the capabilities and limitations of the proximity measurement methods in detecting plagiarism.

3 Experiment

We quantitatively assessed the performance of each proximity measurement method and the BERT-based model. The assessment focused on two primary metrics: Accuracy (ACC) and the F1 Score. A table summarizing these results is presented in the Table 1. The BERT-based model exhibited exceptional performance with an accuracy of 0.9984, while the classification results based on

[1] `https://huggingface.co/bert-base-uncased`

proximity measurements reached approximately 0.95. Although not reaching the same level as BERT, this outcome proves the effectiveness of proximity methods on plagiarism detection.

Table 1: Comparative results of each method on plagiarism detection.

Method	ACC (↑)	F1 (↑)
BERT	0.9984	0.9975
Correlation	0.9547	0.9547
Cosine	0.9531	0.9531
Euclidean	0.9460	0.9460

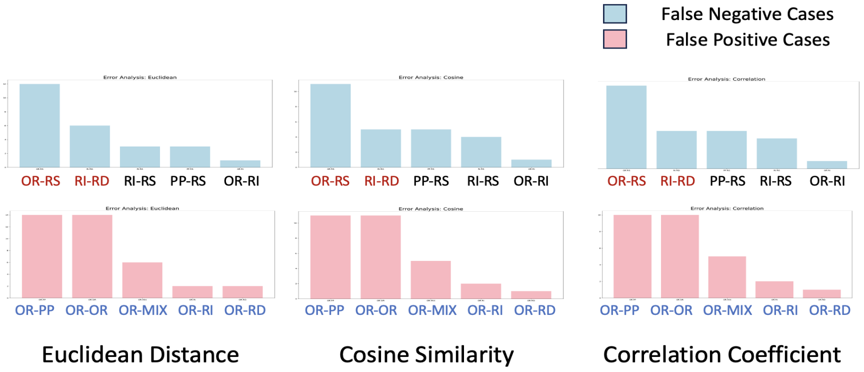


Figure 2: Failure Cases Analysis.

Further Analysis An in-depth analysis of the failure cases for the three metrics revealed insightful patterns. Histograms were used to categorize errors (False Positives and False Negatives) for each augmentation technique, as seen in figure 2. The error types remained consistent despite varying metrics, indicating specific systematic challenges across these methods.

False Negative Error Cases Two primary types of false negatives were identified. The first involved Original and Random Swap pairs, where the random shuffling of words often led to mis-identification by obscuring the context. The second challenge arose from Random Insertion and Random Deletion pairs, where adding or deleting words disrupted the text's semantic integrity, leading to increased false negatives. These findings highlight the subtleties in plagiarism detection when textual context and coherence are altered.

False Positive Error Cases We can see that original-Parrot paraphraser and original-original pairs are often mis-classified as plagiarism, regardless of the proximity measurements. In contrast to EDA techniques, which involve simple word changes, deletions, or insertions, text generated using a Parrot paraphraser maintains semantic similarity with the original but exhibits greater structural diversity. Consequently, these modified pairs pose challenges for proximity measurements in accurately calculating similarity. Further error analysis revealed that misclassifications are most prevalent in the following cases: 1) Sentence lengths were strikingly similar. 2) Topics were similar across different papers. 3) GitHub addresses were present, creating misleading similarities.

4 Conclusion

This project explored plagiarism detection using three proximity measurement methods: correlation, Euclidean distance, and cosine similarity. Our key contribution is the development of a novel dataset, tailored to test these methods, and the demonstration of their high effectiveness, with each approaching 95% accuracy. This research not only validates the utility of these proximity measurements in academic plagiarism detection but also sets a foundation for future advancements in this field by conducting a failure case analysis.

References

[1] Damodaran, P.: Parrot: Paraphrase generation for nlu. (2021)

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[3] Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382–6388 (2019)