Applied Computer Science, vol. 19, no. 2, pp. 126–141
doi: 10.35784/acs-2023-20

Submitted: 2023-05-03 | Revised: 2023-06-20 | Accepted: 2023-06-25

Keywords: NLP, text segmentation, mal-segmentation, BERT

Abdelrahman HALAWA [0009-0004-7107-1049]*, *Shehab GAMALEL-DIN* **, *Abdurrahman NASR****

EXPLOITING BERT FOR MALFORMED SEGMENTATION DETECTION TO IMPROVE SCIENTIFIC WRITINGS

Abstract

Writing a well-structured scientific documents, such as articles and theses, is vital for comprehending the document's argumentation and understanding its messages. Furthermore, it has an impact on the efficiency and time required for studying the document. Proper document segmentation also yields better results when employing automated Natural Language Processing (NLP) manipulation algorithms, including summarization and other information retrieval and analysis functions. Unfortunately, inexperienced writers, such as young researchers and graduate students, often struggle to produce well-structured professional documents. Their writing frequently exhibits improper segmentations or lacks semantically coherent segments, a phenomenon referred to as "mal-segmentation." Examples of mal-segmentation include improper paragraph or section divisions and unsmooth transitions between sentences and paragraphs. This research addresses the issue of mal-segmentation in scientific writing by introducing an automated method for detecting mal-segmentations, and utilizing Sentence Bidirectional Encoder Representations from Transformers (sBERT) as an encoding mechanism. The experimental results section shows a promising results for the detection of mal-segmentation using the sBERT technique.

1. INTRODUCTION

Text segmentation involves dividing text into coherent segments that reflect different topics, with clear points indicating topic transitions. Proper segmentation improves the readability and understanding of a document, and facilitates downstream applications like summarization and information extraction by producing more accurate results. Conversely, incorrect segmentation negatively impacts document understanding (Levy, 2013). In this research, the authors refer to incorrect or improper segmentation as "mal-segmentation," where text is divided in a way that misleads the intended meaning.

Unfortunately, not all writers, especially junior researchers such as postgraduate students and assistant researchers, possess the necessary expertise. Consequently, scientific articles and theses often suffer from mal-segmentation, leading to longer study times, potential

* Al-Azhar University, Faculty of Engineering, Systems and Computer, Egypt, ahalawa@azhar.edu.eg
** Al-Azhar University, Faculty of Engineering, Systems and Computer, Egypt, drshehabg@yahoo.com
*** Al-Azhar University, Faculty of Engineering, Systems and Computer, Egypt, anasr@azhar.edu.eg

misunderstandings of the document's message, and difficulties in following the argumentation. Mal-segmentation can mislead readers and hinder their comprehension of the author's message. Additionally, accurate text segmentation algorithms rely on well-segmented text. Therefore, there is an urgent need to detect and correct improper text segmentation.

There are different types of mal-segmentation, including mal-segmentation between sentences, paragraphs, and sections/subsections (Ugur Akinci, 2012). Section 3 provides more detailed explanations and examples for each type. This article focuses specifically on mal-segmentation between paragraphs, as well-written paragraphs contribute to the overall comprehensibility of the text.

One objective of this research is to automatically detect mal-segmentation in scientific articles and provide authors with suggestions for correction. The proposed mal-segmentation detection model in this article primarily relies on sentenceBERT (Reimers, 2019) to generate semantically meaningful sentence embeddings. The research includes a series of experiments that build upon each other, aiming to enhance accuracy by refining the definition of segments in the training model. Initially, the context was divided into fixed-size sliding windows, then variable-size windows were tested, and later the window size encompassed a full paragraph. Finally, the concept of a threshold, with evolving methods for calculating its value, was introduced to improve model accuracy. Further details regarding these experiments are discussed in Section 4.

Section 2 reviews similar studies and highlights common segmentation features, while Section 3 provides an in-depth discussion of mal-segmentation, including its various types and consequences. Section 4 describes the evolution of the suggested model and methodology through a series of experiments aimed at automatically detecting mal-segmentation. Finally, Section 5 summarizes the findings and suggests future directions.

2. RELATED WORK

Text segmentation plays a crucial role in natural language processing (NLP), and extensive research has been conducted in this field over the years. Text segmentation approaches can be categorized into two main types: supervised and unsupervised algorithms, as discussed below.

2.1. Text segmentation based on unsupervised algorithms.

One branch of unsupervised methods is based on lexical cohesion, which states that similar vocabulary tends to be in the same topic segment and vice versa. Hearst et al. (Hearst, 1997) introduced TextTiling, which is the most famous and earliest algorithm for text segmentation. TextTiling is based on the fact that high vocabulary intersection between two adjacent blocks is taken to mean high coherence. (Ponceleon, 2001) combine content-based methods with boundary-based methods. In which, analyzing the temporal distribution and the rate of arrival of features to compute an initial segmentation in the first pass. Then, detecting changes in content-bearing words by using the content-bearing features in the second pass. (Lin M. a., 2004) introduces a method that combines multiple segmentation features to improve accuracy, which include noun phrases, topic noun phrases, verb classes, word stems, combined features, cue phrases, and pronouns. Whereas (Lin M. a., 2005) uses natural language processing techniques such as noun phrases extraction, beside lexical knowledge sources such as WordNet to segment lecture videos.

Moreover, (Shah, 2015) propose a method for determining segment boundaries by matching blocks of SRT (subtitle resource tracks) and Wikipedia texts of a lecture video's topics. First, he generates feature vectors based on noun phrases in the entire Wikipedia text for Wikipedia blocks (one block for each Wikipedia topic) and SRT blocks (120 words in one SRT block). He then uses cosine similarity to compute the similarity between a Wikipedia block and an SRT block. Finally, a segment boundary is defined as an SRT block that has both the maximum cosine similarity and is greater than a defined similarity threshold. (Soares, 2019) proposes a versatile method for automatic temporal segmentation for video lectures that investigates detectable audio characteristics as well as the semantics of the teacher's words stated. (Solbiati, 2021) offer an unsupervised method that employs a new similarity score based on BERT embeddings (Devlin, 2018) that employ similarity score heuristics that are not based on neural models.

2.2. Text segmentation based on supervised algorithms

Deep neural network-based segmentation models have been developed recently. A two-level hierarchical network is a frequent structure for them. (Wang, 2018) offers a complete neural segmenter based on the BiLSTMCRF framework. While (Barrow, 2020)introduces the Segment Pooling LSTM (SLSTM) model, which can simultaneously segment and label segments in a document. A strategy for teaching the model to recover from errors by aligning the predicted and ground truth segments is also created to facilitate joint training. Moreover, (Almuhareb, 2019) developed a recurrent neural network (RNN)-based deep learning strategy, namely bidirectional long short-term memory (Bi-LSTM), to handle the problem of Arabic word segmentation without and with rewriting. Also, (Lo, 2021) proposes a transformer-over-transformer system, named transformer2, to conduct neural text segmentation. It is made up of two parts: bottom-level sentence encoders that use pre-trained transformers and an upper-level transformer-based segmentation model that uses sentence embeddings.

On the other hand, (Somasundaran, 2020) provides a paradigm that explicitly considers coherence. Coherence-Aware Text Segmentation (CATS), the suggested model, encodes a sentence sequence using two hierarchically connected Transformer networks. Whereas (Maraj, 2021) suggests a method that uses a pre-labeled text corpus in conjunction with an upgraded neural Deep Learning model. BERT is employed as a rich sentence encoder, and it is demonstrated that by using a text segmentation focused data augmentation strategy, state-of-the-art results may be obtained with minimum training.

In this paper we propose a new concept called mal-segmentation which means a wrong in text segmentation of author writing. In addition, we introduce a model to automatically detect the mal-segmentation. Our model mainly based on sentenceBERT (Reimers, 2019) to generate semantically meaningful sentence embeddings that can be compared by calculating cosine similarity between them.

3. WHAT IS MAL-SEGMENTATION?

Text segmentation refers to the process of dividing written text into meaningful units, such as sentences, paragraphs, or topics. This term encompasses both the mental processes employed by humans when reading text and the artificial processes implemented in computers, which are

the focus of natural language processing (Text_segmentation, 2011). In contrast, this research introduces the concept of mal-segmentation, which is defined as the improper segmentation of a piece of text that leads to a misunderstanding of the intended meaning. Mal-segmentation significantly affects the structure of a document. In professional scientific writing, such as articles and theses, the structure enhances readability, understandability, and the clarity of ideas. Moreover, it influences the reader's comfort level and the time required to study the document. Therefore, accurate segmentation is essential.

Mal-segments can take various forms, including non-smooth transitions between paragraphs and sections, as well as incorrect splits or unsplit paragraphs and sections/subsections. For instance, a lengthy paragraph may benefit from being divided into multiple paragraphs, while other paragraphs may need to be combined. The same applies to sections and subsections.

This article discusses the analysis of a document, such as an article or thesis, in terms of its segments or units from two perspectives: "Well Segmenting" and "Smooth Transitioning." Well Segmenting focuses on the integrity of each unit, whether it is a sentence, paragraph, section, or subsection (Hinkel, 2001). It ensures that each unit presents a complete and coherent idea. On the other hand, Smooth Transitioning considers the relationship between consecutive units. It ensures that the idea of each unit leads to the idea of the next unit, thereby enhancing the overall coherence of the argument. Effective transitions make the argumentation presented in the integral units easier to understand and more convincing.

Typically, transitions can be analyzed between sentences, paragraphs, and sections/ subsections, as illustrated in Fig. 1. Further explanation and examples will be provided in the following sections.

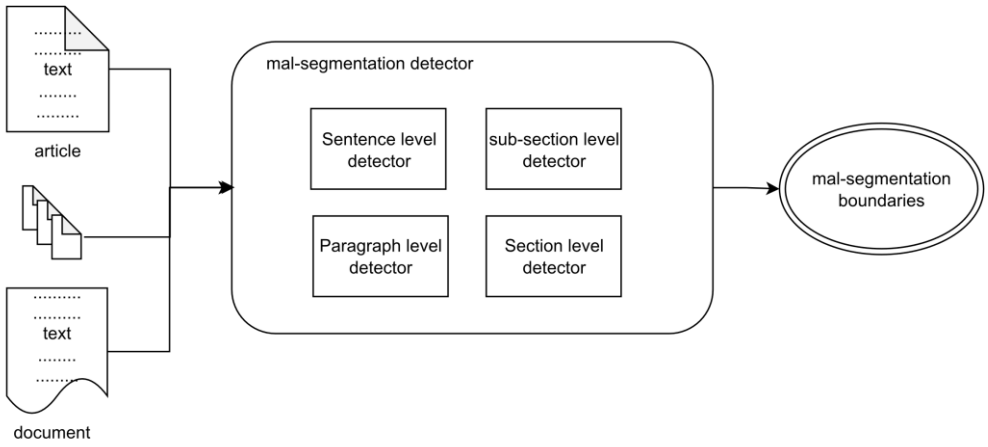


Fig. 1. A Proposed Mal-Segmentation Detection Process

Transitions between Sentences

A single sentence can serve different purposes, such as presenting additional points, providing examples or evidence, or discussing exceptions to previous statements. Consequently, a well-integrated paragraph should maintain strong connections among its sentences, ensuring that each sentence relates to both the preceding and following sentences. To illustrate this, Fig. 2 presents two examples of sentence transitions, with the latter being evaluated as having a better transition compared to the former, despite both examples conveying the same meaning.

- **Poor Transition example**: “Many students are tempted to plagiarize because they run out of time on an assignment. Good time management skills will give students plenty of time to complete assignments.”
- **Better Transition example**: “Many students are tempted to plagiarize because they run out of time on an assignment. However, students can reduce the temptation to plagiarize if they exercise better time management skills to ensure that they will have plenty of time to complete assignments.”

Fig. 2. Two Sentence-Transition Examples (University, UAH)

In the "Better Transition example," the writer establishes a connection between the two sentences by repeating keywords ("temptation" and "plagiarize") and using a transition word ("however"). The writer echoes the phrase "tempted to plagiarize" from the first sentence by using the words "temptation" and "plagiarize" in the second sentence. Additionally, the transition word "however" indicates the contrast between the two sentences. In contrast, in the "Poor Transition example," the writer states the second sentence as a fact and leaves it to the reader to infer the connection between this fact and the first sentence. Similarly, in the previous example, "Poor Transition Example," the writer states the second sentence as a fact and leaves it to the reader to infer the connection between this fact and the first sentence.

Likewise, on the same evaluation scale, the transition between Sentence 2 and Sentence 3 in Fig. 3 is considered a poor transition because the main keywords are not shared between them.

Unlike in many Western countries where companies employ people whose skills can be effective immediately, Japanese companies select applicants with potential who can be trained to become suitable employees. For this reason, recruiting employees is an important exercise for companies, as they invest a lot of time and money in training new staff. This is basically true both for factory workers and for professionals. Professionals who have studied subjects which are of immediate use in the workplace, such as industrial engineers, are very often placed in factories and transferred from one section to another. By gaining experience in several different areas and by working in close contact with workers, the engineers are believed, in the long run, to become more effective members of the company. Workers too feel more involved by working with professionals and by being allowed to voice their opinions. Loyalty is believed to be cultivated in this type of egalitarian working environment.

Fig. 3. A Sample Paragraph Containing Poor Sentence Transitions (ielts-mentor, 2022)

Transitions between Paragraphs

Transitions between paragraphs are essential for demonstrating the relationships between them. Regardless of how well-constructed each paragraph is on its own, they must be logically connected to ensure the essay forms a coherent whole (Ugur Akinci, 2012). Two paragraphs are connected through sentences that incorporate key ideas from each paragraph, along with connector keywords to clarify the relationship between them. Fig. 4 illustrates examples of both poor and better transitions between paragraphs.

- **Poor Paragraph Transition:**
  Paragraph A: Malcolm X uses the rhetorical strategy of logos (logic) to convince his audience. . .. [paragraph about logos]
  Paragraph B: Malcolm X's article also has a lot of pathos (emotion). . .. [paragraph about pathos]

- **Better Paragraph Transition:**
  Paragraph A: One of the main rhetorical strategies employed by Malcom X is logos (logic). . .. [paragraph about logos]
  Paragraph B: In addition to using logos as a rhetorical strategy, Malcolm X also employs pathos (emotion) to persuade his audience; in fact, there are more examples of pathos in the article than logos. . .. [paragraph about pathos]

Fig. 4. Two Paragraph Transition Examples (University, UAH)

In Fig. 4, the example of "Better Paragraph Transition" demonstrates the use of the opening signal phrase "in addition" in Paragraph B, indicating that it presents an additional point to support the argument in Paragraph A. Furthermore, this phrase refers back to the main idea (logos) of the preceding paragraph, Paragraph A. Additionally, the second clause "in fact, there are more examples of pathos..." emphasizes the relationship between the two points in the two paragraphs. It not only illustrates an additional point in the second paragraph but also suggests that it presents a more significant point.

This example has been chosen, since it represents the most commonly used attributes for the upper levels of the Decision Tree. The most used attributes by a substantial margin were the Meteor score and its intermediate results, such as Meteor Mean and Meteor Chunks. The next most important attributes, showing up the most frequently in the top levels of the Decision Tree, are the BLEU score and the difference in translation lengths. It is clearly observable that attributes using reference translations are more influential, than metrics having no access to them.

Furthermore, as shown in table 4.10, using Google Translate as a candidate, results in substantially higher recall values for *automated translation* and therefore higher precision values for *professional translation*. This is due to the fact that Google Translate seems to possess a treat that makes the algorithm predict text fragments more often as *automated translation*. A first estimated reasoning could be, that Google's translating machine produces output that partially shares characteristics of professional translations.

Fig. 5. A Sample Poor Transition between Paragraphs (Luckert, 2016).

Another example of a poor transition between paragraphs is shown in Fig. 5, where two consecutive paragraphs lack coherence between them.

Transitions between Sections/Subsections

In many cases, lengthy papers are divided into various sections consisting of multiple paragraphs. For example, in a problem-solution paper, you may have multiple paragraphs discussing the identified problem, followed by several paragraphs explaining the proposed solution. To ensure smooth flow and enable the reader to easily follow your argument, it is

important to include effective transitions between each section. Similar to transitions between paragraphs, transitional sentences should relate to the main ideas in each section and demonstrate their connection. Fig. 6 illustrates examples of both poor and better transitions between sections.

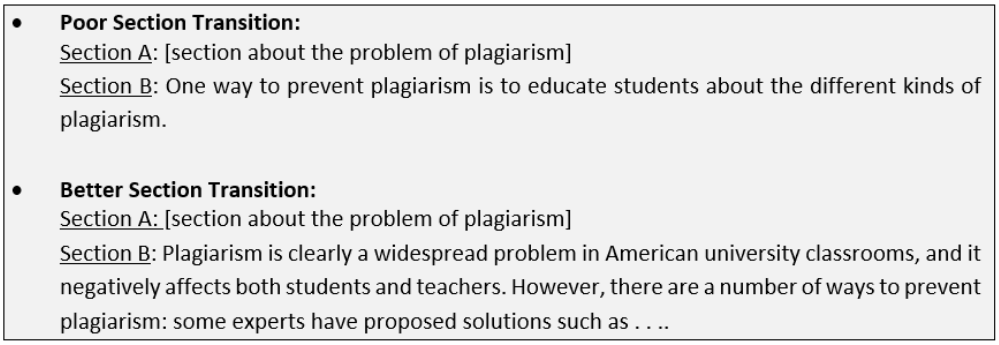


Fig. 6. Two Section Transition Examples (University, UAH)

In the example of "Better Section Transition," the writer refers to the main idea of the previous section (the problem of plagiarism) and the main idea of the upcoming section (solutions to the problem of plagiarism). The writer also indicates a relationship between the two sections by using the connector "however," which highlights a contrast between them.

To summarize, authors' written texts may encounter the problem of poor transitions between sentences, paragraphs, or sections. This issue is referred to as the "mal-segmentation problem," which can mislead readers and hinder their proper understanding of the presented argumentation, thereby affecting the quality of the manuscript.

The next section introduces a proposed model for the automatic detection of mal-segmentation using a new similarity score based on BERT embeddings (Devlin, 2018). This article focuses solely on the detection of mal-segmentation between paragraphs written in English language.

4. MAL-SEGMENTATION DETECTION

Mal-segmentation is a problem that can have a negative impact on the readability of a scientific document, thereby affecting the understandability of its argumentation. Therefore, identifying and correcting mal-segmentation is essential. This section describes a series of experiments that form the methodology for detecting mal-segmentation in this research. Each experiment builds upon the results of the previous one. All experiments utilize sentence embeddings to assess the semantic similarity between input contexts, which is formalized as the degree to which two sentences are semantically equivalent (Cer D. a.-G., 2017). Each experiment defines the input context differently with the aim of improving accuracy. The experiments consider a specific approach for sentence representation, namely, SentenceBERT (sBERT) (Reimers, 2019), as explained below.

SentenceBERT (sBERT) is a modification of BERT designed to derive semantically meaningful sentence embeddings. BERT (Devlin, 2018) is a pre-trained transformer network which reaches state of-the-art-results for many NLP tasks. The technique used by

SentenceBERT to generate embeddings involves the application of Siamese and triplet network structures (Schroff, 2015), which enable comparisons using cosine similarity.

This research conducted a series of experiments to detect mal-segmentation in scientific writings, with each experiment redefining the input context to improve the accuracy of the results. It should be noted that this article solely focuses on detecting mal-segmentation within sections.

4.1. Data Selection and Preparation

The articles used as experimental inputs in this research, e.g., (Solbiati, 2021), (Galanopoulos, 2019), and (Maraj, 2021)], were carefully selected to minimize their impact on the experiment results. The following criteria were used for selecting the experimental articles:

- Articles from highly ranked journals such as IEEE and ACM, assuming they were properly written with adequate segmentation.
- Articles written by native English speakers, assuming they surpass non-native speakers and are more professional writers.

Regarding data preparation, horizontally crossing sections such as "Abstract," "Introduction," "Related Work," and "Conclusion" were excluded, with the focus on the argumentation parts of the articles. Paragraphs were manually identified for the training dataset. NLP preprocessing techniques such as text normalization, stop words removal, elimination of Unicode characters, and others were applied. The article text was divided into blocks of words or sentences to define boundaries that determine the set of fragments comprising the input context for the algorithm.

4.2. Experimental results

4.2.1. Experiment 1: Sliding Window

This experiment employed a segmentation scheme inspired by (Solbiati, 2021), where sections were divided into sentences and then processed for algorithm application. The sentences were cleansed by removing special characters and stop words, as well as converting all words to lowercase. Subsequently, the sentences were grouped into fixed-size windows, and the cosine similarity between each consecutive window pair was calculated. Local minimum points/similarities were identified as segmentation boundaries. Accordingly, mal-segments were defined as the original segments that were not among the identified local minima. More details about the algorithm can be found in Algorithm 1.

Algorithm 1: ***sliding_window***

Input:

context: the whole text

ori_segs: original segments boundaries of the context

window_size: fixed window size

Output:

list of mal-segmentation boundaries

1- *Divide the context into sentences.*

2- *Clean each sentence by removing punctuation, extra spaces and so on.*

3- *Compute sentence embedding Si.*

4- *Divide the context into blocks of size equal to window_size{Si · · · Sk}, and perform a block-wise max pooling operation to extract the embedding for each block. we repeatedly apply a max pooling operation to extract words with high semantic value from a given context.*

5- *Compute pairwise cosine similarity simi between the adjacent blocks. A lower cosine value means the semantic/coherent between the two blocks is small.*

6- *Select the local minimums points as segment boundaries Segi (segment position inside context).*

7- *If the list of Segi … Segn not contains any one from the ori_segs, we consider it as mal-segmentation mal-segs.*

8- *Return mal-segs*

Algorithm 1. The Sliding Window Algorithm

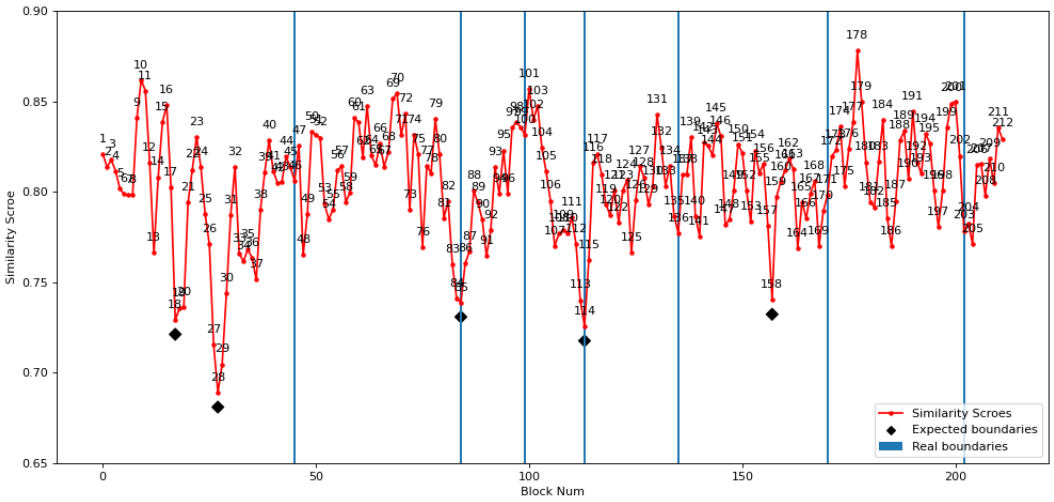


Fig. 7. Similarity Scores for Window size = 5

Fig. 7 illustrates a sample result of the Sliding Window algorithm applied to Section 5 of Luckert's thesis (Luckert, 2016) (access: http://www.diva-portal.org/smash/get/diva2:920202/fulltext01.pdf). In Fig. 7, the bold dots represent the similarity scores between consecutive windows, while the vertical lines indicate the actual section boundaries. The diamond points represent the predicted boundaries.

As shown in Fig. 7, there are 2 true positives (TP) indicated by the diamond points intersecting with the vertical lines. However, according to our algorithm, the other 3 diamond points are considered as mal-segmentation because they were not selected by the author as boundaries, despite having the lowest similarity scores. This violates the principle of smooth transition. The author should reconsider the segmentation blocks and sentence wordings to increase the similarity scores.

4.2.2. Experiment 1 results

Unfortunately, it has been observed that the window size significantly impacts the obtained results. Different window sizes lead to different segment boundaries, as evident when comparing the results in Fig. 7 (window size = 5) with those in Fig. 8 (window size = 4), as summarized in Table 1. This phenomenon emphasizes the need to neutralize the window size, which is the focus of experiment 2. In this experiment, various window sizes were applied to determine the segment boundaries.

Tab. 1. Test results of different window sizes

	precision	accuracy	recall	F1-score
Window size 5	0.4	0.96	0.28	0.33
Window size 4	0.13	0.91	0.26	0.17

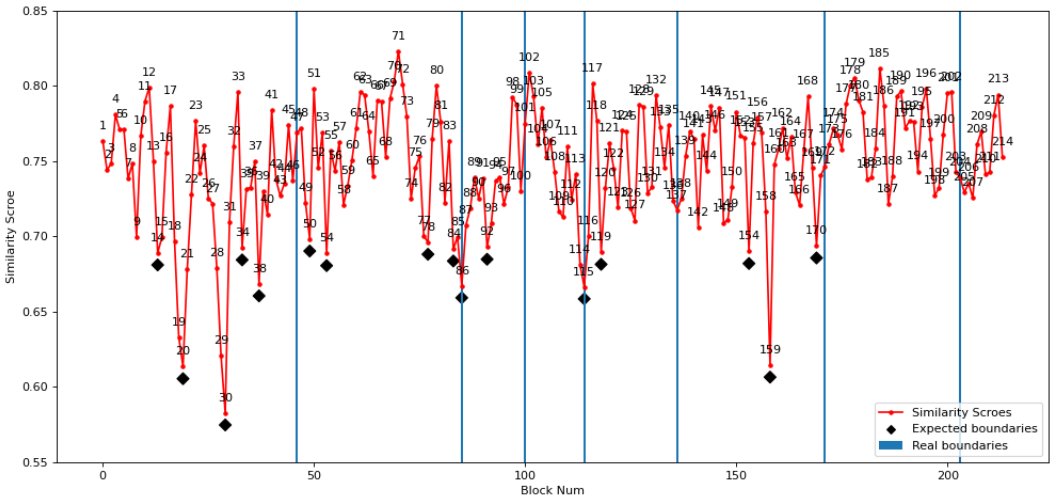


Fig. 8. Similarity Scores for window size = 4

4.2.3. Experiment 2: Sliding window with different window sizes

In this experiment, the sliding window algorithm is applied to different window sizes, resulting in L sets of mal-segments, where L is equal to the number of applied window sizes. To obtain the final list of segment boundaries, the appearance of each mal-segment in all L sets is counted. A mal-segment is considered as a segment boundary if its count exceeds a specific threshold, which depends on the number of used windows. In other words, if a mal-segment appears multiple times across different window sizes and exceeds the acceptable threshold, it is considered a true positive with an acceptable probability. For more details, refer to Algorithm 2.

```
Algorithm2: sliding_window_many_windows
Input:
 context: the whole text
 ori_segs: original segments boundaries of the context
 window_num: number of windows start from 1 to window_num
Output:
 list of mal-segmentation boundaries
 1-  L ← empty_list
 2-  while  window_size ← from 1 to sindow_num
       mal-segs = sliding_window(context, ori_segs, window_size)
          add mal_segs to L list //each item in L list is a set of mal-segs
 3-  Calculate the number of occurrences (count) of each mal-segi.
 4-  Sort these mal-segi based on its count in descending order.
 5-  Return the most top mal-segi as the final segment boundaries.
```

Algorithm 2. The algorithm of how to select the final boundaries when applying different window sizes

4.2.4. Experiment 2 results

The result of this experiment is 0.41, 0.89, 0.32, and 0.36 for precision, accuracy, recall, and F1-score respectively. Also, it was observed that some segment boundaries Seg_i might occur within a paragraph, which is not ideal as it violates the assumption that segment boundaries should be located between paragraphs. To address this issue, Experiment 3 was conducted, which utilized variable-sized paragraphs as windows instead of fixed-size windows of sentences, as discussed in the following section.

4.2.5. Experiment 3: Using paragraphs instead of window sizes

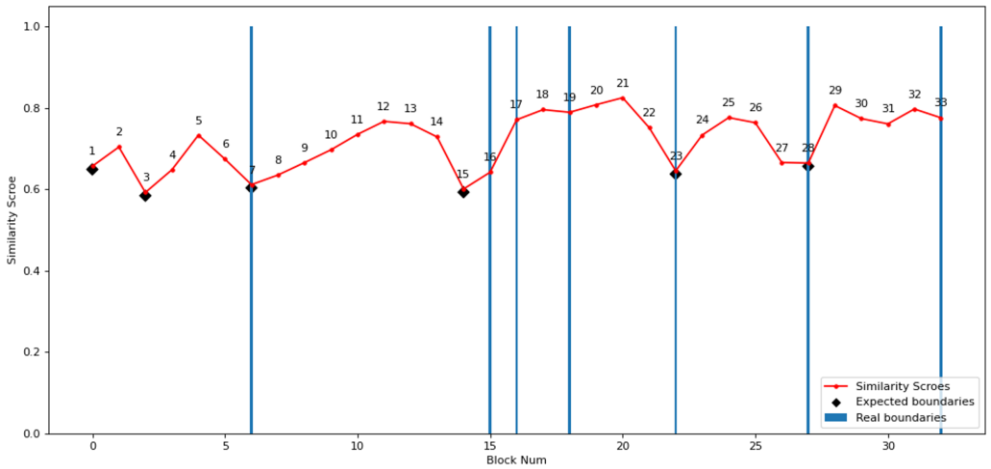
In this approach, it is assumed that the boundaries of the paragraphs are known (marked during data preparation). The window consists of a certain number of paragraphs rather than sentences. The main window size was set to one paragraph, but larger window sizes were also considered. Fig. 9 (a) and Fig. 9 (b) illustrate the results for one-paragraph blocks and three-paragraph blocks, respectively. After combining the final segment boundaries, the segmentation boundaries are guaranteed to be in the correct positions.

4.2.6. Experiment 3 results

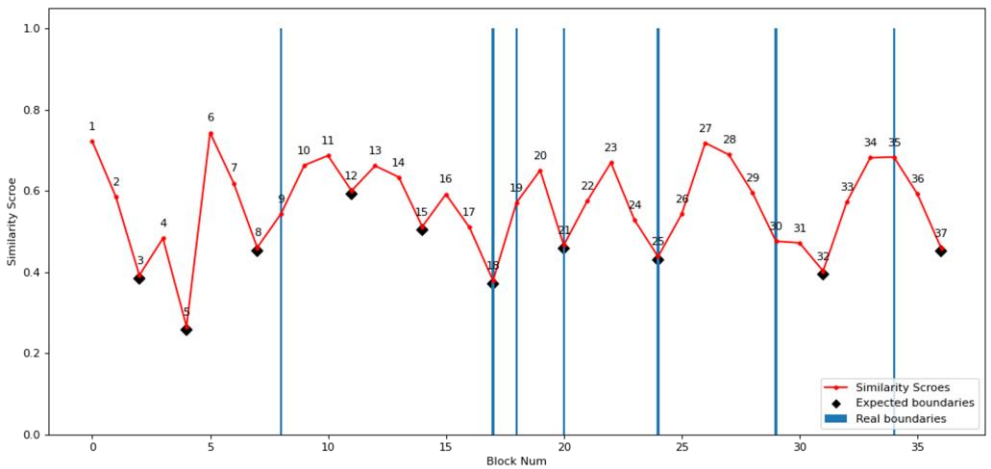
We noticed that the results are on the right way as shown in Table 2.

Tab. 2. Test results of different paragraphs in the block

	precision	accuracy	recall	F1-score
Three paragraphs in the block	0.43	0.81	0.42	0.43
One paragraphs in the block	0.3	0.74	0.40	0.35



(a) Three paragraphs in the block



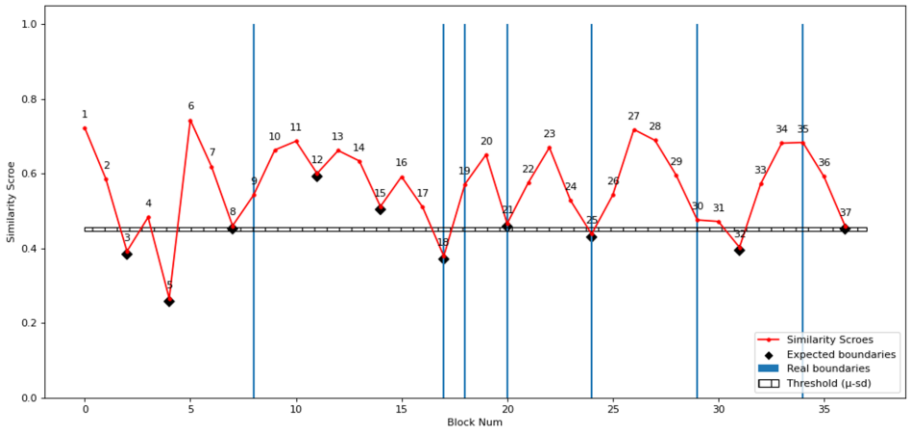
(b) One paragraph in the block

Fig. 9. Similarity Results when Blocks are in Paragraphs

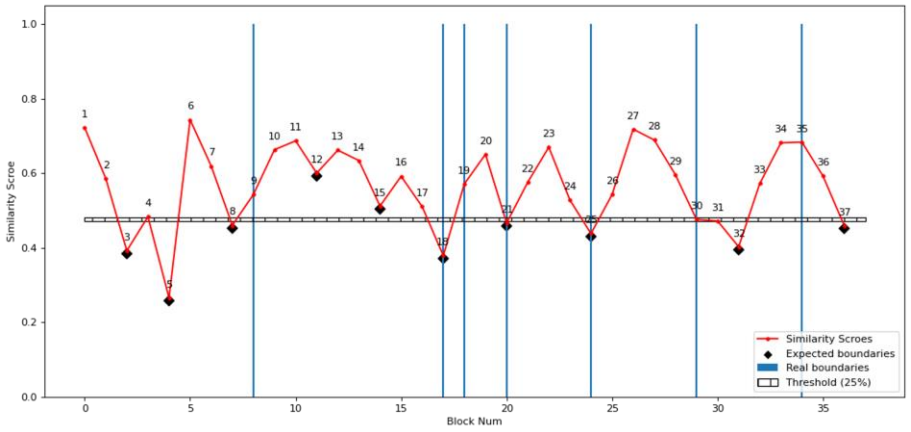
It is worth noting that some boundaries with high similarity scores are falsely identified as segmentation boundaries in the final output. For example, Fig. 8 contains some expected boundaries with high similarity scores of about 0.65, which lead to false positive segmentations. Therefore, it is necessary to reduce the number of expected boundaries by setting an upper bound for similarity scores, which is the focus of Experiment 4.

4.2.7. Experiment 4: using thresholds

The purpose of this threshold is to reduce the count of segments and address the aforementioned issues. In this technique, Algorithm 2 was applied with the modification that only segmentation boundaries below a specific threshold were considered. The value of the threshold is dynamically computed based on the context. Two techniques were used for threshold calculation:



(a) Threshold at μ – SD



(b) Threshold at highest 25%

Fig. 10. Similarity Results under Two Types of Thresholds

Setting the threshold value based on a fixed percentage (25%)

In this technique, the similarity scores were sorted in descending order, and the threshold value was set as the value at the index N/4, where N is the size of the similarities list. This means that the highest 25% of values are ignored. Unfortunately, this approach is not mature enough to eliminate false positives. It would be more accurate to calculate the threshold value relative to the context size.

Setting the threshold value based on μ - SD (mean - standard deviation)

In this alternative technique, the mean (μ) was calculated to represent the average of all similarities between consecutive blocks. Additionally, the standard deviation (SD) was calculated to measure the amount of variation or dispersion in the set of similarities. Based on the mathematical definition of SD, the similarities between blocks are expected to fall within the range of (μ - SD) to (μ + SD). Therefore, segmentation boundaries below (μ - SD) are considered as effective segmentation boundaries.

Fig. 10 (a) illustrates the threshold value (μ - SD), while Fig. 10 (b) represents the threshold using the top 25%. These figures depict some expected boundaries that exceed the threshold line, indicating that they were not included in the final boundary list.

4.2.8. Experiment 4 results

Tab. 3. Test results of different thresholds

	precision	accuracy	recall	F1-score
Threshold at μ – SD	0.49	0.84	0.44	0.46
Threshold at highest 25%	0.43	0.81	0.43	0.42

Table 3 exposes that the threshold value at (μ - SD) value achieved the best results among all experiments. At this stage, the experiments presented in this article conclude. However, we plan to continue conducting further experiments in the same direction to achieve better and more refined results.

5. PERFORMANCE EVALUATION

Based on the previous results obtained from the aforementioned experiments, the average results are presented in Table 4. It is worth noting that the best results were achieved in experiment 4, where the threshold was set at the (μ - SD) value. This is considered the final performance for the proposed methodology.

Tab. 4. The average of experiment's test results

	precision	accuracy	recall	F1-score
Experiment 1	0.265	0.935	0.27	0.25
Experiment 2	0.41	0.89	0.32	0.36
Experiment 3	0.365	0.775	0.41	0.39
Experiment 4	0.46	0.825	0.435	0.44

6. CONCLUSION

This research has introduced the concept of mal-segmentation in scientific writing. Detecting mal-segments in a document can help authors improve the readability and understandability of their writing. Additionally, this research has presented an algorithm that utilizes sBERT to detect and highlight mal-segments in a document, allowing authors to make corrections. The accuracy of the algorithm has been observed to be approximately 50%. The development of this algorithm involved four stages: sliding windows with a fixed size, sliding windows with multiple sizes, sliding windows using full paragraphs, and the application of a threshold function.

Furthermore, this research aims to extend its coverage to include non-English texts and other input contexts such as section-based segmentation. Additionally, exploring unsupervised approaches is being considered, particularly due to the challenges associated with creating a supervised dataset of reasonable size, which requires significant computational power and language experts to review and evaluate the results linguistically. These capabilities and resources are currently beyond the scope of our available means.

REFERENCES

Almuhareb, A. a.-T. (2019). Arabic word segmentation with long short-term memory neural networks and word embedding. *IEEE Access, 7*, 12879-12887. https://doi.org/10.1109/ACCESS.2019.2893460

Barrow, J., Jain, R., Morariu, V., & Manjunatha, V. (2020). A joint model for document segmentation and segment labeling. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 313-322). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.29

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). *Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation*. arXiv. https://doi.org/10.48550/arXiv.1708.00055

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo- Cespedes, M., Yuan, S., Tar, Ch., Sung, Y.-H. Strope, B., & Kurzweil, R. (2018). *Universal sentence encoder*. arXiv. https://doi.org/10.48550/arXiv.1803.11175

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Galanopoulos, D., & Mezaris, V.(2019). Temporal lecture video fragmentation using word embeddings. In Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., & Vrochidis, S. (Eds.) *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8--11, 2019, Proceedings, Part II* (vol. *25*, pp. 254--265). Springer. https://doi.org/10.1007/978-3-030-05716-9_21

Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics, 23*(1), 33-64.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied language learning, 12*(2), 111-132.

ielts-mentor. (2022). Retrieved from https://www.ielts-mentor.com/reading-sample/gt-reading/3162-employment-in-japan ?

Levy, C. M., & Ransdell. S. (1996). *The science of writing: Theories, methods, individual differences and applications.* Routledge. https://doi.org/10.4324/9780203811122

Lin, M., Nunamaker, J.F., Chau, M., & Chen, H. (2004). Segmentation of lecture videos based on text: a method combining multiple linguistic features. *37th Annual Hawaii International Conference on System Sciences.* (pp. 9-9). IEEE. https://doi.org/10.1109/HICSS.2004.1265045

Lin, M., Chau, M., Cao, J., & Nunamaker, J. F. (2005). Automated video segmentation for lecture videos: A linguistics-based approach. *International Journal of Technology and Human Interaction (IJTHI), 1*(2), 27-45. https://doi.org/10.4018/jthi.2005040102

Lo, K., Jin, Y., Tan, W., Liu, M., Du, L., & Buntine, W. (2021). *Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence*. arXiv. https://doi.org/10.48550/arXiv.2110.07160

Luckert, M., & Schaefer- Kehnert, M. (2016). Using machine learning methods for evaluating the quality of technical documents.

Maraj, A., Martin, M. V., & Makrehchi, M. (2021). A More Effective Sentence-Wise Text Segmentation Approach Using BERT. In Llads, J., Lopresti, D., & Uchida, S (Eds.), *Document Analysis and Recognition--ICDAR 2021*, (pp. 236-250). Springer. https://doi.org/10.1007/978-3-030-86337-1_16

Ponceleon, D., & Srinivasan, S. (2001). Automatic discovery of salient segments in imperfect speech transcripts. *Proceedings of the tenth international conference on Information and knowledge management*, 490-497. The ACM Digital Library. https://doi.org/10.1145/502585.502668

Precision_and_recall. (2022). Retrieved from wikipedia: https://en.wikipedia.org/wiki/Precision_and_recall?oldformat=true

Reimers, N., & Gurevyvh, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks.* arXiv. https://doi.org/10.48550/arXiv.1908.10084

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *IEEE conference on computer vision and pattern recognition (CVPR)* (pp.815-823). IEEE. https://doi.org/10.1109/CVPR.2015.7298682

Shah, R. R., Yu, Y., Skaikh, A. D., & Zimmermann, R. (2015). TRACE: linguistic-based approach for automatic lecture video segmentation leveraging Wikipedia texts. *2015 IEEE International Symposium on Multimedia (ISM)* (pp. 217-220). IEEE. https://doi.org/10.1109/ISM.2015.18

Soares, E. R., & Barrére, E. (2019). An optimization model for temporal video lecture segmentation using word2vec and acoustic features. *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, 513-520. The ACM Digital Library. https://doi.org/10.1145/3323503.3349548

Solbiati, A., Heffernan, K., Damaskinos, G., Poddar, S., Modi, S., & Cali, J. (2021). *Unsupervised topic segmentation of meetings with BERT embeddings*. arXiv. https://doi.org/10.48550/arXiv.2106.12978

Glavas, G., & Somasundaran, S. (2020). Two-level transformer and auxiliary coherence modeling for improved text segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 7797-7804. https://doi.org/10.1609/aaai.v34i05.6284

Text_segmentation. (2011). Retrieved from wikipedia: https://en.wikipedia.org/wiki/Text_segmentation

Ugur Akinci, G. K. (2012). *Writing Transition Phrases and Sentences: 12 Types of Sentence and Paragraph Transitions with 112 Examples.*

University, UAH. (n.d.). *WRITING EFFECTIVE TRANSITIONS.* Retrieved from https://www.uah.edu/images/administrative/student-success-center/resources/handouts/handouts_2019/writing_effective_transitions.pdf

Wang, Y., Li, S., & Yang, J. (2018). *Toward fast and accurate neural discourse segmentation*. arXiv. https://doi.org/10.48550/arXiv.1808.09147