# Content Significance Distribution of Sub-Text Blocks in Articles and Its Application to Article-Organization Assessment

You Zhou[a] and Jie Wang[b]

*Richard Miner School of Computer & Information Sciences, University of Massachusetts, Lowell, MA 01854, U.S.A.*

Keywords: Content Significance Distribution, Embedding Similarity, Article Structure, Beta Distribution, Article-Organization Assessment.

Abstract: We explore how to capture the significance of a sub-text block in an article and how it may be used for text mining tasks. A sub-text block is a sub-sequence of sentences in the article. We formulate the notion of content significance distribution (CSD) of sub-text blocks, referred to as CSD of the first kind and denoted by CSD-1. In particular, we leverage Hugging Face's SentenceTransformer to generate contextual sentence embeddings, and use MoverScore over text embeddings to measure how similar a sub-text block is to the entire text. To overcome the exponential blowup on the number of sub-text blocks, we present an approximation algorithm and show that the approximated CSD-1 is almost identical to the exact CSD-1. Under this approximation, we show that the average and median CSD-1's for news, scholarly research, argument, and narrative articles share the same pattern. We also show that under a certain linear transformation, the complement of the cumulative distribution function of the beta distribution with certain values of $\alpha$ and $\beta$ resembles a CSD-1 curve. We then use CSD-1's to extract linguistic features to train an SVC classifier for assessing how well an article is organized. Through experiments, we show that this method achieves high accuracy for assessing student essays. Moreover, we study CSD of sentence locations, referred to as CSD of the second kind and denoted by CSD-2, and show that average CSD-2's for different types of articles possess distinctive patterns, which either conform common perceptions of article structures or provide rectification with minor deviation.



## 1 INTRODUCTION

In articles crafted by skilled writers, certain sentence positions hold greater significance compared to other positions, as do certain sub-sequences of sentences. A prime example is news articles, where the sentences positioned towards the beginning tend to carry greater significance than those towards the end, resulting in an inverted-pyramid-like structure for content significance. In linguistics, article structures are qualitatively classified based on content-significance distributions. Some are classified into self-explanatory geometric shapes, including inverted pyramid, hourglass, diamond, and Christmas tree. Other classifications include narrative, five-boxes, and organic. The narrative presents a straightforward, chronological account of events. For information about five-box and organic structures, the reader is referred to Saleh's guide to article writing (Saleh, 2014).

These qualitative classifications serve as a rule of thumb for writers to organize various types of articles and for readers to grab the significant content. However, content significance in an article has not been quantitatively studied, and its full potentials beyond qualitative classification of article structures is yet to be unlocked.

In a recent study on ranking sentences in an article, we note that an ad hoc location weight is assigned to a sentence to reflect its significance based on intuitive judgments specific to the given type of articles (Zhang and Wang, 2021; Zhang et al., 2021). Despite the rudimentary simplicity of this approach, feature analysis in this study demonstrates that such weights do play a significant role in sentence ranking.

Motivated by this result, we explore how qualitative classifications of article structures may be turned into quantitative descriptions for carrying out certain text mining tasks. In particular, we explore content significance distribution of sub-text blocks in an article that leads to the formulation of CSD of the first kind. For this notion to be useful in practice, we need to deal with the exponential blowup of the number

[a] https://orcid.org/0009-0005-0919-5793
[b] https://orcid.org/0000-0003-1483-2783



Zhou, Y. and Wang, J.
Content Significance Distribution of Sub-Text Blocks in Articles and Its Application to Article-Organization Assessment.
DOI: 10.5220/0012232600003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 121-131
ISBN: 978-989-758-671-2; ISSN: 2184-3228

of sub-text blocks, for it is time consuming to compute the exact CSD-1 for a long article. To overcome this obstacle, we devise an approximation method to compute CSD-1 over a moderate number of text blocks chosen independently at random. We show that, through experiments, the approximated CSD-1 is almost identified to the exact CSD-1.

We investigate four common types of articles: argument, narrative, news, and scholarly research. To do so, we form four datasets using existing datasets, one for each type of articles. We show that the average and median CSD-1 for each type of articles share the same pattern. Moreover, we show that a CSD-1 can be resembled as a linear transformation of the complement of the cumulative distribution function of the beta distribution of parameters $\alpha$ and $\beta$ with $0 < \alpha < 1$ and $0 < \beta < 1$.

To demonstrate the usefulness of CSD-1, we apply CSD-1 to intrinsically determining if an article is well written. In particular, we use CSD-1's to extract features and train an SVC classifier using these features to assess intrinsically how well an article is organized. Experiment results show that this method achieves high accuracy for student essays. article organization.

Next, we investigate CSD of the second kind based on sentence locations in an article. Unlike computing CSD-1 that incurs exponential running time, exact CSD-2 can be computed efficiently. We show that the average and median CSD-2's for each type of articles are close to each other, with distinctive patterns for different types of articles, which conform common perceptions of the structures for news and narrative articles, and rectify with minor deviations an earlier perception of the structures for scholarly research and argument articles in the study of sentence ranking (Zhang et al., 2021).

The rest of the paper is organized as follows: We describe related work on automatic assessment of article qualities in Section 2. We present in Section 3 MoverScore and the datasets. In Section 4 we define CSD of the first kind, describe an approximation algorithm to compute it, and show that CSD-1's for different types of human-written articles all share the same pattern. We also discuss CSD-1 for articles formed by random sentences. In Section 5 we show how to resemble a CSD-1 using the beta distribution with $0 < \alpha < 1$ and $0 < \beta < 1$ by the complement of its cumulative distribution function under a linear transformation. In Section 6 we show how to use CSD-1 to intrinsically assess article organization. In Section 7 we define and discuss CSD of the second kind, denoted by CSD-2. We conclude the paper in Section 8 with remarks and suggestions for further investigations.

## 2 RELATED WORK

To the best of our knowledge, we are not aware of any prior work on quantitative investigations of content significance distribution of sub-text blocks in an article.

Automatic assessment of article qualities, on the other hand, has attracted attention in recent years. The quality of an article is determined by a number of factors, including grammaticality, readability, stylistic attributes, and the depth of expertise presented. Textual analysis can be carried out at the word level, such as identifying verb formation errors, calculating average word frequency, and determining average word length, and at the sentence level. Word-level features have been used to measure word usages and lexical complexity in assessing essays (Attali and Burstein, 2006). Incorporating sentence-level features for assessing article qualities provides a fruitful approach (Cummins et al., 2016).

In an attempt to address the complexity of assessing article qualities, Yang et al (Yang et al., 2018) introduced a modularized hierarchical convolutional neural net, where individual sections of an article are treated as separate modules, with an attentive pooling layer applied to the concatenated representations of these sections, which are fed into a softmax layer for evaluation.

To investigate visual effects of an article such as font choices, the layout of the article, and images included in the article, a multimodal approach (Shen et al., 2020) was presented to capture implicit quality indicators that extend beyond the textual content of an article. The integration of these visual aspects with the textual content enhances the effectiveness of article quality assessment.

The concept of text coherence also plays a pivotal role. For example, a hierarchical coherence model was introduced (Liao et al., 2021), which seeks to leverage local coherence within sentences as well as broader contextual relationships and diverse rhetorical connections. This approach transcends the conventional assessment of sentence similarity, incorporating a richer understanding of the article's coherence.

# 3 SIMILARITY MEASURES AND DATASETS

It is critical to use an appropriate metric to measure how similar semantically a sub-text block is to the entire text. Traditional token-based metrics for measuring text similarity, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and Jaccard's coefficients (Jaccard, 1912), fail to capture similarities between texts in lexical forms that convey the same or similar meanings. To overcome this limitation, we would need a metric on semantic similarity.

## 3.1 MoverScore

We propose to use MoverScore (Zhao et al., 2019) to measure semantic similarities between two texts. In our case, one text is a sub-text block and the other is the entire article. MoverScore uses Earth Mover's Distance (EMD) (Levina and Bickel, 2001) to compute the distance between the contextual embeddings of the two texts to be compared with, where contextual embeddings may be computed by ELMo (Peters et al., 1802), BERT (Devlin et al., 2018), or some other transformers. We use HFace's SentenceTransformer ((Reimers and Gurevych, 2019)) to generate text embeddings. The resulting EMD distance score measures the semantic similarity of the two texts. In so doing, MoverScore provides many-to-one soft alignments to map the candidate word into several most related reference words, producing a more accurate assessment of the semantic similarities between two texts that is more aligned with human judgment.

## 3.2 Datasets

We form four datasets, one for each type of articles, using the following existing datasets (all in English): SummBank (Radev, 2003), Argument Annotated Essays (AAE) (Stab and Gurevych, 2014), Predicting Effective Arguments (PEA) (PEA, ), BookSum (Kryściński et al., 2021), and research papers from arXiv.

- NewsA for news articles. NewsA is a dataset of 200 news articles selected independently at random from SummBank, with on average of 23 sentences in an article. SummBank is a large collection of news articles with sentence rankings annotated by three judges.
- ArguE for argument articles. ArguE is a dataset of 200 essays selected independently at random from the union of AAE and PEA, with an average of 21 sentences in an article. AAE consists of persuasive essays written by students for preparing for standardized tests, and PEA consists of argument essays written by students of grades 6 – 12, annotated by experts for discourse elements in argumentative writing.
- NarrC for narrative articles. NarrC is a dataset of 200 documents selected independently at random from BookSum. BookSum is a large collection of long-form narrative summaries from novels, plays, and stories with a large number of chapter-level documents which covers source documents from the literature domain, such as novels, plays and stories. We randomly select 200 chapters from the datasets which contains 255 sentences each in a chapter.
- SchRP for scholary research papers. SchRP is a dataset of 200 papers selected independently at random from arXiv.org in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics, with approximately an equal number of articles in each subject with an average of 210 sentences in each paper.

# 4 CSD OF THE FIRST KIND

Let $A = \langle S_1, S_2, \ldots, S_n \rangle$ be an article, represented as a sequence of sentences, where $S_i$ is the $i$th sentence. Let $k$ be an integer between 1 and $n$. There are $N = \binom{n}{k}$ many sub-text blocks consisting of $k$ sentences. The number of sentences in a text block is also referred to as the size. For example, let $n = 10$ and $k = \lfloor 0.3n \rfloor = 3$. Then there are $\binom{10}{3} = 120$ sub-text blocks of size 3, listed in lexicographical order as follows: $\langle S_1, S_2, S_3 \rangle, \langle S_1, S_3, S_4 \rangle, \cdots, \langle S_8, S_9, S_{10} \rangle$.

Let $\mathrm{MSc}(X,Y)$ represent the MoverScore of text $X$ and text $Y$. Let $T_1$ and $T_2$ be two text block. We say that $T_1 > T_2$ if either $\mathrm{MSc}(T_1, A) > \mathrm{MSc}(T_2, A)$ or $\mathrm{MSc}(T_1, A) = \mathrm{MSc}(T_2, A)$ and $T_1$ proceeds $T_2$ in lexicographical order. Sort the $N$ sub-text blocks of size $k$ in descending order according to this ordering and let $T_{k,j}$ be the $j$th sub-text block in the sorted list. That is, $T_{k,1} > T_{k,2} > \cdots > T_{k,N}$. Then the CSD-1 for $A$ with size $k$ is a discrete function over $x_j = j/N \in [0,1]$ with $1 \leq j \leq N$, defined by

$$\text{CSD-1}(A, k, x_j) = \mathrm{MSc}(T_{k,j}, A).$$

We are particularly interested in selecting 30% of sentences to form a sub-text block because the previous results indicate that selecting 30% of sentences

Table 1: Comparisons between approximated CSD-1 and exact CSD-1 over NewsA.

| x-value | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c = 0.1 | exact | 0.909 | 0.896 | 0.861 | 0.843 | 0.829 | 0.817 | 0.802 | 0.791 | 0.782 | 0.771 | 0.759 |
| | appr. | 0.903 | 0.896 | 0.861 | 0.843 | 0.829 | 0.817 | 0.802 | 0.791 | 0.782 | 0.771 | 0.754 |
| c = 0.3 | exact | 0.928 | 0.916 | 0.894 | 0.878 | 0.866 | 0.858 | 0.850 | 0.842 | 0.835 | 0.826 | 0.818 |
| | appr. | 0.925 | 0.916 | 0.894 | 0.878 | 0.866 | 0.858 | 0.850 | 0.842 | 0.835 | 0.826 | 0.814 |
| c = 0.5 | exact | 0.948 | 0.938 | 0.923 | 0.914 | 0.907 | 0.902 | 0.898 | 0.892 | 0.887 | 0.883 | 0.877 |
| | appr. | 0.946 | 0.938 | 0.923 | 0.914 | 0.907 | 0.902 | 0.898 | 0.892 | 0.887 | 0.883 | 0.872 |
| c = 0.7 | exact | 0.953 | 0.952 | 0.946 | 0.944 | 0.941 | 0.939 | 0.936 | 0.933 | 0.930 | 0.927 | 0.924 |
| | appr. | 0.950 | 0.952 | 0.946 | 0.944 | 0.941 | 0.939 | 0.936 | 0.933 | 0.930 | 0.927 | 0.922 |

(a) Average CSD-1 over NewsA.

| x-value | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c = 0.1 | exact | 0.897 | 0.865 | 0.810 | 0.784 | 0.763 | 0.750 | 0.737 | 0.724 | 0.710 | 0.697 | 0.686 |
| | appr. | 0.892 | 0.865 | 0.810 | 0.784 | 0.763 | 0.750 | 0.737 | 0.724 | 0.710 | 0.697 | 0.679 |
| c = 0.3 | exact | 0.922 | 0.891 | 0.847 | 0.823 | 0.805 | 0.791 | 0.777 | 0.768 | 0.755 | 0.744 | 0.734 |
| | appr. | 0.918 | 0.891 | 0.847 | 0.823 | 0.805 | 0.791 | 0.777 | 0.768 | 0.755 | 0.744 | 0.727 |
| c = 0.5 | exact | 0.962 | 0.924 | 0.888 | 0.870 | 0.858 | 0.848 | 0.840 | 0.832 | 0.825 | 0.817 | 0.809 |
| | appr. | 0.959 | 0.924 | 0.888 | 0.870 | 0.858 | 0.848 | 0.840 | 0.832 | 0.825 | 0.817 | 0.804 |
| c = 0.7 | exact | 0.972 | 0.970 | 0.956 | 0.947 | 0.938 | 0.932 | 0.926 | 0.918 | 0.912 | 0.906 | 0.897 |
| | appr. | 0.970 | 0.970 | 0.956 | 0.947 | 0.938 | 0.932 | 0.926 | 0.918 | 0.912 | 0.906 | 0.895 |

(b) CSD-1 over a sample news article.

appropriately from an article would typically capture the major points of the article (see, for example, (Zhang and Wang, 2021)).

Note that $\binom{n}{k} \geq (n/k)^k$. For $k = \lfloor 0.3n \rfloor$, we have $(n/k)^k > 3^{0.3n}$, resulting in an exponential blowup. When $n$ is large, computing CSD-1 is intractable. For example, recall that in SchRP of scholarly research papers, the average number of sentences in each article is 210. We have $\binom{n}{k} = \binom{210}{63} > 3 \times 10^{54}$, which is much too big for any computer to handle. Approximation is therefore needed.

Table 2: Average and median approximated CSD-1 for articles of each type, where the x-values are the normalized sequence of text blocks in ascending order of MoverScores compared with the article itself, and the values in the table are the MoverScores.

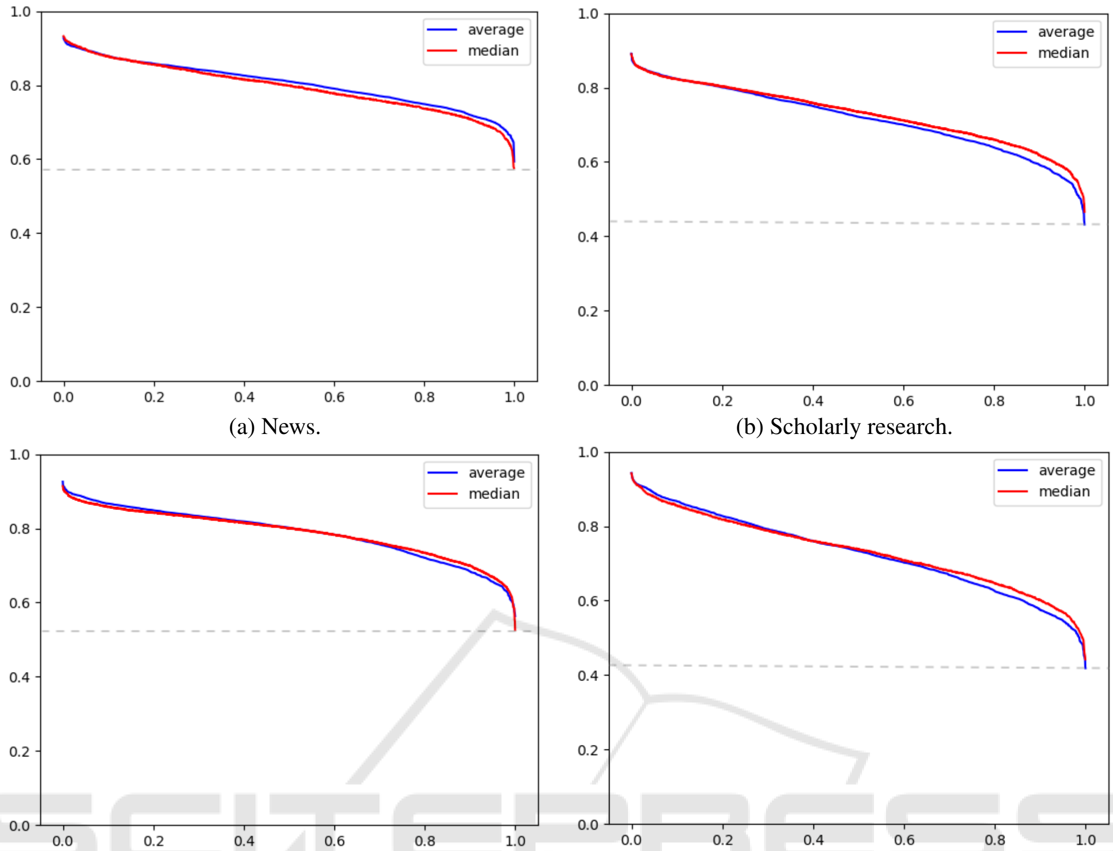
| x-value | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| News | avg | 0.93 | 0.92 | 0.89 | 0.88 | 0.87 | 0.86 | 0.85 | 0.84 | 0.83 | 0.83 | 0.82 |
| | med | 0.93 | 0.92 | 0.90 | 0.88 | 0.87 | 0.86 | 0.85 | 0.84 | 0.83 | 0.81 | 0.81 |
| SchR | avg | 0.94 | 0.92 | 0.89 | 0.87 | 0.85 | 0.83 | 0.81 | 0.79 | 0.78 | 0.76 | 0.75 |
| | med | 0.94 | 0.92 | 0.88 | 0.85 | 0.84 | 0.82 | 0.80 | 0.79 | 0.78 | 0.76 | 0.75 |
| Argu | avg | 0.89 | 0.87 | 0.84 | 0.82 | 0.81 | 0.80 | 0.79 | 0.77 | 0.76 | 0.75 | 0.74 |
| | med | 0.89 | 0.87 | 0.84 | 0.82 | 0.81 | 0.81 | 0.79 | 0.78 | 0.77 | 0.76 | 0.75 |
| Narr | avg | 0.93 | 0.91 | 0.88 | 0.87 | 0.86 | 0.85 | 0.84 | 0.83 | 0.83 | 0.82 | 0.81 |
| | med | 0.92 | 0.90 | 0.87 | 0.86 | 0.85 | 0.84 | 0.84 | 0.83 | 0.82 | 0.82 | 0.81 |

### 4.1 Approximating CSD-1

Approximating CSD-1 for each article proceeds as follows:

1. Generate independently at random 5,000 text blocks.
2. Cluster sentences in $A$ using Affinity Propagation (Frey and Dueck, 2007) based on sentence embeddings generated by SentenceTransformer. Let $C_1, C_2, \ldots, C_m$ be resulted clusters, where $m$ is determined by the clustering algorithm. Let $n_i$ be the number of sentences in cluster $C_i$ ($i = 1, 2, \ldots, m$). Select $\lfloor 0.3n \rfloor \times n_i/n$ sentences from cluster $C_i$ to form a text block and randomly select 5,000 such text blocks.

   The objective of this step is to avoid pure random sampling and try to cover the exact curve as much as possible with limited samples. We achieve this by selecting text blocks at random from different topics just in case the text blocks selected in Step 1 have missed certain topics.
3. Combine the 5,000 text blocks generated in Step 1 and the 5,000 text blocks generated in Step 2. Use these 10,000 text blocks to compute CSD-1 as an approximation to the exact CSD-1 for $A$.




(a) News. (b) Scholarly research.

(c) Argument. (d) Narrative.

Figure 1: Average and median approximated CSD-1 for articles of each type, where the x-axis is the normalized sequence of text blocks in ascending order of MoverScores compared with the article itself, and the y-axis is the MoverScores.



Extensive experiments indicate that CSD-1 under this approximation is almost identical to the exact CSD-1 for various sizes. For example, Table 1 (a) shows the average approximated CSD-1 and the exact CSD-1 for articles in NewsA, while Table 1 (b) shows the approximated CSD-1 and the exact CSD-1 for a random sample from NewsA, where $c$ is a fractional to determine the number $k$ of sentences in a text block with $k = \lfloor cn \rfloor$. Thus, it is sufficient to use approximated CSD-1 to replace exact CSD-1.

Table 2 depicts the average and median approximated CSD-1's with $c = 0.3$ over each dataset, where SchR, Argu, and Narr represent, respectively, Scholarly Research, Argument, and Narrative, while avg and med represent, respectively, average and median. Figure 1 depicts the corresponding curves.

## 4.2 Segments of CSD-1

It is evident that the average approximated CSD-1 and the median approximated CSD-1 for each type of articles are very close to each other, and they all share the same pattern. This pattern can be divided into three segments, referred to as, from left to right, *the left segment* (L-segment), *the middle segment* (M-segment), and *the right segment* (R-segment). The L-segment contains a very small number of text blocks that are substantially more significant than the rest, with the largest value close to 0.9. The union of these text blocks is the most significant content of the article. The R-segment contains a very small number of text blocks with the least significance, with value below 0.6 and above 0.4. These text blocks often contain connection sentences. The M-segment is the majority of text blocks that are gradually decreasing in terms of significance. Figure 2 depicts the three segments of CSD-1.

## 4.3 CSD-1 for Randomly Generated Articles

To establish a baseline for CSD-1 on articles written by under-educated writers, we generate articles by selecting at random unrelated sentences from 20 ex-

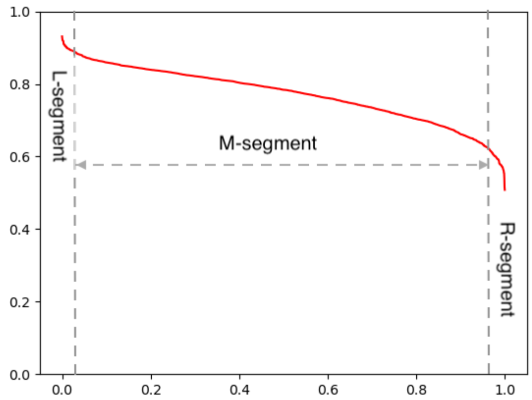



Figure 2: CSD-1 segments.

isting articles, one sentence per article, and placing them in a random order. We call such articles "Random sentences". Existing articles are selected from SummBank in one setting, and Wikipedia (wik, ) in another setting.

We also form articles using 20 identical sentences and 20 sentences with high embedding similarity. For the former, each CSD-1 is simply a straight line. For the latter, we generate similar sentences from a given sentence as follows: Select at random one or two words and replace them with synonyms to form 20 new sentences so that their pairwise embedding similarity is greater than 0.9. Figure 3 depicts the average CSD-1 on random-sentence articles and similar-sentence articles. It can be seen that the average CSD-1 for random-sentence articles exhibits a much larger y-value range from nearly zero to below 0.7.

# 5 TRANSFORMING BETA DISTRIBUTION TO CSD-1

We show that a CSD-1 curve can be resembled using the beta distribution for certain values of parameters $\alpha$ and $\beta$ under a certain linear transformation. In particular, we present the following observations:

1. A typical CSD-1 curve resembles the complement of a cumulative distribution function (CDF) of a U-shape probabilistic density function (PDF).
2. The beta distribution, denoted by $\text{Beta}(\alpha, \beta)$, provides a U-shape PDF with $0 < \alpha < 1$ and $0 < \beta < 1$.

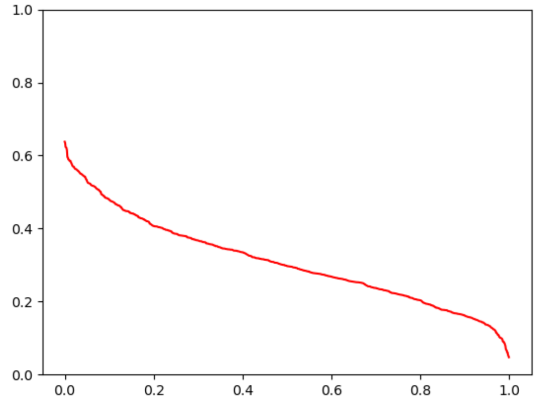
Denote by $I_x(\alpha, \beta)$ the CDF of $\text{Beta}(\alpha, \beta)$. Let

$$C_x(\alpha, \beta) = 1 - I_x(\alpha, \beta).$$

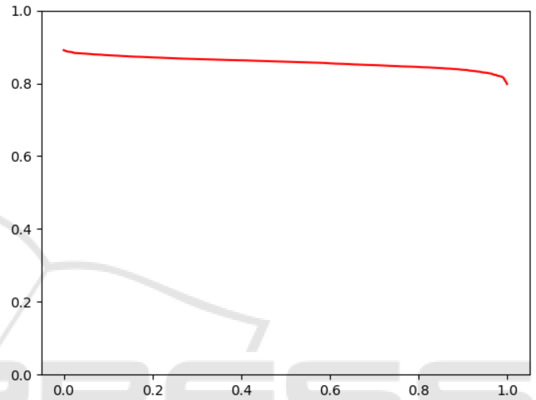
Figure 4 shows the curve of $C_x(0.4, 0.3)$.

We apply a linear transformation to obtain a CSD-1 curve in the desired range. Let

$$LC_x(a, b \mid \alpha, \beta) = a \cdot C_x(\alpha, \beta) + b,$$




a) Random sentences.




b) Similar sentences.

Figure 3: Average CSD-1's with $c = 0.3$ for (a) articles of random sentences and (b) articles of similar sentences, with the same x-axis and y-axis as those in Figure 1.

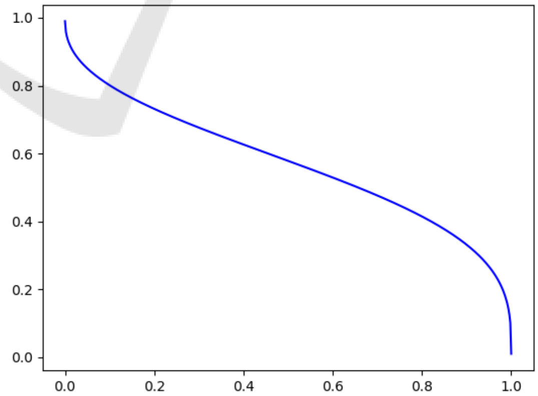
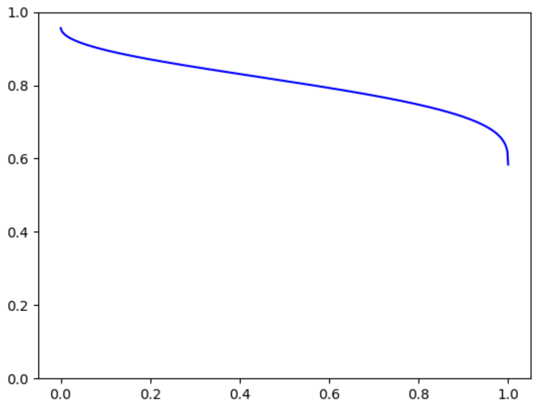



Figure 4: The curve of $C_x(0.4, 0.3)$, which spans the entire y-axis from 0 to 1.

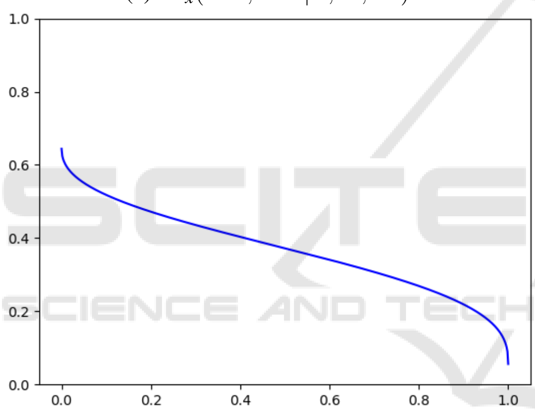
where $a \geq 0$, $b \geq 0$, and $a + b \leq 1$. For easier reading, we may also write $LC_x(a = a_0, b = b_0 \mid \alpha = \alpha_0, \beta = \beta_0)$ as $LC_x(a_0, b_0 \mid \alpha_0, \beta_0)$.

For example, Figure 5(a) depicts the curve of $LC_x(a = 0.38, b = 0.55 \mid \alpha = 0, 45, \beta = 0.3)$, which resembles the curves of Figure 1(a) for NewsA. Figure

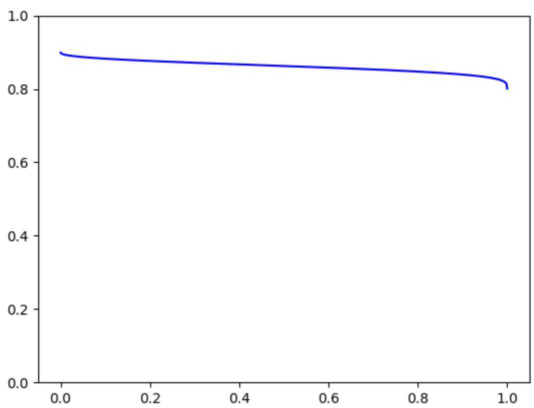
5(b) depicts the curve of $LC_x(a = 0.6, b = 0.05 \mid \alpha = 0,4, \beta = 0.35)$, which resembles the curves of Figure 3(a) for random sentences. Figure 5(c) depicts the curve of $LC_x(a = 0.1, b = 0.8 \mid \alpha = 0,4, \beta = 0.25)$, which resembles the curves of Figure 3(b) for similar sentences.



(a) $LC_x(0.38, 0.55 \mid 0,45, 0.3)$.



(b) $LC_x(a = 0.6, b = 0.05 \mid \alpha = 0,4, \beta = 0.35)$.



(c) $LC_x(a = 0.1, b = 0.8 \mid \alpha = 0,4, \beta = 0.25)$.

Figure 5: The curves resemble, respectively, (a) the average and median curves for news articles in Figure 1(a), (b) the average curve for random sentences in Figure 3(a), and (c) the average curve for similar sentences in Figure 3(b).

We observe that $\alpha$ controls how the L-segment of the CSD-1 curve looks like and $\beta$ controls how the R-segment of the CSD-1 curve looks like. In particular,

1. If $\alpha$ is smaller, then the L-segment is larger both vertically and horizontally.
2. If $\beta$ is smaller, then the R-segment is larger both vertically and horizontally.

We conjecture that by choosing the values of $\alpha$, $\beta$, $a$, and $b$ appropriately, we can resemble CSD-1 for any type of article using $LC_x(a, b \mid \alpha, \beta)$.

## 6 ASSESSING ARTICLE ORGANIZATION

As an application of CSD-1, we show how to assess intrinsically how well an article is organized using features extracted from multiple CSD-1's with various sizes. For this purpose, we use the ASAP (the Automated Student Assessment Prize ) dataset (Stab and Gurevych, 2014), for it contains subsets suitable for carrying out this task. It consists of eight sets of essays, written by students and scored by experienced graders on idea, organization, style, convention, sentence fluency, and word choice.

In particular, we choose Set 7 and Set 8, for they provide an organization score for each essay, assessing whether an essay is organized in a way that enhances the central ideas and its development, focusing on whether the sentence orders are compelling and move the reader through the text easily, and the connections between ideas and events are clear and logically sequenced.

Specifically, Set 7 consists of 1,730 essays with an average of 250 words, scored by two graders with scores from 0 to 3. Averaging the two scores for each essay yields five labels from 1 to 3 with an increment of 0.5, one essay one score. Set 8 consists of 918 essays with an average of 650 words, scored by two or three graders with scores from 1 to 6. Averaging scores for an essay may result in slight discrepancy depending on whether it has two or three scores. For example, an average score could be 3.5 with 2 scores or 3.67 with 3 scores for different papers of about the same quality. We combine such scores into one score, yielding eleven labels from 1 to 6 with an increment of 0.5.

Neither set provides sufficient data to train neural-net classifiers. Instead, we train SVC classifiers by extracting features from approximated CSD-1 of various sizes as percentages of $n$, the number of sentences in an article. For example, we may divide $n$ by 10 or by 5 to yield nine or 19 different sizes. Figure 6 shows approximated CSD-1's of text blocks of three different sizes of a single document in Set 7, which are of essentially the same shape with the one of a larger size above the one of a smaller size.

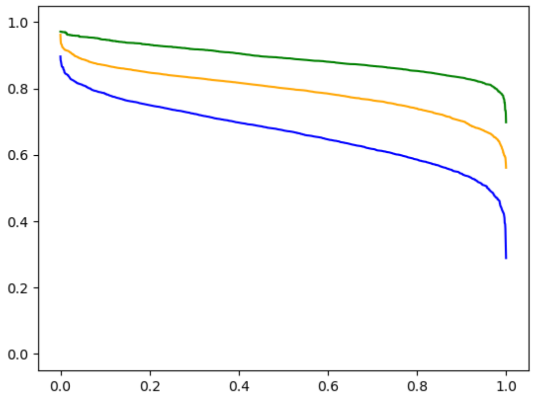



Figure 6: Approximated CSD-1's with sizes determined by $c_i = 0.2 + 3(i-1)/10$ for $i = 1, 2, 3$.

However, even with 10,000 text blocks to obtain an approximated CSD-1, the computation is still too high, which hinders our application of assessing essay organization in real time. To overcome this obstacle, we sample $N$ text blocks uniformly and independently at random to further approximate CSD-1's on each size. Through intensive and extensive experiments, we find that choosing $N = 1,000$ and dividing $n$ by 10 offer a satisfactory trade-off between accuracy and time complexity for our applications. Note that selecting a larger value of $N$ and dividing $n$ by a number smaller than 10, while providing a better accuracy, incurs a tremendous time complexity required to calculate CSD-1's for 9 times from 10% to 90% for each article, making it unpractical. This yields, for each article, nine vectors of 1,000 dimensions from the corresponding CSD-1. We then select 10 equal positions from a 1,000d vector to produce, for each article, nine 10d vectors.

We train a multi-label SVC classifier for Set 7 based on the given training set using the 10d feature vectors with a standard 80-20 split, and do the same for Set 8.

For a test article, we call the predicted result "exact", "adequate", and "acceptable" if, respectively, the predicted label by SVC is identical to its label, $\pm 0.5$ of the exact, and $\pm 1$ of the exact. It is common for experienced graders to have small dependencies among them, with $\pm 0.5$ been adequate and $\pm 1$ acceptable. Table 4 shows the binary F1 scores for "exact", "adequate", and "acceptable".

In developing a practice application, for datasets with a smaller number of labels, such as Set 7 with 5 labels, we may hesitate to consider $\pm 1$ of the exact

Table 3: F1 scores of predicted labels for test articles.

| | Exact (%) | $\pm 0.5$ (%) | $\pm 1$ (%) |
|---|---|---|---|
| Set 7 | 71.26 | 85.21 | 99.08 |
| Set 8 | 67.19 | 76.66 | 98.01 |

acceptable, but it is reasonable to consider $\pm 1$ of the exact acceptable for datasets with a larger number of labels, such as Set 8 with 11 labels. Thus, we would want to adopt the 6-point system to achieve over 98% accuracy of acceptable intrinsic evaluation of article organization.

## 7 CSD OF THE SECOND KIND

We define CSD-2 to reflect the significance of sentence locations. The idea is to identify where the top 30% of sentences with the highest MoverScores are located. The reason of choosing 30% is the same as that in Section 4.

Given an article $A = \langle S_1, S_2, \ldots, S_n \rangle$ of a certain type with $S_i$ being the $i$th sentence for $i = 1, 2, \ldots, n$, select the top $t = \lfloor 0.3n \rfloor$ sentences that are most similar to $A$ as key sentences determined by MoverScores. Let these sentences be $S_{i_1}, S_{i_2}, \ldots, S_{i_t}$ with $i_1 < i_2 < \cdots < i_t$. We normalize the location index of each of these sentences in $A$ by $n$ and define CSD-2 as the following discrete function:

$$\text{CSD-2}(A, i/n) = \begin{cases} \text{MSc}(S_{i_j}, A), & \text{if } i = i_j, \\ 0, & \text{otherwise.} \end{cases}$$

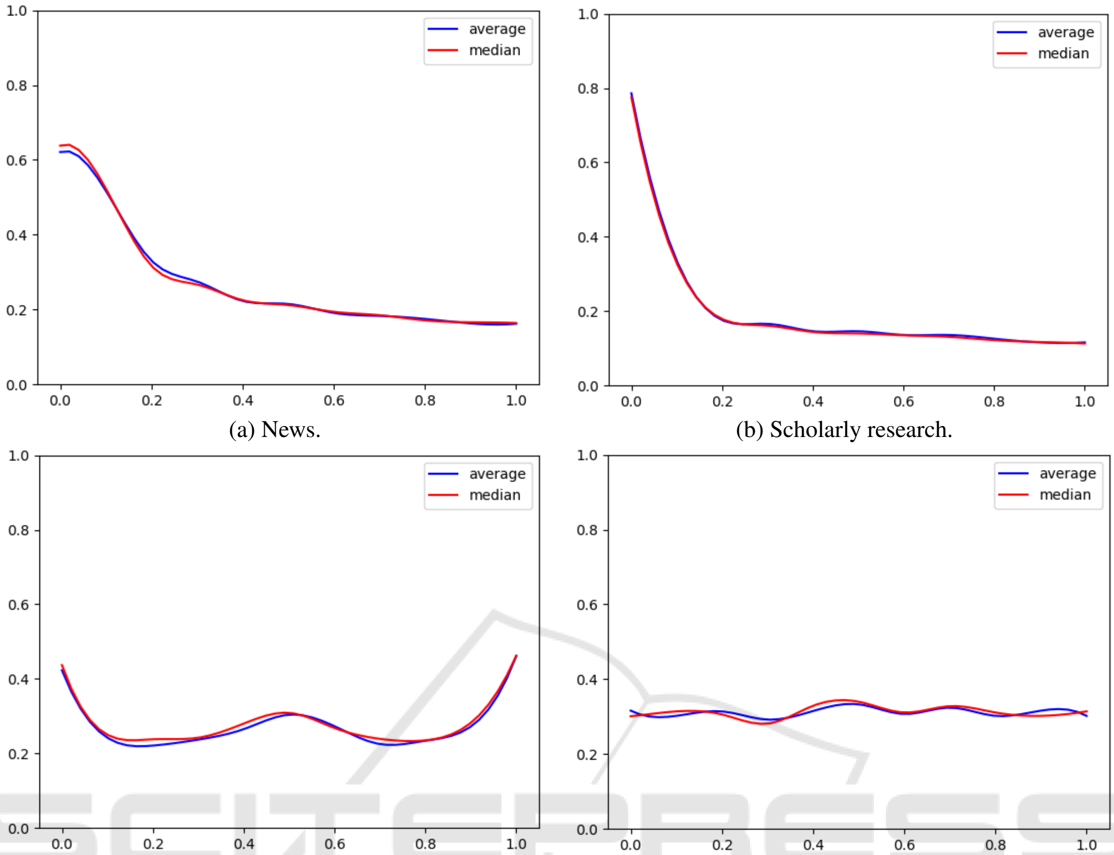
We compute the average and median CSD-2's for each type of articles over, respectively, the ArguE, NewsA, NarrC, and SchRP datasets. Table 4 depicts the average and median CSD-2 for each dataset, and Figure 7 depicts the corresponding CSD-2 curves.

Table 4: Average and median CSD-2 for each dataset.

| x-value | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| News | avg | 0.62 | 0.52 | 0.33 | 0.28 | 0.22 | 0.22 | 0.19 | 0.18 | 0.18 | 0.16 | 0.16 |
| | med | 0.64 | 0.52 | 0.32 | 0.27 | 0.23 | 0.21 | 0.19 | 0.19 | 0.17 | 0.17 | 0.16 |
| SchR | avg | 0.79 | 0.34 | 0.18 | 0.17 | 0.15 | 0.15 | 0.14 | 0.14 | 0.13 | 0.12 | 0.12 |
| | med | 0.78 | 0.33 | 0.18 | 0.16 | 0.14 | 0.14 | 0.14 | 0.13 | 0.12 | 0.12 | 0.11 |
| Argu | avg | 0.42 | 0.24 | 0.22 | 0.24 | 0.27 | 0.30 | 0.27 | 0.23 | 0.23 | 0.27 | 0.46 |
| | med | 0.44 | 0.25 | 0.24 | 0.24 | 0.28 | 0.31 | 0.27 | 0.24 | 0.24 | 0.28 | 0.46 |
| Narr | avg | 0.32 | 0.30 | 0.31 | 0.29 | 0.32 | 0.33 | 0.31 | 0.32 | 0.30 | 0.32 | 0.30 |
| | med | 0.30 | 0.31 | 0.31 | 0.28 | 0.33 | 0.34 | 0.31 | 0.33 | 0.31 | 0.30 | 0.31 |

The following observations are evident.

1. For each dataset, the median CSD-2 is very closed to the average CSD-2, indicating that the average CSD-2 curve provides a good representative for each dataset.
2. For news articles, the average CSD-2 is monotonically decreasing. This is in line with an inverted




(a) News. (b) Scholarly research.

(c) Argument. (d) Narrative.

Figure 7: Average and median CSD-2 for articles of each type, where the x-axis is the normalized sentence indexes and the y-axis is the MoverScores of sentences and the article they are in.



pyramid in the common perception for the structure of news articles. The small bumps over the x-interval $(0.2, 0.6)$ indicate that the structure of a news article may not be a straight inverted pyramid.

3. For scholarly research articles, the average CSD-2 is also monotonically decreasing. Different from the average CSD-2 for news articles, we note that the values over the x-interval $(0, 0.2)$ drops rapidly from about 0.8 to below 0.2. It decreases slowly over the x-interval $(0.2, 1]$. This indicates that in a scholarly research paper, the first 20% of sentences would be the most significant, and the rest of the article would be justifications of these statements. This differs from an earlier perception that presumes an hourglass structure for scholarly research articles (Zhang and Wang, 2021).

4. For argument articles, the average CSD-2 resembles a shallow "W", with two peaks at both ends and one peak in the middle, where the peaks at both ends are higher than the one in the middle. This indicates that the structure of an argument article would, on average, start and end with the most significant arguments that match each other, with other secondary arguments somewhere in the middle. This differs from an earlier perception that presumes a pyramid structure for argument articles (Zhang and Wang, 2021).

5. For narrative articles, the average CSD-2 resembles a shallow wave line centered around a value slightly greater than 0.3. This indicates that the structure of a narrative article would, on average, includes sentences of approximately the same significance throughout the article. This is in line with the common perception.

## 8 CONCLUSIONS AND FINAL REMARKS

We present for the first time a quantitative method to capture content significance distributions of sub-texts within an article and show that it is a promising new approach to unlocking potentials for certain text mining tasks using linguistic knowledge which so far only has qualitative descriptions. In addition to CSD-1 and CSD-2, we believe that CSDs of other kinds may also be possible that are awaiting to be explored.

We have demonstrated how to use CSD-1 to assess article organization with high accuracy, and we believe that other applications may also be possible. For instance, we may use CSD-2 to identify the type of a given article to help obtain more accurate ranking of sentences. Incorporating sentence ranking to a large language model such as GPT-3.5-turbo (Brown et al., 2020), LLaMA (Touvron et al., 2023), and PaLM (Chowdhery et al., 2022) is expected to help generate a better summary for a given article.

Our approach of computing CSDs relies on metrics of comparing semantic similarities of a sub-text block (note that a sentence is a special case of sub-text block) to the article it is in. While MoverScore is arguably the best choice at this time, computing MoverScores incurs a cubic time complexity (Zhao et al., 2019). Fortunately, this task is highly parallelizable and we have implemented a parallel program to carry out this task on a GPU, which provides much more efficient computation of CSD-1. Nevertheless, finding a more effective and efficient measure for content similarity is highly desirable for our tasks, particularly for long articles.

We would also like to seek intuitions and mathematical explanations why the functions $LC_x(a, b \mid \alpha, \beta)$ resemble CSD-1 curves.

Finally, we would like to explore if CSDs may be used to assess the overall quality of an article with a single score with better accuracy than an early attempt (Wang et al., 2022) using a multi-scale essay representation that can be jointly learned, which employs multiple losses and transfer learning from out-of-domain essays.

## ACKNOWLEDGMENT

We would like to thank Jay Belanger for a valuable suggestion on function transformation.

## REFERENCES

Feedback prize - predicting effective arguments. https://www.kaggle.com/competitions/feedback-prize-effectiveness/data. Accessed: 2022.

Wikimedia downloads.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cummins, R., Zhang, M., and Briscoe, T. (2016). Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–799.

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

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11:37—-50.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. (2021). Booksum: A collection of datasets for long-form narrative summarization.

Levina, E. and Bickel, P. (2001). The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE.

Liao, D., Xu, J., Li, G., and Wang, Y. (2021). Hierarchical coherence modeling for document quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13353–13361.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (1802). Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Radev, Dragomir, e. a. (2003). Summbank. Radev, Dragomir, et al. SummBank 1.0 LDC2003T16. Web Download. Philadelphia: Linguistic Data Consortium, 2003.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saleh, N. (2014). *The Complete Guide to Article Writing: How to Write Successful Articles for Online and Print Markets*. Writer's Digest Books; Illustrated edition (January 14, 2014).

Shen, A., Salehi, B., Qi, J., and Baldwin, T. (2020). A multimodal approach to assessing document quality. *Journal of Artificial Intelligence Research*, 68.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, Y., Wang, C., Ruobing, L., and Lin, H. (2022). On the use of BERT for atutomated essay scoring: joint learning of multi-scale essay representation. *https://arxiv.org/pdf/2205.03835v2.pdf*.

Yang, P., Sun, X., Li, W., and Ma, S. (2018). Automatic academic paper rating based on modularized hierarchical convolutional neural network. *arXiv preprint arXiv:1805.03977*.

Zhang, H. and Wang, J. (2021). An unsupervised semantic sentence ranking scheme for text documents. *Integrated Computer-Aided Engineering*, 28:17–33.

Zhang, H., Zhou, Y., and Wang, J. (2021). Contextual networks and unsupervised ranking of sentences. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1126–1131.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.