Computational Analysis of Semantic Connections Between Herman Melville's Reading and Writing

Nudrat Habib[1 [0009−0005−9310−1701]], Steven Olsen-Smith[2 [0009−0004−9213−6956]], and Elisa Barney Smith[1,2 [0000−0003−2039−3844]]

[1] Luleå Technical University, Sweden
{nudrat.habib, elisa.barneysmith}@ltu.se
[2] Boise State University, Idaho, USA
{sosmith,ebarneysmith}@boisestate.edu

Abstract. This study investigates the potential influence of Herman Melville's reading on his own writings through computational semantic similarity analysis. Using documented records of books known to have been owned or read by Melville, we compare selected passages from his works with texts from his library. The methodology involves segmenting texts at both the sentence level and non-overlapping 5-gram level, followed by similarity computation using BERTScore. Rather than applying fixed thresholds to determine reuse, we interpret precision, recall, and F1 scores as indicators of possible semantic alignment that may suggest literary influence. Experimental results demonstrate that the approach successfully captures expert-identified instances of similarity and highlights additional passages warranting further qualitative examination. The findings suggest that semantic similarity methods provide a useful computational framework for supporting source and influence studies in literary scholarship.

Keywords: Semantic similarity · BERTScore · Segmentation · Literary Studies.

1 Introduction

Similarities between literary texts can arise for many different reasons. In some cases they may reflect the influence of earlier authors on later ones, while in others they result from shared ideas, stylistic conventions, or overlapping cultural contexts. Identifying such relationships is a necessary undertaking in literary and computational studies because it helps reveal how authors engage, consciously or unconsciously, with existing works. When dealing with historical and culturally significant documents, however, the sheer quantity of potentially relevant textual data presents serious obstacles to our comprehension of what materials an author might have encountered during their lifetime. Computational and pattern recognition methods make it possible to trace potential textual relationships across large digital corpora, enabling scalable semantic similarity analysis and the discovery of textual patterns within literary and cultural heritage collections. Of

the three most frequently downloaded authors on Project Gutenberg, this study focuses on Herman Melville, a prominent 19th century American author whose materials are widely accessible in digital form. Our investigation examines both the books Melville authored and those he is known to have read. Melville typically autographed his books when he acquired them, often inscribing the date of acquisition along with the city or immediate setting where he recorded the autograph inscription [20]. From the books of his collection that survive today, Melville's inscribed date of acquisition supplies a chronological baseline for dating his reading of books that may have influenced works he went on to write. Even in cases of books from his library that are not known to survive, Melville's practice of referring to book purchases in dated letters and journals can furnish similar evidence of acquisition [20]. When no external or holograph evidence exists to document Melville's ownership or consultation of a potentially influential book, scholars are tasked to demonstrate its influence through methods of text comparison to infer possible relationships between his reading and his writing [2]. Melville's characteristic methods of source appropriation are especially well suited for the study of literary influence in terms both of detectability and of investigative outcome. As Harrison Hayford remarked in his study of Melville's second book Omoo, the passages he derived from sources were "always rewritten and nearly always improved" by the process of appropriation [12]. Yet his methods of borrowing and altering from other books frequently left intact phrasal constructions and wording that make it possible to trace derived passages back to their original sources. This is of particular value in cases where no manuscripts of Melville's works or source books marked and drawn upon by him are known to survive. The combination in Melville's writings of vestigial source text and his creative reworking of that material illuminate his processes of literary composition as well as his ideological departures from his printed sources, exposing depths of authorial intention and varieties of artistic craft that can only be spotted through these investigations [18].

Traditional methods of identifying source influence involve a scholar's close reading and comparison of a created work and source books suspected to have played a role in its composition. While the scholar cannot be replaced, this process can be facilitated by modern computational and pattern recognition tools. We aim to utilize these methods to identify and analyze sources of textual influence in Melville's works and present them to scholars for further interpretation and validation. Computational tools for evaluating textual similarity already exist and are widely used in areas such as author attribution, plagiarism detection, and semantic pattern analysis [10]. We focus on semantic similarity analysis as a means of identifying potential traces of literary influence between Melville's works and books he owned or consulted during the process of composition. By framing the task in terms of automatic text comparison with reference sources, we draw on established computational and pattern recognition techniques while shifting the emphasis from plagiarism detection, typically associated with modern contexts, to the scholarly problem of systematically identifying, analyzing, and evaluating patterns of source influence within literary and cultural heritage

texts. By considering different forms of textual similarity, ranging from word-for-word overlap to paraphrastic resemblance, we argue that semantic similarity measures provide a flexible and powerful approach, particularly when the exact nature of influence is unknown in advance. The research question we address is:

Can semantic similarity analysis identify traces of literary influence in Herman Melville's writings when compared to books he owned or may otherwise have consulted in the course of composition?

Our contributions are as follows

- We position semantic similarity analysis, particularly BERTScore, as a tool for investigating the possible influence on an author's writing.
- We present a feasibility study comparing selected passages from Melville's works with texts from his documented readings, testing whether semantic similarity measures can highlight instances of overlap.
- Our findings provide insight into the strengths and limitations of using BERTScore in this context, laying the groundwork for scaling up to full corpus analyses and comparing our approach with alternative similarity metrics.

The rest of this paper is organized as follows. Section 2 reviews related work on textual similarity, intertextual analysis, and computational approaches to literary influence. Section 3 describes the methodology, including data selection, segmentation strategies, and the use of BERTScore for semantic similarity analysis. Section 4, titled Semantic Similarity Results: Evidence of Possible Influence, presents the experimental results and discussion. Section 5 outlines the limitations of the current study and outlines the directions for future research, and Section 6 concludes this research.

2 Related Work

Text similarity measures how closely two pieces of text match each other. In its extreme it can identify cases of plagiarism and many tools have been developed for this task. A high degree of textual similarity does not mean plagiarism. It can reflect overlapping language or common expressions that require careful human interpretation to assess intent and originality [17]. N-gram analysis has been a fundamental tool in text comparison. A study related to the performance evaluation on number of tokens used in an n-gram was done by [4], concluding that as n increases, the probability of finding commonalities in the documents decreases. They observed the worst results for n-grams above 5. They recommended splitting both the reference and suspicious documents into matching n-grams. [25] further compared variants that removed or retained stop words and explored the use of named-entity n-grams, while [1] highlighted the effect of punctuation and sentence-boundary errors on classification accuracy. These studies emphasize how segmentation and pre-processing choices influence detection outcomes.

Beyond surface matching, researchers have explored statistical and machine-learning approaches to text comparison. [23] employed correlation-based metrics such as overall Pearson and weighted mean for sentence-level similarity scoring. [4] applied Naive Bayes classification and evaluated performance using the F1

score, whereas [1] implemented a Support Vector Machine with lexical, syntactic, and semantic features. The Meteor metric, originating in machine translation, has also been adapted to detect paraphrastic or semantically equivalent passages [1]. Although these studies were primarily designed for plagiarism detection, the underlying methods are equally valuable for investigating literary source use.

Source and influence studies have a long tradition in pre-computational literary scholarship, starting in earnest roughly a century ago with John Livingston Lowe's The Road to Xanadu, a study of Samuel Taylor Coleridge's use of sources in his poetry [15]. Traditional methods of comparison continue to be carried out often with significant results. Yet no work to date has systematically applied machine reading or automated analysis to examine how authors draw upon their sources in composing their own works. Broad digital applications to reading practices—most notably the Archaeology of Reading [7] (which transcribes and exposes marginalia in machine-readable form) and Annotated Books Online [11], have made authorial reading methods and explicit annotations accessible for querying and analysis. Simultaneously, single-author initiatives such as The Walt Whitman Archive [5], The William Blake Archive [8] and Melville's Marginalia Online [21] have built extensive image and transcription corpora that encode the particular books an author actually handled and marked, steadily building the provenance and reading evidence necessary for undertakings like the present study. As these large-scale efforts finally achieve critical masses of machine-readable data, our research is now in a position to demonstrate the potential of computational approaches, providing new insights about source-use that are both textually grounded and empirically testable. These methods allow researchers to assess and verify instances where Melville and other major authors engaged with known sources, and to identify previously unrecognized influences that may have shaped his writing.

The BERTScore was introduced by [28] as an automatic evaluation metric for text generation. It measures semantic similarity between two texts by comparing their contextualized token embeddings produced by pretrained transformer models. BERTScore can capture paraphrastic and meaning-level correspondence that extends beyond surface lexical overlap. In their original work, [28] demonstrated that BERTScore correlates more strongly with human judgments than existing metrics (BLEU and METEOR) and provides improved model selection performance for tasks such as machine translation and image captioning. Building on this foundation, [27] applied BERTScore to the evaluation of translations of Japanese novels, further highlighting its usefulness for assessing semantic fidelity in literary contexts. More recently, [13] noted that current implementations of BERTScore do not expose all of the information the metric can produce, and introduced BERTScoreVisualizer, a tool that provides detailed token-level visualizations beyond the standard precision, recall, and F1 outputs. Together, these studies underscore the importance of BERTScore for semantic similarity analysis and highlight its effectiveness across a range of text comparison tasks. Most research to date has applied BERTScore in machine translation and related generation settings, its ability to quantify semantic similarity makes it a

strong candidate for investigating possible literary influence between an author's reading and writing.

While prior work has demonstrated the effectiveness of embedding-based metrics for semantic similarity, their application to literary source and influence studies remains limited. This study builds on these advances by applying semantic similarity analysis to examine potential textual relationships between Melville's reading and writing. The study necessarily relies upon a limited corpus consisting of texts for which documentary evidence of Melville's reading has been established, and for which machine-readable texts are available in clean versions. Many books Melville consulted are no longer identifiable or easily accessible in digitized forms, and these conditions impose limits on corpus selection. Our objective here is not to reconstruct the entirety of Melville's reading but to evaluate whether semantic-similarity methods can recover and illuminate source relationships already established through existing scholarship. Future work will expand the corpus to propose other passage pairs for expert evaluation.

3 Methodology

The broader goal of this research is to investigate semantic similarity across all works authored by Melville and to compare them with the books he is known to have read or owned, as documented in [2, 24]. Conducting a large-scale similarity analysis across Melville's complete corpus and all of his known library holdings presents significant computational and methodological challenges. Each of these works span hundreds of pages, and the number of possible pairwise comparisons is prohibitively large. As a preliminary step, this study focuses on evaluating whether the proposed similarity analysis method can effectively detect meaningful textual overlaps. Specifically, we aim to determine whether the approach can recognize known instances of semantic correspondence between Melville's writings and the texts he read as a proxy for future work of new discovery.

To achieve this, we first compiled a set of example text pair instances in which potential similarities were believed to exist based on scholarship by experts analyzing source appropriation [16, 19, 3]. From these we selected four source-text pairings as representative case studies drawn from different stages of Melville's fiction-writing career, spanning from Typee (1846) through Mardi and Moby-Dick (1849-1851) to Israel Potter (1855). They were also chosen because scholarship has identified distinct forms of source appropriation within them, including the adaptation of factual information, the reuse and transformation of phraseology, rhetorical amplification and revision of source attitudes, and sustained narrative rewriting. Together these examples provide confirmed and varied test cases for evaluating whether N-gram and BERTScore can detect forms of textual influence that range from localized verbal echoes to broader conceptual and narrative borrowings. The books and the corresponding text segments used in our experiments are detailed in Table 1. These identified correspondences served as our ground truth for evaluation. To ensure robustness, we did not restrict our analysis to only the passages whose correspondence had previously been

reported. Instead, we extended the context by including text segments occurring both before and after each candidate and target passage. This approach allowed for a more realistic test of the similarity model's capacity to recognize relevant relationships in a broader and less constrained textual environment. The structure of the books varies widely, with some short chapters and some long and some have no formal divisions at all. We thus did not extract uniform text lengths or sections across all text pair instances. Instead, we ensured that each extracted passage included sufficient context to allow for a meaningful similarity comparison, whether or not actual similarity was present.

The texts for both Melville's writings and his library holdings were obtained from Project Gutenberg[3] and the Internet Archive[4] in plain text format. No additional OCR correction was performed. Texts were used as provided by the source repositories. After identifying candidate segments, we divided them into smaller units suitable for computational comparison. Two different chunking strategies were applied:

- **N-gram based segmentation:** We use the same n-gram size for both texts; the segments from Melville's writings and the corresponding segments from the books he read, following the approach of [25]. This method captures fine-grained lexical overlap and allows us to detect shorter recurring phrases or patterns that confirm borrowing from sources identified in the scholarship. We choose the n-gram of 5, similar to the use in [26]. We use non-overlapping n-grams. This has the advantage of less computation effort with a trade-off on accuracy as compared to an overlapping window [25]. No stop-word removal, stemming, lemmatization, or punctuation filtering was applied.
- **Sentence level segmentation:** The selected segments were also divided at the sentence boundary. This provides larger semantic units that preserve syntactic and contextual information but will result in uneven chunk lengths. Sentence segmentation was performed using spaCy (en_core_web_sm), followed by a rule-based merging step in which short segments and sentence fragments were combined with neighboring sentences before comparison.

Using both chunking methods allowed us to examine similarity at different granularities, short phrase-level and longer, sentence-level.

For similarity analysis we used the BERTScore, a metric that evaluates the semantic similarity between two pieces of text using contextualized embeddings from pretrained transformer models [28]. Unlike traditional lexical overlap metrics such as BLEU or ROUGE, BERTScore captures the meaning of words in context, making it particularly suitable for identifying subtle or paraphrased similarities across texts [22]. This is especially important for our study, as Melville's rewriting of source passages often produces alterations in vocabulary and phrasing, even when conveying similar ideas. BERTScore allows us to go beyond surface-level alignment and evaluate whether two passages express similar meanings, regardless of exact duplication. Given that the influence of a text may appear through paraphrase, thematic similarity, or conceptual overlap rather than

[3] http://www.gutenberg.org, accessed 12 May 2025
[4] http://www.archive.org, accessed 12 May 2025

Table 1. Reference books and their corresponding mentions in Melville's works with target pages/chapters used in the comparison study.

S#	Reference Book	Author	Pages / Chapters	Melville's Book	Pages / Chapters
1	*A Visit to the South Seas: In the U. States Ship Vincennes, During the Years 1829 and 1830, Volume 1*	C.S. Stewart	Letter XXXIII (pp. 306–318)	*Typee; or, Narrative of a Four Months' Residence*	Chapter XXV
2	*Life and Remarkable Adventures of Israel R. Potter*	Henry Trumbull	pp. 30–45	*Israel Potter; His Fifty Years of Exile*	Chapter 4
3	*Religio Medici*	Sir Thomas Browne	Part I, Sections 6–8	*Mardi; and a Voyage Thither*	3 chapters(97-99)
4	*The Natural History of Sperm Whale*	Thomas Beale	Introductory remarks	*Moby Dick*	Chapter 32

Algorithm 1: Algorithm for Similarity Analysis

Input: Folders:
$M = \{M_1..M_n\}$ (Melville), $R = \{R_1..R_n\}$ (Read),
$S = \{5G, \text{sent}\}$ Modes: 5G = non-overlapping 5-grams; sent = sentences
Output: Scored CSVs for each mode
Text segmentation (SEG)
Function SEG(*file, mode*):
  return segments by mode
foreach *mode* $\in S$ **do**
  for $i \leftarrow 1$ **to** n **do**
    $X \leftarrow$ SEG (M_i, mode)
    $Y \leftarrow$ SEG (R_i, mode)
    $N \leftarrow \max(|X|, |Y|)$
    create CSV with header [ID, text, read_text]
    for $k \leftarrow 1$ **to** N **do**
      write row [k, X_k or "", Y_k or ""]
    end
    save CSV as merged_i_mode
  end
  Computing Precision(p), Recall(r) and F Score(f) using BERTSCORE
  foreach *CSV* $\in$ *{merged_1_mode..merged_n_mode}* **do**
    $T \leftarrow$ CSV[text]; $U \leftarrow$ CSV[read_text]
    for $i \leftarrow 1$ **to** $|T|$ **do**
      for $j \leftarrow 1$ **to** $|U|$ **do**
        $(p, r, f) \leftarrow$ BERTSCORE (T_i, U_j)
        store (p, r, f) for iteration i
      end
      store the similarity scores (p_{ij}, r_{ij}, f_{ij}) for all j as columns p_i, r_i, f_i in the CSV
    end
    save CSV as scored_i_mode
  end
end

direct quotation, BERTScore is a well-suited choice for our task. In addition, BERTScore operates at the token level and provides interpretable alignments through precision, recall, and F1 scores, allowing localized correspondences between passages to be identified. This is particularly valuable for literary influence studies, where partial reuse, fragmented borrowing, and phrase-level semantic overlap are often more significant than overall sentence-level similarity. Since our experiments were conducted at both the sentence level and the 5-gram level, a token-level alignment approach was considered more suitable than sentence-level embedding methods such as SimCSE, which represent entire sentences as single embedding vectors [9] and do not provide localized token-level correspondences or precision, recall, and F1-based interpretability for shorter textual fragments.

BERTScore produces precision (p), recall (r), and F1 scores that quantify the degree of semantic similarity between two texts. In prior work, especially in machine translation evaluation, higher scores are associated with closer semantic correspondence between a candidate and a reference text. However, the literature does not define universal thresholds for copying or reuse, and such

thresholds do not directly translate to studies of literary influence. In this research, we therefore do not treat BERTScore values as binary indicators of reuse. Instead, relatively higher scores are interpreted as signals of potential semantic alignment that may indicate thematic, phrasal, or conceptual influence. Our goal is to identify passages where the model detects meaningful similarity that warrants qualitative attention. We use the Microsoft/deberta-xlarge-mnli variant of BERTScore, which has shown strong performance in natural language understanding tasks [6]. The algorithm of our methodology is shown in Algorithm 1.

4 Semantic Similarity Results: Evidence of Possible Influence

Our experimental results demonstrate that the model reliably recognized patterns of similarity within correspondences previously identified in the scholarship. In the subsections that follow, we examine these results across the different experimental settings and discuss expert identified instances at both 5-gram and sentence level granularities, then model identified instances at both granularities. Finally we compare difference in results at both granularities. The examples presented in the following sections are illustrative rather than exhaustive. For the expert-identified correspondences, the selected passages exemplify different forms of source appropriation and serve as reference points for evaluating whether high similarity scores occur at expected locations within the similarity matrices. For the model-identified correspondences, examples were selected from among higher-scoring sentence and 5-gram pairs identified by BERTScore and subsequently examined for their potential relevance to textual influence.

4.1 Expert Identified Instances: Validation of Semantic Similarity

To illustrate how the model performs on an expert identified instance of potential influence, we consider the first instance in Table 1, where Melville's book *Typee* is compared with Stewart's book *A Visit to the South Seas*. For illustrative purposes, throughout this paper only a selected portion of the experimental output from the text subsection is presented. The remaining results, comprising a broader set of sentence and 5-gram-level comparisons, are omitted to maintain clarity and focus on representative cases and are available upon request.

Results are displayed at two granularities: 5-gram sequences, Figure 1, and sentence-level segments, Figure 2. In both figures, the colored text (green) highlights the portions previously marked by experts as potentially similar. Moreover, in the heatmap red cells represent pairs with stronger similarity, while yellow and green cells indicate weaker similarity. The computed F1-scores at their peak exceed 0.6 for both the 5-gram and sentence-level analyses, indicating a measurable degree of similarity consistent with possible influence.

In Figure 1, one specific 5-gram segment pair ("of a plurality of husbands" versus "a plurality of husbands, instead") produces a notably high F1-score (0.77), reflecting a strong degree of local similarity between the two texts. This

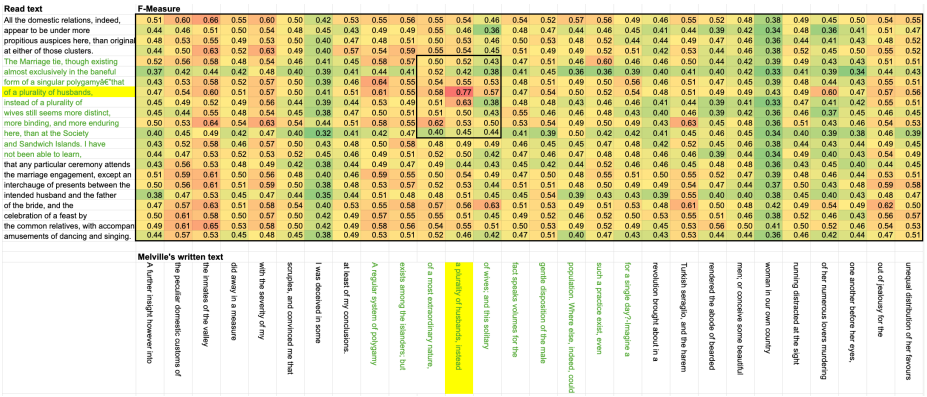

Fig. 1. 5-gram analysis: F1 scores for instance 1 in Table 1 comparing *Typee* with Stewart's *A Visit to the South Seas*. Document wide BERTScore F1 range is 0.21–0.80.

Fig. 2. Sentence level analysis: F1 scores for instance 1 in Table 1 comparing *Typee* with Stewart's *A Visit to the South Seas*. Document wide BERTScore F1 range is 0.38–0.64.

elevated score likely results from overlapping lexical tokens shared between the work Melville read in the book by Stewart and the one he later authored. In Figure 2, the sentences from these same text portions also display comparatively higher similarity values (0.63), although it was slightly lower than in the 5-gram analysis. This difference arises because sentence-level comparisons encompass more tokens, and the similarity tends to be concentrated within particular portions of the sentence rather than uniformly distributed across the entire segment.

The example in Figure 2 illustrates similarity at the sentence level, where the sentence in Melville's writing appears to be influenced by a single sentence from the source text. The text pairs 1 and 2 in Table 1 are both examples of the sentence to sentence similarity scenario. However, other scenarios occur in which

one written sentence is influenced by several sentences in the source. In Table 1 instances 3 and 4 represent such a scenario. Figure 3, shows the result from instance 4 of Table 1 where text from *Moby Dick* is compared with text from Beale's *The Natural History of Sperm Whale*. For this example, the values for all metrics (precision, recall, and F1 score) are presented in Figure 3. The orientation of read vs written text is swapped for better display. It is important to note that sentence-to-sentence influence generally produces high similarity values, as seen in Figure 2. In contrast, when the same idea is distributed across multiple sentences, the similarity signal is diluted because each individual sentence contains only a portion of the shared meaning, which lowers the sentence-level similarity score as seen in Figure 3. Similar sentence to multiple sentence influence was identified in instance 3 from Table 1, showing the same results of diluted similarity signal, but results are not included in this paper. The differences in precision, recall and F1 scores are also important here. The higher precision indicates that most tokens in Melville's sentence matched strongly with those in the source. The comparatively lower recall indicates the source contains more content that was not reflected in Melville's sentence, as the idea read in the source sentence was distributed and explained in multiple sentences in Melville's writing. The example strongly illustrates the importance of subjecting computational results of this procedure to human expert examination. As an utterance inspired by multiple successive sentences in the work Melville read, the passage is deeply indebted to the source. It's rhetoric, moreover, illustrates Melville's frequent tendency in source appropriation to consciously "amplify" or "augment" stylistically the attitude or claim conveyed in the original text–a practice that parallels his related method of reversing or opposing the attitude of a source, as seen in the example illustrated by Figure 1 [19, 18]. Yet the low values in the output of Figure 3 do not on their face indicate a strong correlation.

Melville Text	Precision (p)	Recall (r)	F1 score
And here be it said, that the **Greenland whale** is an **usurper** upon the **throne** of the seas.	0.58	0.44	0.5
He is not even by any means the largest of the whales.	0.57	0.37	0.44
Yet, owing to the long priority of his claims, and the profound ignorance which, till some seventy years back, invested the then fabulous or utterly unknown sperm-whale, and which ignorance to this present day still reigns in all but some few scientific retreats and whale-ports; this **usurpation** has been every way complete.	0.52	0.52	0.52
Reference to nearly all the leviathanic allusions in the great poets of past days will satisfy you that the **Greenland whale,** without one rival, was to them the **monarch** of the seas.	0.58	0.53	0.55
Reference to nearly all the leviathanic allusions in the great poets of past days will satisfy you that the **Greenland whale,** without one rival, was to them the **monarch** of the seas.	0.51	0.46	0.48

Read_text

The **Greenland whale**, or Balcena mysticeius, has so frequently been described in a popular manner, that the public voice has long **enthroned** him as **monarch** of the deep, and perhaps the dread of disturbing such weighty matters as a settled sovereignty and public opinion, may have deterred those best acquainted with the merits of the case from supporting the more legitimate claims of a his southern rival to this pre-eminence.

Fig. 3. Precision, recall and F1 score values for instance 4 from Table 1 comparing Beale's *The Natural History of the Sperm Whale* and *Moby Dick*. Document wide BERTScore ranges are: P (0.31–0.75), R (0.25–0.62), F1 (0.31–0.62).

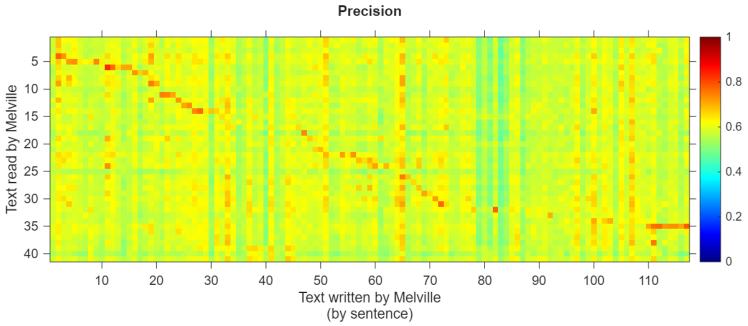

Fig. 4. Precision heat map for complete sentence level text comparison of instance 2 in Table 1 comparing Trumbull's *Life and Remarkable Adventures of Israel R. Potter* with Melville's *Israel Potter; His fifty years of exile*. Document wide BERTScore p range is (0.36–0.80).

4.2 Model-Ranked Correspondences Within a Known Source Relationship

After applying the method against specific correspondences previously identified in scholarship, we next examined the model's rankings across a larger portion of a source relationship already known to scholars. The objective was to determine whether the method could identify additional high-similarity passages beyond the specific correspondences identified in prior studies. We present the complete chapter results in Figure 4 showing a heatmap of precision values at the sentence level comparing Melville's novel *Israel Potter*, a work known to be a rewriting of Henry Trumbull's *Life and Remarkable Adventures of Israel Potter*. Precision was selected because the analysis focuses on passages from Melville's writing and the degree to which they align with texts he is known to have read. In this context, precision provides the most direct indication of how strongly a candidate Melville passage corresponds to a potential source passage. The figure illustrates a diverse range of color intensities. As in the previous figures, darker and red cells represent pairs with stronger similarity, while yellow and green cells indicate weaker similarity. A few features are visible in this figure. The presence of the high value (red) diagonal line comes from Melville literally retelling the story from the source text for a novel conceived and openly acknowledged by Melville to be a rewriting of Trumbull's 1824 *Life* [14]. In the run of early chapters represented by the passages in Figure 4, Melville is believed to have transcribed words and expressions from Trumbull into approximately half the lines of his lost manuscript, occasionally adding "brief speculations on Israel's motives, or his recurrent foreshadowings of coming tribulations" [3]. As composition proceeded beyond the early chapters of Israel Potter, Melville allowed himself greater freedom from his source and introduced a variety of imaginative episodes that do not appear in Trumbull's narrative.

To gain deeper insight into these results, we identified the cells corresponding to higher precision values and analyzed the associated text segments. For

Table 2. Three text pair examples for analysis of parallels from Melville's *Israel Potter; His Fifty Years of Exile* and Trumbull's *Life and Remarkable Adventures of Israel R. Potter*. The first is at the sentence level and the other two at the 5-gram level.

Melville Text	**Read Text**	**Results**
Israel had now been three days without food, except one two-penny loaf. (sentence 11)	I had now been three days without food (with the exception of a single two-penny loaf) and felt myself unable much longer to resist the cravings of nature—my spirits, which until now had armed me with fortitude began to forsake me—indeed I was at this moment on the eve of despair! When, calling to mind that grief would only aggravate my calamity, I endeavoured to arm my soul with patience; and habituate myself as well as I could, to woe. Accordingly I roused my spirits; and banishing for a few moments these gloomy ideas, I began to reflect seriously on the methods by which to extricate myself from this labyrinth of horror. (sentence 6)	p = 0.80 r = 0.43 F1 = 0.56
presenting a smaller supply than (chunk 301)	which containing a small quantity (chunk 88)	p = 0.65 r = 0.70 F1 = 0.68
at that season of the (chunk 120)	at that season of the (chunk 249)	p = 1.00 r = 1.00 F1 = 1.00

instance, from the full output, the highest value in the heatmap occurs at source sentence 6 and Melville's sentence 11 where the precision value reached 0.8. These sentences along with the results are shown as the first text pair example in Table 2. The BERTScore of p=0.80 indicates that most tokens in Melville's sentence matched strongly with those in the source. However, the recall was lower (0.43) due to the source containing significantly more content that was not reflected in Melville's sentence. The resulting F1 score was 0.56, suggesting semantic overlap and possible conceptual influence. Also visible in Figure 4 are some vertical streaks at sentences 65 and 107 in Melville's *Israel Potter; His Fifty Years of Exile*. This highlights one of the limitations: small or generic text tends to yield higher values, which may result in false positives. For example, sentence 65 *"About noon the knight visited his workmen."* produces relatively high similarity values because it is short and semantically generic, containing broad concepts such as time, authority, and labor. These elements occur frequently across many contexts, allowing embedding-based metrics to find plausible matches even when no specific relationship exists. As a result, this sentence has a higher potential to produce false positives. Similarly sentence 107 *"Now the superintendent of the garden was a harsh, overbearing man."* yielded relatively high similarity values throughout because it describes a generic authority figure and evaluative traits that commonly appear in narrative texts.

4.3 Evaluating Similarity at Different Levels of Text Granularity

We have conducted experiments at both the 5-gram and sentence levels to compare how the granularity of text segmentation affects the model's ability to detect semantic similarity. Both approaches successfully identified instances of correspondence between Melville's writings and the source texts. The 5-gram analysis provided a more fine-grained view of similarity patterns. As illustrated in Figure 1 and further supported by Table 2. The 5-gram method was particularly

effective in capturing localized lexical and semantic overlaps, short phrases or recurring expressions that might indicate influence at the phrase level. The second text pair in Table 2 is taken from the 5-gram experiments from instance 2 from Table 1. Here even at the 5-gram level of granularity the p, r and F1 values are high and the semantic similarity is also evident from the tokens as they reflect similar overall semantic meaning. The third text pair illustrates an exact 5-gram match between the source and target texts. In this case, precision, recall, and F1 all reached 1.00, indicating complete overlap between the two textual fragments and demonstrating the model's ability to identify instances of direct lexical reuse in addition to broader semantic correspondences. In contrast, the sentence-level analysis yielded a broader contextual comparison but introduced variability in precision and recall values. This variation primarily stems from differences in sentence length and the number of tokens being compared. When one sentence contains significantly more tokens than the other, precision tends to be higher for shorter sentences, while recall decreases because a smaller portion of the longer sentence aligns semantically. The 5-gram method, by comparison, mitigates this issue by keeping token counts relatively consistent across comparisons, allowing for a more balanced representation of local similarity.

Overall, these results suggest that the 5-gram segmentation approach is better suited for detecting localized lexical or phrasal similarities, while the sentence-level analysis provides complementary insights into broader semantic relationships. Using both levels of analysis together would offer a more comprehensive understanding of textual influence, capturing both micro-level echoes and macro-level thematic parallels.

5 Limitations and Future Work

This study operates under a few practical constraints that shape the scope of the results. First, we used non-overlapping 5 grams that offer an efficient segmentation strategy, although it may overlook similarities that cross unit boundaries. Overlapping five grams would yield more complete coverage of the text, but doing so substantially increases computational cost. This reflects an unavoidable tradeoff between granularity and computational efficiency. Second, a further limitation arose from the behavior of BERTScore when comparing very short and generic sentences yielding high precision values because their broad semantic content can be matched to many different contexts in the reference text, even when no meaningful similarity exists. This can also be seen in Figure 4 and was discussed in detail in Section 4.2. Future work may address this by applying weighting scores by sentence specificity, or incorporating additional models that better distinguish generic similarity from substantive semantic alignment. By selecting a complementary text comparison technique, many false positives could be removed from consideration. Finally, the analysis was conducted at only two segmentation levels: 5-gram and sentence-level. These levels work well when influence occurs at comparable spans of text, such as the sentence-to-sentence as shown in Figure 2. When influence spans mismatched chunks, such as one sen-

tence drawing on several others, sentence-level comparisons may produce lower similarity scores due to misaligned comparison windows. Additional segmentation strategies, including phrase-level and paragraph-level segmentation, may therefore better capture both localized and broader contextual similarities.

6 Conclusion

This study explored whether semantic similarity methods, specifically BERTScore, can help identify potential influence between Herman Melville's writings and the books he read. Using expert-identified examples as a starting point, the analysis showed that BERTScore was able to detect these known similarities and also highlighted additional passages with notable semantic alignment. These findings suggest that semantic similarity measures can provide meaningful insights into possible connections between an author's reading and writing.

Acknowledgments

This work is partly supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by Knut and Alice Wallenberg Foundations and counterpart funding from Lulea University of Technology (LTU).

References

1. Altheneyan, A.S., Menai, M.E.B.: Automatic plagiarism detection in obfuscated text. Pattern Analysis and Applications **23**, 1627–1650 (2020)
2. Bercaw, M.K.: Melville's Sources. Northwestern University Press (1987)
3. Bezanson, W.E.: Historical note. In: Hayford, H., Parker, H., Tanselle, G.T. (eds.) Israel Potter: His Fifty Years of Exile, The Writings of Herman Melville, vol. 8. Northwestern University Press and the Newberry Library, Evanston and Chicago (1982)
4. Clough, P., Stevenson, M.: Developing a corpus of plagiarised short answers. Language resources and evaluation **45**, 5–24 (2011)
5. Cohen, M., Folsom, E., Price, K.M.: The Walt Whitman Archive, https://whitmanarchive.org/
6. Dernbach, S., Agarwal, K., Zuniga, A., Henry, M., Choudhury, S.: Glam: Fine-tuning large language models for domain knowledge graph alignment via neighborhood partitioning and generative subgraph encoding. In: Proceedings of the AAAI Symposium Series. vol. 3, pp. 82–89 (2024)
7. Dijstelberge, P., Grafton, A., et al.: Annotated books online: A digital archive of early modern annotated books (nd), https://annotatedbooksonline.com, accessed: 24 February 2026
8. Essick, R., Viscomi, J.: The William Blake Archive, https://blakearchive.org/
9. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 6894–6910 (2021)

10. Habib, N., Adewumi, T., Liwicki, M., Barney, E.: Trends and Challenges in Authorship Analysis: A Review of ML, DL, and LLM Approaches. arXiv preprint arXiv:2505.15422 (2025)
11. Havens, E., Jardine, L., Symonds, M., et al.: The archaeology of reading (nd), https://archaeologyofreading.org, accessed: 24 February 2026
12. Hayford, H., Blair, W.: Omoo: A Narrative of Adventures in the South Seas. Hendricks House, New York (1969)
13. Jaskowski, S., Chava, S., Shah, A.: BERTScoreVisualizer: A Web Tool for Understanding Simplified Text Evaluation with BERTScore. arXiv preprint arXiv:2409.17160 (2024)
14. Leyda, J.: The Melville Log: A Documentary Life of Herman Melville, 1819–1891. Harcourt Brace, New York (1951), 2 vols.
15. Lowes, J.L.: The Road to Xanadu: A Study in the Ways of the Imagination. Houghton Mifflin (1927)
16. Mathiessen, F.O.: American Renaissance Art and Expression in the Age of Emerson and Whitman. Oxford University Press (1941)
17. Meo, S.A., Talha, M.: Turnitin: Is it a text matching or plagiarism detection tool? Saudi Journal of Anaesthesia **13**(Suppl 1), S48–S51 (2019)
18. Norberg, P., Olsen-Smith, S.: The technical development and expanding scope of melville's marginalia online. Leviathan **25**(2), 61–85 (2023)
19. Olsen-Smith, S.: Melville's copy of Thomas Beale's" The natural history of the sperm whale and the composition of Moby-Dick". Harvard Library Bulletin **21**(3), 1–77 (2011), fall 2010
20. Olsen-Smith, S.: Books and marginalia, real and virtual. A New Companion to Herman Melville p. 283 (2022)
21. Olsen-Smith, S., Norberg, P.C., Marnon, D.C.: Melville's Marginalia Online. Boise State University (2006)
22. Rostam, Z.R.K., Takács, M., Kertész, G.: Evaluating large language models: A review of metrics and benchmarks. In: 2025 IEEE 23rd Jubilee International Symposium on Intelligent Systems and Informatics (SISY). pp. 000233–000240. IEEE (2025)
23. Šarić, F., Glavaš, G., Karan, M., Šnajder, J., Bašić, B.D.: Takelab: Systems for measuring semantic text similarity. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 441–448 (2012)
24. Sealts, Merton M., J.: Melville's Reading. University of South Carolina Press (1988), includes bibliographical references and index
25. Shrestha, P., Solorio, T.: Using a variety of n-grams for the detection of different kinds of plagiarism. Notebook for PAN at CLEF **2013** (2013)
26. Suchomel, S., Kasprzak, J., Brandejs, M., et al.: Three way search engine queries with multi-feature document comparison for plagiarism detection. In: CLEF (Online Working Notes/Labs/Workshop). pp. 1–8 (2012)
27. Tanahashi, Y., Kanzaki, K., Yamamoto, E., Isahara, H.: Improving semantic similarity calculation of Japanese text for MT evaluation. In: Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation. pp. 542–550 (2020)
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BertScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675 (2019)