Benchmarking BERT-based Models for Latin: A Case Study on Biblical References in Ancient Christian Literature

Davide Caffagni[1,†], Federico Cocchi[1,2,†], Anna Mambelli[3,4], Fabio Tutrone[5], Marco Zanella[6,7], Marcella Cornia[3,*] and Rita Cucchiara[1]

[1]Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy
[2]Department of Informatics, University of Pisa, Pisa, Italy
[3]Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy
[4]Fondazione per le scienze religiose (FSCIRE), Bologna, Italy
[5]Department of Cultures and Societies, University of Palermo, Palermo, Italy
[6]Department of Mathematics, University of Padua, Padua, Italy
[7]Department of History and Cultures, University of Bologna, Bologna, Italy

Abstract

Transformer-based language models like BERT have revolutionized Natural Language Processing (NLP) research, but their application to historical languages remains underexplored. This paper investigates the adaptation of BERT-based embedding models for Latin, a language central to the study of the sacred texts of Christianity. Focusing on Jerome's *Vulgate*, pre-*Vulgate* Latin translations of the Bible, and patristic commentaries such as Augustine's *De Genesi ad litteram*, we address the challenges posed by Latin's complex syntax, specialized vocabulary, and historical variations at the orthographic, morphological, and semantic levels. In particular, we propose fine-tuning existing BERT-based embedding models on annotated Latin corpora, using self-generated hard negatives to improve performance in detecting biblical references in early Christian literature in Latin. Experimental results demonstrate the ability of BERT-based models to identify citations of and allusions to the Bible(s) in ancient Christian commentaries while highlighting the complexities and challenges of this field. By integrating NLP techniques with humanistic expertise, this work provides a case study on intertextual analysis in Latin patristic works. It underscores the transformative potential of interdisciplinary approaches, advancing computational tools for sacred text studies and bridging the gap between philology and computational analysis.

Keywords

Sentence Similarity Search, Sentence Embeddings, Ancient Languages

1. Introduction

The advent of Transformer-based language models [1] such as BERT [2, 3, 4, 5] has revolutionized the field of Natural Language Processing (NLP), offering unprecedented capabilities in tasks ranging from text classification to semantic similarity analysis [6, 7, 8, 9] and demonstrating their adaptability to other modalities beyond text [10]. By leveraging self-attention mechanisms and large-scale pre-training, these models capture fine-grained contextual relationships previously unattainable with traditional machine learning. While highly effective for modern languages [11, 12, 13, 14], their application to historical languages remains underexplored [15]. Historical languages pose unique challenges, including scarce high-quality annotated datasets, variations in orthography or morphology, and the need to deal with diachronic linguistic changes that can make finding semantic patterns very difficult [16, 17, 18]. Despite

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20–21, 2025, Udine, Italy

*Corresponding author.
†These authors contributed equally.
davide.caffagni@unimore.it (D. Caffagni); federico.cocchi@unimore.it (F. Cocchi); anna.mambelli@unimore.it (A. Mambelli); fabio.tutrone@unipa.it (F. Tutrone); marco.zanella@unipd.it (M. Zanella); marcella.cornia@unimore.it (M. Cornia); rita.cucchiara@unimore.it (R. Cucchiara)
0009-0002-3279-6480 (D. Caffagni); 0009-0005-1396-9114 (F. Cocchi); 0000-0001-5538-5882 (A. Mambelli); 0000-0002-7063-7782 (F. Tutrone); 0009-0000-1208-6743 (M. Zanella); 0000-0001-9640-9385 (M. Cornia); 0000-0002-2239-283X (R. Cucchiara)

these challenges, understanding historical languages like Latin holds significant promise, not only for enriching NLP methodologies but also for advancing research in fields such as historical linguistics, philology, historical-religious studies, and exegesis.

This work investigates the adaptation of BERT-based models [19, 20, 21] for Latin, a pivotal language in the study of the sacred texts of Christianity and their receptions. Latin's central role in the Christian exegetical tradition, along with its rich corpus of sacred and hermeneutical texts, provides an ideal context for developing NLP models for historical languages. Analyzing ancient Latin biblical texts – Jerome's *Vulgate* and the pre-*Vulgate* Latin translations of the Greek Bible (*Vetus Latina*) – is crucial to understanding the context and history of their formation, as well as the reception history of the Hebrew and Greek Bibles, with their various forms of exegesis and rewriting. Indeed, the issue with authoritative sources lies in their intrinsic textual plurality, which is itself a sign of exegetical plurality. At the same time, sacred texts, as historical objects, may also be reconstructed from their tradition, closely connecting biblical texts to later Christian works that comment on, quote, rework, and allude to them. In particular, Latin patristic commentaries, such as Augustine's *De Genesi ad litteram*, encapsulate intricate intertextual relationships with the biblical texts. These textual corpora pose distinct challenges for NLP due to their complex syntax, specialized vocabulary, and historical variations at the orthographic, morphological, and semantic levels. Further complicating this analysis, biblical references in patristic texts are frequently oblique, involving rephrasings, paraphrases, or allusions rather than quotations.

To address these challenges, this paper explores the potential of BERT-based models trained on Latin textual corpora to improve the identification and analysis of biblical references in Latin patristic commentaries. Our approach includes fine-tuning the models using corresponding passages[1] from the *Vulgate* and pre-*Vulgate* Latin translations of the Bible, leveraging the natural variations (*i.e.*, variant readings) between these biblical versions as a rich source of data for refining the embedding space. We report results on annotated biblical references from ancient Christian Latin commentaries, demonstrating the effectiveness of this methodology. During fine-tuning, we further enhance the model performance by employing self-generated hard negatives, derived from the embedding model itself, to refine its ability to discern subtle distinctions in intertextual relationships. This process supports the development of computational tools capable of detecting both "explicit" citations and "implicit" allusions in Latin texts with a high degree of accuracy.

The contributions of this study are threefold. First, it outlines the methodological integration of humanistic expertise and NLP techniques, particularly the fine-tuning of BERT for sacred texts in Latin. Second, it presents a case study on the identification of biblical references in Latin patristic commentaries, demonstrating the practical applications of these models. Third, it highlights the potential of interdisciplinary approaches to transform the study of sacred texts and their receptions, bridging computational analysis and traditional philology. By advancing the application of Transformer-based models to Latin, this paper contributes to both the technical and scholarly dimensions of biblical text studies. In doing so, it underscores the transformative possibilities of interdisciplinary research at the intersection of computer science and the humanities, fostering new insights into the textual, intellectual, and exegetical heritages of religious communities.

2. Intertextual References in Ancient Christian Commentaries: A Case Study on Biblical Corpora

2.1. Annotating Biblical References

The analysis of biblical references within ancient Christian commentaries relies on manually curated datasets from Latin biblical and patristic texts in their critical reference editions. In particular, the commentary chosen for this case study is Augustine's *De Genesi ad litteram libri duodecim*[2], a pivotal

[1] With a slight abuse of notation, we will use the terms "passage" and "verse" interchangeably, referring to a piece of the biblical text identified by a book, a chapter, and a verse number (*e.g.*, Gen. 3.1).

[2] The edition used in this study is that of J. Zycha [22] (*i.e.*, the most recent critical edition to date), downloaded from the Corpus Corporum database available at https://mlat.uzh.ch/ and manually revised before annotation.

Table 1
Distribution of annotated references across similarity score ranges for the two biblical corpora, W_VULG (*Vulgate*) and S_VL (*Vetus Latina*). The total number of biblical passages in each corpus is also provided.

		# References				
Corpus	**# Passages**	0.0-0.25	0.25-0.5	0.5-0.75	0.75-1.0	**All**
W_VULG	35,057	51	50	46	45	192
S_VL	20,791	44	23	20	83	170

work in the Christian exegetical tradition. This commentary, completed in the early 5th century, provides Augustine's detailed hermeneutical reflections on the Book of Genesis, which inspire and give way to the definition of broader theological motifs. It also explicitly and implicitly interacts with multiple versions of the Latin Bible, that is, Jerome's *Vulgate* and pre-*Vulgate* translations. Given Augustine's intellectual prominence and central role in shaping Christian hermeneutics, his works provide an ideal case for studying biblical references in ancient Christian literature.

As biblical textual corpora, we employ two (at least partially) different versions of the Latin Bible: the *Vulgate* (W_VULG) and the so-called *Vetus Latina* (S_VL). The W_VULG, a critical edition by R. Weber and R. Gryson [23][3], is the standard scholarly edition of Jerome's *Vulgate*. In contrast, the *Vetus Latina*, an older and fragmentary collection of Latin translations reconstructed mostly by indirect tradition, is accessible as a whole through the 18th-century edition of the Benedictine monk P. Sabatier [24][4] (S_VL). Compared to the *Vulgate*, the Sabatier's edition presents challenges due to its lack of digital integration.

Annotating Augustine's commentary involves identifying textual parallels to passages in the Bible, determining whether references are exact quotations, paraphrases, or thematic allusions, and systematically tagging them using the INCEpTION annotation platform [25][5]. This platform facilitates the encoding of detailed information about each reference, including its source (*i.e.*, W_VULG or S_VL), its location (*i.e.*, book, chapter, verse), and a similarity score quantifying the degree of lexical overlap between the annotated passage of the commentary and corresponding biblical verses. The similarity score ranges from 0 to 1, where 0 indicates no lexical overlap and 1 denotes an exact lexical match.

2.2. Benchmark Characteristics

The resulting dataset comprises 192 annotated references to the W_VULG Bible and 170 to the S_VL Bible, classified into four similarity categories based on their lexical overlap scores: 0.0-0.25, 0.25-0.5, 0.5-0.75, and 0.75–1.0. These similarity ranges capture the spectrum of intertextual relations, from loose thematic connections to verbatim citations. Table 1 details the distribution of references across these similarity ranges. Notably, references to W_VULG are distributed relatively evenly, while references to S_VL skew toward high similarity scores, with 83 instances scoring between 0.75 and 1.0. It is also important to note the differing overall sizes of the two biblical corpora. The W_VULG contains 35,057 passages (each corresponding to a biblical verse), whereas the S_VL only comprises 20,791 passages, due to the unavailability of some original books in digital format.

3. Mapping Intertextuality via BERT-based Models for Latin

Our goal is to identify intertextual references between patristic commentaries and biblical passages. For this task, we focus on Augustine's *De Genesi ad litteram* as the query text, influenced by the Latin Bible as a key source, and examine references to the W_VULG and S_VL Latin translations of the Bible, as detailed in Sec. 2. We frame this problem as an information retrieval task: given a query, the objective is to retrieve the most relevant documents from a collection. In our settings, a query q is a passage

[3]Available in digitized form from the Deutsche Bibelgesellschaft at https://www.die-bibel.de/en/bible/VUL/.
[4]Available at the following links: https://archive.org/details/bibliorumsacroru01saba/page/n7/mode/2up, https://archive.org/details/bibliorumsacroru02saba/page/n7/mode/2up, https://archive.org/details/Sabatier3.
[5]https://inception-project.github.io/

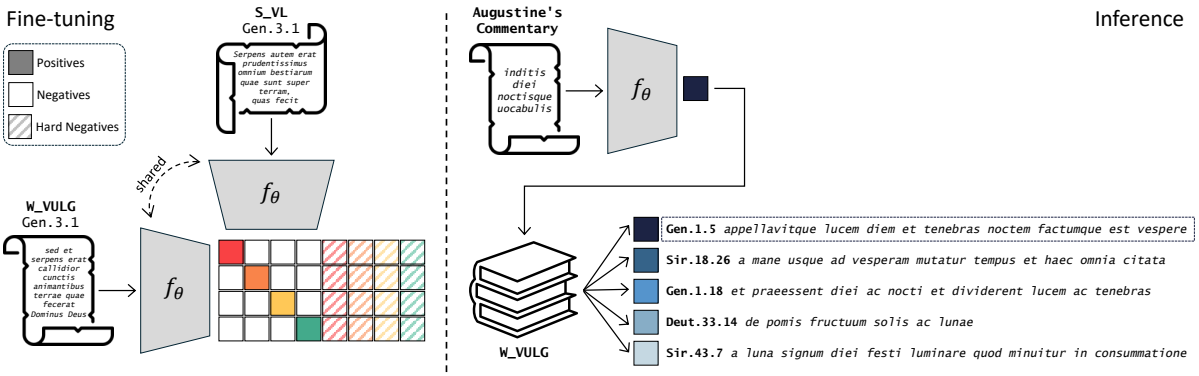

Figure 1: Overview of fine-tuning and inference pipelines.

from Augustine's commentary and documents d are verses from the two considered versions of the Bible. Each query is associated with a positive document d^*, corresponding to an intertextual reference between the commentary and the Bible(s). In practice, q may be a literal citation of the biblical verse d^*, or it may just allude to d^*. The former type of relationship is typically easier to identify by measuring the text overlap between a query and a document. Conversely, allusions to the Bible(s) are hard to detect, as they require complex semantical analysis, a task that is not trivial even for human experts.

3.1. Retrieving Bible Passages from Commentary Sentences

We propose to leverage Transformer-based language models [1], such as BERT [2], to effectively capture the complex intertextual references between patristic commentaries and biblical passages. To this end, let f_θ be a BERT-like pre-trained model. Before processing a query sentence or a document with f_θ, the input is first tokenized. Each token is assigned a unique integer ID, which acts as an index to select the corresponding embedding in the input embedding matrix of f_θ. This sequence of token embeddings is then passed through a stack of twelve Transformer layers, each comprising two main components: the attention operator, which relates each token to all other tokens in the sequence, and a feed-forward network that processes each token independently. The result is a sequence of output embeddings from the final Transformer layer, one for each token in the input. To obtain a single feature vector (*i.e.*, *embedding*) representing the entire input sequence, we experiment with two aggregation strategies:

- **CLS Token Embedding.** BERT-like models prepend a special classification token (*i.e.*, CLS) to the input sequence. The output embedding corresponding to the CLS is often regarded as a condensed and global representation of the entire input sequence.
- **Token Averaging.** An alternative strategy involves aggregating information from all tokens in the sequence to create a more comprehensive representation. This is achieved by taking the average of the embeddings of all tokens, except the CLS, in the input. Unlike the CLS token, which focuses on providing a global summary, token averaging distributes equal importance to each token, potentially capturing finer-grained information about the input sequence.

These embeddings, representing the query and the document, are mathematically expressed as follows:

$$\begin{aligned} \mathbf{q} &= f_\theta(q) \in \mathrm{R}^m, \\ \mathbf{d} &= f_\theta(d) \in \mathrm{R}^m. \end{aligned} \tag{1}$$

At this point, we measure the relevance of $\mathbf{d}$ with respect to $\mathbf{q}$ by calculating the cosine similarity between the two vectors:

$$s(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q}}{\|\mathbf{q}\|} \frac{\mathbf{d}^\top}{\|\mathbf{d}\|} \tag{2}$$

where $\|\cdot\|$ indicates the Euclidean norm. Ideally, the relevance score between a query and its positive document should be maximized. Conversely, the similarity score with respect to any negative document – defined as any document other than the positive one – should be minimized.

3.2. Fine-tuning with Self-Hard Negative Mining

While the model f_θ is pre-trained on general language modeling tasks, it has not been specifically trained for the task of text retrieval. To adapt f_θ for this purpose, we fine-tune it using contrastive learning, a method that has proven effective for retrieval [26, 27] and other multimodal tasks [28, 29, 30]. In detail, given a batch of query-positive document pairs $(q, d^*) \in \mathcal{B}$, we embeds queries and documents with f_θ, and then we compute the InfoNCE loss function [31]:

$$\mathcal{L} = -\sum_{(\mathbf{q},\mathbf{d}^*)\in\mathcal{B}} \log \frac{\exp s(\mathbf{q},\mathbf{d}^*)}{\exp(s(\mathbf{q},\mathbf{d}^*)) + \sum_{\mathbf{d}\in\mathcal{N}} \exp(s(\mathbf{q},\mathbf{d}))} \tag{3}$$

By minimizing Eq. 3, we encourage f_θ to map a query and its positive document (q, d^*) to two points on the unit sphere that are close to each other. Conversely, negative documents unrelated to q, that are represented by $\mathcal{N}$ in the preceding formula, are pushed away from the embedding representation of q.

Overcoming the Lack of Training Data. A key challenge in training f_θ is the limited availability of commentary queries paired with their corresponding biblical passages (cf. Table 1). To mitigate this issue, we draw inspiration from self-supervised contrastive learning [32, 33] and propose a surrogate task for training. Specifically, we sample a verse from the W_VULG Bible as a query q, and pair it with the corresponding verse from the S_VL version as the positive document d^* (or vice versa). At each training step, we sample N negative documents for each query. In addition, we treat the positive and negative documents from other queries within the same batch as further negatives.

Additional Hard Negative Samples. The previously described procedure, commonly employed in contrastive learning [34, 35, 36], enhances model sensitivity to the distinctions between related and unrelated documents by exposing it to a larger number of negative samples. The quality of these negatives is crucial: documents that are similar to the query in the embedding space but not semantically related are referred to as *hard negatives*. These hard negatives are known to improve the robustness of models trained with contrastive loss functions [37, 38, 39, 40] like InfoNCE.

In this work, we propose an effective strategy for mining hard negatives during training. First, we generate document embeddings by processing verses from the W_VULG version of the Bible with the pre-trained model f_θ. Then, for each positive document d^* associated with a query q, we retrieve the top-k most similar documents and use them as hard negatives for q. Fine-tuning the model using hard negative documents coming from the BERT model itself, as opposed to randomly sampling documents, makes the loss function in Eq. 3 more challenging to minimize, ultimately leading to improved performance.

4. Experimental Results

4.1. Experimental Setup

Considered BERT-based Embedding Models. In this study, we model f_θ with three language models sharing the architecture of the BERT model [2], namely Latin RoBERTa [20], Latin BERT [19], and LaBERTa [21]. All considered models have been pre-trained with the masked language modeling objective [2, 3]: the model is asked to predict the missing words that are randomly masked in the input sentence. The main difference between the three models is the Latin corpus chosen for pre-training. Latin RoBERTa [20] was trained on 390M tokens extracted from the Latin portion of CC-100 [41]. Latin BERT [19] used 642M tokens from a variety of sources spanning the Classical era to the 21st century. Lastly, LaBERTa [21] was trained on Corpus Corporum[6] for a total of 167M tokens.

Training Details. All models produce embeddings of size m equal to 768. We fine-tune them with the loss function detailed in Eq. 3, using identical hyperparameters and settings. Specifically, we train with the Adam [42] optimizer, a learning rate fixed to 1×10^{-6}, a batch size of 32 queries, and we sample 7 negative documents for each query. Training typically requires 6 hours on a single NVIDIA A40 GPU.

[6] https://mlat.uzh.ch

Table 2
Performance comparison of existing BERT-based models for Latin on the W_VULG and S_VL corpora, using either the CLS token or the mean of all tokens in the sentence to compute similarities. All results are reported without fine-tuning the embedding model.

		Corpus: **W_VULG**					Corpus: **S_VL**				
Model	**Aggregation**	R@1	R@2	R@3	R@5	R@10	R@1	R@2	R@3	R@5	R@10
Latin RoBERTa [20]	CLS Token	13.0	13.0	13.5	13.5	14.1	4.7	6.5	8.2	8.2	10.0
Latin RoBERTa [20]	Token Averaging	**18.2**	**20.3**	**23.4**	**27.6**	**30.2**	**15.9**	**17.6**	**18.2**	**20.6**	**23.5**
Latin BERT [19]	CLS Token	18.2	24.5	26.6	28.1	29.7	18.8	24.1	28.2	32.9	34.1
Latin BERT [19]	Token Averaging	**33.3**	**38.0**	**41.7**	**44.8**	**47.9**	**35.3**	**39.4**	**42.9**	**45.3**	**48.8**
LaBERTa [21]	CLS Token	31.3	39.1	**44.3**	47.4	**55.7**	29.4	34.7	**40.0**	**45.3**	**50.6**
LaBERTa [21]	Token Averaging	**34.4**	**40.6**	43.8	**47.9**	52.6	**33.5**	**37.6**	**40.0**	43.5	47.6

Table 3
Performance comparison of Latin BERT [19] and LaBERTa [21] with different fine-tuning strategies on the W_VULG and S_VL corpora, including results with and without hard negatives.

		Corpus: **W_VULG**					Corpus: **S_VL**				
Model	**Fine-tuning**	R@1	R@2	R@3	R@5	R@10	R@1	R@2	R@3	R@5	R@10
Latin BERT [19]	-	33.3	38.0	41.7	44.8	47.9	35.3	39.4	42.9	45.3	48.8
Latin BERT [19]	w/o Hard Neg.	38.5	46.9	52.1	**55.2**	58.9	35.9	42.3	**47.1**	49.4	54.1
Latin BERT [19]	w/ Hard Neg.	**47.4**	**51.6**	**54.2**	**55.2**	**59.9**	**38.8**	**42.9**	45.3	**50.0**	**55.3**
LaBERTa [21]	-	34.4	40.6	43.8	47.9	52.6	33.5	37.6	40.0	43.5	47.6
LaBERTa [21]	w/o Hard Neg.	41.1	50.0	**54.2**	**59.4**	**64.6**	37.1	45.3	**48.8**	**57.6**	**62.4**
LaBERTa [21]	w/ Hard Neg.	**43.2**	**50.5**	52.1	56.3	63.5	**41.8**	**45.9**	48.2	55.3	61.2

4.2. Evaluating BERT-based Embedding Models for Latin

Impact of Token Aggregation Strategies. Table 2 provides an in-depth comparison of the three pre-trained BERT-based models for Latin considered in this study, evaluated on the W_VULG and S_VL corpora. These evaluations assess their ability to retrieve the correct biblical passage corresponding to a query without any task-specific fine-tuning. Performance is measured using Recall at top-k (R@k) for $k \in \{1, 2, 3, 5, 10\}$. As described in Sec. 3.1, the analysis explores two distinct strategies for aggregating token embeddings into fixed-size representations: the CLS token and token averaging.

As it can be seen, token averaging consistently demonstrates its utility by capturing finer-grained information distributed across all tokens in a sequence, leading to substantial performance improvements for almost all models on both W_VULG and S_VL. Among the three evaluated models, Latin BERT and LaBERTa are the most effective configurations across both corpora, achieving the highest recall scores in most scenarios and surpassing the performance of Latin RoBERTa by a consistent margin. Therefore, in the rest of the paper, we focus on the Latin BERT and LaBERTa models and report fine-tuning results using token averaging as aggregation strategy.

Effect of Fine-tuning and Self-Hard Negative Mining. Table 3 presents a performance comparison of Latin BERT and LaBERTa models with different fine-tuning strategies. The results clearly demonstrate that fine-tuning significantly enhances retrieval performance, and the addition of hard negatives further boosts effectiveness across all settings, particularly for R@1 which is critical for precise retrieval tasks.

Without fine-tuning, both Latin BERT and LaBERTa show moderate performance, with R@1 values below 35% for both corpora. Fine-tuning without hard negatives consistently improves the retrieval accuracy. For instance, Latin BERT improves from an R@1 of 33.3% to 38.5% on W_VULG, while LaBERTa increases from 34.4% to 41.1%. Similar trends are observed on S_VL, with notable gains across other recall metrics as well. This highlights the importance of adapting pre-trained models to the specific task of retrieving intertextual references.

The inclusion of hard negatives during fine-tuning further enhances performance across all metrics, confirming the effectiveness of this strategy. Latin BERT achieves the highest gains, with R@1 reaching

Table 4
Performance comparison of BERT-based models, with and without fine-tuning, across various subsets of the W_VULG and S_VL corpora based on annotated similarity scores.

		All			0.0-0.25			0.25-0.5			0.5-0.75			0.75-1.0		
Model	**Fine-tuning**	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Corpus: **W_VULG**																
Latin RoBERTa [20]	✗	18.2	27.6	30.2	0.0	0.0	0.0	8.0	18.0	18.0	34.7	50.0	54.3	33.3	46.7	53.3
Latin BERT [19]	✗	33.3	44.8	47.9	1.9	1.9	5.9	20.0	34.0	38.0	60.8	69.6	73.9	55.5	**80.0**	80.0
LaBERTa [21]	✗	34.4	47.9	52.6	0.0	5.8	9.8	20.0	42.0	48.0	63.0	71.7	78.3	60.0	77.7	80.0
Latin BERT [19]	✓	**47.4**	55.2	59.9	5.9	15.7	25.5	**40.0**	**52.0**	**52.0**	**73.9**	**78.3**	**82.6**	**75.6**	**80.0**	**84.4**
LaBERTa [21]	✓	43.2	**56.3**	**63.5**	**15.7**	**31.4**	**41.2**	26.0	46.0	50.0	69.6	73.9	**82.6**	66.7	77.8	**84.4**
Corpus: **S_VL**																
Latin RoBERTa [20]	✗	15.9	20.6	23.5	0.0	0.0	0.0	0.0	4.3	4.3	5.0	10.0	15.0	31.3	38.6	42.2
Latin BERT [19]	✗	35.3	45.3	48.8	0.0	2.3	2.3	4.3	8.7	13.0	40.0	45.0	50.0	61.4	78.3	83.1
LaBERTa [21]	✗	33.5	43.5	47.6	0.0	0.0	4.5	8.7	26.1	30.4	**45.0**	50.0	**55.0**	55.4	69.9	73.5
Latin BERT [19]	✓	38.8	50.0	55.3	0.0	6.8	15.9	8.7	17.4	21.7	40.0	45.0	**55.0**	**67.5**	**83.1**	**85.5**
LaBERTa [21]	✓	**41.8**	**55.3**	**61.2**	**9.1**	**27.3**	**40.9**	**13.0**	**34.8**	**34.8**	40.0	**55.0**	**55.0**	**67.5**	75.9	80.7

47.4% on W_VULG and 38.8% on S_VL. LaBERTa also benefits significantly, improving R@1 to 43.2% on W_VULG and 41.8% on S_VL. These results underline the role of hard negatives in refining the ability of the models to distinguish between closely related and unrelated documents.

Analyzing Performance at Higher Reference Difficulty Levels. Table 4 reports the performance of models with and without fine-tuning at varying levels of difficulty, quantified as the similarity between a query and its referred biblical passage. The lowest similarity range (*i.e.*, 0.0-0.25) corresponds to the hardest queries with low text overlap concerning the biblical passage. In this range, models struggle to identify corresponding passages, with recall scores close to zero when not fine-tuned. These results underscore the challenge of detecting loosely referred passages. However, fine-tuning significantly improves the models, particularly LaBERTa, which achieves a recall of 15.7% on the W_VULG corpus and 9.1% on S_VL. In the mid-similarity ranges (*i.e.*, 0.25-0.5 and 0.5-0.75), performance sees a substantial boost, with fine-tuned versions of Latin BERT and LaBERTa achieving notably higher recall scores. For instance, in the 0.5-0.75 range, LaBERTa reaches 69.6% on W_VULG and 40.0% on S_VL. In the highest similarity range (*i.e.*, 0.75-1.0), models perform the best, with fine-tuned versions of Latin BERT and LaBERTa achieving R@1 scores close to or above 70% for both corpora. This analysis suggests that while models excel at identifying exact or near-exact matches, their performance decreases significantly as the references become less direct, though fine-tuning helps mitigate this challenge.

5. Conclusion

In this paper, we demonstrated the effectiveness of BERT-based models in capturing intertextual references within Latin texts, with a particular focus on patristic commentaries. By employing a fine-tuning strategy that incorporates hard-negative mining, we achieved significant performance improvements across both the W_VULG and S_VL corpora. The experimental results showcase the ability of models fine-tuned with the proposed strategy to handle references with varying degrees of lexical overlap, including implicit allusions that present particular challenges. These results underscore the potential of Transformer-based approaches for Latin NLP tasks and provide a solid foundation for future research in historical text analysis, bridging computational methods with philological expertise.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. This work was supported by the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU (CUP B53C22001770006).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[4] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: Advances in Neural Information Processing Systems, 2019.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942 (2019).

[6] R. Nogueira, K. Cho, Passage Re-ranking with BERT, arXiv preprint arXiv:1901.04085 (2019).

[7] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: Proceedings of the International Conference on Learning Representations, 2020.

[9] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, Transactions of the Association for Computational Linguistics 8 (2021) 842–866.

[10] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, R. Cucchiara, The Revolution of Multimodal Large Language Models: A Survey, in: Findings of the Annual Meeting of the Association for Computational Linguistics, 2024.

[11] T. Pires, E. Schlinger, D. Garrette, How Multilingual is Multilingual BERT, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2019.

[12] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for Finnish, arXiv preprint arXiv:1912.07076 (2019).

[13] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: Proceedings of the Italian Conference on Computational Linguistics, 2019.

[14] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020.

[15] T. Sommerschield, Y. Assael, J. Pavlopoulos, V. Stefanak, A. Senior, C. Dyer, J. Bodel, J. Prag, I. Androutsopoulos, N. de Freitas, Machine learning for ancient languages: A survey, Computational Linguistics 49 (2023) 703–747.

[16] E. Manjavacas, L. Fonteyn, Adapting vs. Pre-Training Language Models for Historical Languages, Journal of Data Mining & Digital Humanities (2022).

[17] A. Palmero Aprosio, S. Menini, S. Tonelli, BERToldo, the Historical BERT for Italian, in: Proceedings of the Workshop on Language Technologies for Historical and Ancient Languages, 2022.

[18] B. Hutchinson, Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing, in: Findings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2024.

[19] D. Bamman, P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, arXiv preprint arXiv:2009.10053 (2020).

[20] P. B. Ströbel, RoBERTa Base Latin Cased v1, 2022. URL: https://huggingface.co/pstroe/roberta-base-latin-cased.

[21] F. Riemenschneider, A. Frank, Exploring Large Language Models for Classical Philology, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2023.

[22] J. Zycha (ed.), Sancti Aureli Augustini: De Genesi ad litteram libri duodecim eiusdem libri capitula. De Genesi ad litteram imperfectus liber. Locutionum in Heptateuchum libri septem, Pragae-Vindobonae-Lipsiae, Tempsky-Freyta, 1894.

[23] R. Weber, R. Gryson (eds.), Biblia Sacra iuxta Vulgatam Versionem, Stuttgart, Deutsche Bibelgesellschaft, 2007[5] (R. Weber, 1969[1]).
[24] P. Sabatier (ed.), Bibliorum Sacrorum latinae versiones antiquae seu Vetus Italica (3 vols.), Reims, Reginaldus Florentain, 1743–1751.
[25] J.-C. Klie, M. Bugert, B. Boullosa, R. E. De Castilho, I. Gurevych, The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation, in: Proceedings of System Demonstrations of the International Conference on Computational Linguistics, 2018.
[26] M. Cornia, M. Stefanini, L. Baraldi, M. Corsini, R. Cucchiara, Explaining Digital Humanities by Aligning Images and Textual Descriptions, Pattern Recognition Letters 129 (2020) 166–172.
[27] N. Messina, M. Stefanini, M. Cornia, L. Baraldi, F. Falchi, G. Amato, R. Cucchiara, ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval, in: Proceedings of the International Conference on Content-based Multimedia Indexing, 2022.
[28] S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara, Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023.
[29] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara, Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models, in: Proceedings of the European Conference on Computer Vision, 2024.
[30] N. Moratelli, D. Caffagni, M. Cornia, L. Baraldi, R. Cucchiara, Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization, in: Proceedings of the British Machine Vision Conference, 2024.
[31] A. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv preprint arXiv:1807.03748 (2018).
[32] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised Dense Information Retrieval with Contrastive Learning, Transactions on Machine Learning Research (2022).
[33] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, et al., Text and Code Embeddings by Contrastive Pre-Training, arXiv preprint arXiv:2201.10005 (2022).
[34] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, in: Proceedings of the International Conference on Machine Learning, 2020.
[35] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, in: Advances in Neural Information Processing Systems, 2020.
[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the International Conference on Machine Learning, 2021.
[37] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, in: Proceedings of the British Machine Vision Conference, 2018.
[38] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard Negative Mixing for Contrastive Learning, in: Advances in Neural Information Processing Systems, 2020.
[39] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, S. Ma, Optimizing dense retrieval model training with hard negatives, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.
[40] L. Baraldi, M. Cornia, C. Grana, R. Cucchiara, Aligning Text and Document Illustrations: Towards Visually Explainable Digital Humanities, in: Proceedings of the International Conference on Pattern Recognition, 2018.
[41] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020.
[42] D. P. Kingma, J. L. Ba, ADAM: a Method for Stochastic Optimization, in: Proceedings of the International Conference on Machine Learning, 2015.