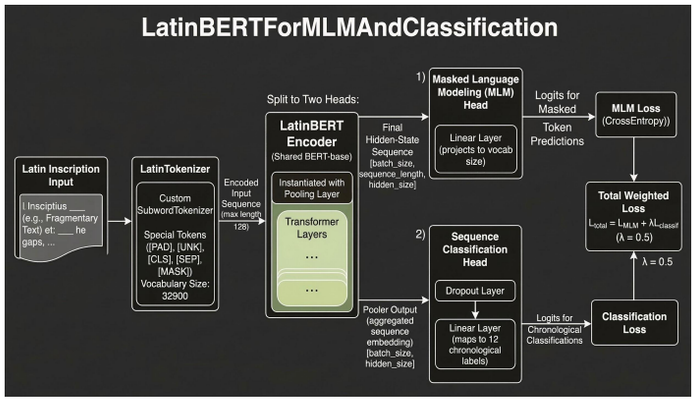
Language Modeling for Epigraphs: a BERT model for EDR's Latin Epigraphs text completion

Olmo Ceriotti (ceriotti.2193258@studenti.uniroma1.it)
Federico Gerardi (gerardi.1982783@studenti.uniroma1.it)
Saverio Giulio Malatesta (saveriogiulio.malatesta@uniroma1.it)
Silvia Orlandi (silvia.orlandi@uniroma1.it)

Introduction The reconstruction of ancient Latin inscriptions is a fundamental task in epigraphy, yet it remains challenging due to the physical degradation of stone and metal over centuries. Traditionally, epigraphers have relied on manual methods to fill "lacunae"—missing segments of text—which are time-consuming and prone to human error. This paper introduces a specialized deep learning model, LatinBERTForMLMAndClassification, developed by researchers at Sapienza University of Rome. Built upon the foundational LatinBERT model, it leverages the Epigraphic Database Roma (EDR), a comprehensive collection of over 100,000 Roman inscriptions. By employing a multi-task learning objective, the model simultaneously performs Masked Language Modeling (MLM) to predict missing words and chronological classification to estimate the date of an inscription. This dual approach allows the model to capture diachronic linguistic variations, achieving a Top-1 accuracy of 63.7%—a significant improvement over previous benchmarks.



Conclusion and Future Directions The LatinBERTForMLMAndClassification model marks a substantial advancement in computational epigraphy. By fine-tuning on the EDR dataset, the researchers achieved a perplexity of 8.3 and a Top-1 accuracy of 63.7% for text completion. This significantly outperforms previous attempts, such as the 4.02% accuracy reported in prior case studies. A vital finding was the success of the multi-task setup; the auxiliary dating task (68.8% accuracy) provided a synergistic effect that enhanced linguistic reconstruction. While challenges remain in predicting exact numerical values, future work will focus on incorporating granular metadata, such as provenance, and expanding the model to handle longer missing sequences. Ultimately, this research provides a robust framework for automating the preservation and study of invaluable historical documents.

Methodologies and Data Preprocessing The methodology involves a rigorous integration of data curation, custom tokenization, and a multi-headed neural network architecture.

Dataset and Corpus Manipulation The researchers utilized the EDR, cleaning an initial 114,365 inscriptions down to a refined dataset of 82,534 records. Preprocessing was essential to transform raw annotations into clean input:

- Lacunae: Unrestorable text was replaced with an [UNK] token.
- Reconstructions: Tags were removed while retaining reconstructed words to train the model on accurate restorations.
- Abbreviations: Shorthand was expanded into full word forms (e.g., "P(ublius)" to "publius") to reduce task complexity.
- Standardization: Gross misspellings were corrected to standardized counterparts to avoid unnecessary noise.

Architecture The model integrates two task-specific heads atop a shared LatinBERT encoder:

- MLM Head: A linear layer that predicts masked tokens by projecting the encoder's final hidden states to the vocabulary size.
- Sequence Classification Head: This head processes the pooler output (the [CLS] token's hidden state) to assign the epigraph to one of 12 chronological periods, ranging from 5 BC to 5 AC.

Training Procedure Input text was processed via a custom LatinTokenizer using subword units to handle complex morphology. The model was fine-tuned using the AdamW optimizer with a learning rate of 7e-5 over 9.0 epochs. The total loss was a weighted sum of the individual task losses, using a \lambda value of 0.5 to balance the chronological classification with the primary MLM objective.