Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Master Thesis

Enhancing Knowledge Extraction from Violent Ancient Historical Texts through Fine-Tuned Large Language Models and Historical Databases

submitted by

Alhassan Ahmed Said Abdelhalim

MIN-Faculty

Department of Informatics

Study program: Intelligent Adaptive Systems

Matriculation Number: 7594783

Submission Date: 18.11.2024

First Reviewer: Dr. Michaela Regneri

Second Reviewer: Prof. Dr. Sören Laue

Abstract

This thesis explores the application of fine-tuned large language models (LLMs) for automating the detection and classification of violent events in ancient historical texts. The primary research objectives were to determine whether LLMs could accurately identify violent versus non-violent texts and classify them across multiple dimensions of violence, such as level of violence, contextual background, underlying motives, and long-term consequences. Given the significant manual effort involved in annotating historical texts for violence, this research aims to alleviate that burden by automating the process using machine learning models.

The thesis is structured around two core experiments. In the first experiment, we fine-tuned BERT and RoBERTa models on manually annotated examples of violent and non-violent texts. The datasets utilized were derived from historical texts curated in the ERIS and Perseus digital humanities databases. The fine-tuned models demonstrated high accuracy in identifying violent passages, significantly outperforming general-purpose models like GPT-4omini. Fine-tuning these models on domain-specific data led to superior precision and recall scores. In the second experiment, we expanded the scope to multi-class classification, categorizing violent events into four dimensions: level of violence (interpersonal, intrapersonal, intersocial, intrasocial), context (political, military, social, etc.), motive (strategic, emotional, religious, etc.), and long-term consequences (death, conquest, plunder, etc.). The results highlighted the models' ability to handle complex categories with high performance in frequently occurring classes, though challenges persisted in distinguishing more nuanced or low-frequency categories.

Despite the overall success, limitations arose due to the inherent complexity of historical texts. Historians often rely on extra-textual knowledge, that is, insights beyond the text itself, to interpret nuanced contexts. For example, understanding the Roman Senate allows historians to infer that a debate in this setting involves senators, even if not explicitly stated. Similarly, knowledge of cultural norms, like the symbolic meanings of Greek rituals, or awareness of an author's biases, helps historians add depth to their interpretations. Social hierarchies also inform inferences, such as identifying a high-ranking official when a text mentions someone giving orders. These contextual layers are challenging for LLMs, which primarily rely on the explicit content of the text without access to such background knowledge. Such subtle interpretations remain challenging for LLMs, which primarily rely on the text provided. Additionally, the classification of violent acts is difficult even for experts, given the strict criteria used in historical analysis. Computational constraints also restricted the use of models larger than RoBERTa Large, and limited access to expansive annotated datasets posed further challenges.

The thesis concludes that while automation through LLMs can significantly accelerate the annotation process, these models are not replacements for expert human analysis but rather serve as complementary tools. Future work will aim to expand the classification framework to include additional categories, explore larger and more advanced models, and address the cultural and contextual nuances inherent in ancient texts.

Acknowledgment

I would like to express my deepest gratitude to Dr. Michaela Regneri and Prof. Dr. Sören Laue for their exceptional supervision and mentorship throughout this journey. Michaela, thank you for being a great supervisor, a patient listener, and an inspiring mentor. The valuable lessons I've learned from you will stay with me for years to come. Sören, thank you for the numerous opportunities, unwavering support, and the invaluable skills I acquired in both teaching and research. I am sincerely grateful to both of you for everything you have done.

I would also like to extend my heartfelt thanks to the current and former members of the Machine Learning team with whom I had the pleasure of working. Thank you, Tomislav, Hanna, Guanlue, Robin, Merle, and Nina. The Machine Learning group will always hold a special place in my heart. And of course, my sincere appreciation to Anne for all the administrative and personal support you provided.

A special thanks to Prof. Dr. Werner Rieß for generously providing the data and taking the time to meet and discuss history with us. I am also deeply grateful to Justine Diemke for her invaluable guidance with annotations and for agreeing to review my chapter on historical analysis.

Additionally, Thanks to Sri Gowry Sritharan for her diligent work in expanding the dataset from ERIS, which significantly contributed to this research.

Finally, I want to extend my deepest gratitude to my family. Your unwavering support made it possible for me to travel to Germany and pursue this journey. Your encouragement kept me going until the very end. None of this would have been possible without you.

Contents

List of Tables

List of Figures

1. Introduction

In our introductory chapter, we introduce the motivation for our research. We outline the significance of this work, our specific research objectives, and the methodological approaches we employed. Finally, we provide an overview of the thesis structure to guide the reader through the subsequent chapters.

1.1. Motivation

Violence, as a pervasive element in human history, has left indelible marks on societies, shaping cultural values, political structures, and social norms. Understanding the role of violence in shaping ancient civilizations offers valuable insights into how societies evolved, how power was negotiated, and how conflicts were resolved. However, the task of analyzing historical texts to extract information on violent events is both labor-intensive and time-consuming. Traditionally, historians have relied on manual analysis, which involves intensively reading and annotating vast amounts of text to identify instances of violence and extract knowledge from those instances. This process can take several months or even years to complete for some historical works, especially when dealing with extensive collections of ancient manuscripts.
While traditional manual annotation remains the gold standard for extracting nuanced interpretations from historical texts, it is increasingly evident that the sheer volume of ancient manuscripts and the complexities inherent in historical language make this approach highly inefficient. Given the rapid growth of digital archives and the availability of vast historical corpora, there is a pressing need to develop automated methods that can assist historians in extracting relevant information more efficiently.
Large Language Models (LLMs) offer a promising solution to this problem. By leveraging the capabilities of state-of-the-art neural networks, such as BERT, RoBERTa and chatGPT, we can significantly expedite the process of identifying violent events and extracting contextual information from ancient texts. These models not only reduce the time required for analysis but also enable researchers to uncover patterns and insights that may go unnoticed during manual annotation.
This research, therefore, seeks to bridge the gap between historical Research and computational linguistics by automating the annotation of violence in ancient texts. The primary motivation behind this thesis is to develop and evaluate methodologies that can enhance the efficiency of historical analysis while maintaining the depth of understanding traditionally achieved through manual methods. In doing so, we aim to complement, rather than replace, the expertise of historians, allowing them to focus on deeper interpretative tasks rather than spending extensive time on initial data processing.

1.2. Research objectives

This thesis aims to leverage recent advancements in Large Language Models (LLMs) to automate the detection and analysis of violence in ancient texts. By focusing on historical datasets such as ERIS and Perseus, this research explores the capabilities of models like BERT, RoBERTa, and GPT-based models to enhance the efficiency of text annotation processes traditionally performed by historians.

Research Questions

- How can fine-tuned LLMs be employed to detect and classify instances of violence in ancient historical texts?
- To what extent can these models match or complement the accuracy of expert historians in identifying violence and contextual information?

- What are the key challenges and limitations of applying LLMs to ancient datasets, especially given the nuances of historical language and cultural context?
- Can automated techniques uncover patterns or insights that might be overlooked in manual historical analysis?

Objectives

- Develop and fine-tune existing LLMs to perform binary classification of violent and non-violent texts, followed by multi-class classification to extract context, motive, and long-term consequences of violent acts.
- Evaluate the performance of these models on historical datasets, measuring their precision, recall, and overall F1 scores to assess their accuracy compared to expert human annotations.
- Address the challenges of applying LLMs to ancient texts, such as handling limited annotated data and understanding the cultural and linguistic nuances of historical sources.
- Propose methodologies that could enhance the integration of automated tools with traditional historical analysis, aiming to complement the work of expert historians.

Methodological Approach To achieve these objectives, the research is structured around two core experiments:

1. Binary Classification: Fine-tuning BERT and RoBERTa models to distinguish between violent and non-violent passages in ancient texts.
2. Multi-Class Classification: Using the best-performing models to further classify instances of violence into categories based on context, motive, and long-term consequences.

The experiments will include data preparation, model fine-tuning, and thorough evaluation using metrics such as precision, recall, and F1 scores. These methodologies are aimed at addressing the outlined research questions and achieving the stated objectives.

1.3. Thesis Structure Overview

This thesis is organized into six main chapters, each building upon the previous to explore the automation of violence detection in ancient texts using large language models. The following overview outlines the structure of the thesis:

1. **Introduction** The Introductory chapter sets the stage by introducing the motivation behind this research, which aims to bridge the gap between computational linguistics and historical analysis. It presents the research objectives, the methodological approaches taken, and an overview of the thesis structure.

2. **Modern Network Structures and Large Language Models**
   This chapter provides a foundational background in neural network architectures, focusing on the evolution of the Transformer network. It delves into the development and significance of models like GPT, BERT, and RoBERTa, in this chapter we aim to introduce the technical aspects of the experiments conducted in this research.

3. **Ancient Violence from a Historical and Computational Perspective**
   The third chapter explores the historical context of violence as depicted in ancient texts. It examines cultural norms, power dynamics, and conflict resolution, emphasizing the importance of detecting and analyzing violence for historical research. Additionally, it introduces the challenges of annotating ancient texts and highlights how large language models can assist historians in this labor-intensive task.

4. **Violence Detection Using Large Language Models**
   The fourth chapter focuses on the first core experiment: binary classification of violent versus non-violent texts. It covers the dataset preparation, model selection, and fine-tuning procedures. The evaluation metrics and results are presented to demonstrate the effectiveness of the models in automating the detection of violent content within historical texts.

5. **Knowledge Extraction Using Large Language Models**
   Building upon the results of the binary classification task, this chapter extends the research to multi-class classification. It involves categorizing detected violent texts into specific subcategories, such as the level of violence, context, motive, and long-term consequences. This section explores the models' performance in extracting deeper insights from ancient texts, showcasing their ability to capture complex historical information.

6. **Conclusion**
   The final chapter summarizes the key findings, discusses the limitations of the current research, and proposes directions for future work. It reflects on the potential of large language models to significantly reduce the time and effort required for manual annotation, ultimately supporting historians in their analytical endeavors rather than replacing them.

The thesis is designed to guide the reader progressively, from understanding the technical foundations of large language models to applying them in the historical analysis of ancient texts.

2. Modern Network Structures and Large Language Models

The rise of deep learning has revolutionized numerous fields, with Natural Language Processing (NLP) standing out as a key beneficiary of these advancements. Over the past decade, the development of neural network architectures has significantly transformed our ability to process and understand human language. Among these advancements, the Transformer network has emerged as the foundation for contemporary NLP, providing the structural basis for large language models (LLMs). These models, capable of capturing nuanced linguistic patterns, have set a new standard in tasks ranging from translation to sentiment analysis. The development of Transformer-based architectures represents a paradigm shift in NLP, moving from earlier neural networks like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to models that leverage self-attention mechanisms. This evolution has enabled models to overcome the limitations associated with sequence length and contextual understanding, challenges that once constrained earlier generations of NLP systems.
This chapter provides a comprehensive overview of the fundamental architectures and innovations that paved the way for modern LLMs. It begins by tracing the historical development of deep learning models, leading to the introduction of the Transformer network and its widespread adoption. We will then examine key LLMs, including GPT, BERT, and RoBERTa, which have pushed the boundaries of what is possible in NLP. These models will serve as the technical foundation for our experiments on violence detection in ancient texts. In the end of the chapter, an overview of the thesis is provided, outlining the structure and flow of the research. This is intended to serve as a roadmap for readers to guide them through the technical foundations laid in this chapter to the historical and computational analysis presented in later chapters.

2.1. The Emerging and Evolution of the Transformer Network

Transformers [4] are the new state-of-the-art network structure for solving natural language processing. Before 2017, research areas in deep learning were dominated by two types of neural network structures which are recurrent neural networks(RNNs) [14] and Convolutional neural networks (CNN) [1].

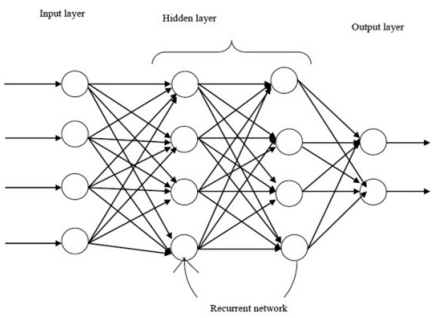


Figure 2.1.: RNN structure[1].

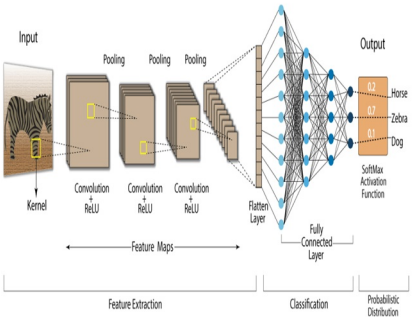


Figure 2.2.: CNN structure[2].

Variants of these network structures, such as Long Short-Term Memory (LSTM) networks, addressed some of the inherent limitations of earlier neural networks. However, Recurrent Neural Networks (RNNs) consistently struggled with natural language processing (NLP), particularly when handling long sequences of data [15]. This difficulty stems from the vanishing or exploding gradient problem, which occurs when the recursive output is either too large or too small [16]. While LSTMs performed effectively with shorter data sequences, they encountered significant challenges with longer sequences. Training LSTMs on extensive text data is

considerably harder, and transfer learning becomes ineffective without task-specific labeled datasets [17]. Another approach was the adaptation of Convolutional Neural Networks (CNNs) for NLP, offering advantages such as parallelism and the ability to leverage the ReLU activation function [18]. However, CNNs have a notable limitation in NLP tasks, as they process a fixed-sized input window, limiting their ability to capture context across an entire document.

2.1.1. Machine Translation: First Approach to Attention

In 2013, Kalchbrenner et al.[19] introduced a machine translation model that laid the groundwork for Bahdanau et al.'s[3] 2014 innovation: the concept of content-based neural attention. Attention is a mechanism positioned between the encoder and decoder, enabling the decoder to access information from every hidden state of the encoder. This allows the model to "attend" to the most relevant parts of the input sequence, improving its ability to learn from inputs and address the challenge of processing long sequences[20]. The key idea behind this approach is that each word in the input sequence can attend to all other elements in the output sequence, as depicted in Figure2.3.

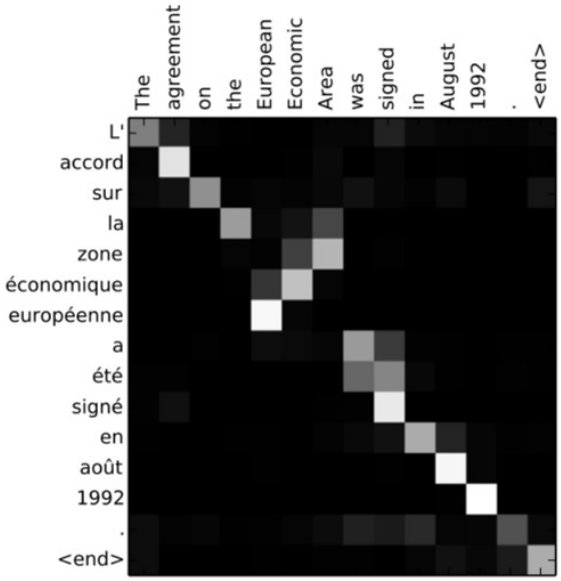


Figure 2.3.: How words are attended [3].

The attention mechanism works as follows:

- For every output position i, you generate a query q_i.
- So we have a set of queries $\{q_1, \ldots, q_{n_q}\}$, each of size d_k.
- For every input position j, you generate a key k_j.
- So we have a set of keys $\{k_1, \ldots, k_{n_k}\}$, also of size d_k (same size as the queries).
- We also have a set of values $\{v_1, \ldots, v_{n_k}\}$, each of size d_v. For simplicity, d_v has the same dimension as d_k.

Stacking all of these vectors into matrices gives:

$$Q = \underbrace{\begin{bmatrix} q_1^\top \\ \vdots \\ q_{n_q}^\top \end{bmatrix}}_{n_q \times d_k}, \quad K = \underbrace{\begin{bmatrix} k_1^\top \\ \vdots \\ k_{n_k}^\top \end{bmatrix}}_{n_k \times d_k}, \quad V = \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v} \tag{2.1}$$

We compute the relevance scores by taking the dot product between queries and keys. The output is calculated as:

$$\text{out} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.2}$$

We perform a weighted average on the values V using the attention weights derived from the relevance scores:

$$\underbrace{\begin{bmatrix} w_1^{(1)} & \dots & w_{n_k}^{(1)} \end{bmatrix}}_{1\times n_k} \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v} = \sum_{i=1}^{n_k} w_i^{(1)} v_i^\top \tag{2.3}$$

2.1.2. Attention Nets: The Emergence of the Transformer

In 2017, Vaswani et al. published a breakthrough paper that expanded the concept of attention by adding the element of self-attention [4]. The new model introduced in this paper is called the Transformer, which became the basis for many future foundation models [21]. In their proposed model, the Query, Key, and Value vectors are split into 8 heads (64-dimensional vectors), as shown in Figure 2.4, and the attention layer is applied to each head in the same way as before. This allows the network to learn 8 different semantic aspects of attention, such as grammar, vocabulary, conjugation, synonyms, and so on. This means the Transformer can build up an internal understanding of words.

First, the output of the self-attention is given by the equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.4}$$

Expanding the above equation gives:

$$\begin{aligned}
\text{Attn}(Q, K, V) &= \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \\
&= \text{softmax}\left(\frac{1}{\sqrt{d_k}} \underbrace{\begin{bmatrix} q_1^\top \\ \vdots \\ q_{n_q}^\top \end{bmatrix}}_{n_q \times d_k} \underbrace{\begin{bmatrix} k_1 & \dots & k_{n_k} \end{bmatrix}}_{d_k \times n_k}\right) \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v} \\
&= \text{softmax}\left(\frac{1}{\sqrt{d_k}} \underbrace{\begin{bmatrix} q_1^\top k_1 & \dots & q_1^\top k_{n_k} \\ \vdots & \ddots & \vdots \\ q_{n_q}^\top k_1 & \dots & q_{n_q}^\top k_{n_k} \end{bmatrix}}_{n_q \times n_k}\right) \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v} \\
&= \underbrace{\begin{bmatrix} \text{softmax}\left(\frac{q_1^\top k_1}{\sqrt{d_k}}, \dots, \frac{q_1^\top k_{n_k}}{\sqrt{d_k}}\right) \\ \vdots \\ \text{softmax}\left(\frac{q_{n_q}^\top k_1}{\sqrt{d_k}}, \dots, \frac{q_{n_q}^\top k_{n_k}}{\sqrt{d_k}}\right) \end{bmatrix}}_{n_q \times n_k} \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v}
\end{aligned} \tag{2.5}$$

Here, the softmax operator is applied row-wise. The dot products between queries and keys are scaled by $1/\sqrt{d_k}$, which is why this is referred to as scaled dot-product attention.

It's important to note that the weighted sum in this sub-module is derived from the matrix-matrix product mentioned earlier. To better understand this, consider focusing on the first row of the query-key dot product matrix. By expressing the elements of the softmax row vector as $\{w_1^{(1)}, \ldots, w_{n_k}^{(1)}\}$, we can simplify the notation and arrive at the following formula:

$$\underbrace{\begin{bmatrix} w_1^{(1)} & \ldots & w_{n_k}^{(1)} \end{bmatrix}}_{1 \times n_k} \underbrace{\begin{bmatrix} v_1^\top \\ \vdots \\ v_{n_k}^\top \end{bmatrix}}_{n_k \times d_v} = \sum_{i=1}^{n_k} w_i^{(1)} v_i^\top \tag{2.6}$$

Here, we get the weighted sums. It should be noted that the values $w_i^{(1)}$ are scalars, while the values $v_i^\top$ are vectors of size d_v.

The basic concept is that each query is represented as a single row in a matrix. To handle multiple queries, additional rows are added to the matrix. The extensions build on this basic concept by using trainable parameters to linearly project the Q, K, and V matrices, similar to a dense layer. Additionally, several of these operations are performed in parallel and then combined along one axis.

The first component of an encoder or decoder layer is as described above. The second component is a simple fully connected layer given by:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.7}$$

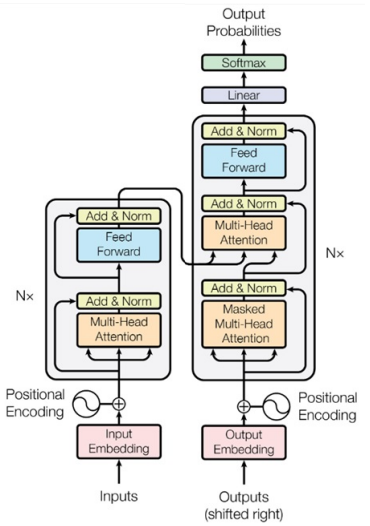


Figure 2.4.: Architecture of the vanilla Transformer [4].

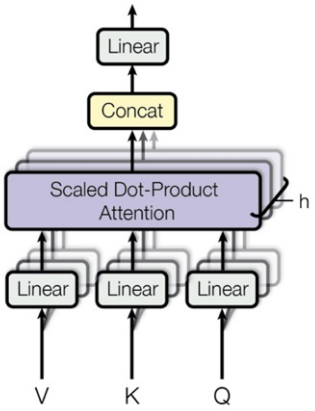


Figure 2.5.: Input, output, and value are split into multiple heads [4].

The Transformer combines the self-attention component and the fully connected layer with residual connections and layer normalization. This process is repeated for several layers, with six layers being used for both the encoder and decoder in the original paper.

2.1.3. Variants of Transformer

The Transformer architecture marked a significant innovation that various fields in AI rapidly sought to adopt. Initial adaptations occurred in computer vision, with research groups attempting to integrate attention mechanisms into convolutional neural networks (CNNs) [22], [23], or to remove CNNs entirely [24]. However, these approaches often modified the Transformer architecture too drastically, resulting in poor scalability. A key breakthrough came with the introduction of the Vision Transformer (ViT) [5], which was designed specifically for image processing and yielded results comparable to state-of-the-art computer vision algorithms. The ViT works by dividing an image into patches, flattening those patches, applying positional

embeddings, feeding them into the encoder, and pre-training the model. These steps are illustrated in Figure 2.6.

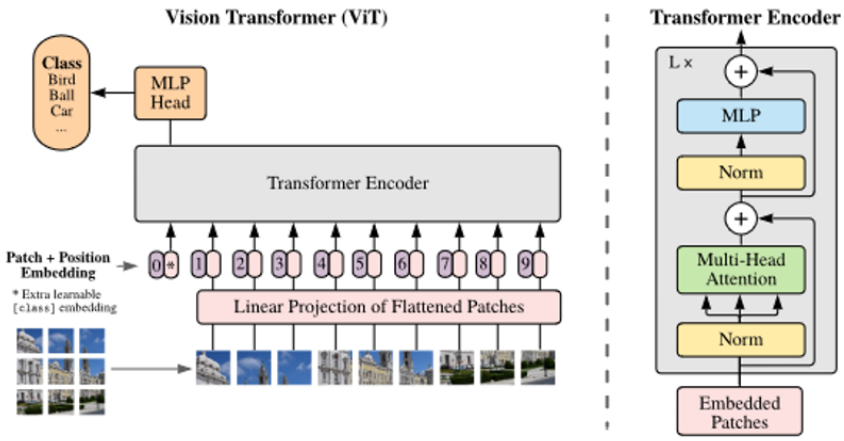


Figure 2.6.: Vision Transformer model structure [5].

The ViT was initially trained on the ImageNet-21K dataset [25] and fine-tuned on multiple datasets such as COCO (Common Objects in Context) [26] and the Visual Genome dataset [27]. An example of ViT's performance can be seen in Figure 2.8, where the transformer attends to key features of the input image.

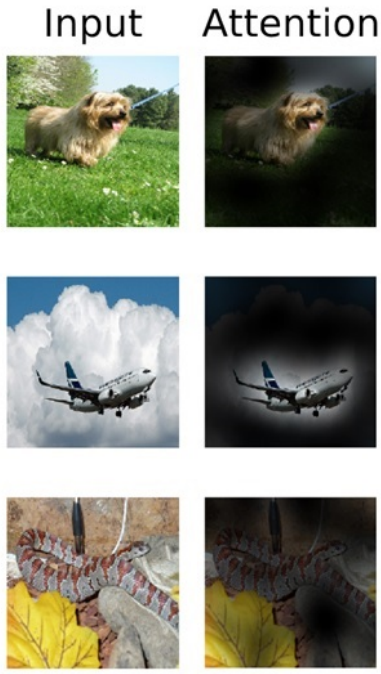


Figure 2.7.: Results of the ViT Transformer [5].

One downside of the ViT is that it requires an enormous amount of data for effective training. However, with sufficient data, the ViT has been shown to outperform state-of-the-art CNNs [28]. A variety of other transformer models are highlighted in Figure 2.8.

2.2. Large Language Models

Large Language Models (LLMs) have become a cornerstone of natural language processing, enabling sophisticated text understanding and generation capabilities. They rely on transformers trained on vast amounts of text data, allowing them to grasp linguistic patterns and contextual meanings across diverse text corpora.

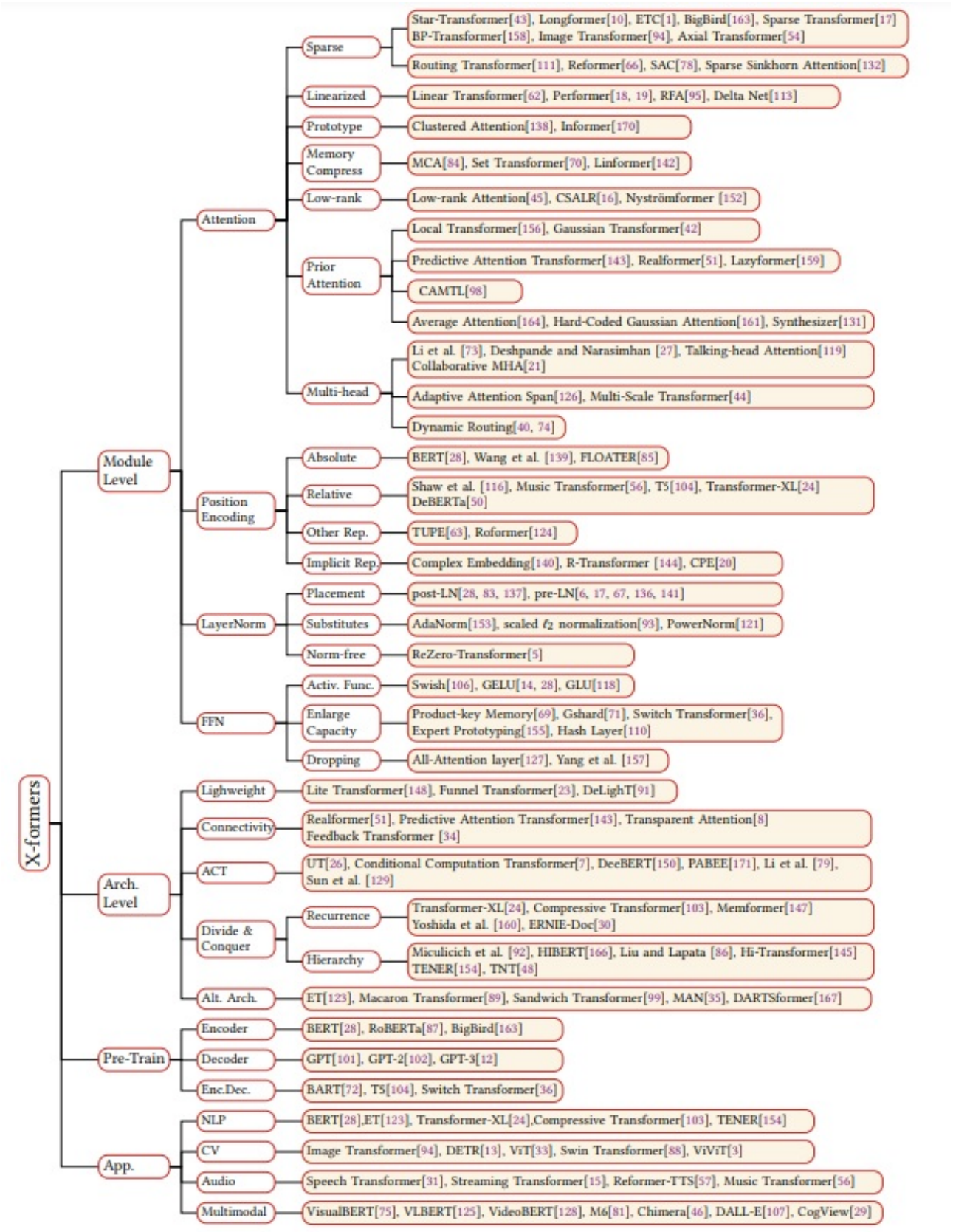


Figure 2.8.: Different types of transformers from a 2022 survey [6]

This section focuses on three influential LLMs; GPT, BERT, and RoBERTa. These models have transformed the field of NLP and are now the standard for NLP tasks.

2.2.1. GPT: Generative Pre-trained Transformer

The GPT series, developed by OpenAI, represents a significant leap in language modeling by introducing a transformer-based architecture optimized for generative tasks. GPT [29] and its successors (GPT-2 and GPT-3) employ a unidirectional language model that processes text from left to right. This approach enables GPT to predict the next word in a sequence, making it particularly effective for tasks like text generation, machine translation, and question answering.

GPT's underlying architecture is based on the transformer model. However, unlike models such as BERT, GPT focuses only on one direction—predicting future words based on prior context. This has proven useful in creative applications where generating coherent, contextually accurate text is crucial [30].

While GPT's generative capabilities are impressive, its lack of bidirectionality can limit its understanding of ambiguous or complex text structures. This led to the development of models like BERT, which are designed to overcome some of these limitations by incorporating bidirectional context.

Later Developments in GPT: With the release of GPT-3 in July 2020, OpenAI introduced three models of increasing size, with parameter counts of 1B, 6.7B, and 175B, respectively named Babbage, Curie, and Davinci[31]. Each of these models exhibited remarkable scaling capabilities, making GPT-3 the most powerful model at the time. The sheer size of GPT-3 allowed it to perform tasks with minimal fine-tuning, showcasing its few-shot learning capabilities.

In 2021, OpenAI published Codex, a specialized GPT model targeted for programming applications. Codex was developed by fine-tuning a 12B parameter version of GPT-3 using code from GitHub. This model powers tools like GitHub Copilot, which assists programmers by generating code based on natural language input.

In March 2022, OpenAI released two instruction-tuned versions of GPT-3: *davinci-instruct-beta* and *text-davinci-001*, which were fine-tuned to follow human instructions. These models represented a shift in the development of GPT models, focusing on improved interaction with human prompts and yielding better alignment with user needs.

GPT-4 and Multimodal Capabilities: GPT-4, released in March 2023, marked the next milestone in the evolution of generative models. While the size and training details of GPT-4 were not disclosed, it introduced significant improvements, particularly its ability to process both text and image inputs, making it a multimodal model. GPT-4's ability to handle different data types opened up new possibilities for applications in fields such as image recognition, language translation, and chatbot interactions.

The release of GPT-4 also coincided with the development of reinforcement learning from human feedback (RLHF) to refine its conversational abilities. This technology was integrated into OpenAI's ChatGPT, which builds on the capabilities of GPT-3.5 and GPT-4 for a more conversational, human-like interaction experience.

o1-Model and A New Breakthrough in AI Efficiency: In 2024, OpenAI introduced the **o1** model[32], a transformative step in AI architecture aimed at reducing both computational cost and energy consumption. The o1-Model is designed to achieve constant time inference, regardless of model size, a significant departure from traditional models where inference time scales with the number of parameters.

o1 achieves this by leveraging innovative techniques in model compression, tokenization, and sparse attention mechanisms. This allows the model to maintain high performance on generative tasks while dramatically improving efficiency. The model is expected to open new doors for deploying large language models in resource-constrained environments, such as mobile devices or edge computing.

o1's introduction signals a new era for AI scalability, where the trade-offs between model size and computational efficiency can be minimized without sacrificing performance. This breakthrough could have far-reaching implications for democratizing access to large language models, enabling broader adoption across industries and making advanced AI tools more accessible.

2.2.2. BERT: Bidirectional Encoder Representations from Transformers

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. [33], revolutionized the field of natural language processing (NLP) by introducing a bidirectional mechanism to understand text, which marked a significant departure from previous unidirectional models. Prior to BERT, models like GPT (Generative Pre-trained Transformer) read text in a unidirectional manner, either left-to-right or right-to-left. This unidirectional processing inherently limited the model's ability to fully capture the context of a word based on its surrounding text. In contrast, BERT processes words in both directions simultaneously, allowing it to consider the entire context before predicting or classifying a word. This bidirectional feature is particularly powerful for tasks where the meaning of a word or phrase depends heavily on the surrounding words, such as named entity recognition[34], question answering[35], and text classification[36], where disambiguating meaning is crucial.

BERT's architecture is built upon the Transformer encoder, which allows the model to attend to different parts of the input text using self-attention mechanisms. This self-attention mechanism helps BERT efficiently capture long-range dependencies between words in a sentence, making it highly effective for understanding context-rich language tasks. The encoder's ability to consider relationships between words at various distances in the text allows BERT to better understand nuances in language, making it well-suited for complex linguistic tasks.

The pre-training of BERT relies on two tasks: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, a percentage of the input tokens are randomly masked, and the model is trained to predict the original masked words based on the surrounding context. This task forces BERT to build a deep understanding of the relationships between words and sentences. For example, if the word "cat" is masked in the sentence "The ... sat on the mat," BERT must learn from the context to predict that the missing word is "cat." This bidirectional learning through MLM contrasts with models like GPT, which predict the next word sequentially and do not have access to both past and future context during training.

Next Sentence Prediction (NSP) is another critical component of BERT's pre-training. In NSP, BERT is given pairs of sentences and is tasked with determining whether the second sentence follows the first in a natural sequence. This enables the model to capture a higher-level understanding of sentence relationships and coherence. For example, if presented with the sentences "John went to the store" and "He bought some apples", BERT learns to infer the logical sequence between them. NSP helps in tasks where sentence order and flow are important, such as in document-level understanding and question-answering tasks.

BERT's pre-training on these dual tasks results in a model that has a robust grasp of language structure and semantics, making it highly effective across a wide range of NLP applications. Fine-tuning BERT on task-specific datasets enables it to achieve state-of-the-art performance in various downstream tasks, such as sentiment analysis, machine translation, and summarization. The fine-tuning process is efficient because BERT's pre-trained language representations can be adapted to new tasks with relatively small task-specific datasets, compared to training a model from scratch.

Since its introduction, BERT has been widely adopted in both academic research and industry applications. One reason for its widespread impact is its ability to significantly improve performance in tasks requiring deep contextual understanding. For instance, in question-answering systems, BERT's ability to grasp the context of the question and the document from which the answer is derived leads to more accurate answers. Additionally, in named entity recognition (NER), BERT can disambiguate entities based on the surrounding text, improving accuracy in identifying persons, locations, organizations, and other named entities.

Moreover, BERT's success led to the development of various BERT variants and improvements, such as RoBERTa (Robustly Optimized BERT Approach)[37] and ALBERT (A Lite BERT)[38], which optimize training procedures or reduce the computational resources required by the model. These variants make BERT's architecture more accessible for large-scale deployments, expanding its utility across diverse computational environments. Additionally, multilingual BERT models have enabled applications of the model to languages beyond English, making it an essential tool for international NLP applications.

However, despite its significant contributions, BERT also has limitations. One major drawback is its computational expense[39]. The large number of parameters in BERT, especially in its larger configurations (BERT-Large), requires substantial computational resources for both pre-training and fine-tuning. This has led to efforts to make BERT more efficient, including distilling BERT into smaller models that retain much of its performance while being easier to deploy in resource-constrained environments[40].

Overall, BERT has transformed the landscape of NLP by introducing a model that not only understands text bidirectionally but also excels in a variety of tasks without requiring task-specific architectures. Its introduction sparked a new era of research into Transformer-based models, which continue to advance the state-of-the-art in language understanding and generation tasks.

2.2.3. RoBERTa: A Robustly optimized BERT pre-training appraoch

RoBERTa (Robustly Optimized BERT Pretraining Approach), introduced by Liu et al. [37], is a significant refinement of BERT, designed to enhance its performance and robustness across a variety of NLP tasks. While RoBERTa retains the foundational Transformer architecture introduced in BERT, it focuses on optimizing the pre-training process to address several limitations and inefficiencies observed in the original BERT framework.

One of the most notable changes in RoBERTa is the removal of the Next Sentence Prediction (NSP) task. The creators of RoBERTa found that the NSP objective, initially intended to help BERT understand sentence relationships, was not as crucial for performance on downstream tasks as previously thought. By eliminating this task, RoBERTa reallocates resources to more intensive Masked Language Modeling (MLM), increasing the model's exposure to training data. This adjustment allows RoBERTa to learn more comprehensive language patterns without the constraints imposed by sentence pair classification.

Another key innovation in RoBERTa is the significant increase in the volume and diversity of pre-training data. RoBERTa was trained on larger datasets than BERT, including sources like the Common Crawl and news articles, resulting in a corpus size nearly ten times that used for BERT's pre-training. This extensive data exposure provides RoBERTa with a broader understanding of language, enhancing its ability to generalize across different domains and linguistic contexts. The increased dataset size is coupled with longer training durations, allowing the model to converge more effectively and learn more nuanced language representations.

In addition to a larger training corpus, RoBERTa employs adjustments to the pre-training hyperparameters, optimizing the training process. It uses larger batch sizes, longer sequences, and a higher learning rate, which enable more stable and faster convergence. This makes the training process more efficient, allowing the model to better capture long-range dependencies and contextual information within the text. These optimizations contribute to RoBERTa's superior performance in various NLP benchmarks, particularly in tasks requiring deep contextual understanding, such as reading comprehension and language inference.

RoBERTa also modifies the Masked Language Modeling (MLM) task by removing the constraint that 15% of tokens must be masked in each training sequence, allowing it to sample dynamically from the corpus. This variation enables RoBERTa to utilize a more diverse range of language contexts, leading to improved language modeling and a better understanding of sentence structure and semantics. The model's flexibility in the MLM task contributes to its robustness when applied to complex language tasks.

Due to these optimizations, RoBERTa achieves state-of-the-art results across several benchmark datasets, including GLUE (General Language Understanding Evaluation)[41], SQuAD (Stanford Question Answering Dataset)[42], and RACE (Reading Comprehension Dataset from Examinations)[43]. In particular, RoBERTa consistently outperforms BERT in scenarios requiring comprehensive language understanding and nuance, making it a preferred choice for many advanced NLP applications.

Beyond pre-training, RoBERTa is fine-tuned on task-specific datasets in a similar manner to BERT. The fine-tuning process benefits from the richer and more nuanced pre-trained language representations that RoBERTa develops, allowing it to excel in diverse tasks such as text classification, sentiment analysis, and entity recognition. This makes RoBERTa particularly effective for applications in both research and industry, where precise language understanding is crucial.

The success of RoBERTa has inspired the development of several derivatives and enhancements, each aiming to refine or extend its capabilities. Notable examples include models like XLM-R (Cross-lingual RoBERTa)[44], which extends the model's functionality to multilingual contexts.

However, despite its advancements, RoBERTa shares some limitations with BERT, such as the high computational costs associated with training and deployment. Efforts to optimize RoBERTa continue, with research focusing on making the model more efficient without sacrificing performance. This includes exploring techniques like model pruning[45], quantization[46], and distillation[47] to reduce the resource burden while maintaining accuracy.

3. Ancient Violence from a Historical and Computational Perspective

Violence, in all its forms, has long played a critical role in shaping human societies, their values, and their norms. While violence is a fundamental aspect of every human society, its meaning is shaped by cultural context. Due to its diverse and multifaceted nature, violence is difficult to define, with various academic disciplines offering differing interpretations. For the purposes of this work, we adopt the definition provided by Werner Rieß, who states that ***"violence is a physical act, a process in which a human being inflicts harm on another human being via physical strength"***[48]. Additionally, Mercy et al. define interpersonal violence as ***"the intentional use of physical force or power against other persons by an individual or small group of individuals"***, further dividing it into family or partner violence and community violence [49].
Interpersonal violence in ancient texts is not merely a reflection of individual conflicts; it also mirrors broader societal structures, providing insights into the functioning of power, justice, and social order. This analysis offers critical perspectives on the cultural, legal, and social dynamics of early civilizations.
This chapter will explore the multifaceted role of interpersonal violence in ancient texts, examining its connections to cultural values, social hierarchies, and psychological motivations. It will also address the importance of detecting violent incidents, explore the Perseus and ERIS datasets, and discuss the challenges faced in this field of research.

3.1. Violence in Ancient History

3.1.1. Cultural Norms and the Role of Violence

In many ancient societies, violence was often tied to cultural concepts of honor, revenge, and justice. Interpersonal violence frequently emerged in situations where personal or familial honor was at stake. The idea of honor was not just personal but deeply entrenched in societal expectations. In texts like The Iliad[50], the entire narrative revolves around concepts of honor, glory, and the consequences of personal affronts. Achilles' wrath, triggered by the loss of his war prize, Briseis, reflects how deeply intertwined honor and violence were in ancient Greek society[51, 52].

Figure 3.1.: Battle of Achilles and Hector as depicted by The Iliad[7].

Similarly, in the ancient Near East, honor-based violence was integral to social and familial relationships.

The Epic of Gilgamesh illustrates this when Gilgamesh and Enkidu engage in violent confrontations that reflect broader societal values about bravery and heroism[53].

Violence was also heavily regulated by legal codes in many ancient civilizations. The famous Code of Hammurabi, dating from 1754 BCE, provides a detailed insight into how violence was viewed and controlled. It formalized retributive justice with its principle of lex talionis—the law of retaliation, or "an eye for an eye" [54, 55].

Figure 3.2.: The Hammurabi Codex[8].

In ancient Rome, violence within the family was a domain of the paterfamilias, the male head of the household. Roman law granted him significant authority over his family members, including the use of physical force in cases where honor, discipline, or familial duty was at stake[56]. The Twelve Tables, the early code of Roman law, formalized many of these practices, outlining permissible forms of punishment and the legal consequences of violent actions[57].

Figure 3.3.: Augustus as the paterfamilias of the royal family [9].

3.1.2. Power Dynamics and Gendered Violence

In many ancient societies, violence was an essential mechanism for asserting dominance, both in personal relationships and on the broader political stage. From duels between warriors to the enforcement of royal decrees, violence functioned as a tool for maintaining power and authority. Ancient texts often depict rulers or heroes using violence to reaffirm their social status or control subordinates.
In The Iliad, the conflicts between Achilles, Agamemnon, and Hector are not just about individual grievances but about power struggles and maintaining dominance among warriors. Violence here is portrayed as a method of asserting superiority and ensuring one's place in the hierarchy. Similarly, in Roman history, figures such as Julius Caesar and Augustus used political violence as a means to consolidate power and neutralize rivals[58].

Figure 3.4.: The death of Julius Caesar as depicted by the neoclassic painter Vincenzo Camuccini[10].

Moreover, violence wasn't always physical. Threats of violence or punitive measures were employed by kings and rulers to keep their subjects in line. For example, in the ancient Near East, the Assyrian kings famously used the threat of extreme violence—such as mutilation or public executions—as a psychological tool to suppress rebellion and assert their dominion[59].
Interpersonal violence between genders in ancient societies often reflected deep-seated power imbalances. Ancient texts frequently depicted men using violence to control or subjugate women, reflecting the patriarchal structures of these societies[60]. This form of gendered violence not only maintained social hierarchies but also reinforced ideas of masculinity and femininity.
In Greek mythology, many stories depict male gods exerting violence against female characters. For instance, the abduction and rape of Persephone by Hades, or Zeus's repeated assaults on women, reflect a social order in which male dominance and the use of violence were normalized. Similarly, in Roman literature, stories of women like Lucretia, who committed suicide after being raped, underscore how violence was a tool not only for enforcing male dominance but also for defining the boundaries of female virtue and honor[61].
The legal frameworks of ancient societies also codified this gender imbalance. Roman law (as previously mentioned) granted the paterfamilias the authority to exercise physical punishment on female family members, reinforcing the subjugation of women within the family. These violent dynamics were not limited to personal relationships but were deeply embedded in social and political structures.

3.1.3. Patterns of Conflict Resolution in Ancient Texts

In ancient societies, conflict resolution often took a variety of forms, ranging from violent confrontations to non-violent settlements. One common form of conflict resolution was through physical violence, including

duels or battles, as seen in many epic tales such as the Iliad and The Odyssey.
However, non-violent forms of conflict resolution were also prominent. In certain societies, compensation or reparations were preferred over direct violence. The wergild system[62] in early Germanic law, for instance, allowed for financial compensation for injuries or deaths, as an alternative to blood feuds. In other cases, legal systems formalized conflict resolution through courts and legal processes, such as the Gortyn Code of ancient Crete[63], which regulated personal disputes and prescribed fines for certain acts of violence.

Figure 3.5.: A murder and subsequent Wergild payment. From the Heidelberger Sachsenspiegel[11].

Mediation and diplomacy played critical roles in resolving interpersonal and societal conflicts in ancient texts. Mediators were often key figures such as kings, gods, or community elders, who sought to restore peace by intervening in disputes. In The Epic of Gilgamesh[64], for instance, the conflict between Gilgamesh and Enkidu is resolved through divine intervention, which leads to their friendship and subsequent adventures. This shows that mediation, even when involving supernatural forces, was a crucial way to defuse interpersonal tensions.
In The Odyssey, Odysseus[65] often uses diplomacy rather than violence to navigate his way home, whether dealing with hostile forces or competing suitors for Penelope's hand. These examples demonstrate how ancient texts valorized not only the warrior but also the diplomat, illustrating the importance of wise counsel and negotiation as tools of conflict resolution.
Diplomacy was equally important in political contexts. Ancient kings and rulers often used marriage alliances, treaties, and negotiations to resolve potential conflicts or avert war. The famous peace treaty between the Hittites and the Egyptians following the Battle of Kadesh (around 1259 BCE) is one of the earliest recorded diplomatic resolutions in history[66], setting a precedent for formal peace agreements.

3.1.4. Psychological and Emotional Drivers Behind Violence

The portrayal of violence in ancient texts is not just about physical conflict but also about the deep psychological forces that drive individuals to commit violent acts. Ancient authors often used violent interactions to explore the emotional struggles of their characters. Common psychological drivers include anger, jealousy, fear, and pride, which are reflected in both personal and societal conflicts.
In The Iliad, Achilles' rage is not just a reaction to an external slight but a profound emotional wound tied to his sense of honor and identity. This connection between personal honor and violent response was a central part of the warrior ethos in many ancient cultures. Achilles' refusal to fight after being dishonored by Agamemnon and his subsequent violent outbursts against Hector after the death of Patroclus illustrate how ancient texts often framed violence as a manifestation of deep psychological and emotional turmoil[67].
Furthermore, The Oresteia[68] explores the psychological complexity of vengeance. Clytemnestra's murder

of Agamemnon is motivated not only by political considerations but also by her grief over the death of their daughter Iphigenia, which Agamemnon sanctioned. In her, we see how personal loss and emotional pain drive violent actions. The connection between psychological states and violence can also be seen in ancient Roman texts. In works like Seneca's Phaedra[69], the destructive nature of unrequited love and forbidden desire drives the titular character toward violent consequences. Phaedra's lust for her stepson, Hippolytus, leads to a chain of events resulting in his death.

Beyond personal identity, violence was also tied to one's place in the social hierarchy. For example, the Roman practice of damnatio memoriae[70] where the state sought to erase the memory of someone condemned for treason—was itself an act of symbolic violence, showing how state-sanctioned violence could erase not only lives but also legacies. This reflects how violence was wielded not only as a tool for personal vengeance but also for controlling memory and identity at the societal level.

Vengeance was not just a personal motivation but also a societal expectation in many ancient cultures. Vengeful violence often led to cycles of retaliation, which were difficult to break. These cycles are most prominently explored in Greek tragedies, where violent acts demand further retribution, creating an endless loop of violence[71].

For example, in the Roman world, revenge killings often escalated into full-scale political conflicts. The assassination of Julius Caesar, motivated by personal and political betrayals, sparked a series of retaliations that plunged Rome into civil war[72].

Similary, in The Oresteia, the murder of Agamemnon by Clytemnestra leads to the revenge killing of Clytemnestra by her son, Orestes. This cycle of vengeance reflects a broader societal belief that violence must be met with violence to restore honor and balance. However, the conclusion of the trilogy, where Athena intervenes to establish the rule of law and bring an end to the bloodshed[73], marks a critical moment in ancient thought, which is the recognition of the need for judicial mechanisms to break these destructive cycles.

In many ancient texts, cycles of vengeance are not just human constructs but are intertwined with the idea of fate or divine will. In Greek tragedies, violent acts are often seen as inevitable, driven by the gods or by the inescapable force of fate. In Oedipus Rex[74], for instance, Oedipus' violent actions—killing his father and marrying his mother—are fated, and his attempts to avoid this destiny only lead to its fulfillment. This interplay between human emotion, psychological motivation, and the belief in an inexorable fate complicates the narrative of violence, making it both a personal and cosmic force. In Roman texts, too, the concept of fatum (fate) plays a role in justifying cycles of vengeance. For example, in Virgil's Aeneid, Aeneas' violent actions are portrayed as divinely ordained, part of his destiny to found Rome[75]. Here, violence is not merely the result of human emotion but a necessary step in fulfilling a greater destiny.

In some ancient texts, violence is used to symbolize internal struggles, where characters wrestle with conflicting desires, duties, and emotions. The violent actions they commit reflect the chaos and conflict within themselves. This theme is especially prominent in tragedies, where characters are often torn between their desires and their obligations to society or the gods. In Euripides' Medea, Medea's violent murder of her children is both an act of vengeance against her unfaithful husband, Jason, and a reflection of her own internal torment[76]. Medea's conflict between her love for her children and her desire to hurt Jason leads her to commit an unspeakable act of violence, symbolizing the tragic consequences of unresolved inner conflict.

3.1.5. Historical Realities and humanizing Figures through Violence

In many ancient texts, interpersonal violence is not simply a literary or mythological device; it often reflects real historical events and social realities. Violent acts described in historical texts provide valuable insights into the lived experiences of individuals in ancient societies. Whether in the form of political assassinations, military conflicts, or familial disputes, violence in these texts often serves as a window into the power dynamics, social norms, and cultural values of the time.

For example, the assassination of Julius Caesar, detailed in historical accounts such as those by Suetonius and Plutarch, illustrates how political violence was both a method of power consolidation and a reflection of personal betrayal. This event, steeped in personal and political motives, marked a pivotal moment in Roman history, triggering the fall of the Roman Republic and the rise of the Empire[77]. Similarly, the writings of Herodotus and Thucydides recount various acts of interpersonal violence in the context of war

and politics[78], giving historians valuable records of how violence was used to shape political landscapes in the ancient world.

Interpersonal violence in ancient historical texts often provides a direct link to real historical events. These violent acts were sometimes highly publicized, both as political statements and as records of power struggles. In this way, violence serves not only as a narrative device but also as a record of social realities, giving historians concrete examples of how disputes were settled through violent means. Therefore serving as a tool for humanize larger-than-life historical figures, allowing readers to engage with their personal flaws, emotional states, and interpersonal relationships. While many ancient figures are depicted as god-like or heroic, acts of violence can expose their vulnerabilities and bring them down to a more human level.

For example, Julius Caesar's assassination was not just a political event but also a personal betrayal. Accounts by ancient historians such as Suetonius and Cassius Dio emphasize how the violence of the assassination reveals personal rivalries and conflicting ambitions within the Roman elite[79, 80]. This portrayal of Caesar as a victim of interpersonal violence adds depth to his character, showing him not only as a powerful leader but also as a man whose downfall was rooted in personal enmity.

Similarly, in the case of Alexander the Great, the violence that marked his rise to power—his execution of rivals and commanders, his bloody military campaigns—gives us insight into his personality and the pressures of leadership. Ancient biographers like Arrian and Plutarch describe moments of intense interpersonal violence that Alexander himself perpetrated or ordered[81]. These acts of violence humanize him, revealing the darker, more emotional sides of a leader often revered for his genius. His famous murder of Cleitus the Black, one of his closest friends, during a drunken rage shows the fragility of even the greatest of leaders in the face of personal conflict.

This ability of violence to humanize historical figures is also evident in the case of Roman emperors such as Nero or Caligula, where acts of interpersonal violence, often driven by paranoia or vengeance, reflect the emotional and psychological turmoil of these rulers[82, 83].

Figure 3.6.: The death of emperor Caligula at the hands of the Praetorian[12].

Ancient texts demonstrate how violence could make or break a leader's legacy, providing historians with invaluable insights into how ancient societies understood leadership, power, and reputation. The portrayal

of violence in these texts also allows modern readers to connect with the emotional and psychological dimensions of historical figures, adding layers of complexity to their legacies.

3.2. The Importance of Detecting Violence in Ancient Texts

The study of violence in ancient texts is not merely an exercise in historical curiosity—it plays a crucial role in helping scholars understand the complex societal structures, evolving legal frameworks, and human relationships of past civilizations. researchers can gain insights into how ancient societies functioned, how they dealt with conflict, and how these insights can inform contemporary legal and sociological studies

3.2.1. Understanding Societal Structures and Evolution

One of the key reasons for detecting violence in ancient texts is to understand the societal structures and hierarchies of past civilizations. Violence often reflects the underlying power dynamics, social norms, and legal systems that governed ancient societies. For example, interpersonal violence depicted in The Iliad shows how honor and power were central to the warrior ethos in Greek society. Similarly, Roman law codified interpersonal violence through strict legal frameworks that regulated duels, punishments, and vengeance killings. Violence in texts such as Hammurabi's Code or the Twelve Tables provides evidence of how early legal systems sought to control personal disputes and maintain societal order[84]. Moreover, it helps to illuminate how these ancient systems contributed to the foundation of modern legal and political institutions, showing a continuous evolution from violent retribution to more formalized legal processes. Identifying these elements allows researchers to trace the evolution of societal norms and structures over time.

3.2.2. Contributing to Comparative Historical Studies

Violence in ancient texts offers valuable material for comparative historical studies. scholars can make cross-cultural comparisons that highlight both universal human experiences and unique societal characteristics by alyzing how different ancient civilizations handled interpersonal violence.
For instance, comparing the depiction of violence in ancient Greek literature with texts from ancient China or Mesopotamia reveals both similarities and differences in how societies dealt with conflict[85]. While some cultures placed a heavy emphasis on personal honor and revenge, others developed early systems of compensation or legal adjudication to mitigate cycles of violence. These comparisons are essential for understanding the diversity of human responses to violence, helping historians and anthropologists identify patterns in the development of conflict resolution and legal systems.
Such cross-cultural studies also provide a broader perspective on human history, showing how violence—and the attempts to control or resolve it—shaped the development of societies across the world. Comparative studies can also shed light on modern issues, helping to contextualize contemporary conflicts through an understanding of how ancient societies evolved from violent to non-violent methods of conflict resolution.[86]

3.2.3. Informing Modern Concepts of Violence and Law

Ancient texts provide the foundational material for many modern legal and philosophical concepts of violence, justice, and personal rights. Researchers can trace the development of modern legal frameworks that regulate personal conflict by detecting and analyzing violence in these texts.
The lex talionis ("an eye for an eye") principle from Hammurabi's Code laid the groundwork for retributive justice systems that still influence legal systems today. Similarly, Roman law, which regulated familial and societal violence, has contributed to modern notions of crime, punishment, and personal rights[87]. The paterfamilias concept, which granted male heads of households significant control over family members, reflects early legal structures that echo in modern discussions about family law and authority.
Moreover, these ancient legal codes and narratives about violence provide context for ongoing debates about the ethics of violence, retribution, and justice in modern legal systems[88]. Scholars who study ancient texts are better able to understand the roots of these issues, offering historical insights that can inform contemporary policy-making and legal reforms[89].

3.3. Ancient Historical Databases

In the modern study of ancient texts, digital tools and datasets have become invaluable resources for researchers seeking to analyze themes like violence. In this thesis, we focus on two key datasets that play a significant role in the computational analysis of ancient texts, Perseus and ERIS. These platforms provide access to vast collections of ancient literature and historical documents, allowing researchers to study them thoroughly, analyze the texts and also use techniques to detect patterns of violence and extract knowledge from such databases.

3.3.1. Perseus

The Perseus Digital Library[90] is one of the most comprehensive and widely used resources for the study of ancient texts. Developed by Tufts University, Perseus contains a large corpus of Greek and Roman literature, including works from authors such as Homer, Sophocles, and Cicero. It also includes English translations and various linguistic annotations, making it a versatile tool for scholars from different disciplines.
Since its inception, Perseus has contributed to the fields of classics, philology, and history by making previously inaccessible texts readily available for analysis. It has also played a pivotal role in the advancement of digital humanities, serving as a model for other digital text projects.
Perseus offers several features that make it particularly valuable for the study of violence in ancient texts:

- It provides morphological and syntactic analyses of ancient Greek and Latin texts.
- The library allows users to annotate texts and apply computational tools to identify recurring themes.
- As an open-access resource, Perseus is widely available and supports integration with other tools, including machine learning models. Researchers already used Perseus to develop a deep neural network model that recovers missing characters from damaged text [91] and to train a Latin version of BERT [92].

Figure 3.7.: Perseus homepage interface[13].

The Perseus library has faced criticism for its ergonomic, accessibility and and usability issues[93, 94]. The layout of the website is often seen as unintuitive, making it difficult for new users to navigate the resources effectively.

3.3.2. ERIS

ERIS (Hamburg Information System on Greek and Roman Violence) is a project developed since 2012 at the University of Hamburg by Werner Rieß and Michael Zerjadtke[95], aimed at providing a comprehensive,

searchable database focused on the depiction of violence in ancient Greek and Roman texts. The system is built on the MyCore platform and serves as a repository for all descriptions of violence found in the works of Greek and Latin authors, annotated with sociological and contextual information to facilitate in-depth research.

The primary aim of ERIS is to offer researchers a detailed and accessible way to explore interpersonal violence across a wide variety of ancient texts. While violence is ubiquitous in Greek and Roman culture, overarching structures and patterns of violent behavior are often difficult to discern. ERIS addresses this challenge by systematically collecting and annotating texts that describe violent acts, allowing users to search for violence by various criteria, including context, motivation, and outcomes. The user can choose to search for conflicts, persons, groups, authors, works, topologies or even plain English, latin or greek text.

ERIS also tackles the challenge of modeling violence in ancient societies by simplifying complex social processes while maintaining scholarly depth. The system's annotation allows users to search not only for basic information like the author and chronological setting of the work but also for characteristics of violent acts such as context, motives, geographical locations, and socio-economic status of the involved parties. This allows for a multi-dimensional understanding of the consequences of violence, ranging from immediate reactions to long-term effects of the events.

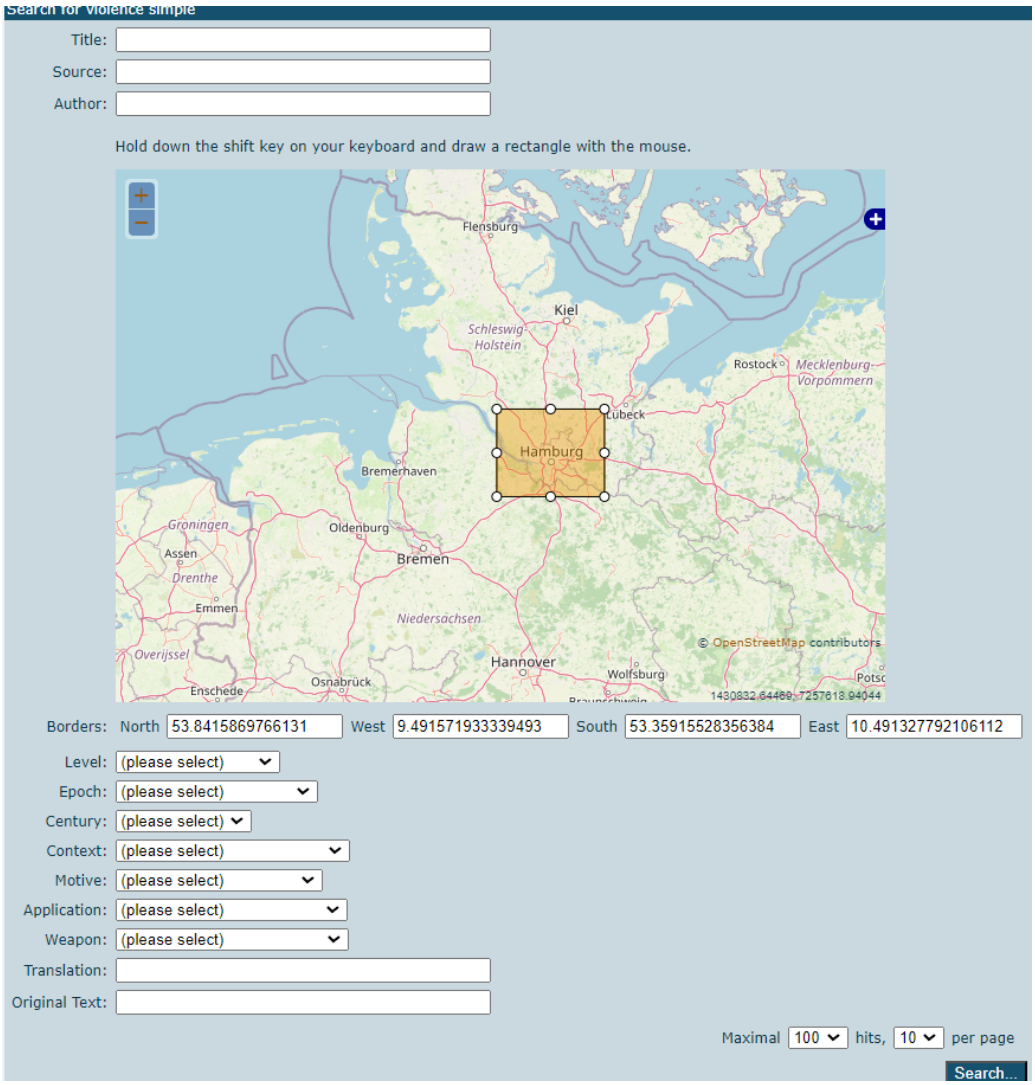


Figure 3.8.: Map search functionality in the ERIS website.

Key features:

- ERIS annotates acts of violence based on multiple sociological factors, such as the context of the act, motives, geographical location, and the socio-economic status and age of the perpetrators and

Figure 3.9.: Corresponding result for Figure 3.8.

victims. This level of detail enables targeted searches, providing users with insights into the broader social and historical implications of violent acts.

- The system links violent acts with related factors, such as legislative reactions, immediate consequences, and broader societal impacts, helping researchers trace how violence was perceived and managed in different historical periods.
- ERIS does not just record physical acts of violence but also considers the legal and social responses to these acts, including legislative measures and changes in societal attitudes. This helps contextualize the role of violence within ancient societies.
- The platform allows for both simple and advanced searches. Users can search for violence within specific regions (using integrated maps), by time period, and by type of violence (e.g., familial, political, or martial). There is also a sophisticated search feature that enables more granular filtering, including by the social status or ethnicity of the perpetrators and victims.
- ERIS offers source texts in both the original Greek or Latin and in English translation, making it accessible to a wider audience, including those outside of classical philology.

In addition to these features, ERIS provides new digital ways of analyzing, visualizing, and representing data about violence in antiquity.Combining text analysis with geospatial tools allows for innovative studies of violence from a socio-political and regional perspective, making it easier to understand how violence evolved, was curtailed, or escalated in specific regions over time.

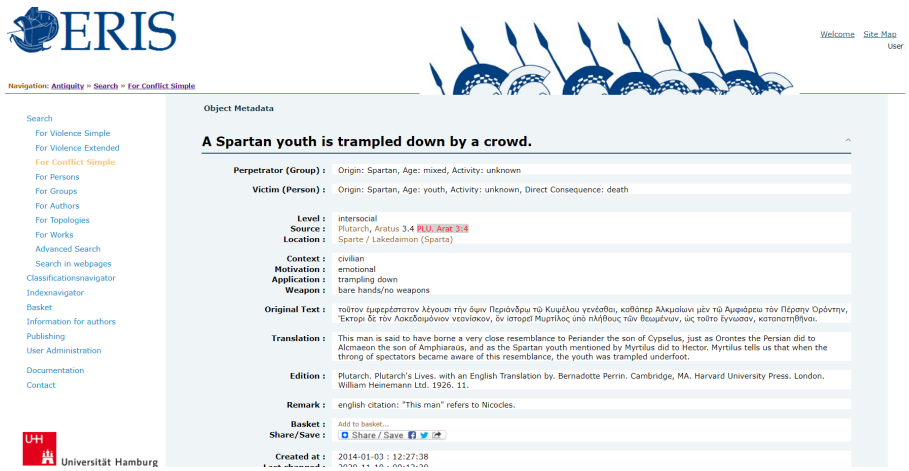


Figure 3.10.: An example of a violent text with its text analysis.

The system is designed to support scholars from different interdisciplinary research areas. Being an open-access model and its comprehensive scope, It has the potential to bring the study of violence in antiquity into broader academic and public discourse. Thus contributing to the understanding of how violence shaped communities, affected power dynamics and legal frameworks. Furthermore, it provides a valuable resource for comparing ancient and modern responses to violence.

3.4. Challenges of Semantic Annotation and Categorization in Ancient Texts

One of the most significant challenges in studying violence in ancient texts is the difficulty of accurately annotating and categorizing violent acts. In texts from Greek, Roman, and other ancient civilizations, violence can be depicted explicitly through direct physical actions or implicitly through metaphorical or symbolic language. Identifying and labeling these instances of violence is essential for extracting meaningful insights about societal structures, power dynamics, and legal frameworks, but the process is fraught with challenges.

3.4.1. Implicit and Symbolic Violence in Language

Ancient texts often embed violent acts within complex cultural and narrative frameworks. Violence might be implied through symbolic gestures, metaphors, or allusions that require deep cultural and contextual understanding to interpret. For example, in Greek tragedies, violence is often described obliquely, taking place offstage, or represented symbolically through language that conveys emotional or psychological harm rather than physical injury.

This presents a significant challenge when creating a digital humanities database, where the goal is to categorize acts of violence across a wide range of texts. **How do we define violence?** Does it include only physical harm, or **do psychological, emotional, and symbolic representations also qualify?**

To illustrate these challenges, we present two examples from our experiment using texts from Perseus:

> *"Before this, none had ventured there. But now they burst into an unravaged and inviolate land, and burned and plundered as far as the river and the city."*
>
> — Plutarch, *Agesilaus* 31

Correct Classification: Non-Violent

Explanation: While the passage describes actions like burning and plundering, which may seem violent, our classification criteria exclude damage to property (including fires in buildings) from being labeled as violent. Therefore, despite the aggressive actions described, this passage is classified as non-violent according to our definitions.

The second example demonstrates how context and subsequent events influence the classification:

> *"Cleitus sprang to his feet and said, 'It was this cowardice of mine, however, that saved thy life, god-born as thou art, when thou wast already turning thy back upon the spear of Spithridates; and it is by the blood of Macedonians, and by these wounds, that thou art become so great as to disown Philip and make thyself son to Ammon.'"*
>
> — Plutarch, *Alexander* 50

Correct Classification: Violent

Explanation: Although Cleitus' words may not depict physical violence directly, the passage is loaded with aggressive confrontation and personal accusations. Cleitus reminds Alexander of his previous cowardice and criticizes his abandonment of Macedonian roots. This intense verbal altercation escalates the situation, leading to Alexander's enraged response, which ultimately results in him fatally stabbing Cleitus during the banquet. The underlying tension and the fatal outcome classify this passage as violent.

These examples highlight the complexities of classifying violence in ancient texts. Actions that may appear violent on the surface may not meet the criteria for violence in our classification system, while passages that seem non-violent may lead to violent outcomes due to contextual factors. This underscores the importance of cultural and narrative context in accurately identifying and categorizing violent events.

3.4.2. Labor-Intensive Nature of Manual Annotation

Another substantial challenge is the sheer amount of labor required to manually annotate violent acts in ancient texts. Annotating just one chapter of a historical book can take months due to the need to carefully read and interpret the text, ensuring that each violent act—explicit or implicit—is correctly identified and categorized.

The process involves multiple layers of analysis:

- Identifying the context in which violence occurs (e.g., personal conflict, war, or ritual).
- Understanding the motives behind the violence (e.g., revenge, self-defense, power assertion).
- Recognizing the social and legal consequences of the violence (e.g., retribution, punishment, societal shifts).

This requires not only expertise in ancient languages and cultures but also an understanding of the socio-political frameworks in which these texts were produced. Due to the specialized knowledge required and the laborious nature of the work, annotating even a small corpus of texts can take an extensive amount of time.

3.4.3. Automating the Process with Large Language Models

Given the complexity and labor-intensive nature of semantic annotation, automating this process using large language models (LLMs) offers an exciting solution. By fine-tuning LLMs specifically on ancient texts, we can train them to detect both explicit and implicit instances of violence, reducing the manual effort required for annotation.

Large language models like BERT, GPT, and other transformer-based models have shown remarkable ability in understanding context and nuance in modern language[96]. Fine-tuning these models could allow them to recognize not only the explicit descriptions of violence but also the subtler, symbolic representations that might be missed in a simple keyword-based search.

For example, if a passage in an ancient text describes a character "offering sacrifice," a well-trained model could determine whether the act involves violence or represents a symbolic gesture based on the context of the narrative and broader cultural patterns.

Therefore, LLMs can learn to differentiate between descriptions of violence that are literal versus those that are metaphorical or part of a ritual practice. The model could also flag ambiguous cases for human review, ensuring the human element is still within the loop and contributing in classifying those odd cases.

3.4.4. Why Automation is Crucial

Automating the annotation process not only speeds up the time required to create digital humanitarian libraries but also opens up new avenues of research. LLMs can detect patterns and relationships between violent acts that might go unnoticed in manual analysis by processing large volumes of text in a fraction of the time it would take a human.

Automating the annotation process would significantly expand the corpus of texts available for study, allowing for more extensive and nuanced analyses of violence in the ancient world. This approach could facilitate the inclusion of a broader range of works, including those from different and even more recent time periods, offering deeper insights into how violence shaped societies across history. this would allow researchers to gain a more comprehensive understanding of the cultural, social, and political roles violence has played throughout human civilization.

3.4.5. Remaining Challenges of Automation

While the potential benefits of automation are clear, there are still challenges in training large language models to handle ancient texts effectively:

- **Limited Training Data:** The available annotated corpora of ancient texts are much smaller compared to those in modern languages, which could slightly limit the effectiveness of model training.

- **Cultural and Contextual Nuances:** Ancient texts are often deeply tied to specific cultural contexts that may be difficult for a machine learning model to fully grasp. Fine-tuning LLMs to understand these contexts will require careful calibration and, likely, ongoing human oversight.
- **Balancing Automation and Manual Review:** Even with advanced LLMs, there will always be cases where human expertise is needed to interpret subtle or ambiguous instances of violence. *The goal of automation is not to replace human scholars but to complement and enhance their work.*

4. Violence Detection Using Large Language Models

The previous chapters have provided the necessary background and context for the experiments conducted in this research. In the following two chapters, we will focus on the methodology, experimental setup, and results of the experiments.

This chapter presents the first of our two experiments: **Violence Detection Using Large Language Models**. We provide an overview of the methodology, dataset, and models employed, describe the experimental setup in detail, present the results, and conclude with a discussion of the findings and their implications. A GitHub repository with code and data is provided in **Appendix A**

4.1. Methodology and Experimental Setup

There are several key components for our methodology in both experiments, including the dataset preparation, model selection, training and evaluation setup, and criteria for violence detection.

4.1.1. Overview

The primary goal of this experiment is to detect instances of violence in ancient texts using fine-tuned large language models (LLMs). Previous works have demonstrated the superiority of LLMs in various classification tasks, such as sentiment analysis [97], text classification [98], and natural language interfaces [99]. A recent survey by Chang et al. provides a comprehensive evaluation of LLMs across different tasks [100].
Detecting violent instances in ancient texts presents unique challenges due to the implicit and symbolic nature of violence in historical narratives. Even human annotators often struggle to accurately identify these subtleties. To address this complexity, we employ a multi-step approach that leverages pre-trained LLMs, fine-tuning them on domain-specific historical data to improve detection accuracy.

4.1.2. Dataset Preparation

Sri Gowry Sritharan [101] reconstructed the ERIS entries (presented in **Chapter 3**) by mapping them back to their original chapters and paragraphs in Perseus. For each ERIS-labeled violent passage, Sritharan retrieved the full sections from which these excerpts were derived. Any paragraph not explicitly marked as violent in ERIS was treated as non-violent, forming the negative examples for the dataset.
This approach enabled us to compile a balanced dataset consisting of both violent and non-violent entries. However, it is important to acknowledge some limitations in this process. Since the non-violent class was defined by assuming any text not explicitly labeled in ERIS as non-violent, there is a risk of misclassification. For instance, if the original annotators missed labeling a violent passage, it would have been inadvertently included in the non-violent category. Despite this limitation, the method was efficient for generating a large enough dataset for training our models.
After receiving the dataset compiled by Sritharan, we performed additional data cleaning to remove duplicates and incomplete entries. This final refinement resulted in a dataset of 2,564 texts, with both violent and non-violent classifications. This dataset served as the foundation for training our models to automate the detection of violence in ancient texts.
A snippit of the data is shown on Figure 4.1.
To extend and leverage the data further, we performed data augmentation to enhance the dataset using a synonym replacement technique. The process is detailed below:

	A	B	C	D	E
1	Book	Chapter	Section	Text	Violence
476	ALEXANDER	63	4	Such was the force of the blow that Alexander recoiled and sank to his knees, where	1
477	ALEXANDER	63	4	but Peucestas held out, and at last Alexander killed the Barbarian.	1
478	ALEXANDER	63	4	But he himself received many wounds, and at last was smitten on the neck with a cuc	1
479	ALEXANDER	63	6	And after sacrificing to the gods he went on board ship again and dropped down the r	1
480	ALEXANDER	64	1	He captured ten of the Gymnosophists who had done most to get Sabbas to revolt, ar	1
481	ALEXANDER	68	4	One of the sons of Abuletes, Oxyartes, he slew with his own hand, running him throug	1
482	ALEXANDER	68	4	and when Abuletes failed to furnish him with the necessary provisions, but brought hi	1
483	ALEXANDER	69	2	In the second place, having discovered that the tomb of Cyrus had been rifled, he put	1
484	ALEXANDER	1	1	It is the life of Alexander the king, and of Caesar, who overthrew Pompey, that I am wr	0
485	ALEXANDER	1	2	For it is not Histories that I am writing, but Lives; and in the most illustrious deeds the	0
486	ALEXANDER	1	3	Accordingly, just as painters get the likenesses in their portraits from the face and the	0
487	ALEXANDER	2	1	As for the lineage of Alexander, on his father's side he was a descendant of Heracles	0
488	ALEXANDER	2	2	Well, then, the night before that on which the marriage was consummated, the bride	0
489	ALEXANDER	2	3	The other seers, now, were led by the vision to suspect that Philip needed to put a clo	0
490	ALEXANDER	2	4	Moreover, a serpent was once seen lying stretched out by the side of Olympias as she	0
491	ALEXANDER	2	5	But concerning these matters there is another story to this effect: all the women of th	0
492	ALEXANDER	2	6	Now Olympias, who affected these divine possessions more zealously than other wo	0
493	ALEXANDER	3	1	However, after his vision, as we are told, Philip sent Chaeron of Megalopolis to Delph	0
494	ALEXANDER	3	2	Moreover, Olympias, as Eratosthenes says, when she sent Alexander forth upon his g	0
495	ALEXANDER	3	3	Be that as it may, Alexander was born early in the month Hecatombaeon, 356 B.C. Th	0
496	ALEXANDER	3	4	But all the Magi who were then at Ephesus, looking upon the temple's disaster as a sig	0

Figure 4.1.: A snippit of the cleaned Sritharan data.

- **Library Imports:** We imported necessary libraries such as `pandas` for data manipulation, `transformers` to utilize a pre-trained BERT model, and `tqdm` for progress visualization.
- **Dataset Loading:** The dataset was read from a CSV file.
- **Pipeline Initialization:** A fill-mask pipeline was created using the base BERT model to generate synonyms for words in the text.
- **Synonym Replacement Function:** We defined a function, called

  `replace_with_synonym`, which splits text into words and replaces selected words with their top predicted synonyms. The first prediction from the BERT model is used for replacement.
- **New Entry Generation:** For each row in the original dataset, three new entries were created by invoking the synonym replacement function. The results were stored in a new list.
- **Data Concatenation:** The original dataset was combined with the newly generated entries to form an augmented dataset.
- **Saving the Augmented Data:** Finally, the augmented dataset was saved to a new CSV file.

This approach effectively quadruples our dataset, providing a richer foundation for training and testing.

4.1.3. Model Selection and Fine-tuning

For this experiment, we employed two distinct approaches. The first involved fine-tuning LLMs on our dataset, while the second utilized the ChatGPT API to classify sentences and compare the results. For the fine-tuning approach, we selected BERT and RoBERTa Large. Due to technical constraints, we were unable to use models larger than RoBERTa Large, which contains over 300 million parameters.
The models were fine-tuned using the labeled Perseus dataset, as mentioned in 4.1.2. We followed a standard fine-tuning procedure, where 80% of the dataset was used for training, and the remaining 20% was reserved for testing. This split allowed us to evaluate the model's performance on unseen data.

4.1.4. Evaluation Metrics and the Importance of the F1 Score

To assess the performance of the models, we used several evaluation metrics, including precision, recall, F1 score, and accuracy. These metrics allowed us to evaluate both the accuracy of violence detection and the model's ability to minimize false positives and false negatives.
Consider a binary classification problem with the following outcomes:

- **True Positives (TP)**: Correctly predicted positive instances.
- **False Positives (FP)**: Negative instances incorrectly predicted as positive.
- **False Negatives (FN)**: Positive instances incorrectly predicted as negative.
- **True Negatives (TN)**: Correctly predicted negative instances.

The key metrics are defined as follows:

1. **Precision** (P):
$$P = \frac{TP}{TP + FP} \tag{4.1}$$
Precision measures the proportion of positive predictions that are correct.

2. **Recall** (R):
$$R = \frac{TP}{TP + FN} \tag{4.2}$$
Recall measures the proportion of actual positives that are correctly identified.

3. **F1 Score** (F_1):
$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{4.3}$$
The F1 score is the harmonic mean of precision and recall.

In the **discussion** section, we discuss why this evaluation metric is superior to only picking only precision or recall.

4.2. Results

Here, we present the results for the violence detection experiment, compare it with chatgpt API and also test with some custom-created examples. Further results will also be included in **Appendix B**.

4.2.1. BERT Model with No Fine-Tuning

Before presenting our results for the fine-tuned models, it is worth exploring how the BERT model performs when applied to the dataset without any fine-tuning. This baseline approach allows us to evaluate the model's behavior based purely on pre-training and to observe its raw performance on violence detection. For this experiment, we used BERT-large and tested it on 20% of the dataset and the whole dataset. The results for the whole dataset are presented in Table 4.1. The rest of the results are shown in **Appendix B**.

	Precision	Recall	F1-Score	Support
Non-Violent	0.53	0.01	0.01	2013
Violent	0.17	0.98	0.30	416
Overall	0.20	0.98	0.30	2564
Macro Avg	0.35	0.49	0.16	2564
Weighted Avg	0.46	0.18	0.06	2564
Accuracy		0.18		

Table 4.1.: Results for BERT Large without Fine-Tuning

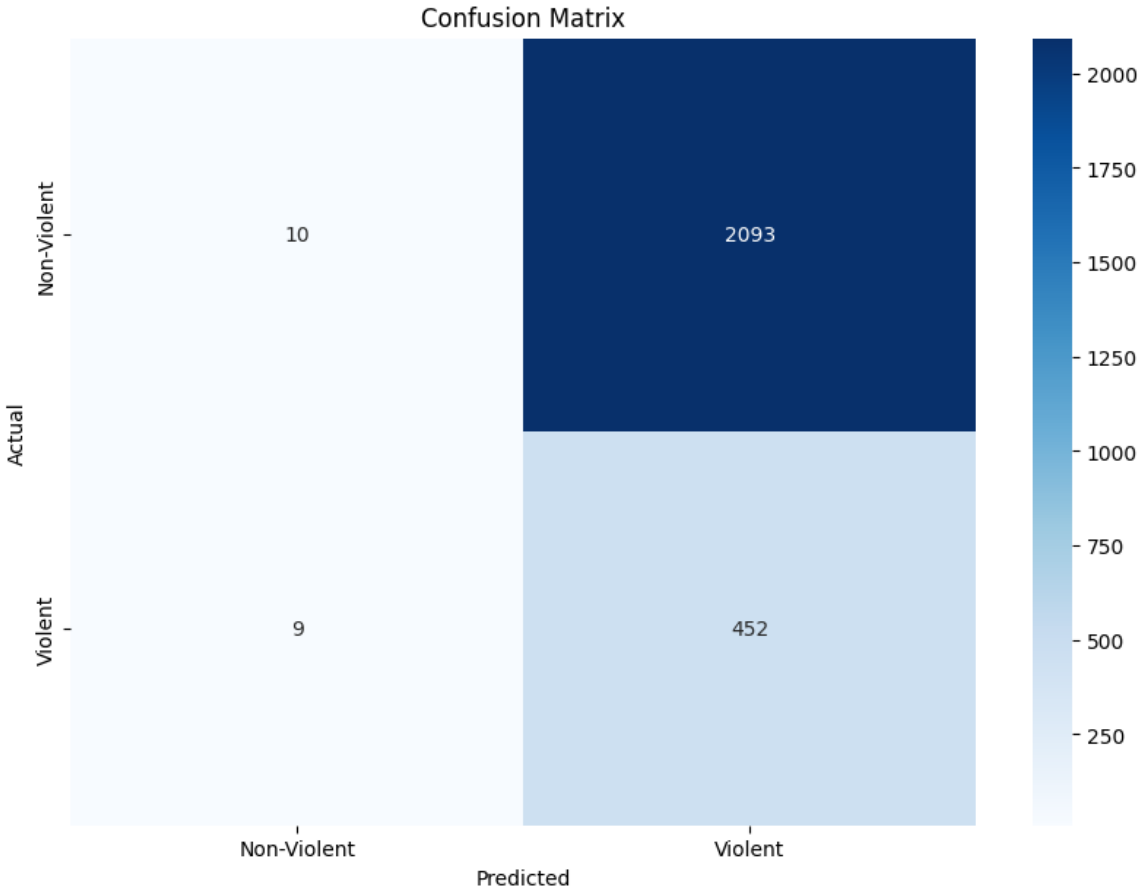


Figure 4.2.: Confusion Matrix for BERT Large without Fine-Tuning

From Table 4.1 and Figure 4.2, it is evident that the pre-trained BERT model tends to classify most sentences as violent, irrespective of whether the violence is implicit or explicit. This behavior leads to the misclassification of the majority of non-violent sentences.These results are expected, as the BERT model has not been fine-tuned for violence detection in historical texts. The pre-trained model is not specialized for this domain, which leads to the poor performance observed in the classification of non-violent sentences. This baseline reinforces the importance of fine-tuning the model to the specific task.

The importance of reporting both precision and recall becomes clear in this case. While precision for non-violent sentences seems relatively high, it is misleading because it captures only a small fraction of the true non-violent cases, likely by chance. The recall for non-violent sentences is extremely low, indicating the model's inability to correctly identify non-violent instances. This imbalance demonstrates how the model struggles without fine-tuning, and underscores the necessity of adjusting the model to the specific task at hand.

4.2.2. RoBERTa Large Model

In this section, we present results for the RoBERTa Large model in Table 4.2

	Precision	Recall	F1-Score	Support
Non-Violent	0.95	0.96	0.96	371
Violent	0.89	0.86	0.87	129
Overall	0.89	0.86	0.87	500
Macro Avg	0.92	0.91	0.92	500
Weighted Avg	0.94	0.94	0.94	500
Accuracy		0.94		

Table 4.2.: Results for RoBERTA Large

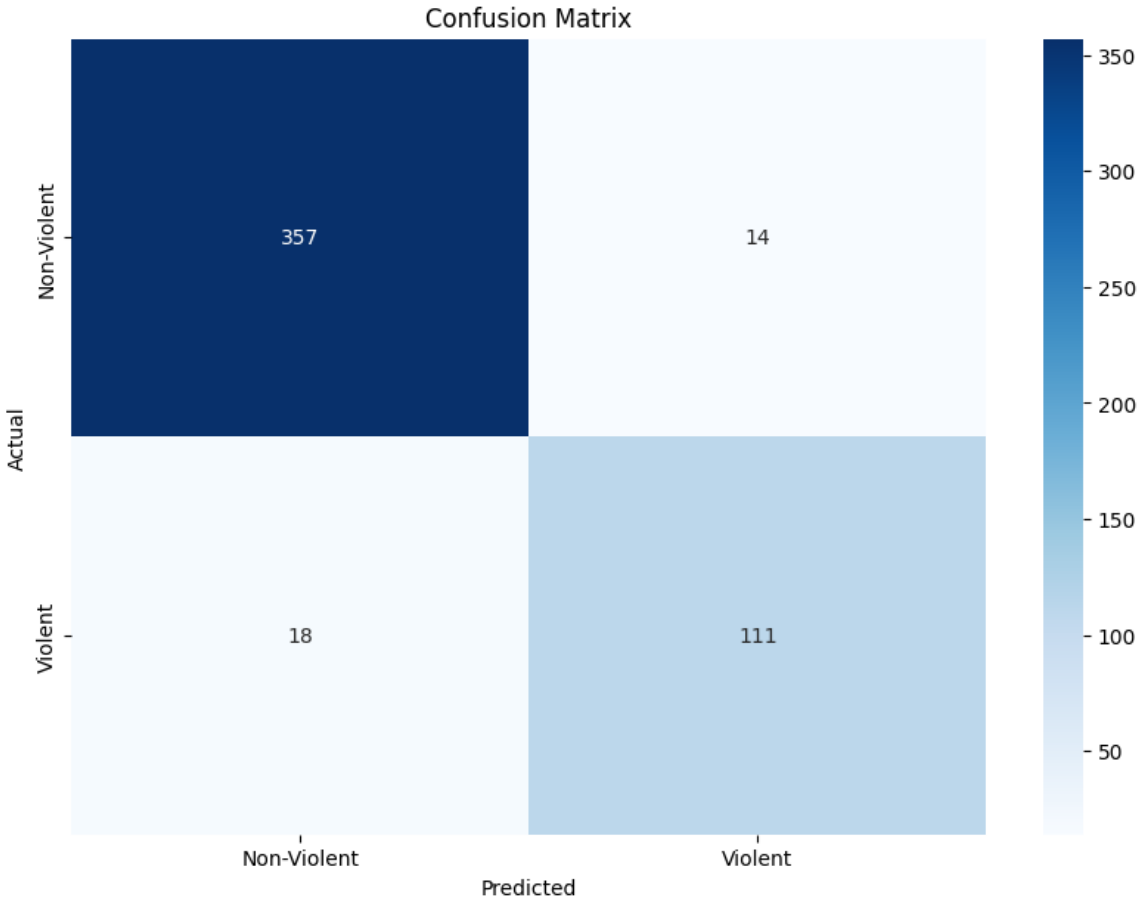


Figure 4.3.: Confusion Matrix for the RoBERTa Large model

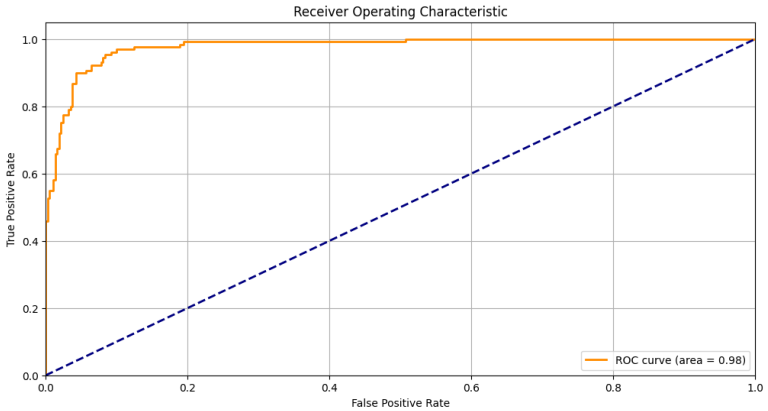


Figure 4.4.: ROC curve for the RoBERTa Large model

4.2.3. RoBERTA Large Model with Augmentation

As explained earlier in data augmentation, we extended the data by performing data augmentation. Here, we provide the results for this approach in Table 4.3:

	Precision	Recall	F1-Score	Support
Non-Violent	1.00	0.99	0.99	1714
Violent	0.95	0.98	0.96	338
Overall	0.9538	0.9763	0.9649	2052
Macro Avg	0.97	0.98	0.98	2052
Weighted Avg	0.99	0.99	0.99	2052
Accuracy		0.9883		

Table 4.3.: Results for RoBERTA Large with Augmentation

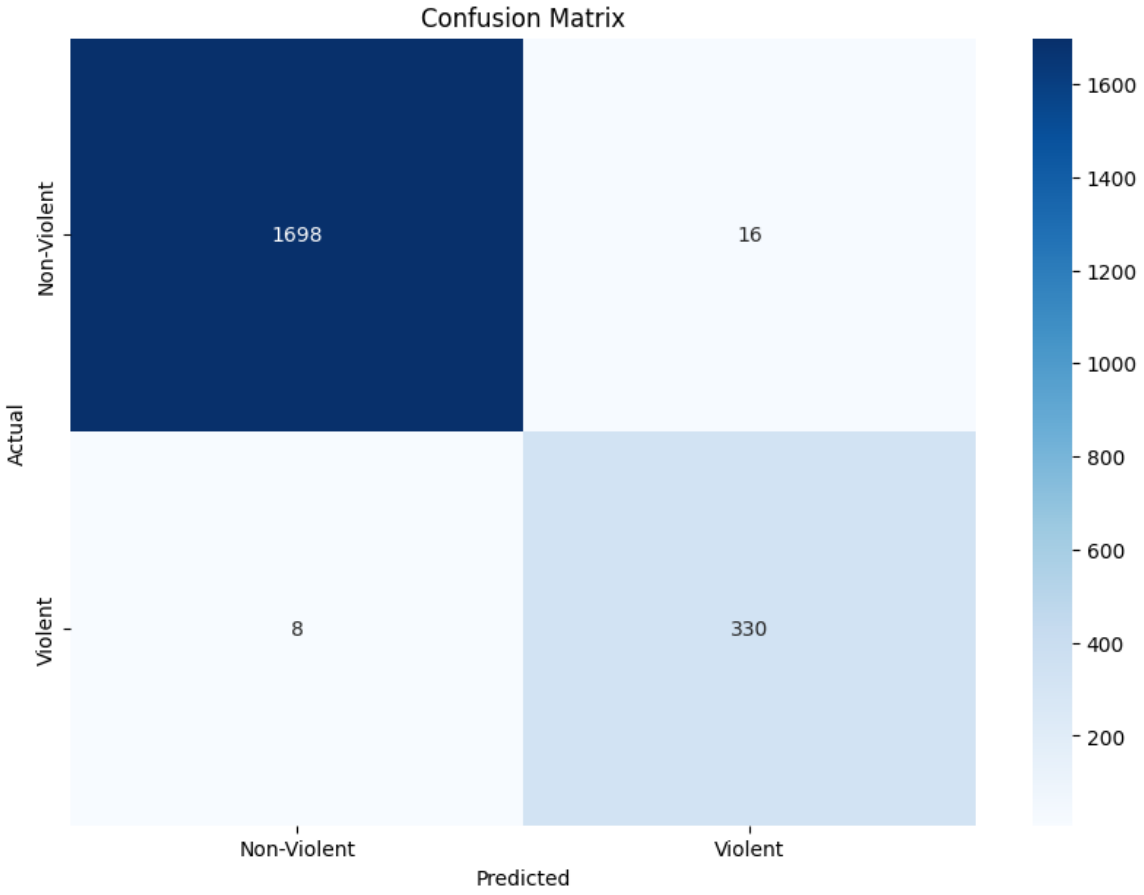


Figure 4.5.: Confusion Matrix for the RoBERTa Large model with augmentation

As we can see, the augmented model provided the best results. It also provided good results when applied to some custom-created data (shown in **Appendix B**). Data augmentation played a significant role in boosting the performance of the RoBERTa model, as it allowed the model to be exposed to a wider variety of violent and non-violent examples. By increasing the dataset size, the model could generalize better to unseen examples, which is reflected in the higher precision and recall for both classes.

4.2.4. Using chatGPT API for Violence Detection

OpenAI offers users the chance to use a chatGPT API that can be included in code and perform tasks that chatGPT can perform[102]. Mainly, we use the GPT 4o-mini[103] to classify sentences to violent or

non-violent and then we check the results

In this part of the experiment, we employed the GPT-4o mini model from OpenAI to classify historical sentences as either violent or non-violent. The goal of this approach was to test the generalization capability of a large pre-trained language model without any fine-tuning on our specific dataset. By leveraging the GPT-4 API, we aimed to evaluate how well the model could detect violence in ancient texts using only its pre-trained knowledge and a carefully crafted prompt.

Setup and Approach The GPT-4o mini model was accessed through the OpenAI API. To guide the model in classifying sentences, we provided a set of examples featuring both violent and non-violent sentences from historical texts. These examples, designed to reflect the complexity of violence in ancient narratives, included explicit acts of violence such as battles and executions, alongside non-violent descriptions of cultural or daily life.

To ensure the model understood the nuances of classifying violence in historical texts, we crafted a detailed prompt. The prompt clearly outlined the criteria for classifying sentences and addressed potential ambiguities around symbolic, verbal, or indirect forms of violence. The full prompt used in the experiment is provided below:

> *You are a historian that classifies historical texts into violent or non-violent based on the provided examples. Respond with only* ***[VIOLENT]*** *or* ***[NON-VIOLENT]*** *for each classification.*
>
> *The following principles apply to the classification of violent acts:*
>
> - *Arrests of people and banishments are initially recorded as acts of violence and discussed with the team before being activated.*
> - *Fictional narratives, such as the conquest of Troy, are included.*
> - *Establishment of colonies, verbal violence (insults), and damage to property (including fires in buildings, etc.) are excluded.*
>
> *Your task is to classify each passage based on the criteria above. Classify an instance as* ***[VIOLENT]*** *if it involves physical harm, threats, or coercion; classify it as* ***[NON-VIOLENT]*** *if it describes cultural practices, speeches, or symbolic gestures.*

In addition to this prompt, a few-shot learning approach was also done by providing the model with few examples and their classification. This extended prompt was designed to provide the model with clear, unambiguous guidance on what constitutes violent versus non-violent acts in historical contexts. This ensures the model will be consistent during classification. We tested the data on our specified test set to compare it with BERT and RoBERTa and also on the whole data to measure the performance across all available texts. The results are shown in Table 4.4 and Table 4.5 and the corresponding confusion matrices in Figure 4.6 and Figure 4.7

	Precision	**Recall**	**F1-Score**	**Support**
Non-Violent	0.91	0.88	0.89	371
Violent	0.69	0.74	0.71	129
Overall	0.6857	0.7442	0.7138	500
Macro Avg	0.80	0.81	0.80	500
Weighted Avg	0.85	0.85	0.85	500
Accuracy		0.8460		

Table 4.4.: Results for GPT 4o-mini on the test set

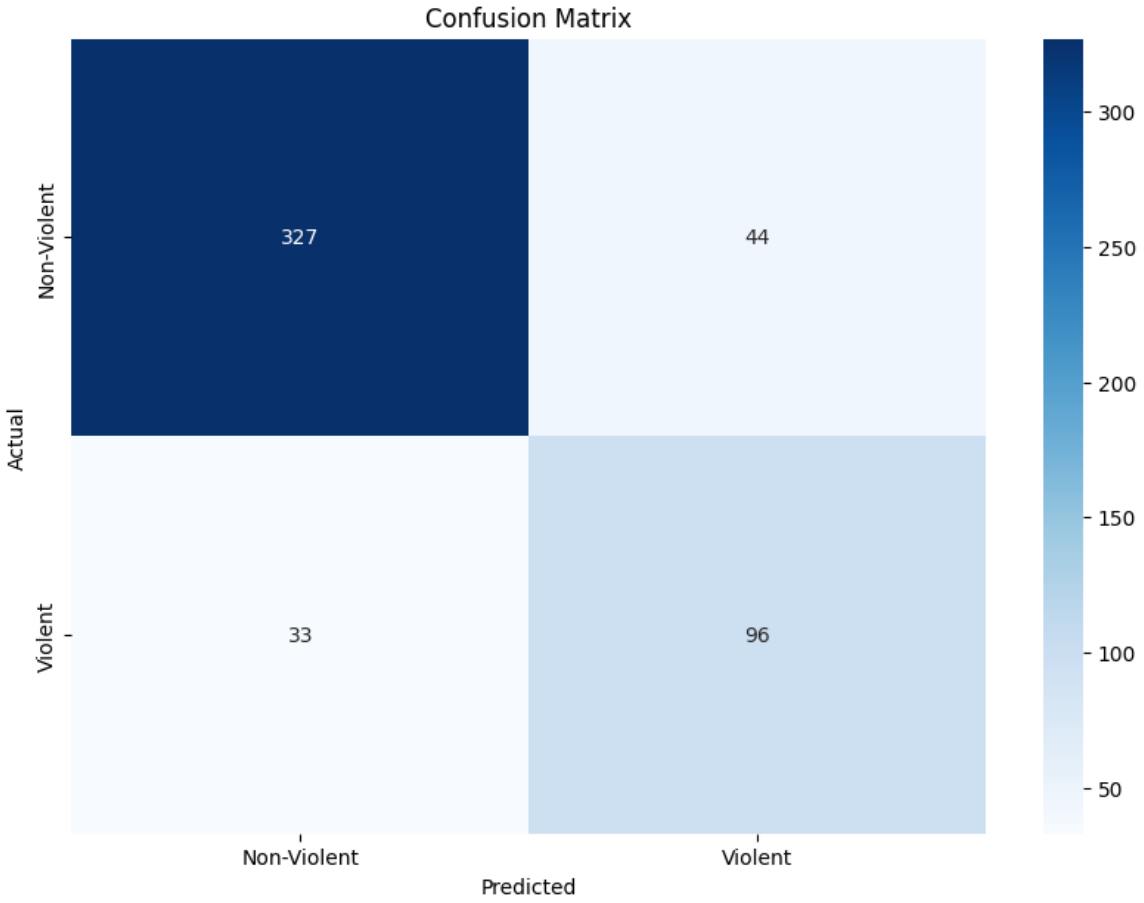


Figure 4.6.: Confusion Matrix for chatGPT 4-o API on the test set

	Precision	Recall	F1-Score	Support
Non-Violent	0.94	0.86	0.90	2103
Violent	0.54	0.75	0.62	461
Overall	0.5365	0.7484	0.6250	2564
Macro Avg	0.74	0.80	0.76	2564
Weighted Avg	0.87	0.84	0.85	2564
Accuracy		0.8385		

Table 4.5.: Results for GPT 4o-mini on all data

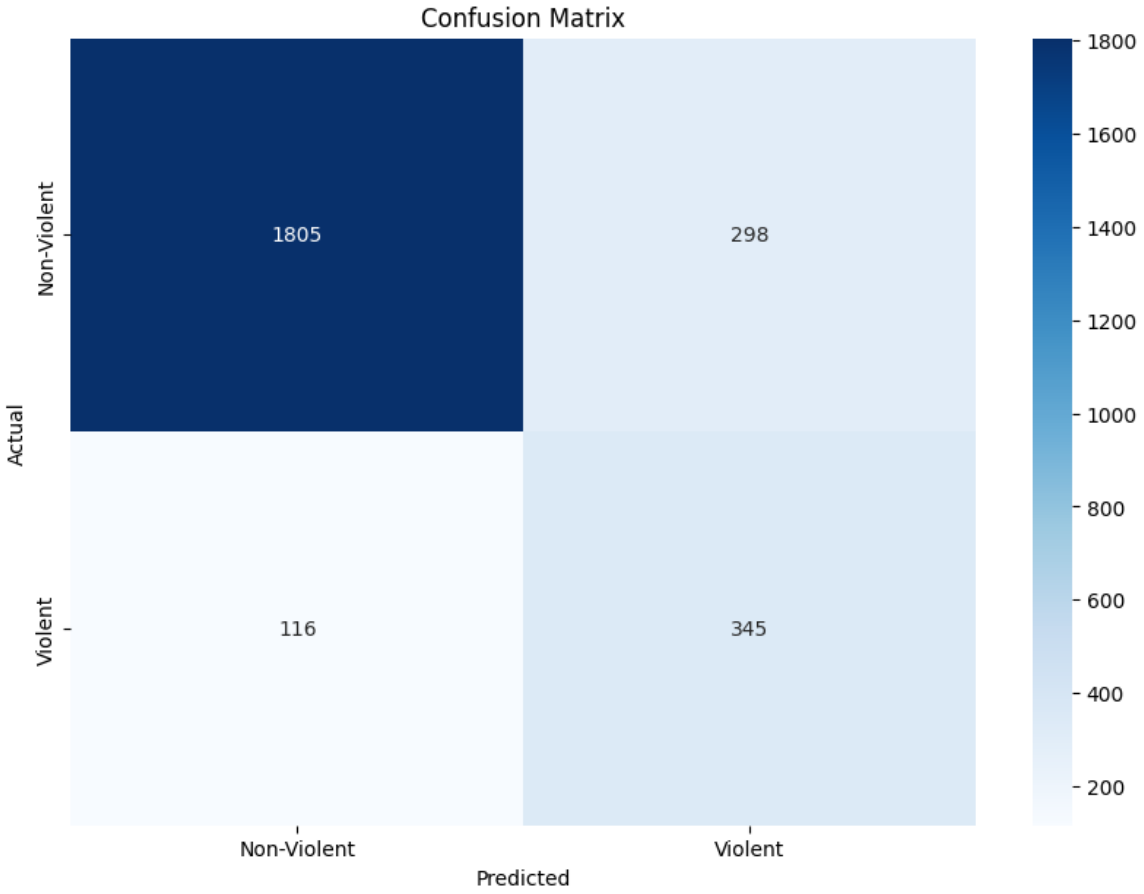


Figure 4.7.: Confusion Matrix for chatGPT 4-o API on all data

One key observation is that while GPT-4o-mini performed reasonably well in detecting violent sentences, it did not surpass the fine-tuned models in overall performance. The GPT-4o-mini model particularly struggled with identifying non-violent instances, misclassifying a higher number of these cases compared to the fine-tuned models. This discrepancy likely arises from the fact that GPT-4o-mini is a general-purpose model, lacking exposure to the specific nuances of ancient texts and the implicit forms of violence often embedded within them.
The fine-tuned models, having been trained on domain-specific data, were better equipped to handle these subtleties. This highlights the importance of domain adaptation in tasks like violence detection in historical texts, where context and nuance play a critical role. A deeper analysis of these findings will be explored in the subsequent discussion section.

4.3. Discussion

In this section, we start by explaining why reporting both precision and recall is superior to only reporting one of them. We also discuss the performance of the models in detecting violence in ancient texts and analyze the strengths, limitations, and key insights from our experiments. Additionally, we compare the results of the fine-tuned models with the performance of the GPT-4o-mini model.

4.3.1. Why F1 score is superior

Recalling from Equation 4.3 that

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

The harmonic mean in the F1 score is sensitive to disparities between precision and recall. Mathematically:

$$F_1 = \frac{2PR}{P+R} = \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{4.4}$$

This property ensures that the F1 score is low if either precision or recall is low, this would reflect the model's overall performance accurately.

Why F1 Score is Superior Reporting all three metrics—precision, recall, and F1 score—provides a comprehensive understanding of the model's performance because precision and recall capture different types of errors:

- **Precision** focuses on the correctness of positive predictions, reducing false positives.
- **Recall** emphasizes the model's ability to identify all positive instances, reducing false negatives.

A high precision with low recall indicates that while the model's positive predictions are generally correct, it misses many actual positives. Conversely, high recall with low precision means the model captures most positives but includes many false positives.
The F1 score combines precision and recall into a single metric, providing a balance between the two:

- It penalizes extreme values where one metric is high and the other is low.
- It is especially useful in imbalanced datasets, where the positive class is rare.

Limitations of Using Only Precision or Recall Relying on a single metric can be misleading:

- A model may achieve perfect precision ($P = 1$) by making very few positive predictions, resulting in low recall.
- Conversely, a model may have perfect recall ($R = 1$) by predicting all instances as positive, resulting in low precision.

Reporting all three metrics prevents such misinterpretations and this would in turn provide us with a holistic evaluation of the model.

Application-Specific Metric Prioritization Different applications prioritize different metrics:

- In medical diagnostics, **recall** is crucial to ensure that all cases are identified[104].
- In spam detection, **precision** is more important to avoid misclassifying legitimate emails as spam[105].

Thus, reporting precision, recall, and F1 score together provides a scientifically rigorous evaluation of classification models. This approach ensures that both types of errors (false positives and false negatives) are accounted for, offering a balanced assessment that is crucial for effective model deployment in practical applications.

4.3.2. Performance of Fine-Tuned Models

The fine-tuned models, particularly RoBERTa Large with data augmentation, provided the best results in terms of both precision and recall, achieving an overall F1-score of 0.96. The key observations are as follows:

- **RoBERTa Large Performance**: Fine-tuning both models on the domain-specific historical dataset yielded significant improvements compared to their non-fine-tuned counterparts.RoBERTa Large had an F1-score of 0.87, showing that it was adept at capturing the complexities of violent instances in historical narratives.
- **Impact of Data Augmentation**: The data augmentation technique we employed played a crucial role in enhancing the model's performance. By expanding the dataset through synonym replacement, we were able to expose the model to a wider variety of sentence structures, which improved generalization. The F1-score of 0.96 for RoBERTa Large with augmentation demonstrates that augmenting small datasets can yield substantial performance gains.

4.3.3. Comparison with GPT-4o-mini Model

While the fine-tuned models consistently outperformed GPT-4o-mini, the results from the ChatGPT API were still noteworthy:

- GPT-4o-mini, a general-purpose pre-trained model, showed a reasonable ability to detect violent sentences but struggled with non-violent instances. This is reflected in its lower F1-score, particularly for the non-violent class. The performance gap highlights the importance of domain-specific fine-tuning, as GPT-4o-mini lacked the nuances of historical violence that were captured by our fine-tuned models.
- Despite its lower performance, GPT-4o-mini still achieved an F1-score of 0.71 on the test set. This demonstrates its potential as a baseline model for quick classification tasks where fine-tuning may not be possible or resource-effective.

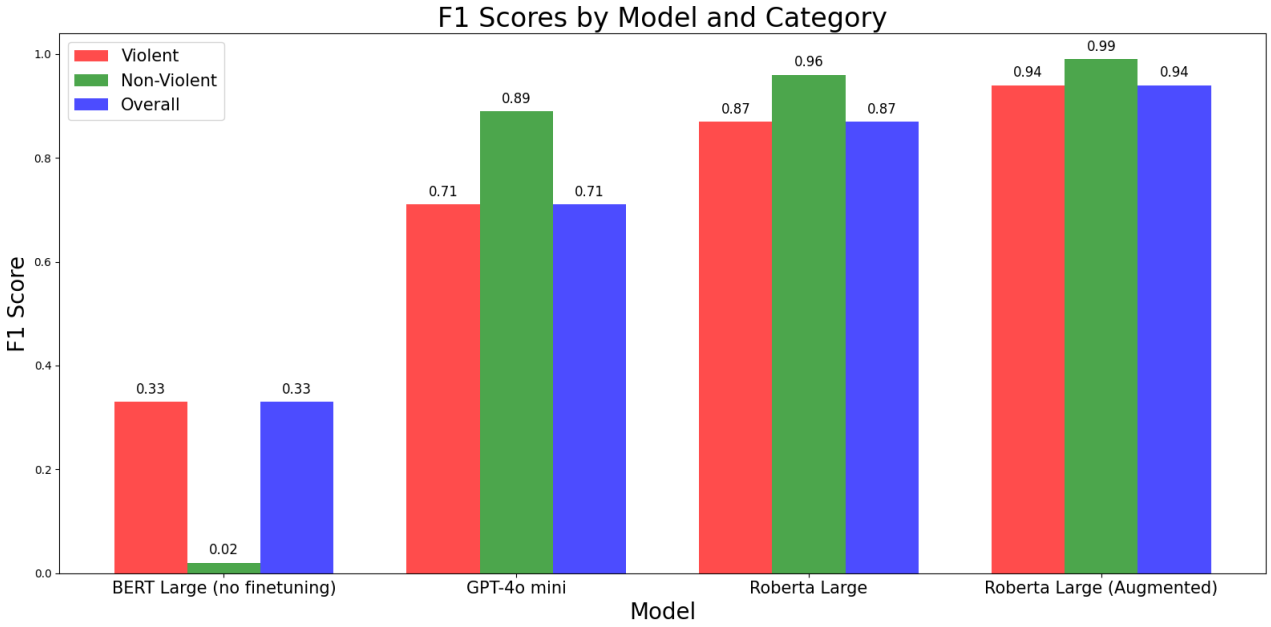


Figure 4.8.: F1 score of the different models

5. Knowledge Extraction Using Large Language Models

In the previous experiment, we fine-tuned language models to perform binary classification (detecting violent and non-violent texts). This chapter presents the second of our experiments: **Knowledge Extraction Using Large Language Models**. Here, we extend the scope to perform multi-class classification across four dimensions: level of violence, context, motive, and long-term consequence. These dimensions were chosen to provide a deeper, more granular understanding of violent events in ancient texts.

We trained four separate models, each focused on classifying one of the four dimensions. This allowed us to extract detailed information beyond simple binary classification, providing a richer framework for analyzing the socio-political dynamics, cultural context, and long-term impacts of violence in historical texts. A GitHub repository with code and data is provided in **Appendix A**

5.1. Methodology and Experimental Setup

Following the same pattern from the previous chapter, we present the methodology, including dataset preparation, model selection, training, and evaluation setup.

5.1.1. Overview

The primary goal of this experiment is to extract specific categories of knowledge from ancient texts using fine-tuned large language models. Several publications demonstrate the use of LLMs for knowledge extraction or in similar domains, such as information extraction from scientific texts [106], biomedical retrieval [107], material science information extraction [108], relation extraction [109], and cross-lingual language retrieval [110]. Additionally, surveys such as those by Zhu et al. [111], Wang et al. [112], and Zhai et al. [113] provide comprehensive overviews of how large language models have been applied in various knowledge extraction tasks.

5.1.2. Dataset Preparation

In this experiment, we use the ERIS dataset [95], which we presented in Chapter 2. The dataset contains 3251 ancient texts from different ancient eras. A snippet of the data is shown in Figure 5.1. One issue with the dataset was that some texts from more recent eras were untranslated, leaving blank entries that would have negatively affected model training. We addressed this by cleaning these entries, resulting in a final dataset of 3165 texts. In Table 5.1, we show the classes of the Motive category. The remaining categories are provided in Appendix B.

	A	B	C	D	E	F	G	H	I	K
		title	level	period	context	motive	applicati	weapon	citation.citationsenglish	longtermconsequence
677		Pericles ra	intersocial	Classical Greece	military	political	Unknow	Unknown	For the expedition around the Peloponnesus ravaged much terr	destruction/devastation
678		The Pelopc	intersocial	Classical Greece	military	political	Unknow	Unknown	Wherein also it was evident that though their enemies did the At	destruction/devastation
679		The Pelopc	intersocial	Classical Greece	military	political	Unknow	Unknown	Wherein also it was evident that though their enemies did the At	destruction/devastation
680		Pericles be	intersocial	Classical Greece	military	political	Unknow	Unknown	Well, then, on sailing forth, Pericles seems to have accomplish	issuing of law/decrees
681		A pentathle	interpersonal	Classical Greece	civilian	none/accident	throwing	javelin	For instance, a certain athlete had hit Epitimus the Pharsalian w	other
682		Less than f	intersocial	Classical Greece	jurisdictional	ambition	Unknow	Unknown	As a result, a little less than five thousand were convicted and s	Unknown
683		Pericles th	intersocial	Classical Greece	military	political	Unknow	Unknown	This was the son who afterwards conquered the Peloponnesian	death
684		The Atheni	intersocial	Classical Greece	jurisdictional	emotional	Unknow	Unknown	This was the son who afterwards conquered the Peloponnesian	Unknown
685		Pericles is	intersocial	Classical Greece	military	political	Unknow	Unknown	Being now near his end, the best of the citizens and those of his	bestowing of honors
686		Alexander	interpersonal	Hellenistic Greece	regicide	emotional	poisonin	poison	They say that Antipater, who had been left by Alexander as vicer	Unknown
687		Olympias b	interpersonal	Classical Greece	familicide	ambition	other	other/com	[7] On the death of Philip, his infant son by Cleopatra, the niece	death
688		Olympias k	interpersonal	Hellenistic Greece	familicide	ambition	Unknow	Unknown	On the death of Philip, his infant son by Cleopatra, the niece of A	death
689		Cassander	intrasocial	Hellenistic Greece	institutional	political	throwing	stone	My own view is that in building Thebes Cassander was mainly int	victory
690		Cassander	interpersonal	Hellenistic Greece	regicide	ambition	poisonin	poison	My own view is that in building Thebes Cassander was mainly int	victory
691		Cassander	interpersonal	Hellenistic Greece	familicide	ambition	poisonin	poison	[2] My own view is that in building Thebes Cassander was mainly	victory
692		Antipater k	interpersonal	Hellenistic Greece	matricide	ambition	Unknow	Unknown	[3] Philip, the eldest of his sons, shortly after coming to the thro	other
693		Demetrius	interpersonal	Hellenistic Greece	regicide	ambition	Unknow	Unknown	[3] Philip, the eldest of his sons, shortly after coming to the thro	Unknown
694		Arsinoe pl	interpersonal	Hellenistic Greece	conspiracy	ambition	Unknow	Unknown	[3] Love is wont to bring many calamities upon men. Lysimachu	Unknown
695		Ptolemaeu	interpersonal	Hellenistic Greece	war/military campaign	ambition	Unknow	Unknown	[2] After these successes, which were shortly followed by the fa	Unknown
696		Perseus or	interpersonal	Hellenistic Greece	conspiracy	ambition	Unknow	Unknown	Then a disagreement arose, some thinking that they should retu	Unknown
697		Ptolemaeu	interpersonal	Hellenistic Greece	fratricide	political	Unknow	Unknown	This Ptolemy fell in love with Arsinoe, his full sister, and married	Unknown

Figure 5.1.: A snippet of the ERIS dataset.

Motive	Records
Tactical/Strategical	1054
Political	598
Following Orders	308
Emotional	210
Economical	176
Ambition	170
Unknown	82
Self-defence	62
Social	44
Other	38
Religious	18
None/Accident	18
Unknown	2

Table 5.1.: Records of Motive. The remaining categories are found in Appendix A.

5.1.3. Model Selection and Fine-tuning

In this experiment, we directly used RoBERTa Large, as it provided the best results in the previous experiment. We followed the same training and testing paradigm, with a standard 80-20 data split for training and testing.

5.1.4. Evaluation Metrics

In this experiment, we also employed the same evaluation metrics—Precision, Recall, and F1-score—as detailed in Chapter 4.

5.2. Results

In this section, we present the results of the multi-class classification tasks for each of the four dimensions: level of violence, context, motive, and long-term consequence. Each model was trained separately for each dimension, and the results include a breakdown of performance metrics across the respective categories.

5.2.1. Level of Violence Classification

In this task, the model was designed to classify instances of violence into four categories: interpersonal, intrapersonal, intersocial, and intrasocial. These categories reflect different relational dimensions of violent

acts, with interpersonal violence occurring between individuals, intrapersonal violence occurring within an individual, intersocial violence involving conflict between social groups, and intrasocial violence occurring within a single social group. Results are shown in Table 5.2. The confusion matrix is shown in Figure 5.2

	Precision	Recall	F1-Score	Support
Interpersonal	0.92	0.91	0.91	96
Intrasocial	0.95	0.83	0.89	72
Intersocial	0.96	0.98	0.97	371
Intrapersonal	0.84	0.94	0.89	17
Overall	0.9463	0.9460	0.9455	556
Macro Avg	0.92	0.91	0.91	556
Weighted Avg	0.95	0.95	0.95	556
Accuracy		0.9460		

Table 5.2.: Level Results

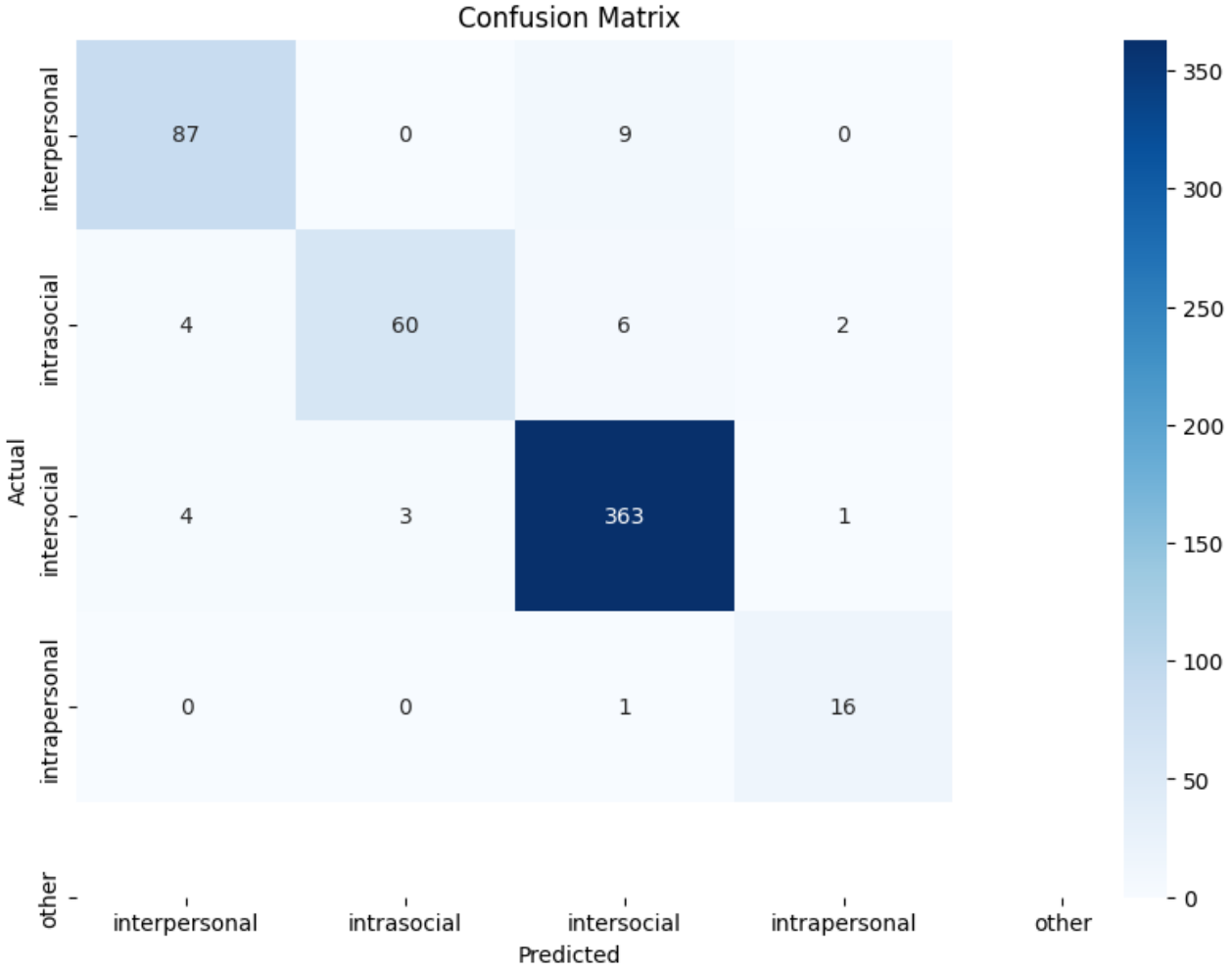


Figure 5.2.: Confusion Matrix for Level of Violence Classification

5.2.2. Context Classification

In this task, the model was required to classify the context in which violent events occurred. The contexts were divided into political, military, and social settings. A political context refers to events driven by governmental, legislative, or diplomatic actions, while a military context involves armed conflicts, invasions,

or other war-related events. The social context is important because it encompasses interactions between individuals or groups within society that are shaped by social norms, cultural practices, and daily life. Results are shown in Table 5.3 and Figure 5.3

	Precision	**Recall**	**F1-Score**	**Support**
Civilian	1.00	0.69	0.82	29
Jurisdictional	0.86	0.80	0.83	30
War/Military Campaign	0.80	0.94	0.87	181
Battle	0.93	0.81	0.87	69
Plunder	0.69	0.53	0.60	17
Ambush	0.85	0.73	0.79	15
Conspiracy	0.82	0.82	0.82	11
Revolt	1.00	1.00	1.00	21
Conquest	0.50	0.57	0.53	7
Naval Battle	1.00	1.00	1.00	2
Religious	1.00	1.00	1.00	6
Institutional	0.60	0.75	0.67	4
Sack	0.00	0.00	0.00	1
Single Combat	1.00	0.50	0.67	4
Siege	0.83	0.81	0.82	31
Unknown	1.00	1.00	1.00	5
Regicide	0.69	1.00	0.81	11
Military	0.90	0.87	0.89	93
Entertaining	0.60	0.43	0.50	7
Mutiny	1.00	0.75	0.86	8
Familicide	1.00	1.00	1.00	2
Fratricide	0.00	0.00	0.00	1
Paramilitary	1.00	1.00	1.00	1
Overall	0.8524	0.8471	0.8437	556
Macro Avg	0.79	0.74	0.76	556
Weighted Avg	0.85	0.85	0.85	556
Accuracy		0.8471		

Table 5.3.: Context Results

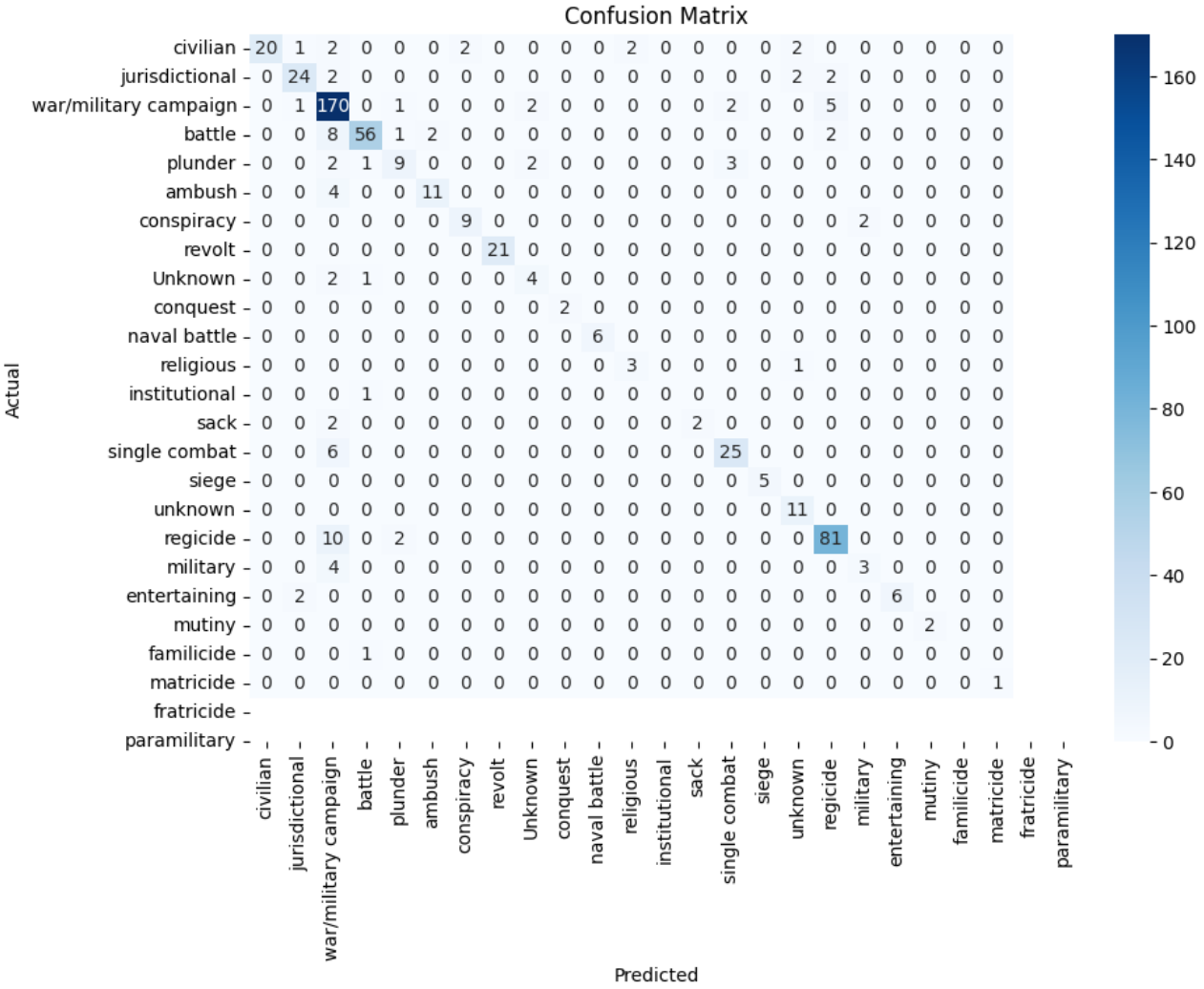


Figure 5.3.: Confusion Matrix for context Classification

5.2.3. Motive Classification

The motive classification aims to categorize the underlying reasons behind violent actions. The model was tasked with distinguishing between various motives such as tactical or strategic considerations, political ambitions, orders followed from a higher authority, emotional impulses, and economic gains. Each motive represents a different driving force behind the violent act. For example, a tactical or strategic motive may relate to achieving a military objective, while an emotional motive may stem from personal anger or revenge. Other motives include self-defense, ambition, religious causes, or accidental and unintended actions. Classifying motives helps us to elucidate the rationale behind the violent events. Results are shown in Table 5.4 and Figure 5.4

	Precision	Recall	F1-Score	Support
Unknown	1.00	0.80	0.89	20
Political	0.84	0.86	0.85	122
Tactical/Strategical	0.87	0.88	0.87	197
Economical	0.74	0.82	0.78	28
Following Orders	0.90	0.86	0.88	77
Self-Defence	0.75	0.69	0.72	13
Emotional	0.97	0.77	0.86	43
Ambition	0.71	0.83	0.76	35
Social	0.71	1.00	0.83	5
Religious	0.83	0.83	0.83	6
Other	1.00	1.00	1.00	6
None/Accident	0.75	0.75	0.75	4
Overall	0.8583	0.8525	0.8533	556
Macro Avg	0.84	0.74	0.75	556
Weighted Avg	0.86	0.85	0.85	556
Accuracy		0.8525		

Table 5.4.: Motive Results

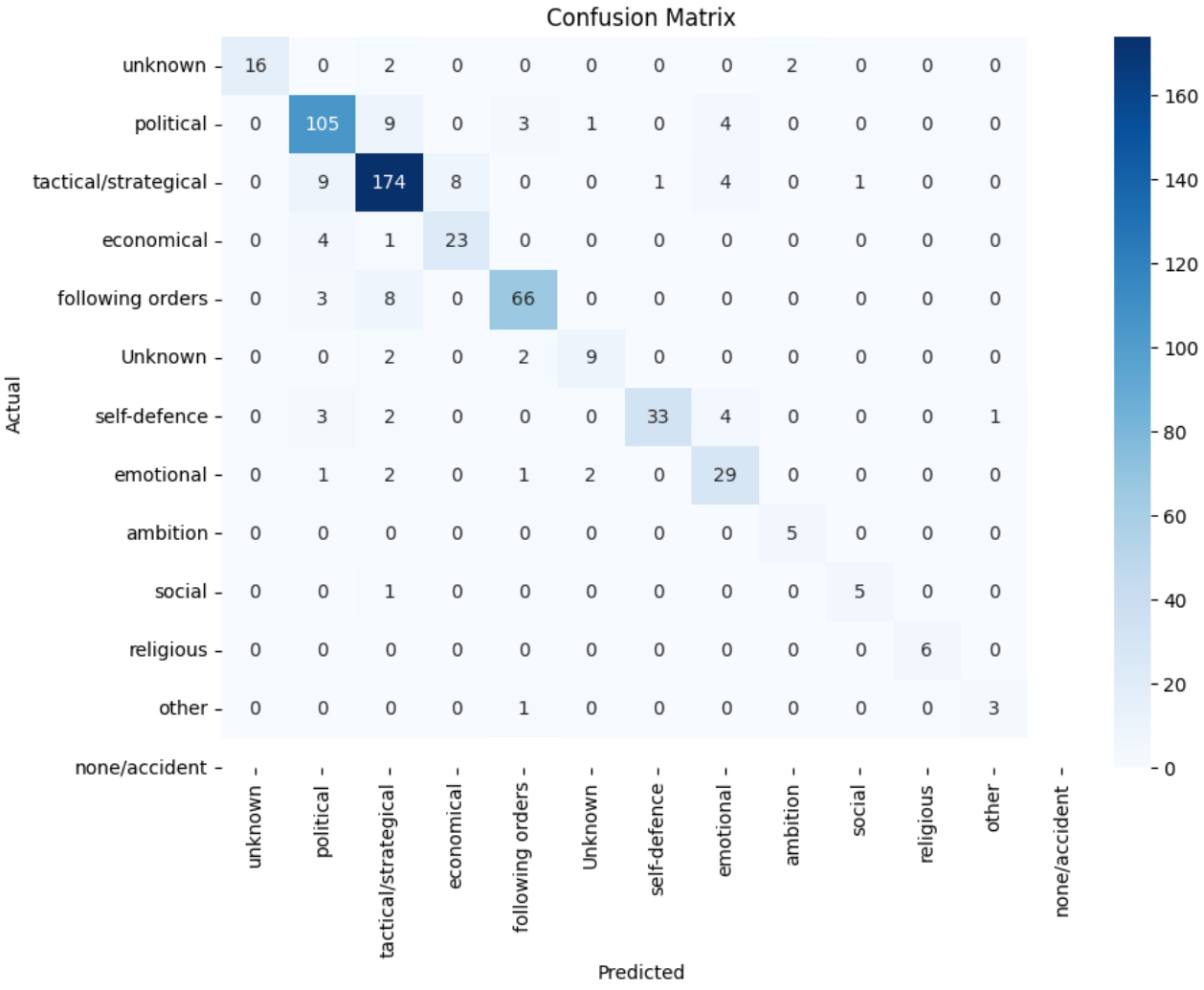


Figure 5.4.: Confusion Matrix for motive Classification

5.2.4. Long-Term Consequence Classification

In this task, the model categorized the long-term consequences of violent events. The long-term consequences were divided into categories such as social disruption, political change, and personal consequences. Social disruption refers to the breakdown of social structures or relationships following violent acts, often leading to instability within a community. Political change includes shifts in governance, power dynamics, or territorial control that result from violent conflict. Personal consequences focus on the individual level, such as psychological trauma, injury, or death. This classification highlights the broader impact of violence beyond the immediate event. Results are shown in Table 5.5 and Figure 5.5

Table 5.5.: Long-term consequence results

Category	Precision	Recall	F1-Score	Support
Unknown	0.78	0.89	0.83	199
Campaign	0.81	0.87	0.85	28
Conquest	0.83	0.83	0.83	24
Coronation/Inauguration	1.00	0.67	0.80	12
Exile	1.00	0.67	0.80	6
Death	0.81	0.72	0.72	32
Other	0.72	0.72	0.72	32
Victory	1.00	1.00	1.00	16
Bestowing of Honors	0.67	0.33	0.44	6
Issuing of Law/Decrees	1.00	0.33	0.50	3
Injury	0.71	1.00	0.83	5
Battle	0.80	0.53	0.64	15
Declaration of War	1.00	1.00	1.00	2
Retreat	0.67	0.80	0.73	10
Mutiny	1.00	0.50	0.67	2
Sending of Envoys	0.93	1.00	0.96	13
Civil Conflict/Civil War	0.00	0.00	0.00	1
Tyranny	0.50	1.00	0.67	2
Capture	0.71	0.71	0.71	14
Destruction/Devastation	0.84	0.81	0.82	26
Repopulation	1.00	1.00	1.00	2
Declaration of Peace/Truce	1.00	0.44	0.62	9
Release of Prisoners	1.00	1.00	1.00	2
Garrisoning of Troops	1.00	0.67	0.80	6
Famine	1.00	1.00	1.00	1
Siege	0.95	0.70	0.81	30
Deportation	1.00	0.25	0.40	4
Treaty/Agreement/Pact	1.00	0.33	0.50	3
Surrender	0.67	1.00	0.80	2
Financial Reward	0.75	1.00	0.86	3
Seclusion	0.33	1.00	0.50	2
Plunder	0.86	1.00	0.92	6
Mutilation	1.00	1.00	1.00	1
Revenge	1.00	1.00	1.00	6
Execution	0.40	0.50	0.44	4
Torture	0.75	1.00	0.86	3
Applause	1.00	0.50	0.67	2
Overall	0.8190	0.8040	0.7986	556
Macro Avg	0.82	0.75	0.75	556
Weighted Avg	0.82	0.80	0.80	556

Continued on next page

Table 5.5 – continued from previous page

Category	Precision	Recall	F1-Score	Support
Accuracy		0.8040		

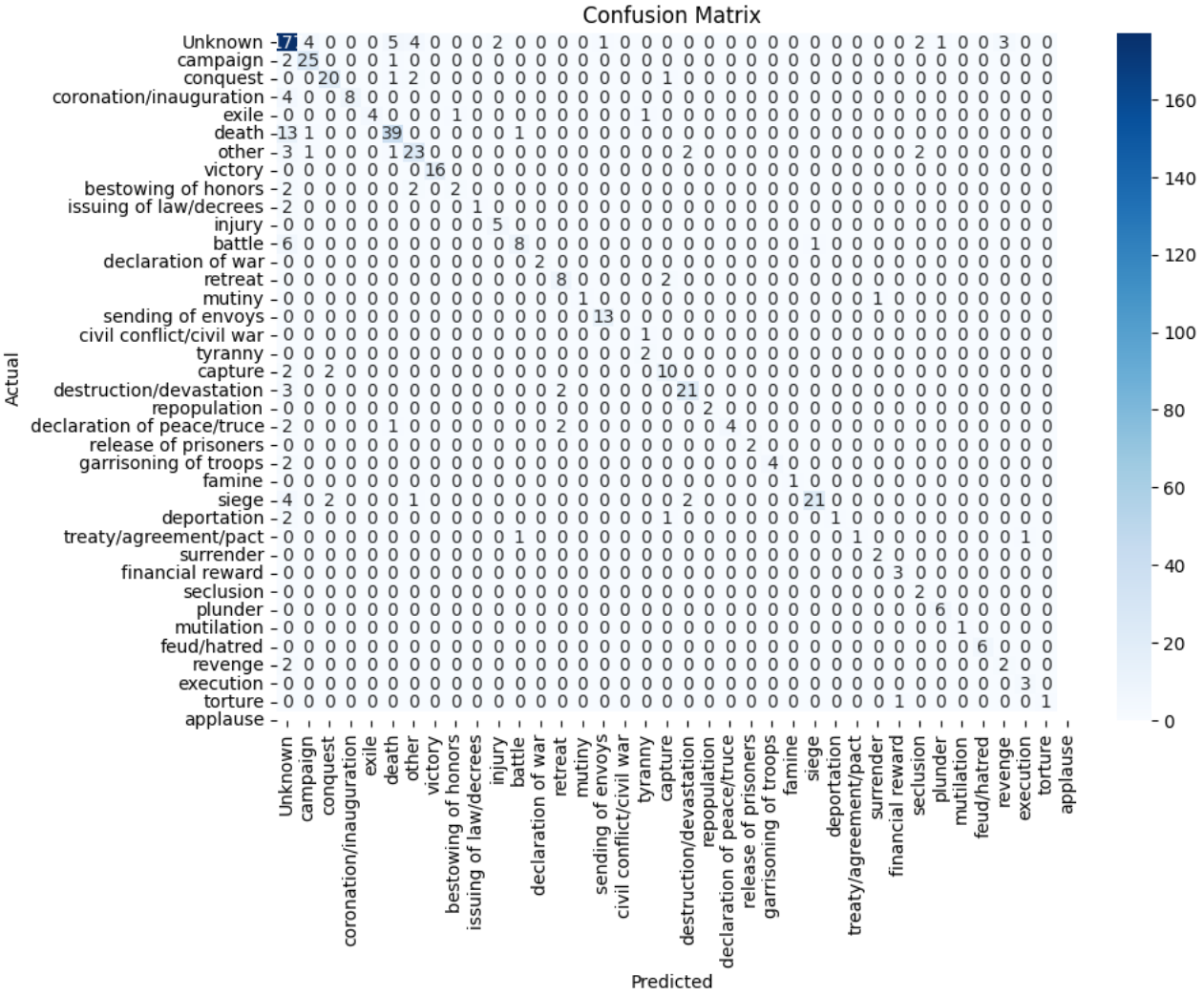


Figure 5.5.: Confusion Matrix for long-term consequence Classification

5.3. Discussion

In this experiment, we extended our analysis beyond binary classification by using large language models to classify four dimensions of violence: level of violence, context, motive, and long-term consequence. The multi-class classification tasks presented challenges and opportunities, revealing insights into how well fine-tuned models can handle complex historical texts. In this section, we will discuss the overall performance, strengths, and limitations of the models, with particular attention to category-specific challenges and areas for improvement.

5.3.1. Performance Across Classifications

The fine-tuned models demonstrated robust performance across most categories, especially in tasks where the training data was well-represented and clearly defined. However, several areas posed difficulties, particularly when dealing with conceptually overlapping categories or those with fewer instances in the dataset.

Level of Violence Classification The model achieved high precision and recall in distinguishing between different levels of violence, particularly in the interpersonal and intersocial categories. This reflects the model's ability to capture the dynamics of violence that occur between individuals and social groups, which are common themes in historical texts. However, the classification of intrapersonal violence—violence occurring within an individual—proved more challenging, likely due to its lower representation in the dataset. Intrapersonal violence, often psychological or internal, requires subtle contextual understanding, which may have contributed to the model's lower precision in this category. Future improvements could focus on expanding the dataset to provide more examples of intrapersonal violence, as well as refining the model's ability to identify less overt forms of violence.

Context Classification In the context classification task, the model excelled in identifying prominent and distinct categories, such as "War/Military Campaign" and "Battle." However, confusion arose in distinguishing between closely related contexts, such as large-scale military campaigns and more focused combat scenarios like battles or single combat. This suggests that while the model is proficient in recognizing broad categories, it struggles to differentiate between nuanced variations within these categories. Moreover, the misclassification of "Regicide" as "War/Military Campaign" highlights the difficulty in separating specific events involving leadership (e.g., the assassination of a ruler) from broader military conflicts. This points to the need for more granular contextual cues and annotations to help the model differentiate between distinct but related historical events.

Motive Classification The classification of motives showed strong performance in identifying broad categories like "Tactical/Strategical" and "Political" motives. However, the model encountered difficulties with more nuanced and less frequently occurring motives, such as "Emotional" and "Ambition." These categories often involve personal motivations that are harder to distinguish from strategic or tactical objectives, leading to variability in precision and recall. Additionally, there was notable confusion between "Political," "Following Orders," and "Tactical/Strategical" motives, reflecting the conceptual overlap between these categories. Political actions, especially in historical military contexts, often serve tactical or strategic goals, making it difficult for the model to separate them clearly. Similarly, orders followed in hierarchical organizations, like the military, are typically part of broader tactical plans, further blurring the lines between these categories. To address this, future iterations of the model could benefit from hierarchical classification techniques or enhanced annotation guidelines that better define the boundaries between overlapping motives.

Long-Term Consequence Classification The long-term consequence classification task produced solid overall results, with the model performing well in identifying categories such as "Destruction/Devastation" and "Victory." These categories, being more concrete and frequently referenced in historical texts, were easier for the model to classify accurately. However, as seen in the motive classification, categories with fewer examples in the dataset, such as "Exile" or "Coronation," presented challenges. Lower precision and recall in these categories can be attributed to their smaller representation, which limits the model's ability to generalize. The complexity of some long-term consequences, which may be abstract or harder to define (e.g., political changes, psychological impacts), further added to the variability in performance. Addressing this issue would require expanding the dataset to include more instances of these less common outcomes, as well as enhancing the model's contextual understanding of abstract consequences.

5.3.2. Key Findings and Limitations

The results of this multi-class classification experiment highlight several important findings:

- **Class Imbalance and Data Scarcity**: Categories with fewer examples, such as "Intrapersonal Violence" and "Exile," were more difficult for the model to classify, leading to lower precision and recall. Increasing the representation of these underrepresented categories in the dataset would likely improve model performance by enabling better generalization.
- **Conceptual Overlap Between Categories**: Overlap between categories—such as "Political" and "Tactical/Strategical" in motive classification, or "War/Military Campaign" and "Battle" in context

classification—led to frequent misclassifications. This suggests that while the model can distinguish between broad categories, it requires more specific guidance to differentiate between closely related or overlapping concepts. Future improvements could include adding more detailed contextual information or employing hierarchical classification techniques.

- **Performance in Concrete vs. Abstract Categories**: The model performed well in classifying well-defined, concrete categories like "Victory" or "Destruction/Devastation" in the long-term consequence classification. However, it struggled with more abstract or less frequent categories, such as "Coronation" or "Political Change." This suggests that the model could benefit from integrating external knowledge sources to better understand the nuances of less concrete outcomes.

5.3.3. Future Directions and Improvements

Several enhancements could improve the model's performance:

- **Data Augmentation for Underrepresented Categories**: Increasing the dataset size for underrepresented categories, such as "Intrapersonal Violence" or "Exile," would help balance the model's training and improve its ability to classify these less frequent occurrences.
- **Refining Conceptual Boundaries**: Developing more refined annotation guidelines or implementing hierarchical classification[114] could help the model distinguish between categories that share conceptual similarities, such as political versus tactical motivations or different forms of military conflict. A recent work by Regneri et al.[115] shows that Large Language Models can detect conceptual abstractions and have some form of understanding conceptual clashes.
- **Incorporating External Knowledge Sources**: Integrating external historical knowledge bases or leveraging multi-modal approaches could enhance the model's ability to understand and classify abstract consequences or motives that are more difficult to define based solely on text.

In conclusion, the experiment underscores the capability of large language models to handle complex, multi-class classification tasks within the domain of ancient historical texts. While the models performed well in several areas, addressing the challenges of data scarcity, conceptual overlap, and abstract category classification will be critical in further refining their performance.

6. Conclusion

This thesis set out to enhance the extraction and classification of violent instances from ancient historical texts through the use of large language models (LLMs). Across two experiments, we demonstrated how fine-tuning state-of-the-art models, such as BERT and RoBERTa Large, can substantially improve the detection and classification of violence, a process that traditionally has been time-consuming and labor-intensive for human annotators. The experiments were structured to address both binary classification (violence detection) and more granular multi-class classification (knowledge extraction across various dimensions).

In the first experiment, we focused on detecting violent and non-violent texts using binary classification. By fine-tuning large language models on domain-specific datasets, we achieved high precision and recall. The results highlighted the advantages of leveraging pre-trained LLMs for specialized tasks, as the models significantly outperformed general-purpose models like GPT-4o-mini in detecting violence, especially within the nuances of ancient historical narratives. This experiment successfully automated what would otherwise be a tedious task of manual annotation, while also ensuring high accuracy in classifying instances of violence.

The second experiment extended this capability by extracting more detailed information through multi-class classification. Here, the models were tasked with classifying violent events across four critical dimensions: level of violence, context, motive, and long-term consequence. The results demonstrated that our fine-tuned models performed exceptionally well in identifying major categories, with robust precision and recall, particularly in well-represented classes. Although challenges were encountered with conceptually overlapping categories and those with fewer examples, our models still achieved notable success in automating this intricate task. This step forward showcases the potential of LLMs to facilitate deeper, more granular analyses of historical texts, allowing researchers to extract a wealth of knowledge that would otherwise require extensive manual effort.

A key achievement of this work lies in the successful automation of tasks that traditionally demand significant human labor. The models we developed were able to handle both the detection of violence and the classification of its various dimensions, which significantly speeds up the process of analyzing large historical datasets. The methodology we employed, including dataset preparation, fine-tuning, and model evaluation, offers a blueprint for future research in both the historical domain and similar fields requiring large-scale text classification. Our work successfully complements the expertise of historians, enabling them to analyze texts faster while retaining the nuance that human scholars bring to their research.

6.1. Limitations

While the results of this research demonstrate significant progress, several limitations remain. First, detecting violent texts is a difficult task even for human experts. Some sentences that might initially seem violent may not be classified as such due to the strict criteria used by historians. While our models outperform non-experts, matching the precision of historical scholars remains a challenge. Additionally, historians can identify specific actors, such as senators in political texts, through contextual clues that machines often miss. This underscores the fact that our work is designed to complement, not replace, expert analysis.

Further limitations include computational constraints. Our experiments were conducted using models no larger than RoBERTa Large due to the resource demands of training larger models, and this restricted our ability to explore even more complex architectures. Overfitting, a common issue when fine-tuning models on relatively small datasets, also posed a challenge, although data augmentation techniques helped mitigate this to some extent. Lastly, the limited amount of annotated ancient text data hindered the model's capacity for generalization, and more extensive datasets could potentially lead to even better performance.

6.2. Future Work

Looking ahead, there are several promising directions for extending this research. First, automating the extraction process for additional categories of knowledge would deepen the analysis of ancient texts and provide even more granular insights into the social, cultural, and political dynamics surrounding violent events. Larger models, such as Llama or GPT-4, could also be explored to enhance performance, especially for complex classification tasks. Incorporating reinforcement learning techniques, such as Reinforcement Learning from Human Feedback (RLHF), could further improve model accuracy by allowing continuous learning from expert historians.
Moreover, expanding the corpus of ancient texts available for training and fine-tuning, especially through collaboration with historians, would help address the challenges of limited data. Finally, a hybrid approach that combines automated classification with human review could be implemented, ensuring that models continue to complement and enhance the expertise of historians while achieving the necessary scale to analyze vast quantities of historical texts.

In conclusion, this thesis demonstrates the efficacy of large language models in automating complex classification tasks in ancient historical texts. By fine-tuning LLMs on domain-specific data, we have laid the groundwork for a faster, more efficient approach to analyzing violence in historical narratives. Although challenges remain, the potential for continued advancements in this area is immense, with exciting prospects for both the historical and machine learning communities.

Bibliography

[1] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[2] An article explaining convolutional neural networks, https://developersbreach.com/convolution-neural-network-deep-learning/, accessed 18. Nov. 2024.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

[7] Brooks Beaulieu. Dieux et mortels: Les thèmes homériques dans les collections de l'école nationale supérieure des beaux-arts de paris. *Nineteenth-Century Art Worldwide*, 4(1), 2005.

[8] Allen D Spiegel and Christopher R Springer. Babylonian medicine, managed care and codex hammurabi, circa 1700 bc. *Journal of Community Health*, 22(1):69–89, 1997.

[9] The father in roman society, https://eaglesanddragonspublishing.com/ancient-everyday-paterfamilias-the-father-in-roman-society/, accessed 18. Nov. 2024.

[10] Daniela Carpi. Caesar's body in shakespeare's julius caesar: Sacralization and de-sacralization of power. *Pólemos*, 9(2):281–294, 2015.

[11] Wergild, compensation, and penance, https://rmblf.be/2014/08/18/colloque-wergild-compensation-and-penance-the-monetary-logic-of-early-medieval-conflict-resolution/, accessed 18. Nov. 2024.

[12] Caligula - der wahnsinnigen kaiser, https://medium.com/@ancient.rome/5-school-misconceptions-about-ancient-rome-40023fd11ce2, accessed 18. Nov. 2024.

[13] Perseus homepage, https://www.perseus.tufts.edu/hopper/, accessed 18. Nov. 2024.

[14] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[15] Jianqiong Xiao and Zhiyong Zhou. Research progress of rnn language model. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 1285–1288. IEEE, 2020.

[16] Jin Wang, Bo Peng, and Xuejie Zhang. Using a stacked residual lstm model for sentiment intensity prediction. *Neurocomputing*, 322:93–101, 2018.

[17] Junyou Shi, Qingjie He, and Zili Wang. A transfer learning lstm network-based severity evaluation for intermittent faults of an electrical connector. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 11(1):71–82, 2020.

[18] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[19] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

[20] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[24] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[28] An article about vision transformer, https://theaisummer.com/vision-transformer/, accessed 18. Nov. 2024.

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[31] Openai models, https://platform.openai.com/docs/models, accessed 18. Nov. 2024.

[32] o1-model, https://openai.com/index/introducing-openai-o1-preview/, accessed 18. Nov. 2024.

[33] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[34] Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. Chinese named entity recognition method based on bert. In *2021 IEEE international conference on data science and computer application (ICDSCA)*, pages 294–299. IEEE, 2021.

[35] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*, 2019.

[36] S Abarna, JI Sheeba, and S Pradeep Devaneyan. An ensemble model for idioms and literal text classification using knowledge-enabled bert in deep learning. *Measurement: Sensors*, 24:100434, 2022.

[37] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] Z Lan. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[39] Zejian Liu, Gang Li, and Jian Cheng. Hardware acceleration of fully quantized bert for efficient natural language processing. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 513–516. IEEE, 2021.

[40] V Sanh. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[41] Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint arXiv:1908.06725*, 2019.

[42] S Minaee, N Kalchbrenner, E Cambria, N Nikzad, and M Chenaghlu. J. gao deep learning-based text classification: a comprehensive review. *arXiv preprint arXiv:2004.03705*, 2020.

[43] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[44] A Conneau. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[45] JS McCarley, Rishav Chakravarti, and Avirup Sil. Structured pruning of a bert-based question answering model. *arXiv preprint arXiv:1910.06360*, 2019.

[46] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021.

[47] Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14284–14291, 2021.

[48] Werner Riess. *Performing Interpersonal Violence Court, Curse, and Comedy in Fourth-Century BCE Athens*. de Gruyter, 2012.

[49] James A Mercy, Susan D Hillis, Alexander Butchart, Mark A Bellis, Catherine L Ward, Xiangming Fang, and Mark L Rosenberg. Interpersonal violence: global impact and paths to prevention. *Injury prevention and environmental health. 3rd edition*, 2017.

[50] Justine Diemke, Felix K Maier, Superhero Comics, Eleonora Sereni, Ulrich Bröckling, Barbara Korte, and Ulrike Zimmermann. helden. heroes. héros.

[51] David Konstan. *The emotions of the ancient Greeks: Studies in Aristotle and classical literature*. University of Toronto Press, 2007.

[52] Douglas L Cairns. Aidōs: The psychology and ethics of honour and shame in ancient greek literature. 1993.

[53] Maureen Gallery Kovacs. *The epic of Gilgamesh*. Stanford University Press, 1989.

[54] Martha T Roth. *Law collections from Mesopotamia and Asia minor*. Scholars Press, 1995.

[55] J Dyneley Prince. The code of hammurabi, 1904.

[56] Bruce W Frier. *The rise of the Roman jurists: studies in Cicero's Pro Caecina*, volume 28. Princeton University Press, 2014.

[57] Michael Steinberg. The twelve tables and their origins: An eighteenth-century debate. *Journal of the History of Ideas*, 43(3):379–396, 1982.

[58] Garrett G Fagan. *The lure of the arena: Social psychology and the crowd at the Roman games*. Cambridge University Press, 2011.

[59] Kurt A Raaflaub et al. *War and peace in the ancient world*. Wiley Online Library, 2007.

[60] Gerda Lerner. The creation of patriarchy. *U of Oxford P*, 1986.

[61] Sarah Pomeroy. *Goddesses, whores, wives, and slaves: Women in classical antiquity*. Schocken, 2011.

[62] Geoffrey MacCormack. Inheritance and wergild in early germanic law—i. *Irish Jurist (1966-)*, 8(1):143–163, 1973.

[63] Michael Gagarin. The organization of the gortyn law code. *Greek, Roman, and Byzantine Studies*, 23(2):129–146, 1982.

[64] Bruce Louden. The gods in epic, or the divine economy. *A Companion to Ancient Epic*, pages 90–104, 2005.

[65] Douglas Cairns. The first odysseus: Iliad, odyssey, and the ideology of kingship. *GAIA. Revue interdisciplinaire sur la Grèce ancienne*, 18(1):51–66, 2015.

[66] Dorothy Natalie Witham. The battle of kadesh: Its causes and consequences. *Master Of Arts, University Of South Africa*, 2020.

[67] S Douglas Olson. The stories of agamemnon in homer's odyssey. *Transactions of the American Philological Association (1974-)*, 120:57–71, 1990.

[68] Aeschylus Aeschylus. *The Oresteia*. BoD–Books on Demand, 2019.

[69] Charles Segal. *Language and desire in Seneca's Phaedra*, volume 5074. Princeton University Press, 2017.

[70] Pieter J Sijpesteijn. Macrinus' damnatio memoriae und die papyri. *Zeitschrift für Papyrologie und Epigraphik*, 13:219–227, 1974.

[71] David Konstan. Shame in ancient greece. *Social Research: An International Quarterly*, 70(4):1031–1060, 2003.

[72] Stephen Dando-Collins. *The Ides: Caesar's Murder and the War for Rome*. Turner Publishing Company, 2010.

[73] David Cohen. The theodicy of aeschylus: Justice and tyranny in the oresteia. *Greece & Rome*, 33(2):129–141, 1986.

[74] Roger D Dawe. *Oedipus rex*. Cambridge University Press, 1982.

[75] Wendell Clausen. An interpretation of the aeneid. *Harvard Studies in Classical Philology*, 68:139–147, 1964.

[76] Donald J Mastronarde et al. *Euripides: Medea*. Cambridge University Press, 2002.

[77] John Dryden et al. *Plutarch's Lives*, volume 2. S. Low, 1859.

[78] Robin Waterfield, Carolyn Dewald, et al. *The histories*. OUP Oxford, 2008.

[79] Caius Suetonius Tranquillus and Robert Graves. *The Twelve Caesars... Translated by Robert Graves*. Cassell; printed in Czechoslovakia, 1962.

[80] Cassius Dio. Works. translation by earnest cary. loeb classical library. 9 vols, 1914.

[81] Brooke Allen. Alexander the great: Or the terrible? *The Hudson Review*, 58(2):220–230, 2005.

[82] Miriam Griffin. *Nero: the end of a dynasty*. Routledge, 2002.

[83] Stephen Dando-Collins. *Caligula: The Mad Emperor of Rome*. Turner Publishing Company, 2019.

[84] Raymond Westbrook. *A History of Ancient Near Eastern Law (2 vols): Volumes 1 and 2*, volume 72. Brill, 2003.

[85] Bruce G Trigger. *Understanding early civilizations: a comparative study*. Cambridge University Press, 2003.

[86] James M Redfield. *Nature and Culture in the Iliad: the Tragedy of Hector*. Duke University Press, 1994.

[87] Bernard S Jackson. Evolution and foreign influence in ancient law. *The American Journal of Comparative Law*, 16(3):372–390, 1968.

[88] George Bizos. Ethics, politics and law in ancient greece and contemporary south africa. *Phronimon*, 9(2):5–15, 2008.

[89] Barry L Eichler. Law and morality in ancient near eastern thought. 2009.

[90] David A Smith, Jeffrey A Rydberg-Cox, and Gregory R Crane. The perseus project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[91] Yannis Assael, Thea Sommerschield, and Jonathan Prag. Restoring ancient text using deep learning: a case study on greek epigraphy. *arXiv preprint arXiv:1910.06262*, 2019.

[92] David Bamman and Patrick J Burns. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*, 2020.

[93] Sarah Lang. Review of perseus digital library. 2018.

[94] Emily Preece and Christine Zepeda. The perseus digital library: A case study. 2009.

[95] Werner Riess and Michael Zerjadtke. Eris: Hamburg information system on greek and roman violence. *Digital Classics Online*, pages 70–75, 2015.

[96] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[97] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[98] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[99] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

[100] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[101] Sri Gowry Sritharan. *Text Mining historischer Daten: Identifikation von konflikthaften Inhalten mit Machine Learning*. Master Thesis, University of Hamburg, 2024.

[102] chatgpt api, https://openai.com/index/introducing-chatgpt-and-whisper-apis/, accessed 18. Nov. 2024.

[103] chatgpt 4o-mini, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, accessed 18. Nov. 2024.

[104] Abdul Ghaaliq Lalkhen and Anthony McCluskey. Clinical tests: sensitivity and specificity. *Continuing education in anaesthesia, critical care & pain*, 8(6):221–223, 2008.

[105] Gordon V Cormack and Thomas R Lynam. Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS)*, 25(3):11–es, 2007.

[106] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

[107] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

[108] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[109] Somin Wadhwa, Silvio Amir, and Byron C Wallace. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access, 2023.

[110] Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–596, 2024.

[111] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

[112] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

[113] ChengXiang Zhai. Large language models and future of information retrieval: Opportunities and challenges. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 481–490, 2024.

[114] Alexander Brinkmann and Christian Bizer. Improving hierarchical product classification using domain-specific language modelling. *IEEE Data Eng. Bull.*, 44(2):14–25, 2021.

[115] Michaela Regneri, Alhassan Abdelhalim, and Soeren Laue. Detecting conceptual abstraction in llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4697–4704, 2024.

A. Code and Data

The code and data used in this thesis are publicly available on GitHub through this link:

https://github.com/AlhassanMady/Violence-classification

B. Extended results section

B.0.1. (non-finetuned models)

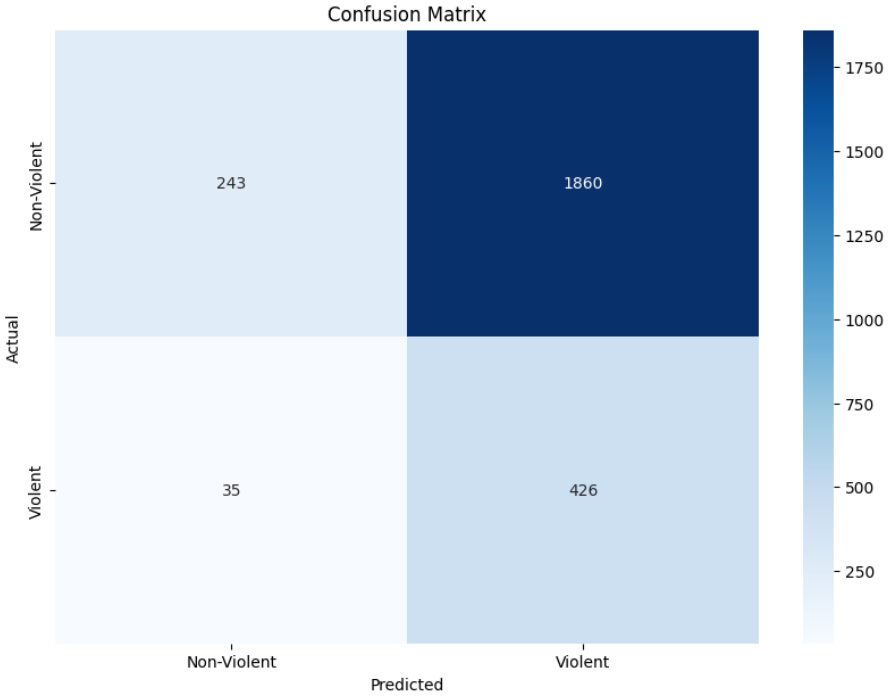


Figure B.1.: Confusion Matrix for BERT base without Fine-Tuning

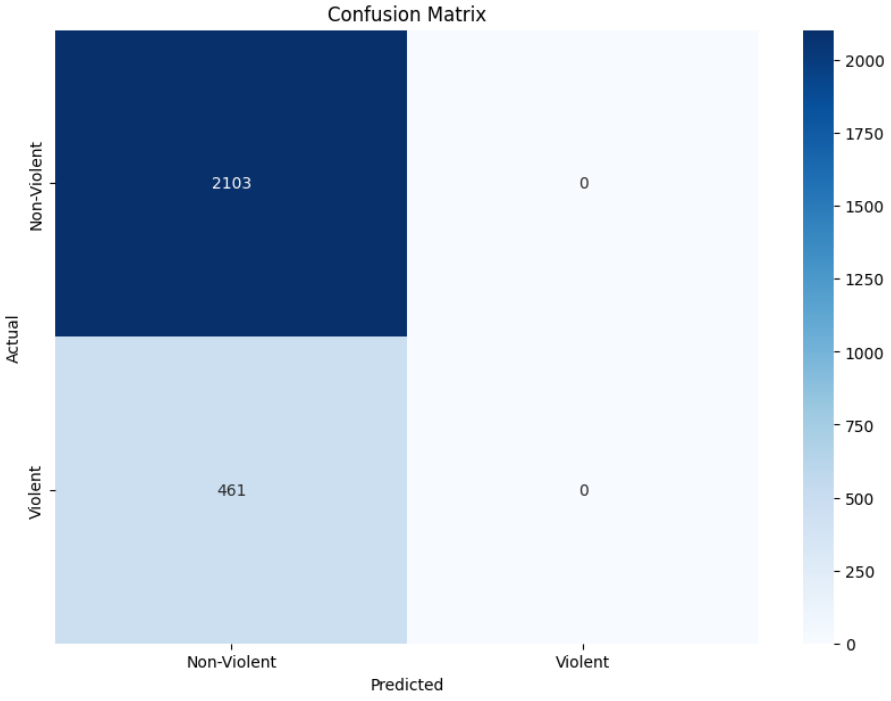


Figure B.2.: Confusion Matrix for RoBERTa Base without Fine-Tuning

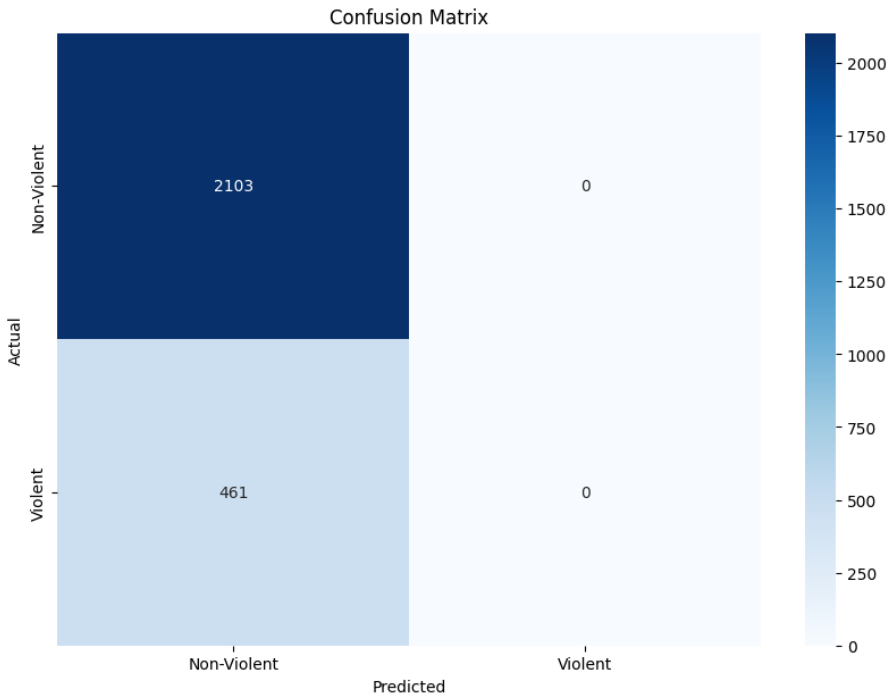


Figure B.3.: Confusion Matrix for RoBERTa Large without Fine-Tuning

B.0.2. BERT results with Finetuning

THis is tested with Random seed 42 for the test set

	Precision	Recall	F1-Score	Support
Non-Violent	0.96	0.97	0.97	396
Violent	0.91	0.89	0.90	131
Overall	0.90	0.89	0.90	527
Macro Avg	0.94	0.93	0.93	527
Weighted Avg	0.93	0.93	0.93	527
Accuracy		0.95		

Table B.1.: BERT Model results

B.0.3. Further categories and classes

Table B.2.: Records of Level of Violence.

Level of Violence	Records
Intersocial	1940
Interpersonal	384
Intrasocial	376
Intrapersonal	76
Other	4

Table B.3.: Records of Context

Context	Records
War/Military Campaign	968
Military	524
Battle	330
Siege	164
Civilian	124
Jurisdictional	106
Revolt	100
Plunder	74
Conspiracy	62
Mutiny	48
Conquest	44
Regicide	44
Ambush	38
Entertaining	22
Institutional	20
Sack	20
Naval Battle	20
Unknown	18
Single Combat	14
Religious	14
Familicide	10
Fratricide	6
Matricide	4
Paramilitary	4
Unknown	2

Table B.4.: Records of Long-term Consequence

Long-term Consequence	Records
Unknown	1064
Death	250
Other	148
Siege	142
Conquest	136
Destruction/Devastation	112
Campaign	110
Victory	100
Battle	82
Plunder	76
Capture	64
Retreat	46
Sending of Envoys	44
Coronation/Inauguration	42
Declaration of Peace/Truce	32
Injury	30
Exile	28
Bestowing of Honors	24
Mutiny	22
Deportation	22

Continued on next page

Table B.4 – continued from previous page

Long-term Consequence	Records
Issuing of Law/Decrees	20
Revenge	16
Execution	16
Financial Reward	16
Release of Prisoners	16
Garrisoning of Troops	16
Surrender	14
Repopulation	14
Mutilation	14
Declaration of War	10
Torture	10
Famine	8
Treaty/Agreement/Pact	8
Seclusion	8
Tyranny	8
Applause	6
Feud/Hatred	4
Civil Conflict/Civil War	2

```
Sentence: A lover stabs their partner in a fit of jealousy after discovering infidelity
Context: conspiracy
Motive: emotional
Consequence: death
Level: interpersonal
--------------------------------------------------
Sentence: Fanatics burn down a rival sect's temple with the priests inside.
Context: religious
Motive: political
Consequence: destruction/devastation
Level: intrasocial
--------------------------------------------------
Sentence: A gang of thieves kills a shop owner during a robbery for gold and goods.
Context: sack
Motive: economical
Consequence: plunder
Level: intersocial
--------------------------------------------------
Sentence: Ramses the second unleashed the full might of the chariots on Mageddo. This was recorded as the largest chariot battle in the bronze age
Context: war/military campaign
Motive: tactical/strategical
Consequence: victory
Level: intersocial
--------------------------------------------------
Sentence: Abel was brutally killed by his brother Cain so that Cain can marry his wife in his stead
Context: conspiracy
Motive: emotional
Consequence: death
Level: interpersonal
```

Figure B.4.: Custom-created examples and their classification

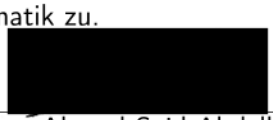
Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Intelligent Adaptive Systems selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel — insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen — benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

Hamburg, den 18.11.2024 Alhassan Ahmed Said Abdelhalim

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.



Hamburg, den 18.11.2024 Alhassan Ahmed Said Abdelhalim