Analyzing Military Doctrines Across Centuries

Project Report

Roy Zhang

Ramiro Lobo

Stats 425 Final Project

University of California, Los Angeles

Department of Statistics & Data Science

1 Introduction

This project examines the evolution of military thought by examining thematic and stylistic differences across seven classic military texts, spanning from ancient Greece and China to 19th-century Europe and the United States. By leveraging modern NLP methods and large language models, we aimed to summarize and compare military doctrines over time and uncover how strategic thought reflects historical context and specific cultures. The selected texts were downloaded from Project Gutenberg and include works from Thucydides, Julius Caesar, Sun Tzu, Arrian, Antoine-Henri Jomini, Carl von Clausewitz, and Alfred Thayer Mahan. These works range from firsthand accounts of war to more theoretical treatises, creating an interesting corpus for analysis.

Our primary objectives were to use traditional NLP techniques along with modern LLMs to explore thematic differences using topic modeling and evaluate these differences through classification. We are particularly interested in differences across centuries, distinctions between theory and historical narrative, as well as intra-period comparisons.

2 Data Collection and Preprocessing

The first step in our project involved obtaining and preparing the text corpus for analysis. We selected seven works related to military thought and strategy that are publicly available through Project Gutenberg. The selected authors represent a mixture of perspectives providing an opportunity for rich comparisons across genres and philosophies.

The texts included in our study were:

- *The Art of War* by Sun Tzu (6th century BC, China)
- *The History of the Peloponnesian War* by Thucydides (5th century BC, Greece)
- *Commentarii de Bello Gallico* by Julius Caesar (1st century BC, Rome)
- *Anabasis of Alexander* by Arrian (2nd century AD, Greece/Rome)
- *On War* by Carl von Clausewitz (19th century, Prussia)
- *Summary of the Art of War* by Antoine-Henri Jomini (19th century, Switzerland)
- *The Influence of Sea Power upon History: 1660–1783* by Alfred Thayer Mahan (19th century, United States)

To download these texts, we used the `gutenbergr` package in `R`, which allowed us to retrieve the body of the text and associated metadata, such as author and title. Once we obtained the texts, we performed several preprocessing steps to clean the texts for analysis. Clausewitz's On War was downloaded separately due to the missing chapters of the `gutenbergr` version.

To begin, we removed parts of the text that were not part of the author's content, such as publishing information, footnotes, and translator's notes, using regular expressions and manual inspection of the text. Since all of the texts, excluding Mahan's, are translations from their original language, there were often significant amounts of footnotes to be removed. Next, we standardized the texts across authors by converting characters to lowercase and removing punctuation, numeric digits, and extraneous whitespace. This step is essential to ensure uniform tokenization across the corpus of texts. We also removed common English stopwords, supplemented by a custom set of stopwords for topic modelling using BERTopic. This supplemental list included words like "chapter", Roman numerals, and labels within the book like "chapter" or "section," which had a large influence over the topic extraction process but hold little semantic value.

We tokenized the cleaned texts using standard word tokenization to support including frequency analyses and topic modeling. Additionally, because modern language models such as BERT have token limits, we also segmented the works into manageable chunks of approximately five lines of text. This chunk size was used to retain the meaning of the text sections while staying with the input size for embedding models. For analysis, we retained each chunk's author, title, and chunk number.

3 Exploratory Analysis

To better understand stylistic and structural differences across our selected texts we began with exploratory data analysis. We computed a variety of descriptive statistics, including word counts and lexical diversity. One of our metrics was type-token ratio (TTR), which measures the proportion of unique words to total words. An author with a higher TTR value suggests greater lexical variability, while lower values may indicate more repetitive language. Sun Tzu's TTR is particularly high compared to the other authors; this can be attributed to Sun Tzu's short text length. Sun Tzu's piece contains a total of 3,912 tokens, 1,423 of which are unique. In comparison, Caesar's, the second shortest of the selection, has 29,254 tokens, 3,754 of which are unique. This significant gap in TTR may suggest a cultural difference in writing style between the Western world and ancient Asia, where conciseness is appreciated. Another fact to note here is that older pieces tend to have a higher TTR. The three pieces with the highest TTR are all from ancient times. This can be attributed to the scarcity of writing media, which forces authors to express their thoughts with a minimal amount of text.

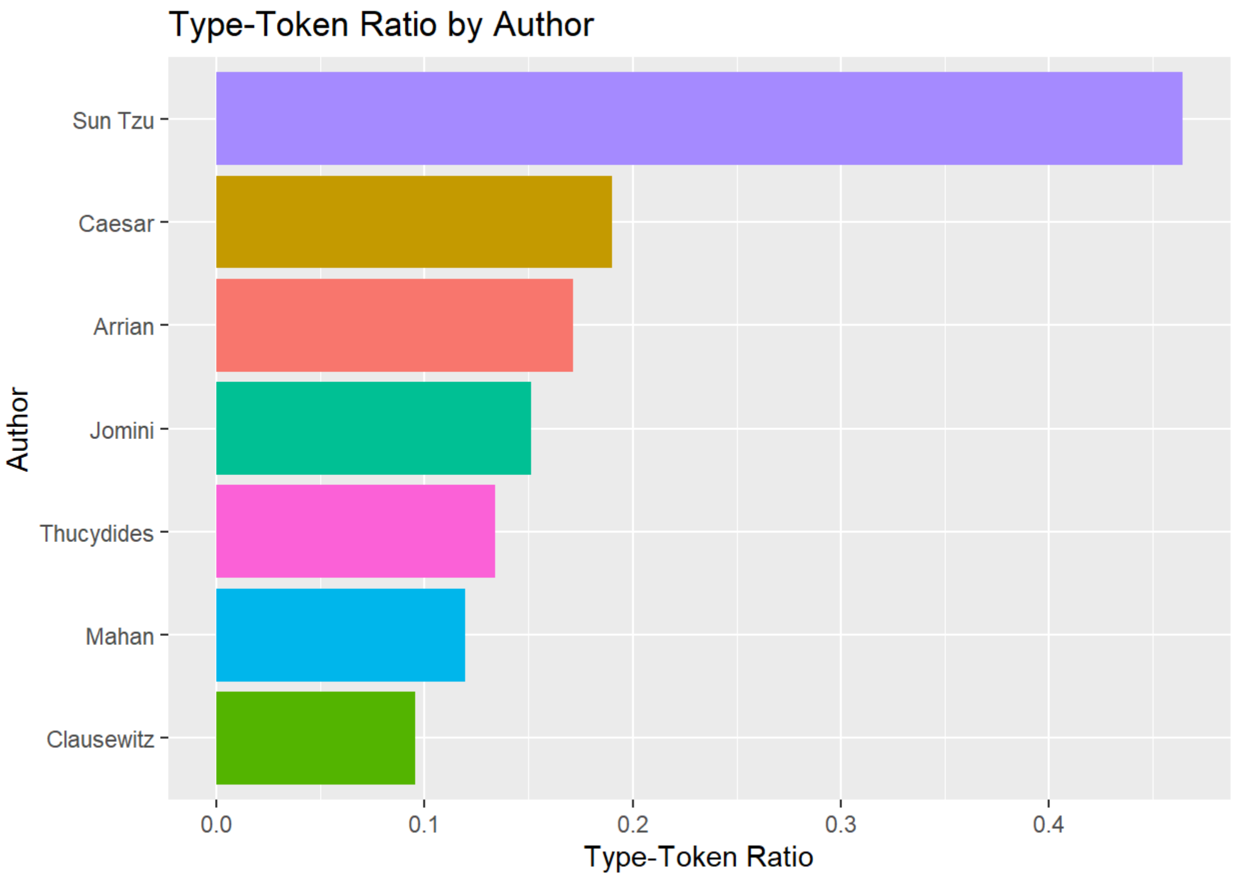

Figure 1: Authors sorted by type-token ratio. Clausewitz and Mahan showed the lowest lexical diversity, while Sun Tzu style resulted in higher TTR scores. However, this may be to the text's brevity rather than an indicator of lexical richness.

We also generated plots with the most frequent word for each author, both as raw counts and as TF-IDF-weighted terms. These visualizations helped us identify distinguishing vocabulary. The raw counts show that narrative authors such as Caesar, Arrian, and Thucydides, focus primarily on specific nouns and military terms such as, ”alexander”, ”gaul”, and ”athenians. On the other hand, more theoretical authors, such as Jomini, Clausewitz, and Sun Tzu tend to use generic or abstract terms more frequently, e.g. ”war”, ”army”, and ”force.” Mahan's vocabulary stands out from the authors' due to his emphasis on naval warfare, reflecting his focus on sea power and battles between the French and English.

When examining term frequencies weighted by TF-IDF, we observe similar distinctions between authors focused on historical narratives and those oriented toward strategic theory. For the historical authors—Arrian, Caesar, and Thucydides—the top TF-IDF terms emphasize specific people, places, and cultural encounters central to their accounts. Arrian's vocabulary reflects his focus on Alexander's campaigns along the Indus River and interactions with diverse peoples. Thucydides emphasizes key figures and factions from the Peloponnesian War, particularly those related to Athens. In Caesar's case, the Gallic campaigns are foregrounded through references to tribes, generals, and legions. In contrast, for strategic thinkers like Clausewitz and Jomini, TF-IDF weighting surfaces a shared emphasis on the Napoleonic era, with references to ’Buonaparte’, ’maneuver’, ’defense’, along with the vocabulary that underscores their theoretical framework. Similarly, for Mahan, his focus on naval power remains prominent, but now there are references to specific individuals as well.

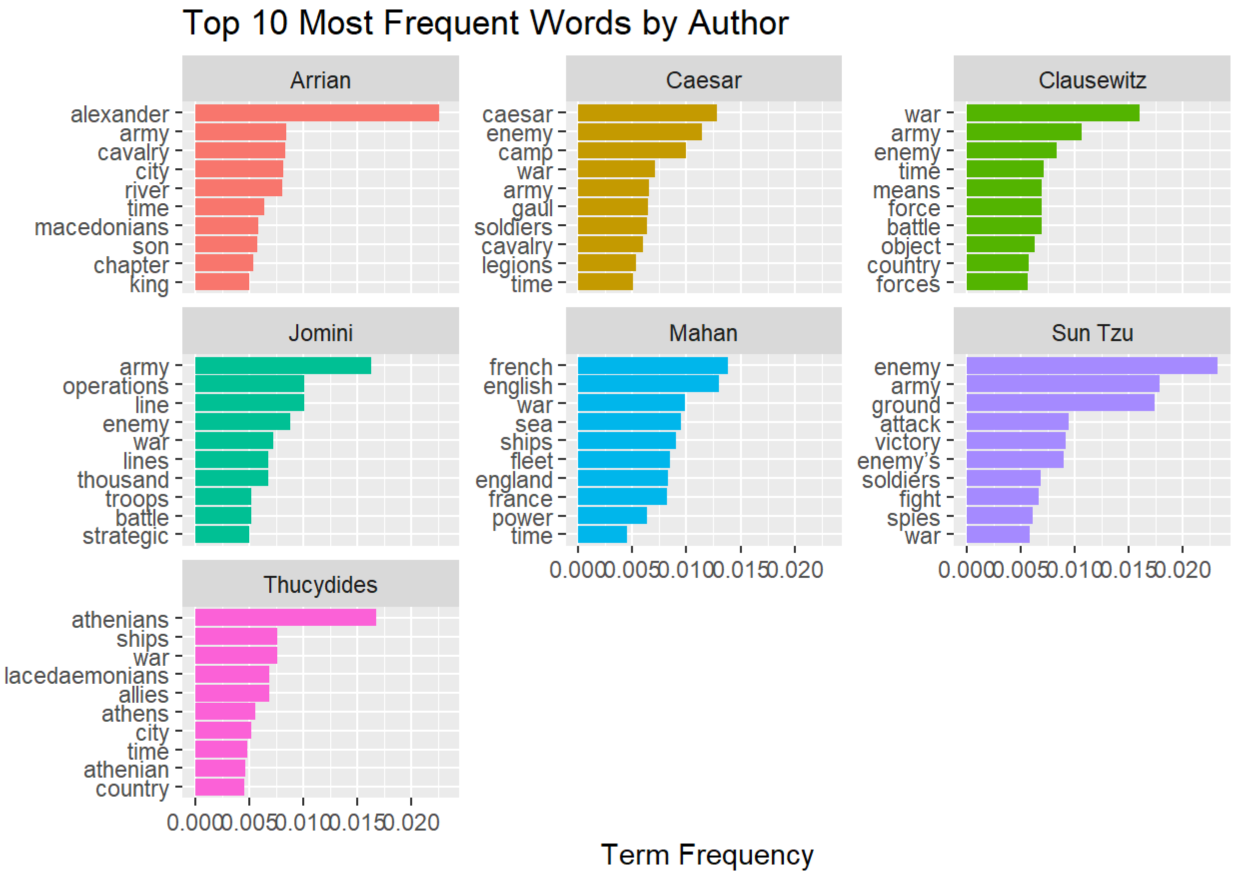

Figure 2: Top 10 words by author. Authors focusing on historical narratives tend to mention specific places and people more frequently, while more theoretical authors highlight more abstract concepts. Mahan’s focus on naval warfare is also apparent through his frequent use of the words ”sea”, ”fleet”, and ”ships.”

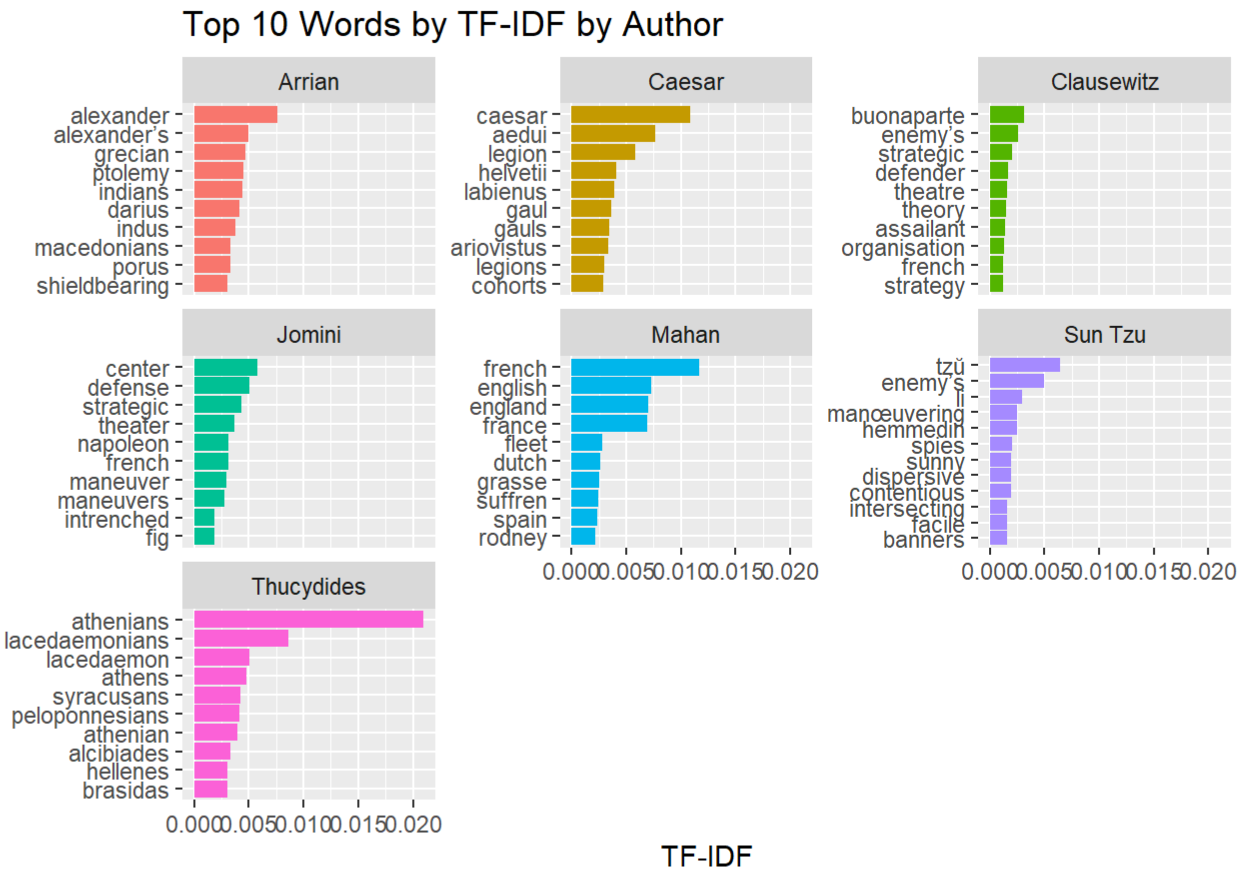

Figure 3: Top 10 TF-IDF-weighted words by author. Building on the raw frequency analysis, this plot highlights the most distinctive terms for each author after adjusting for overall corpus frequency. While raw counts showed general thematic focus, TF-IDF emphasizes what sets each author apart. Historical authors are distinguished by references to specific people and places, while strategic thinkers like Clausewitz and Jomini feature doctrinal terms related to Napoleon and military operations. Mahan's naval emphasis remains apparent, but also includes references to historical figures like Suffren, de Grasse (shown as grasse in the plot), and Rodney.

In summary, the exploratory analysis confirmed the variation in writing style, rhetorical focus, and lexical density throughout the corpus. These findings motivated our subsequent use of topic modeling techniques to deepen our understanding of strategic discourse across time and culture.

4 Author Differentiation with LDA

To confirm whether the authors' views and topics of interest differ, we first apply Latent Dirichlet Allocation (LDA). We categorized the chapters of the books into seven topics, equivalent to the number of authors. LDA breaks the documents down to combinations of topics, which are then further broken down into combinations of tokens. It assumes the documents are results of unordered samplings from various topics, which are sampled from individual words; this assumption also means the order of the words and topics has no significance. [1][3]

If there exist significant differences between the authors' topics, we should find that most of their chapters are grouped into the same topic, and each author has their distinct topic. We performed the LDA 10 times and summarized the results.

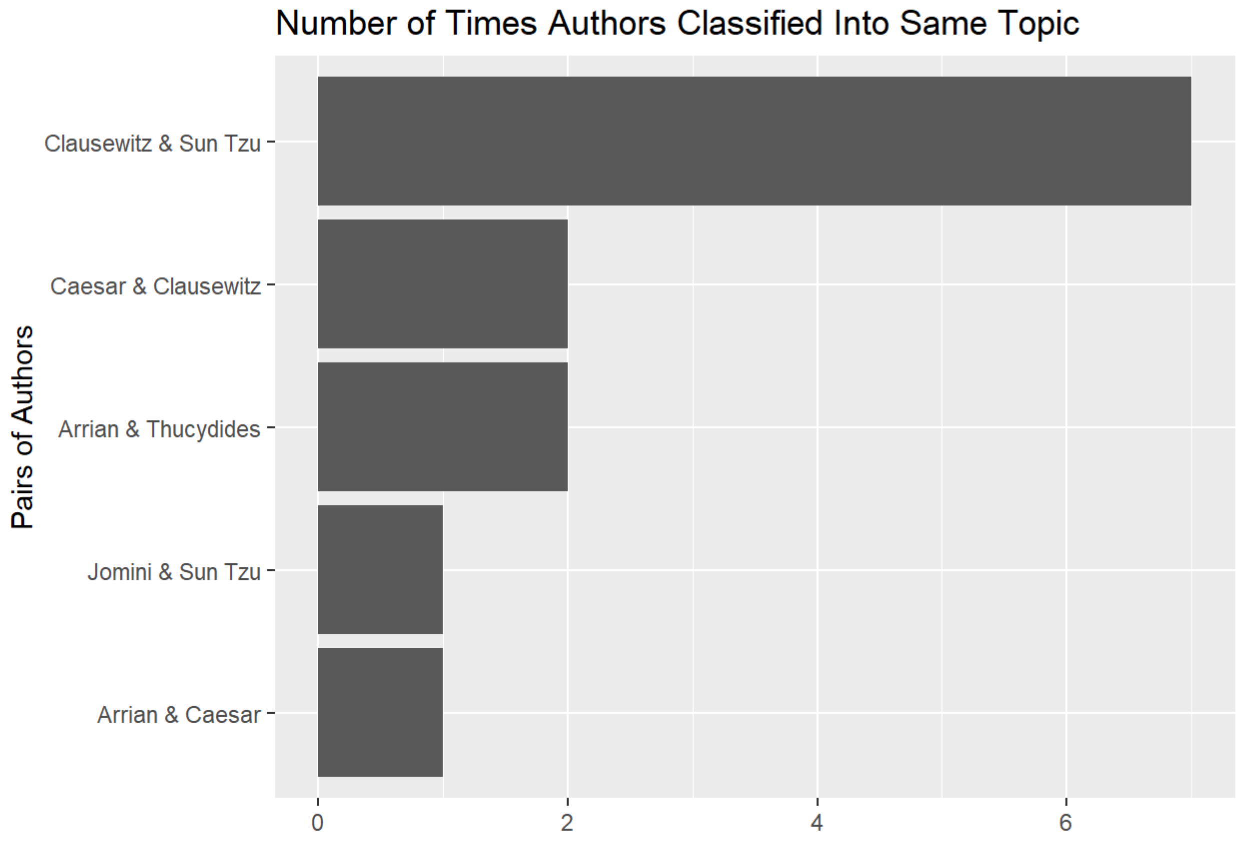

Figure 4: Bar plot of the number of times two different works get classified into the same topic. A higher number indicates that the LDA model is less capable of differentiating between these two authors, suggesting that they may share more similarities.

Figure 4 shows the frequency of LDA grouping two separate pieces in the same topic. According to the results, Clausewitz's *On War* and Sun Tzu's *Art of War* are grouped into the same topic 7 times out of 10, meaning that Sun Tzu and Clausewitz likely have highly similar topics of interest. Overall, the results align with our definition of the pieces' characteristics, as texts on military theory (Clausewitz, Sun Tzu, Jomini) and historical accounts (Arrian, Thucydides, Caesar) are grouped together. Mahan and his navy-oriented philosophy appear unique enough that the LDA never groups his work with any other piece. An interesting point to note here is that Caesar and Clausewitz are grouped together twice, despite being defined as different genres. As Caesar and Rome are hardly mentioned in Clausewitz's work, this raises the question of whether Caesar's account is merely a documentation of his career in Gaul or also a summarization of his military wisdom.

To better understand how the pieces share similarities and answer the question we now have about Caesar, we took a step further into the individual chapters that were misclassified under other people's topics.

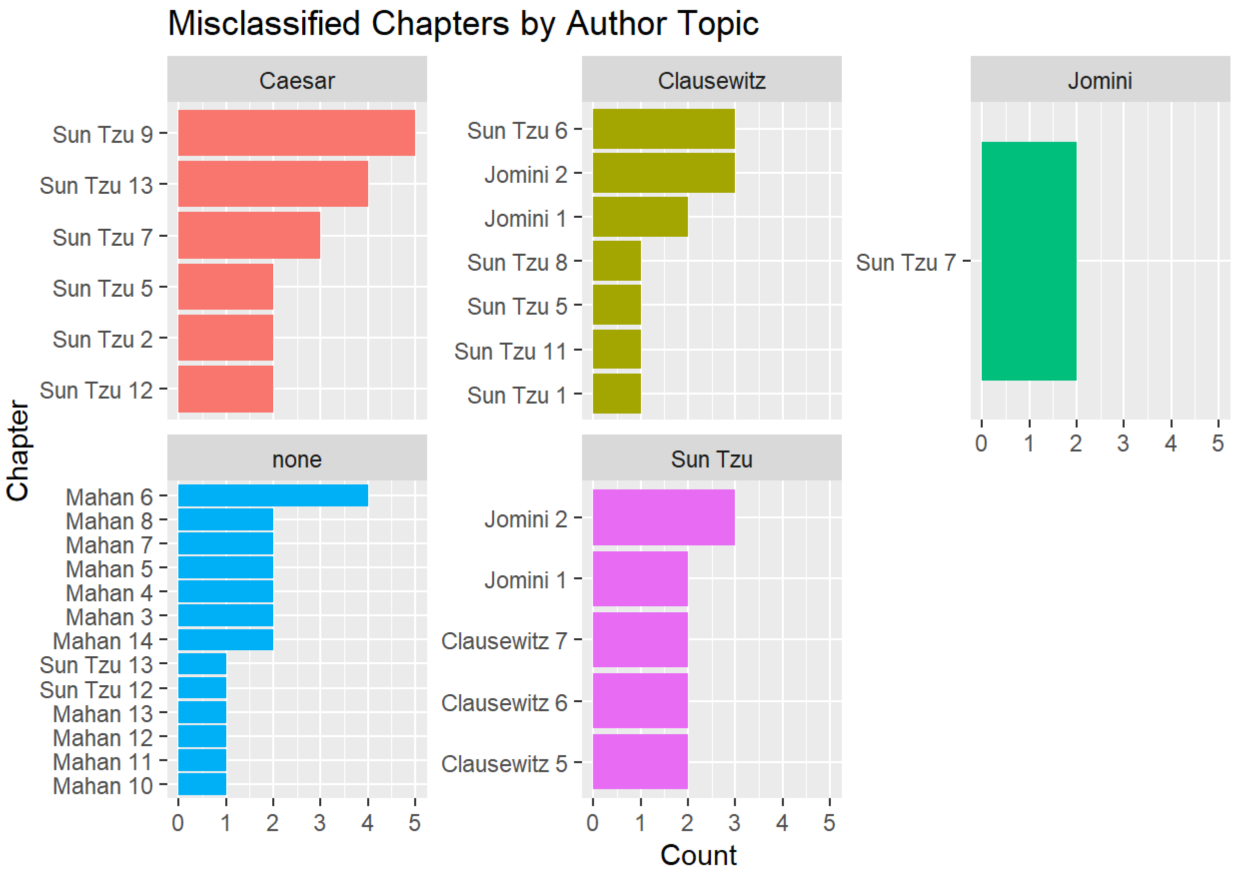

Figure 5: Chapters defined as misclassified as being grouped to a topic that is a consensus topic of other authors. This shows the alignment of individual chapters with authors instead of an overall agreement on genre. As grouping two authors into the same topic is rather frequent, it also demonstrates pieces that have significant variance in topic within themselves.

Plot 5 shows the frequency of such misclassified chapters. Arrian's and Thucydides' chapters are never misclassified, suggesting that their works are purely historical accounts of their respective topics, with few deviations. Sun Tzu, Clausewitz, and Jomini all have some of their chapters misclassified as those of the other two. This suggests that military theories share many similarities that transcend time and space. Mahan's work is broken up into two halves on multiple occasions, and the split appears to be around Chapter 9, which covers the post-American Revolution era. This means that for Mahan, the last decade of the 124 years he covered is vastly different from the preceding century. Mahan's allegiance might also be part of the reason, as this last decade was when the United States began to emerge as a major power in the Atlantic. Sun Tzu's chapters are misclassified under Caesar's topic with high frequency; this provides an answer to our previous question about Caesar: As Caesar's commentaries have high similarity to two of the military theory pieces, Caesar likely went into detail on the rationale behind his decisions and his military philosophy.

5 Author Comparison with N-Grams

At the end of the previous section, we confirmed the presence of military theory in Caesar's work. However, Jomini's work appears to have no relation to Caesar's according to the LDA. This raises the question of whether the focus topic of Jomini differs significantly from that of Sun Tzu and Clausewitz. Since both Clausewitz and Jomini came from the same era and participated in the Napoleonic Wars, and given that Sun Tzu's text is too compact, we will compare Jomini to Clausewitz only.

To obtain more information, we used bigrams to find the most common pairs of tokens for each piece. Compared to the unigram approaches like unigram TF-IDF and word cloud, bigrams provide more insight to the context of the word's usage in the text and the relationship between words. [5] As we found during the EDA, their texts have terms with limited meaning without context, like ”enemy's” and ”thousand.” Extending from unigrams to bigrams could help us better understand the authors' purpose when mentioning these words.

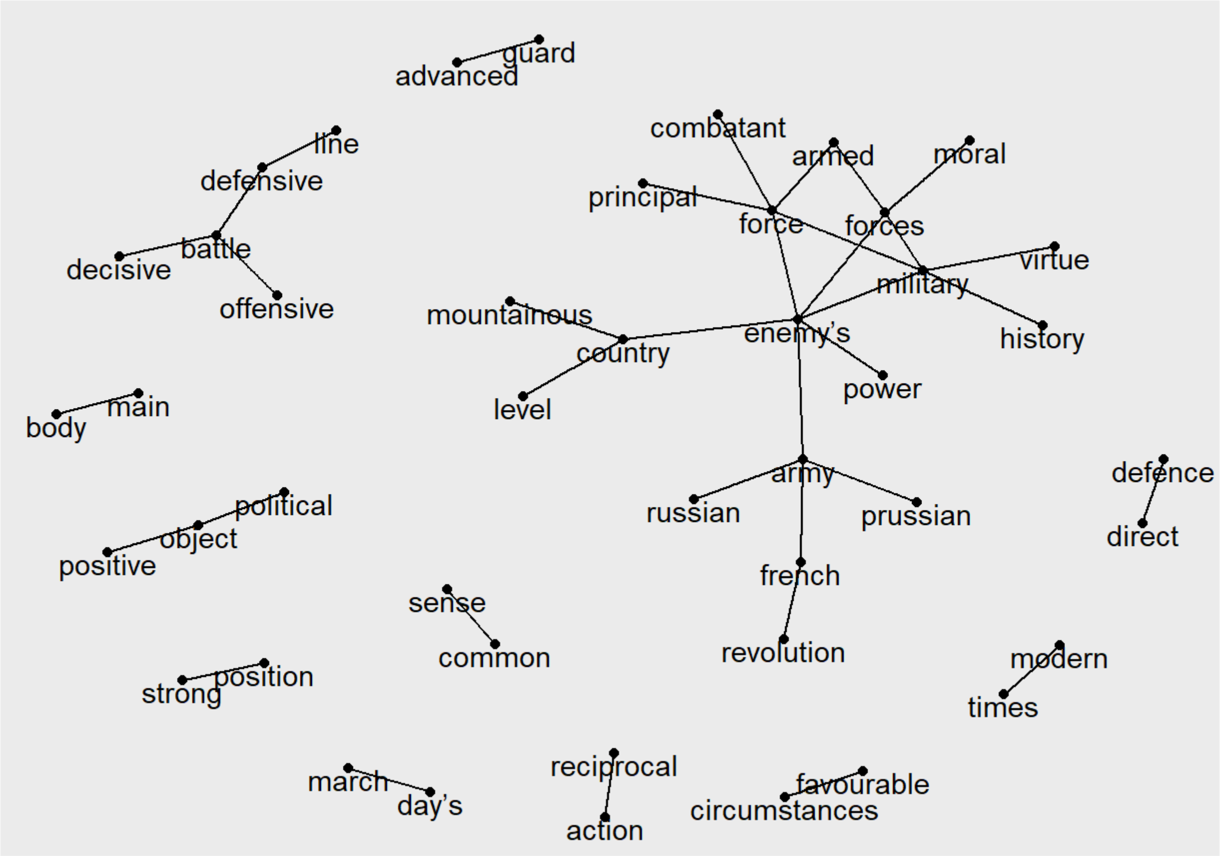

Figure 6: Bigram graph of 30 most frequent bigrams in Clausewitz's work. "force(s)," "enemy's," and "military" are the focal terms in the bigrams.

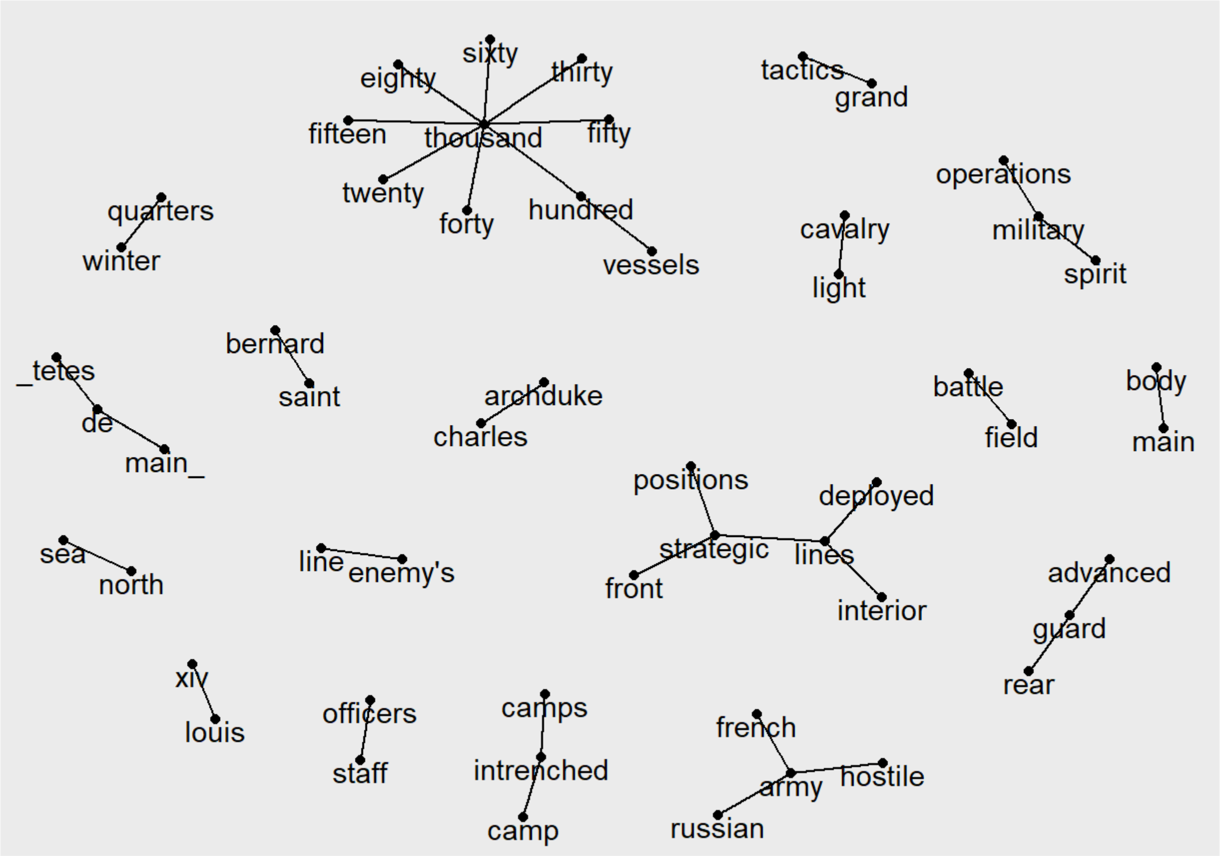

Figure 7: Bigram graph of 30 most frequent bigrams in Jomini's work. "thousand" is the only focal term.

By plotting the bigrams in plots 6 and 7, we could now find the more focal terms in Clausewitz's and Jomini's works. Many of Clausewitz's bigrams are rather abstract and philosophical, like "moral forces," "political object," and "military virtue." Whereas Jomini mentions terms on actual combat and real-life examples more often with numbers ending in thousands, combat terminologies like "interior lines" and "tetes-de-main," and specific individuals like Archduke Charles and Louis XIV. By observing these bigrams, we could conclude on the difference between Clausewitz and Jomini in terms of the topic of interest: Clausewitz is more focused on the high-level strategy and military philosophy, whereas Jomini is more focused on battlefield commands and tactical implementations. This explains Clausewitz's closer connection to his predecessors, Sun Tzu and Caesar, as the fundamentals of war remain consistent over the centuries. On the other hand, it is reasonable that Jomini, with his book geared more towards the actual application of his principles on early nineteenth-century battlefields, has little in common with Caesar, who fought his wars with swords and shields.

6 Topic Modeling with BERTopic

As the final step in our analysis, we applied BERTopic to uncover themes across the corpus of military documents. BERTopic is a modern topic modeling framework that uses transformer-based sentence embeddings and clustering to identify semantically coherent topics. Unlike traditional topic models like LDA, BERTopic preserves contextual relationships between words and enables more interpretable topics, which is particularly useful in our diverse texts.

We initially applied BERTopic using its default settings, which employs the `all-MiniLM-L6-v2` model to generate document-level embeddings. These embeddings were produced from our five-line text chunks discussed in the preprocessing section, which allowed for a consistent input size that stayed within the token limit imposed by BERTopic. This model yielded 46 total topics; however, many were sparsely populated. For analysis, we focused on the top 9 topics, each of which contained over 100 assigned text chunks.

We reviewed and relabeled these top topics using their representative terms, sample passages, and our interpretation. Thematic patterns emerged that corresponded both to specific authors and recurring concepts. For example, Mahan's writings populated the topic we labeled "Naval Warfare and European Powers," while Sun Tzu and Clausewitz were prominent in the topic "War Aims and Political Theory."

Topic #	Suggested Name	Representative Keywords
0	Naval Warfare and European Powers	french, english, france, england, sea, ships, fleets, battle
1	Athenian Diplomacy and Warfare	athenians, athens, lacedaemonians, athenian, treaty
2	Ancient Field Strategy and Campaigns	caesar, cavalry, men, camp, alexander, gaul, enemy, army
3	Defensive Strategy and Force Structure	defensive, divisions, object, corps, attack, line, strength
4	Theory of War and Strategic Thinking	mind, theory, must, truth, criticism, principles, ideas
5	Geography and Terrain	river, mountains, defence, ground, mountain, forest, terrain
6	War Aims and Political Theory	war, political, theory, object, nature, purpose, ends
7	Leadership and Military Virtue	military, general, generals, spirit, staff, influence, character
8	Siege Warfare and Fortifications	fortress, fortresses, siege, may, enemys, castle, towers

Table 1: Top 9 BERTopic topics with names and representative keywords. Analysis of the most prominent BERTopic themes reveals distinct conceptual emphases across authors, from naval and field operations to political and moral dimensions of war.

To understand how authors varied in their engagement with each theme, we calculated the normalized distribution of the top 9 topics across all authors and visualized the results using bar plots. Figure 8 shows the proportion of each topic composed of text from each author. Topic 0, "Naval Warfare and European Powers," is overwhelmingly dominated by Mahan, with minor contributions from Clausewitz, Jomini, and Thucydides. Topic 1, "Athenian Diplomacy and Warfare," is composed almost entirely of documents from

Thucydides, consistent with his detailed account of Athenian military and political affairs. Topic 2, ”Ancient Field Strategy and Campaigns,” draws primarily from the classical authors Caesar, Arrian, and Thucydides, reflecting their focus on battlefield narratives and leadership in antiquity.

Topics 3 through 8 are largely dominated by Clausewitz, who contributes over 50% of the text in each, underscoring his influence across strategic and theoretical themes. Jomini also plays a substantial role in Topics 3, 5, and 7, which align with shared Napoleonic and operational concerns. Although Sun Tzu, Mahan, and other authors appear in several of these topics, their contributions are comparatively modest.

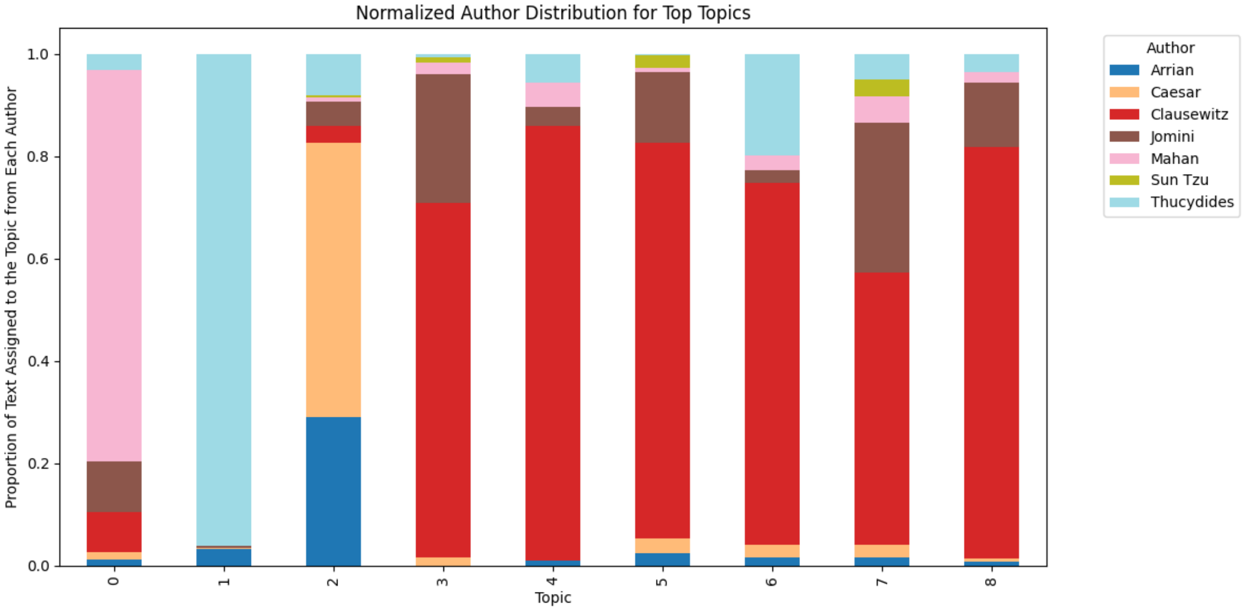

Figure 8: Normalized author distribution for each of the top 9 topics. Each bar shows the proportion of text within a topic attributed to each author, highlighting the dominance of certain authors in particular topics.

Because this distribution reflects the total number of five-line chunks attributed to each author, it may be skewed by differences in document length. To complement this view, we also examined how each author distributes their own writing across the top 9 topics, which we visualize in Figure 9.

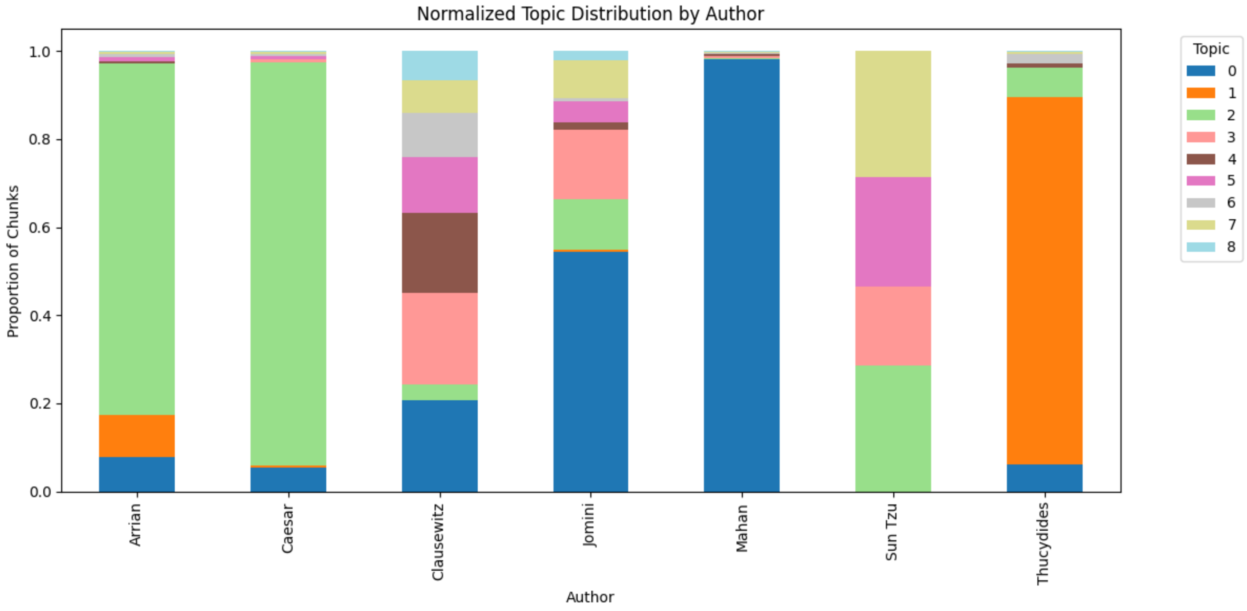

Figure 9: Proportion of each author's text assigned to the top 9 topics. This view highlights the topic distribution within each author's corpus, showing thematic breadth for Clausewitz and Jomini, and narrower focus for authors like Mahan, Thucydides, and Caesar.

This figure highlights each author's topic distribution. Clausewitz and Jomini show engagement with a number of different topics, contributing to topics ranging from military structure and terrain to political aims. Mahan, Thucydides, and Caesar, on the other hand, show a more concentrated focus. As expected, Mahan's unique focus on naval warfare is highlighted, while Thucydides focuses on "Athenian Diplomacy and Warfare." Sun Tzu shows an association with more abstract themes around leadership and strategy.

To further investigate the relationships between topics and authors, we reembedded the text chunks using the `all-mpnet-base-v2` model in place of the default BERTopic embedding model. This alternate embedding was used for visualization and did not affect the topic assignments. We then applied UMAP with cosine distance to project the high-dimensional embeddings into two dimensions. This visualization enabled us to explore how text chunks cluster by topic and author.

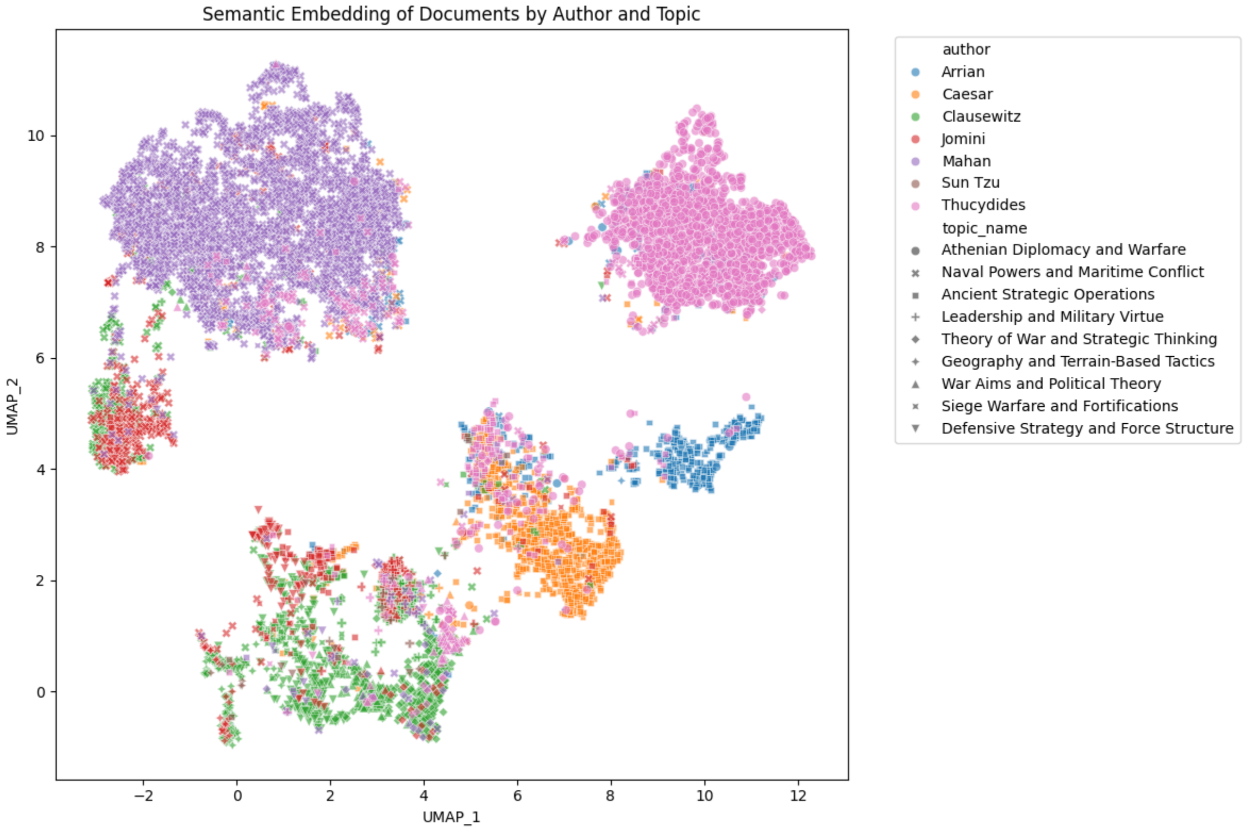

Figure 10: UMAP projection of document embeddings colored by author and shaped by topic. Clusters reveal distinct themes and styles between authors, with clear separation for Mahan and Thucydides and overlapping regions for Clausewitz and Jomini.

Using the author-topic distribution plot (Figure 8), we identified Topic 2: "Ancient Field Strategy and Campaigns" and Topic 7: "Leadership and Military Virtue" as promising for further visualization given their diversity in author contributions. Additional UMAP plots of these topics allow us to examine how semantically similar contributions from different authors are positioned to each other.

Figure 11 shows the UMAP projection of documents within Topic 2: "Ancient Field Strategy and Campaigns." This visualization reveals distinct author-specific clusters for the ancient Greek and Roman authors, Caesar, Arrian, and Thucydides, reflecting their detailed accounts of various military campaigns. In contrast, more modern theorists like Clausewitz and Jomini appear more dispersed within the topic, likely due to overlapping operational vocabulary such as "camp" and "army" rather than narrative cohesion. Overall, this plot highlights semantic convergence among ancient authors while also illustrating clear distinctions in how different writers describe battlefield operations.

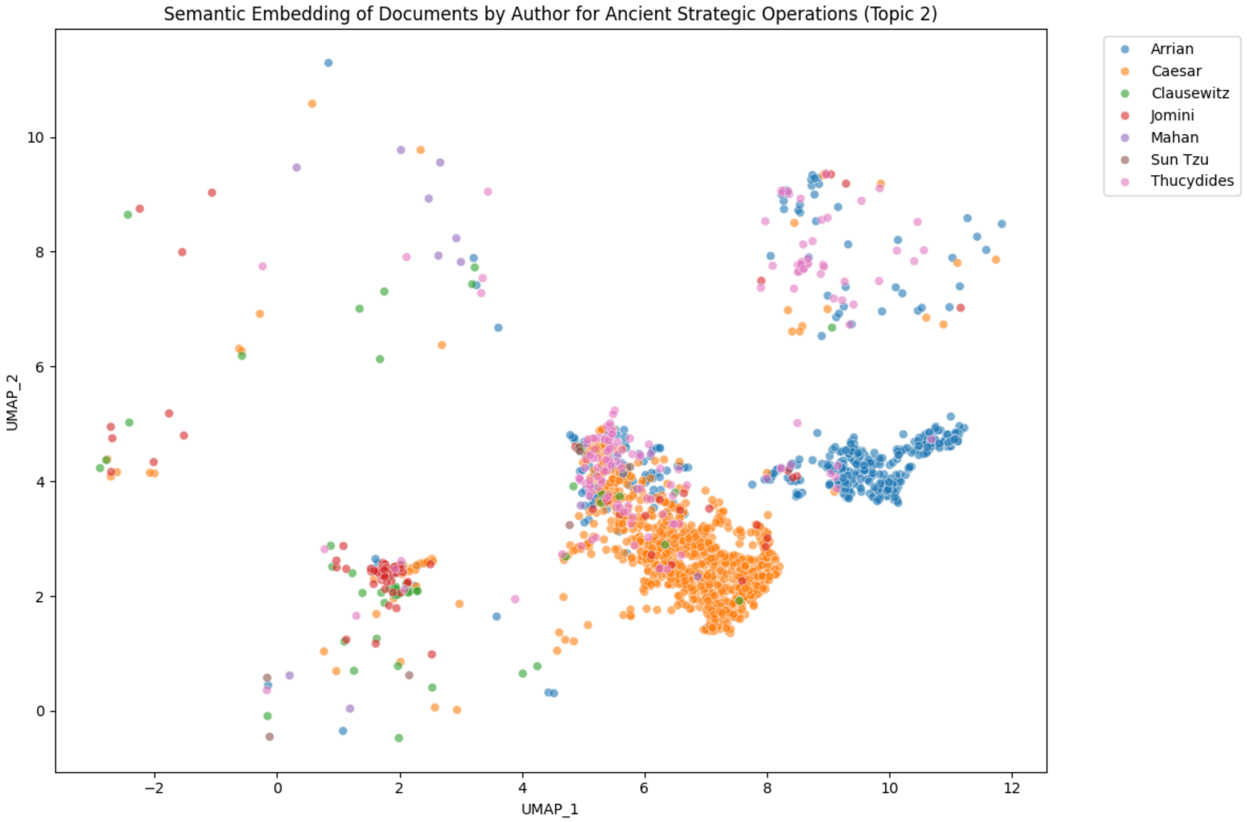

Figure 11: UMAP projection of Topic 2 (”Ancient Field Strategy and Campaigns”) documents, colored by author. Caesar, Arrian, and Thucydides form tight clusters, while Clausewitz and Jomini are more dispersed, reflecting shared operational language.

Now, 12 displays the UMAP projection of documents within Topic 7: ”Leadership and Military Virtue.” Unlike Topic 2, this plot shows a high degree of overlap across some of the authors, with most chunks forming a single dense cluster. This suggests there is convergence in how certain authors, namely Sun Tzu, Clausewitz, Jomini, and Arrian, discuss concepts related leadership, discipline, and command. The absence of clear author-specific groupings indicates that this theme is treated in more consistent way across historical and cultural contexts.

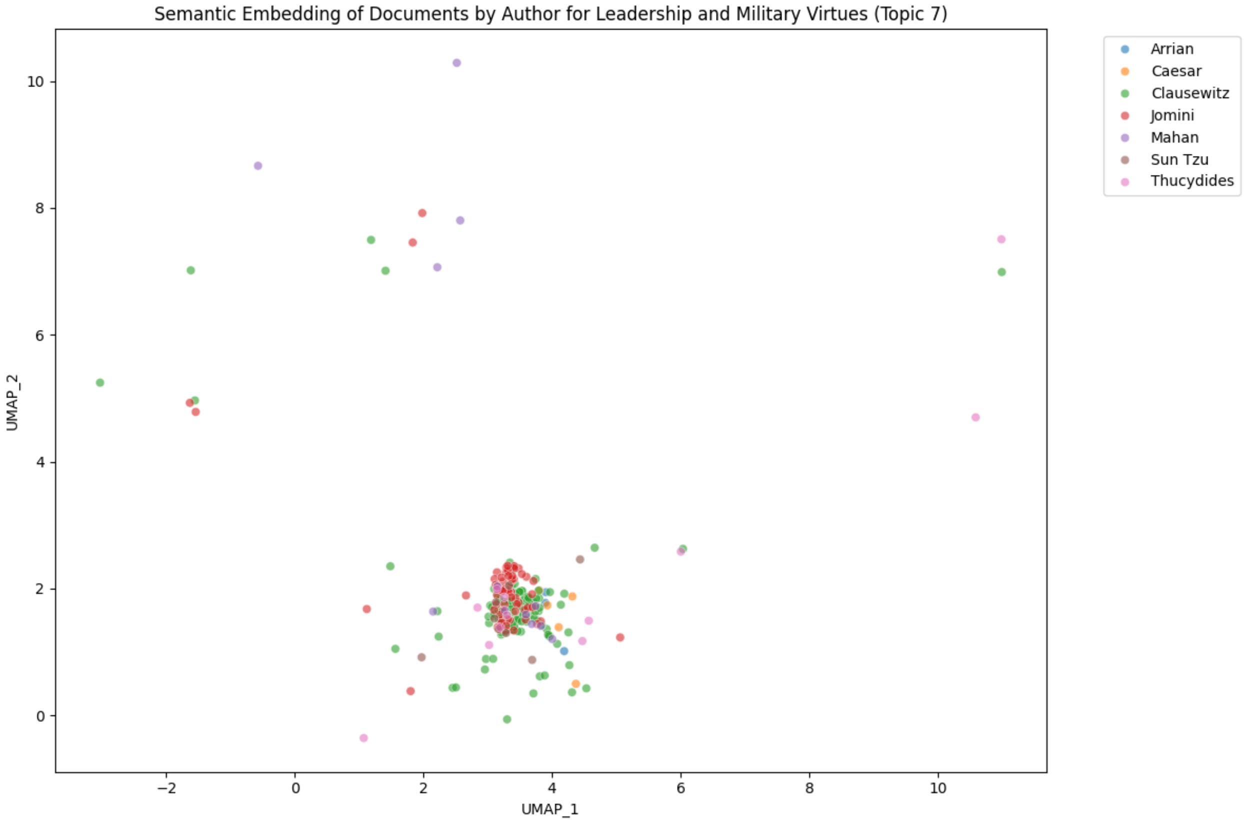

Figure 12: UMAP projection of Topic 7 (”Leadership and Military Virtue”) documents, colored by author. Most texts form a dense, overlapping cluster, suggesting broad semantic agreement across authors on leadership themes.

Together, the BERTopic and UMAP visualizations allow us to analyze thematic, structural, and stylistic differences across the corpus. By focusing on the top topics generated using this model, we were able to identify areas where authors converge across centuries and space, such as in discussions around leadership, and identify unique, defining characteristics, such as specific ancient battlefield accounts. These analyses reinforced earlier observations and offer a semantic map of military though through different times and cultures.

7 Discussion and Next Steps

A takeaway from the LDA and author comparison is that strategic philosophy transcends time, space, and culture. Sun Tzu, Caesar, and Clausewitz lived in different centuries, came from different regions, and had different cultures; however, their works shared a lot of strategic vision in common, such that the LDA treated them as the same topic. The BERTopic modeling also supports this claim, as the topic ”Leadership and Military Virtue” is jointly contributed by Sun Tzu, Jomini, Clausewitz, and Arrian, also coming from vastly different backgrounds. This is the reason why Sun Tzu's *Art of War* and Clausewitz's *On War* remain popular till this day, while Jomini's *Summary of the Art of War* is less well known as its 19th-century battlefield techniques become obsolete nowadays.

A significant limitation to our work is that all selected texts, except Mahan's, which was written in English, were translated from their original languages. All translations took place between the late 19th and early 20th centuries. During translation processes, the authors' semantics might be lost from the original language. The difference in time between composition and translation could also lead to losses of the original rhetoric. Therefore, one of the next steps we will take is to use the original texts instead of the translated versions to retain the maximal semantics and rhetoric. To achieve this, we will likely have to rely more heavily on large language models than traditional techniques for processing and analyzing multiple languages simultaneously.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[3] IBM. What is latent dirichlet allocation?, 2024.

[4] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[5] Julia Silge and David Robinson. *Text Mining with R.* O'Reilly, 1st edition, 2017.

[6] SLTinfo. Type-token ratio. `https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf`, 2014. Accessed: 2025-06-07.