ProjectK: Leveraging Topic Modeling and Sentiment Analysis in r/AskHistorians

Niteesh Reddy Adavelli and Madhuraj Kunta

Introduction. The r/AskHistorians is an online community, serving as a vital hub for historical enthusiasts and experts engaging in rigorous discussions. Understanding the topics and sentiments within this community is crucial for multiple reasons. Firstly, it provides valuable insights into the historical interests and inquiries of a diverse audience, offering a window into public engagement with history. Secondly, the provision of accurate and well-sourced answers within this community contributes to the dissemination of reliable historical information, combating misinformation and promoting historical literacy. Therefore, our project aims to illuminate the prevalent topics and emotions within r/AskHistorians, facilitating a deeper understanding of historical discourse on online platforms.

In terms of NLP, our project spans several key areas. Firstly, data collection and cleaning employ text processing techniques to extract and preprocess posts and comments from the r/AskHistorians subreddit. This encompasses tasks such as tokenization, stop-word removal, and lemmatization to prepare the data for analysis. Secondly, we utilize topic modeling techniques such as Latent Dirichlet Allocation (LDA) and BERTopic to identify prominent historical themes within the discussions. These algorithms uncover hidden structures and patterns within the text data, revealing the major topics of interest within the community. Additionally, sentiment analysis techniques are applied to understand the emotional tone of the conversations, providing insights into how users engage with historical narratives and events. Through addressing the scalability and accuracy of these algorithms, we aim to provide comprehensive insights into the historical discussions within the r/AskHistorians community, enriching our understanding of historical discourse in online forums.

Background. Our work builds upon foundational research in the fields of topic modeling and sentiment analysis, particularly within the context of historical discourse on digital platforms. Latent Dirichlet Allocation (LDA), introduced in pioneering work [1], has been instrumental in uncovering latent thematic structures within textual datasets, providing a framework for understanding the underlying topics within large volumes of text. Subsequent research [2] extended these methodologies to historical texts, paving the way for exploring the application of topic modeling specifically within historical communities such as r/AskHistorians. Additionally, sentiment analysis has emerged as a valuable tool for elucidating user emotions in online discussions. Previous studies [3] have highlighted various sentiment analysis techniques' versatility across diverse domains, laying the groundwork for our research's focus on unraveling the emotional nuances embedded in historical discussions within online communities.

To effectively analyze a large corpus of text data, we draw on various text mining approaches [5], with a particular emphasis on topic modeling. Topic modeling, a form of statistical modeling widely used in machine learning and natural language processing (NLP), identifies hidden topical patterns within a collection of texts. LDA and latent semantic analysis (LSA) are among the most established techniques in this domain. However, it's essential to recognize that while LDA is commonly used in social science research, its efficacy in analyzing social media data has been subject to criticism [7]. By integrating insights from these methodologies and addressing their limitations, our research aims to provide a nuanced understanding of historical discourse within online communities like r/AskHistorians.

Dataset. Hugging Face provides datasets related to Reddit, including the "reddit_submissions", "reddit_comments" datasets for the r/AskHistorians subreddit. It has around 600k submissions out of which we have considered 200k submissions over 3 years (2020-2022).

This dataset encompasses a diverse range of topics and discussions contributed by users within the community. Initially, we performed data preprocessing to refine the dataset for analysis, filtering out rows containing posts tagged as [deleted] or [removed], extracting self-posts, and addressing missing values. Subsequently, to prepare the data for topic modeling, we focused on pre-processing the title and self-text columns, encountering challenges such as short and irrelevant words, inconsistent representations of similar terms, and the need to prioritize important nouns. To overcome these challenges, we implemented strategies such as filtering out short words, performing text normalization, leveraging parts of speech tagging, and generating bigrams and trigrams using the Phrases model to enhance context and meaning capture within the text.

Metric	Value
Total Submissions	200,000
Average Words per Submission	100
Period Covered	2020-2022
Main Language	English

Table 1: Dataset.

Methods.

Latent Semantic Analysis (LSA): We initiated our analysis by employing Latent Semantic Analysis (LSA) to extract topics from the textual data. LSA involves converting the text corpus into a document-term matrix and applying truncated Singular Value Decomposition (SVD) to uncover latent topics. Initially, we converted the textual data into a document-term matrix, representing the frequency of terms in each document. Subsequently, we performed truncated SVD to reduce the dimensionality of the matrix and extract latent topics. Despite its simplicity, LSA struggles with capturing complex relationships and polysemy, as evidenced by its limited interpretability. For instance, it fails to distinguish between different wars, such as WWI and WWII, as seen in Table 1.

Topic	Words
4	'wwii', 'japan', 'germany', 'empire', 'rome'
2	'empire', 'rome', 'kingdom', 'france', 'greece'

Table 2: Keywords generated by LSA.

Latent Dirichlet Allocation (LDA): For our topic modeling strategy, we implemented Latent Dirichlet Allocation (LDA) using Gibbs Sampling, a well-known technique in probabilistic modeling. The GibbsSamplingLDA class encapsulates this approach, allowing us to infer latent topics from textual data. Essential hyperparameters such as the number of topics, Dirichlet priors for document-topic and word-topic distributions, and the number of iterations govern the model's behavior. During the fitting process, the class initializes topic assignments randomly for words in documents and iteratively updates these assignments through Gibbs sampling, a Markov chain Monte Carlo method. By iteratively sampling topic assignments based on observed data, the model gradually converges to a stable distribution of topics, revealing underlying thematic structures in the corpus.

Our LDA model with Gibbs Sampling represents a probabilistic approach to uncovering latent topics within textual data. By iteratively updating topic assignments based on observed word-document co-occurrences, the model effectively captures thematic patterns within the dataset. This technique offers a versatile and robust framework for topic modeling, applicable across various domains and datasets, providing valuable insights into the underlying structure of textual corpora.

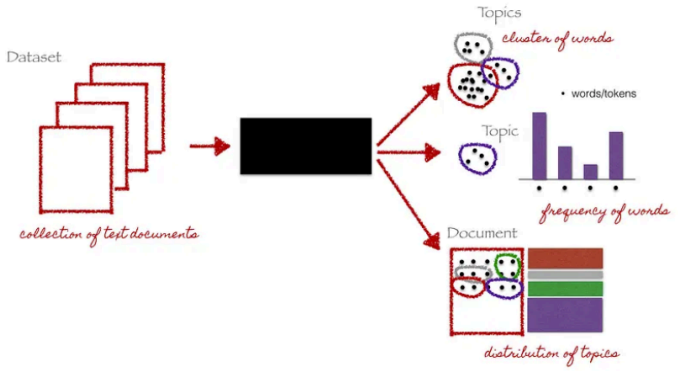

Figure 1: LDA using Gibbs Sampling

BERTopic (Bidirectional Encoder Representations from Transformers):

Transitioning to BERTopic, we utilized Bidirectional Encoder Representations from Transformers (BERT) with the sentence-transformers model "all MiniLM-L6-v2" to generate 384-dimensional sentence embeddings. These embeddings were then compressed into a lower-dimensional space using Uniform Manifold Approximation and Projection (UMAP). In this reduced space, we applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify topic clusters. Finding the optimal cluster sizes required testing various configurations, which involved adjusting both cluster and sample sizes until the method was refined.

Once the clustering was optimized, BERTopic was used to extract and define topics from each cluster utilizing the c-TFIDF method. This approach, while effective in distinguishing topics, shares some limitations with traditional bag-of-words techniques such as LSA and pLSA, notably its disregard for the semantic representations of words which can lead to less meaningful topic distinctions. Despite these challenges, BERTopic provided a robust framework for topic extraction, helping to identify distinct and relevant themes within our dataset.

Model Card - BERTopic

Model Details
- The BERTopic model is an unsupervised topic modeling tool that leverages transformer models and HDBSCAN clustering to analyze and extract topics from large text datasets
- Unsupervised Topic Modeling
- Developed by Maarten Grootendorst in 2021

Intended Use
- Intended for textual data analysis, especially extracting topics from large volumes of text
- Designed for use by data scientists, historians, digital humanities researchers
- Not intended for real-time topic modeling due to computational demands

Factors
- Best performs on large textual datasets with well-defined topics
- Coherence scores, topic diversity, and stability evaluated over multiple runs

Metrics
- Coherence score (c_v and u_mass) used to measure model performance
- No specific decision thresholds
- Adjustments include number of topics, dimensionality reduction parameters, and clustering parameters

Ethical Considerations
- Potential biases due to the user-generated content of the dataset
- No personally identifiable information used, reflects public user posts

Training Data
- Utilizes unsupervised learning, hence no explicit training dataset
- Leverages pre-trained BERT embeddings from the HuggingFace model repository

Evaluation Data
- Data sourced from r/AskHistorians subreddit via Hugging Face datasets
- Aimed at analyzing discussions and questions related to historical topics
- Data preprocessing included text normalization and removal of deleted or removed comments

Caveats and Recommendations
- Effectiveness of topic extraction varies with data quality and parameter settings
- For optimal results, tuning model parameters based on dataset characteristics is recommended
- Further studies should address the impact of data biases on topic modeling

Results. In addition to evaluating human interpretability and topic separation,the coherence scores, including c_v and u_mass, were utilized to assess the interpretability of topics in topic modeling. LDA and NMF exhibited relatively standard results, while BERTopic, leveraging embedding approaches, offered novel insights with higher coherence scores. Higher c_v scores indicate better interpretability, while closer u_mass scores to 0 signify higher coherence..

Table:

Model	c_v_coherence	u_mass_coherence
LDA	**0.49**	**-5.48**
BERTopic BAAI/bge-small	**0.61**	**-1.3**
BERTopic miniLLM	**0.72**	**-0.2**

Table 3: c_v_coherence and u_mass_coherence scores between LDA and BERTopic.

Topic Distribution and Prevalence: The LDA model highlighted specific historical events and themes with varying degrees of engagement. For instance, World War II topics exhibited high engagement levels, whereas discussions on ancient civilizations were less frequent but highly informative. The BERTopic model, which refined the embeddings further, provided a more nuanced understanding by identifying subtler subtopics within these broader categories.

Topic Label	Keywords
Roman Empire	roman, empire, rome, caesar, byzantine
Black Death	plague, black, death,bubonic,pneumonic
9/11 attack	11, alqeda, osama,200, terrorism, attack
Israel Palestine	israel, palestine, jewish, israeli, arab
World War 2	hitler, nazi, holocaust, germany, camps

Table 4: Topic labeling.

After we got the final topics, we combined all reddit posts associated with a topic and visualized the most important words for a few topics using a word cloud

Figure 2: Most important words for ‘World War 2’

Community Analysis Based on Topics: As we can see in (Figure 3), World War 2 is the most popular topic. Aside from wars, there is an interest in broader subjects such as mythology, European colonialism indicating that discussions in the subreddit are not limited to military events. There could be a bias towards more recent or American-centric conflicts (like World War II and the American Civil War) over older or non-American topics (like the Spanish Civil War) which could indicate a more American userbase.

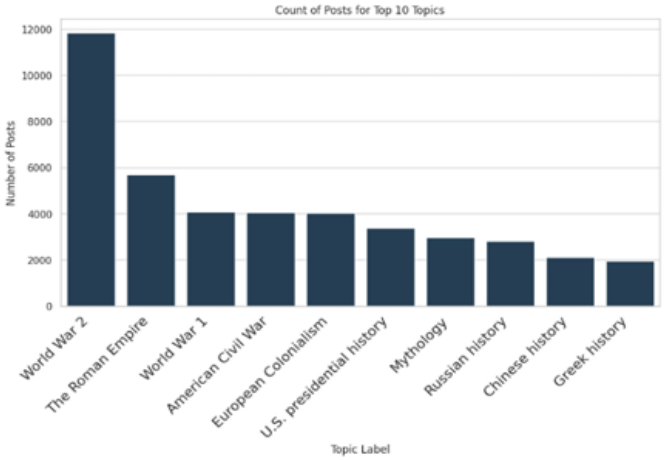

Figure 3: Frequency of posts for top 10 topics

Sentiment Analysis: Sentiment analysis revealed varying emotional tones associated with different historical topics. For example, discussions around wars typically showed a negative sentiment, whereas topics related to cultural achievements or historical figures often had a positive sentiment. This analysis helps in understanding how different events are perceived emotionally by the community.

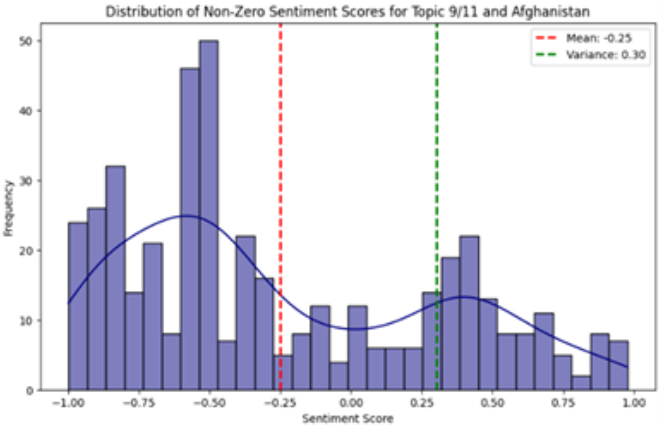

Figure 4:Sentiment scores for 9/11 attacks and Afghanistan

Our evaluation of the r/AskHistorians subreddit using LDA and BERTopic models revealed significant insights, with BERTopic achieving a higher coherence score of 0.72, indicating more refined topic extraction compared to LDA's 0.49. Sentiment analysis showed varying emotional responses, highlighting negative sentiments in war discussions and positive ones in cultural topics. Challenges included BERTopic's tendency to over-segment topics, occasionally leading to a diluted thematic focus. Overall, the advanced analytic techniques employed provided a comprehensive understanding of historical discourse dynamics, blending quantitative and qualitative insights effectively.

Conclusion.

In conclusion, our analysis of historical discussions within the r/AskHistorians subreddit uncovered diverse and dynamic trends in topic modeling, temporal patterns, and sentiment analysis. By leveraging advanced techniques such as LDA, BERTopic, and sentiment analysis, we gained valuable insights into the community's engagement with historical narratives. From the identification of significant historical events to the exploration of emotional responses and temporal dynamics, our study highlights the rich tapestry of historical discourse within online communities and underscores the importance of leveraging NLP methods to uncover and understand historical trends in the digital age.

References.

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[2] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W.(2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. In BMC bioinformatics (Vol. 16, pp.1-10). BioMed Central.

[3] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2), 1-135.

[4] Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism management, 59, 467-483.

[5] Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. Tourism Review, 77(4),1234-1246.

[6] Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management, 57(2), 102034.

[7] Mountford, J. B. (2018). Topic modeling the red pill. Social Sciences, 7(3), 42.

[8] Ankur Tomar (2018). Topic modeling using LDA and Gibbs Sampling Explained.