Improving Usenet Archives: How to Process Metadataless Materials

Kurt Lemai-Nguyen
Princeton University
kl4102@princeton.edu

Brian Mmari
Princeton University
bm6442@princeton.edu

May 2, 2025

Abstract

We introduce a visualization and search tool for archived Usenet materials. Leveraging vectorization we generate a vector database to search for messages based on semantic similarity, and also visualize these vectors with UMAP dimensionality reduction. These tools give researchers new ways of intuitively browsing usenet archives, making historical research of pre- and early internet communication more robust.

Table of Contents

Introduction

USENET was a way of using the Unix-to Unix Copy (UUCP) protocols and programs to send messages to other people in forums known as "Usegroups."[5] Duke University Students, Tom Truscott and Steve Bellovin came up with it as a way to distribute "electronic newsletters" and in 1979 they connected a system to change messages between Duke and University of North Carolina using dial up connections.[5] Newsgroups were categorized by a wide range of topics from science fiction and cooking, to religion and politics.

According to the historian Janet Abbate, USENET took off because of its ability to connect people with shared interests all across the US, one of the earliest forms of real digital communication before the internet.[5] Other historians have taken an interest in USENET for this exact reason of reading into a snapshot of early digital culture, analyzing groups such as women's forums and astrophysicist forums to understand how different social groups used online communication to further their goals.[12, 24] USENET messages then become a vital source of insight into events of the 80s and 90s through the perspective of different digital communities.

Yet despite the important role the USENET plays in history, the archives for these materials are limited in functionality. In 2020 Jozef Jarosciak, an enterprise architect, took the UTZOO NetNews archive of USENET posts and turned it into a website that accessed the materials through a PostgreSQL backend with python.[9] His website, usenetarchives.com, enabled access to reading posts by usegroup, but the search functionality by posts does not work, keeping users in a loading loop.[4] This meant that in order to research using the archives, you had to already know the name of a usegroup, and had to cross reference usegroup posts by hand rather than searching all of them simultaneously. For someone without personal experience with usegroups, this meant that researching USENET could be exceedingly tedious. Our goal with this project was to take inspiration from Jozef's work and to create a new generation of search and visualization of USENET materials .

Related Work

2.1 Ways of Accessing Usegroups

Other than Jozef's USENET archive, online searchable USENET materials are sparse and are scattered through different sites.[3] Google Groups used to have support for USENET posts, which meant that they archived large amounts of materials, but in recent years due to linkrot and decreased support for Google Groups, many USENET posts are hard to access.[1] In terms of just the files, the Internet Archive hosts the Usenet Archive, which contains roughly 7.7 terrabytes worth of materials.[2] These files come in different formats with CSV being one of the most common, but considering the immense size of the materials, independently processing the archives can be challenging. Bartosz Taudul created the Usenet Archive Toolkit which allows users to more easily work with usenet files, including features like data processing, data filtering, data search, and data access.[20] Although this resource makes processing far easier, it still does not provide the full interactive experience for regular historical researchers without technical knowledge to browse the USENET Archives.

2.2 Previous Computational Work on Usegroups

Most of the prominent computational work on USENET comes from the early 2000s meaning that they have not taken advantage of recent developments in natural language processing. Warren Sack developed a conversation map of usegroup interactions to give a social analytical perspective to these interactions and track conversations.[18] Mari Wang developed

NewsView, which applied search engine techniques like FAST Query to access newsgroups through crawlers and indexing.[21] J.C. Paolillo used a factor analytic approach to visualize content by generating word frequency lists to identify topics of discussion.[16] We take inspiration from these systems, but are most interested in Paolillo's work who attempts to show semantic content. Rather than word frequency though we believe that vectorization has greater potential to give a better representation of content.

2.3 Forms of Search

One of the key issues of the USENET materials is the lack of metadata. Keyword search is standard practice in databases as it enables searching without knowing the exact terms in a controlled vocabulary.[11] The issue with forms of search is that there still needs to be ways of constraining search results. Faceted search is a popular form of refining search results by changing facets - ways of categorizing and splitting list in a taxonomy - to easily parse large volumes of information.[23] As mentioned before, because of the lack of organization for USENET materials, facets would have limited functionality, only able to constrain dates and usegroups. Recent work has tried to deal with these issues through developing vector based methods, allowing for more flexible natural language search through LLMs.[15] Our work will aim to emulate some of these strategies to better visualize and search USENET.

Methods

Our project aims to use techniques in the Nguyen et al. paper to develop a smart search interface that can visualize newsgroups, search and retrieve semantically relevant posts, and generate concise summaries of such posts. Our approach involves three key components:

3.1 Indexing and Embedding Archived Usenet Posts

The first step towards achieving our goal is to first use dense vector representations for indexing and embedding archived Usenet posts. We do this using a Hugging Face sentence transformer model called all-MiniLM-L6-v2.

This first method involves first selecting the Hugging Face dataset, which contains a subset of the entire Usenet Archive. We selected 20 newsgroups comprising around 18,000 newsgroup posts on 20 topics from the Hugging Face dataset. All of these data were pre-cleaned, meaning that they contained no headers, signature blocks, or quotations from news articles.

For vectorization, we chose to use MiniLM, a sentence transformer developed by Microsoft Research[22].This sentence transformer, also available through Hugging Face, maps sentences & paragraphs to a 384-dimensional dense vector space, which is useful for semantic search. Various alternatives were explored before coming to a conclusion about this model. Some of these alternatives are outlined below with their potential drawbacks:

Model	Drawbacks
BERT	Computationally expensive; not ideal for very large datasets[7].
Word2Vec	Outdated; prone to gender and societal biases[14].
GloVe	Generates non-contextual embeddings; sensitive to corpus quality[17].
Doc2Vec	Difficult to fine-tune effectively; poor performance on short or informal texts like forum posts[10].
DistilBERT	Faster than BERT but slower than MiniLM; lower accuracy in semantic similarity tasks[19].
TF-IDF / CountVectorizer	No semantic understanding; high-dimensional and sparse representations[6].

Table 3.1: Drawbacks of Common Alternative Models

3.2 Retrieving Contextually Relevant Messages Using Semantic Search

After vectorization, we chose to store and index the vectors in an index-based library called FAISS (Facebook AI Similarity Search). FAISS is a library designed for fast similarity search and clustering of dense vectors. It includes algorithms capable of handling vector datasets of any size, including those that may exceed the available RAM [8]. This library allowed us to quickly search for embeddings of documents that are similar to each other. It is important to note that FAISS also enables comparison between vectors using either L2 (Euclidean) distance or cosine similarity. For our purpose, we settled on the following steps:

- Normalize the vectors to have an L2 norm = 1
- After normalization, the inner product between two vectors becomes equivalent to cosine similarity

The cosine similarity is defined as follows:

$$\text{Cosine similarity} = \cos(\theta) = \frac{v_1 \cdot v_2}{\|v_1\| \, \|v_2\|}$$

and now $\|v\| = 1$, so:

$$\text{Cosine similarity} = v_1 \cdot v_2$$

For similarity search we look for the top 5 most similar documents to the query vector and those that have the highest dot product become outputted.

3.3 Dimensionality Reduction Using UMAP

UMAP is a dimensionality reduction technique that converts high-dimensional data into 2D or 3D space [13]. Before UMAP, t-SNE was the state-of-the-art for dimensionality reduction for visualization [13]. However, UMAP offers several advantages over its predecessor. UMAP is more suited to significantly larger dataset sizes than t-SNE and offers no computational restrictions on embedding dimension [13]. Because of the said advantages, we therefore chose UMAP over t-SNE for visualization purposes.

After the vectorization step, we deploy our model UMAP so as to reduce our high-dimensional dataset into 2D. We set our model with the following parameters:

- Number of neighbors = 15
- Minimum distance = 0.1
- Cosine similarity as metric

We then visualize our dataset as a latent space with 2 axes and the 20 newsgroups as labels in different colors. The above parameters can be changed based on the user's preferences and project necessity. Attached below is an example image of the latent space visualization.

Figure 3.1: Newsgroups Latent Space.

3.4 Leveraging an LLM for Retrieval Augmented Generation (RAG)

We use the Princeton Sandbox API in order to access the AzureOpenAI GPT-4 models, specifically the `GPT_4o` model. The AzureOpenAI client helps connect the API calls to the `GPT_4o` model for embedding. The user is then tasked with first inputting the search query, which gets embedded into the Hugging Face dataset using our model. After that, our model retrieves the relevant messages and sources from the Hugging Face dataset. The retrieved messages are then augmented with the search query, and generated using GPT-4 via the Sandbox API.

Experiments and Results

4.1 Discourse Tracking

By looking at overlaps in the UMAP graph, we can see places where usegroups overlap in topics, indicating that similar conversations may be occurring in different places. An obvious and prominent example of this would be the overlaps between soc.religion.christian, alt.athiesm, and talk.religion.misc. We can see in figure 4.1 that these three usegroups are grouped together, so by clicking through different posts we can get a broad overview of discussions on a topic that happen in parallel in separate usegroups. We can also then use these discussions to identify key terms or topics being discussed and search them in the search box below. This layered interaction helps researchers more quickly identify and cross reference USENET discourse.

Figure 4.1: Overlap about religion.

In a very similar way we can observe not just discussion but events that occur through

USENET. For example, looking at figure 4.2, we can see car sales occurring not only in misc.forsale but also in rec.auto. These included both requests to purchase specific vehicles and listings offering cars for sale. Our visualization system grouped these related posts together, enabling users to more intuitively identify such exchanges and explore all relevant threads where individuals might attempt to buy or sell vehicles.

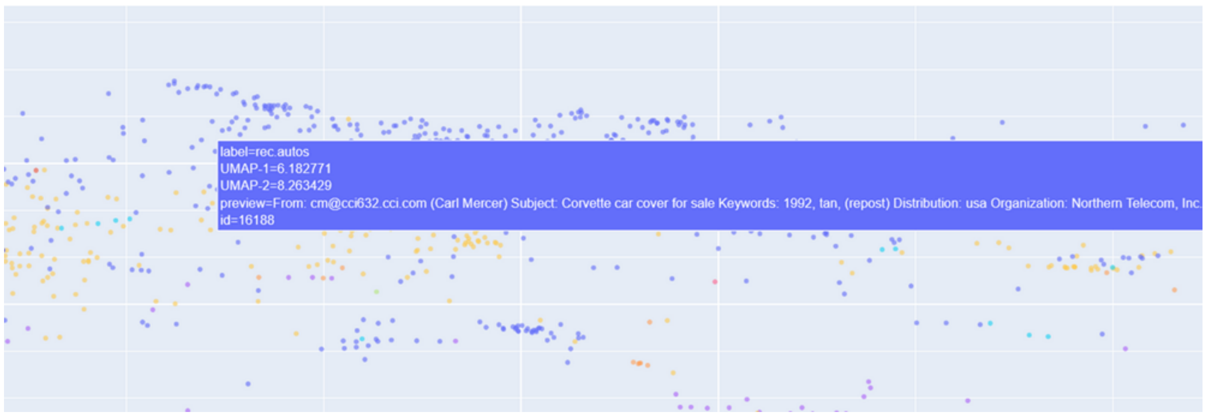

Figure 4.2: Sales of cars.

4.2 Natural Language Search

By using RAG with an LLM our system enables natural language search that has the additional advantage of summarizing posts as well. In figure 4.3, we are able to do an advanced search, looking for posts about buying guns. Whereas usual systems with key word search may be able to pick up on these posts through some flexible and robust searching, with vectorization, we are able to ask it in common parlance "what are posts about the legality of buying guns?" This allows for more complex questions that are more attuned to researchers interests. We acknowledge that these searches may be noncomprehensive, but it is a strong first step for many researchers attempting to parse the immense volume of content in USENET.

One peculiar thing we discovered about our RAG model was non-determinism. We explored several posts and noticed that for the same input submission, the model sometimes produced

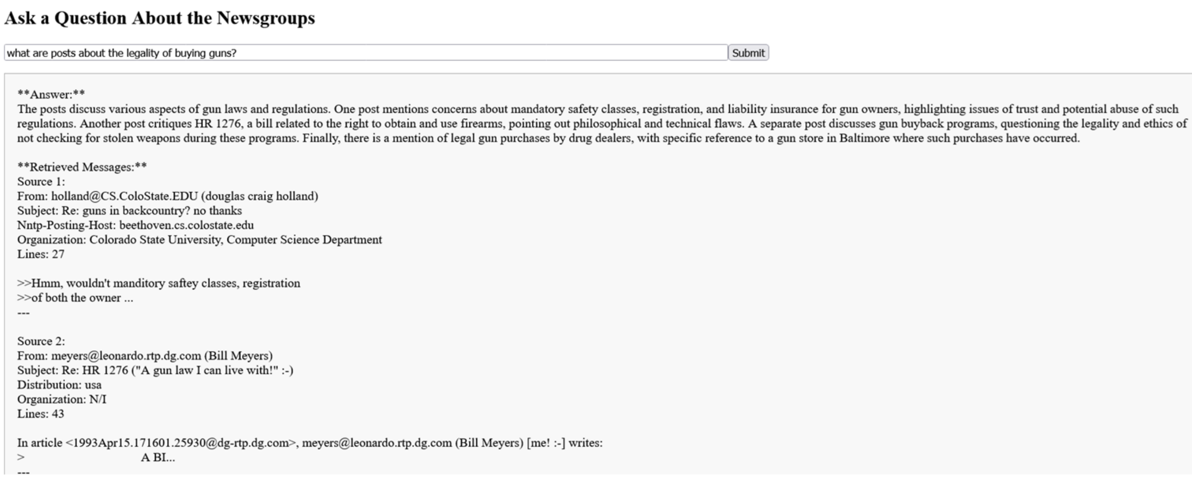

Figure 4.3: Searching about gun laws.

different sets of similar documents. An example of this behavior is outlined below(note: output has been truncated for readability):

First Trial

Input: *tell me more about motorcycles?*

Output: The Usenet posts provide some insights and opinions about motorcycles. Here are a few points gathered from the posts: ...

Source 1:

- From: `bclarke@galaxy.gov.bc.ca`
- Subject: Re: First Bike??
- Organization: BC Systems Corporation
- Lines: 8

In article `<0forqFa00iUzMATnMz@andrew.cmu.edu>`, James Leo Belliveau `<jbc9+@andrew.cmu.edu>` writes:

> I am a serious motorcycle enthusiast without a motorcycle, and to
> put it bluntly, it suck...

Second Trial

Input: *tell me more about motorcycles?*

Output: Insufficient information retrieved.

Source 1:

- From: `azw@aber.ac.uk` (Andy Woodward)
- Subject: Re: Cultural Enquiries
- Organization: University College of Wales, Aberystwyth
- Lines: 35

The above examples demonstrate that our model is probabilistic, generating responses by sampling from a probability distribution over words. While the probabilistic nature of our model enables greater diversity and discovery in retrieval, it also poses challenges for reproducibility and consistency. Future researchers using our interface should be aware that repeated queries may yield slightly different results, depending on model state and stochastic elements in the retrieval or generation pipeline. In contexts where repeatability is crucial, additional techniques may be necessary.

We also ensured that our RAG model was grounded solely on the retrieved Usenet posts. This was important to guarantee that the model did not rely on its internal training data, such as ChatGPT's pretraining data, especially when relevant information was unavailable. Additionally, because the model is grounded on explicit evidence, it is less likely to hallucinate, thereby improving its robustness. A snippet illustrating this behavior is shown below:

Input: *Who won the FIFA World Cup in 2022?*

Output: Insufficient information retrieved.

Source 1: From: etxonss@ufsa.ericsson.se (Staffan Axelsson) Subject: WC: Scores and standings, April 18 Organization: Ericsson Telecom, Stockholm, Sweden Lines: 71 Nntp-Posting-Host: uipc104.ericsson.se

1993 World Championships in Germany: ==

Group A standings (Munich) ... —

Limitations and Future Work

Despite the successes of our project, there's various limitations to our methods. For instance the choice of our vectorization technique introduced biases in our results. MiniLM was chosen over BERT and Doc2Vec primarily for its balance between inference speed and scalability to large datasets. BERT, while more powerful in capturing contextual nuance, was significantly slower and less feasible for our large-scale retrieval setting. Doc2Vec was also considered, but it has known instability during training and underperforms on shorter or informal text(common in USENET posts). Each of these models was carefully evaluated during our design process, and although MiniLM aligned best with our goals, alternative choices could yield different, and potentially richer, retrieval results.

One limitation that we initially hard but later solved by fine-tuning was the amount of time it took to vectorize the messages. Because we're vectorizing 20 newsgroups each comprising about 18000 newsgroup posts, the amount of time it took to vectorize was long. One approach we used was halving the number of batches and utilizing Google Colab's GPU. This significantly decreased our running time.

Another challenge as discussed previously was non-determinism. As mentioned before, the model sometimes outputs different submissions even with the same input command. This is likely because large language models (e.g GPT) are inherently probabilistic. This behavior can also vary depending on the temperature settings and the internal retrieval step (RAG), which may involve sampling from a probability distribution over possible next words.

In the future, expanding to a larger dataset could yield more diverse and interesting results. It is also worthwhile to investigate alternative mapping techniques not explored in this project. For instance, t-SNE could be applied to compare its clustering behavior and visual clarity against the UMAP-based methods used here. Furthermore, we recommend adopting a newer version of the MiniLM sentence transformer, `all-MiniLM-L12-v2`, which offers faster performance and the potential for improved results.

Conclusion

With the wealth of historical resources hidden in the USENET archives, our project aimed to improve access leveraging recent innovations in natural language process.We developed a system that vectorizes posts, and uses UMAP to visualize the relationships between messages. We then allowed users to ask natural language questions and search for posts through RAG with GPT-4o and FAISS, a vector database. We find that our program allows for cross referencing content across different usegroups to see discussions unbound by the categories the service is divided into. Searching becomes more intuitive with natural language in our chat, and the combination of these tools provides users with advanced browsing capabilities to explore USENET to find new and fascinating discussions. We hope that more historians will have less technical barriers to research and spawn interesting findings.

Bibliography

[1] Google Groups Broke Usenet Links. Here's How to View Them. | Pith & Pip.

[2] Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music & Wayback Machine.

[3] Usenet Newsgroup Archives - Google Groups, Deja News, Search, Archives.

[4] UsenetArchives.com.

[5] Janet Abbate. *Inventing the Internet*. Inside technology. MIT Press, Cambridge, Mass., 3rd printing edition, 2000.

[6] Mamata Das, Selvakumar K., and P. J. A. Alphonse. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025.

[9] Jozef Jarosciak. Converting UTZOO-Wiseman Usenet Tapes to Website with PostgreSQL backend using Python 3.8 – Jozef Jarosciak Blog, October 2020.

[10] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.

[11] Fang Liu, Clement Yu, Weiyi Meng, and Abdur Chowdhury. Effective keyword search in relational databases. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 563–574, New York, NY, USA, June 2006. Association for Computing Machinery.

[12] Monica Marra. *On some Pioneering Usenet Newsgroups in Astrophysics*. February 2025.

[13] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[15] Ha Dung Nguyen, Thi-Hoang Anh Nguyen, and Thanh Binh Nguyen. A Proposed Large Language Model-Based Smart Search for Archive System, January 2025. arXiv:2501.07024 [cs].

[16] J.C. Paolillo. Visualizing Usenet: a factor-analytic approach. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pages 10 pp.–, January 2000.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[18] Warren Sack. Conversation map: a content-based Usenet newsgroup browser. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 233–240, New Orleans Louisiana USA, January 2000. ACM.

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[20] Bartosz Taudul. wolfpld/usenetarchive, April 2025. original-date: 2019-06-03T16:34:50Z.

[21] Mari Wang. NewsView: A Recommender System for Usenet based on FAST Data Search. Master's thesis, 2004. Accepted: 2013-03-12T08:14:06Z.

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[23] Bifan Wei, Jun Liu, Qinghua Zheng, Wei Zhang, Xiaoyu Fu, and Boqin Feng. A SURVEY OF FACETED SEARCH. *Journal of Web Engineering*, pages 041–064, November 2013.

[24] Hannah Zeavin. "This is Womenspace": USENET and the Fight for a Digital Backroom, 1983–86. *Technology and Culture*, 63(3):634–664, 2022. Publisher: Johns Hopkins University Press.

Acknowledgments and Honor Code

Our project was supported by Professor Bangalore in office hours who helped walk us through the idea and alternatives. ChatGPT was consulted for learning different python libraries' functions and parameters.

This work represents our own work in accordance with university guidelines.

Kurt Lemai-Nguyen

Brian Mmari