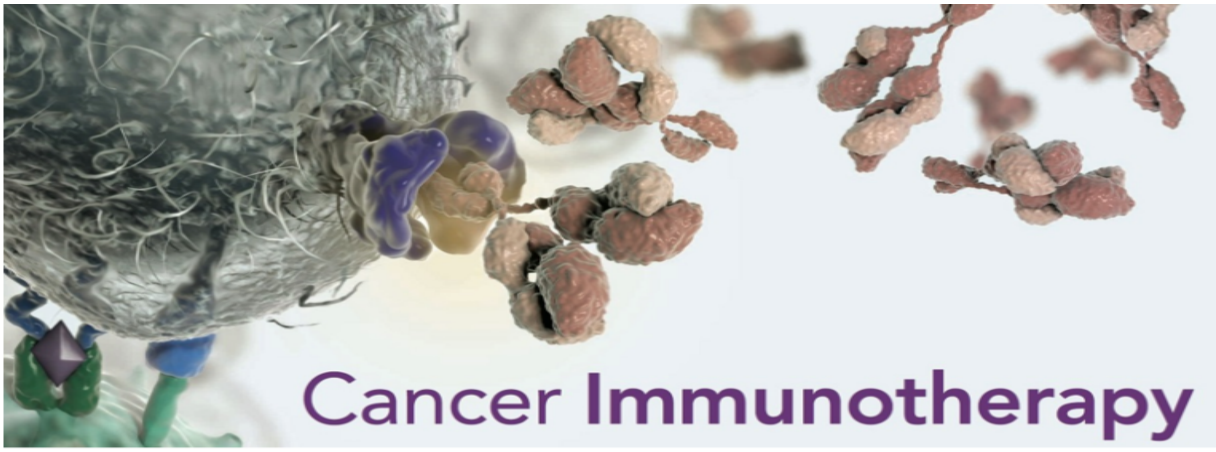
A BERTopic Platform for Perceiving Patients' Views Of Cancer Immunotherapy on Social Media and Online Forums



WPI

A BERTopic Platform for Perceiving Patients' Views Of Cancer Immunotherapy on Social Media and Online Forums

An Major Qualifying Project Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In Partial Fulfillment of the Requirements for the Degree of Bachelor of Science

Submitted By:

Sidney Goldinger (CS)

Florkenthia Jolibois (DS)

Sierra Mangini (CS/DS)

Mago Sheehy (CS/DS)

Tiffany Wee Sit (CS)

Date:

April 27, 2023

Submitted to:

Professor Chun-Kit Ngan

Abstract

Cancer immunotherapy is a promising treatment that harnesses the body's immune system to combat cancer, but not all patients respond to the treatment. To understand patients' views on cancer immunotherapy, we propose a BERTopic Modeling and Sentiment Analysis platform. We were given 4.9 million user-generated posts from various online cancer and health forums and performed text pre-processing on each of them. With 3.6 million cleaned messages, our team chose BERTopic as our methodology to extract the topics people were discussing most frequently. We computed the sentiment of every post in each topic to determine their public perception, and we also compared BERTopic with other methods, showing that it outperformed them. Lastly, we developed a web-based dashboard to support medical professionals and show the topics and their sentiment analysis results through data visualizations. Our proposed platform has the potential to provide valuable insights to medical professionals, giving them a tool to improve patient outcomes and experiences.

Executive Summary

Cancer immunotherapy is a treatment that uses the body's immune system to fight cancer cells. Real-world examples and statistics demonstrate that cancer immunotherapy has been successful in treating different types of cancer patients. To understand patients' views on cancer immunotherapy, we take advantage of topic modeling, and sentiment analysis, popular tools in the NLP domain. Topic modeling helps to extract the underlying topics in large datasets, while sentiment analysis helps to determine the sentiment of each document, and by extension, each topic.

To address this healthcare problem, our team has looked into various topic modeling and sentiment analysis methods. For topic modeling we have explored methods such as LDA, LSA, NMF, PCA, Random Project, and more. For sentiment analysis we explored different libraries such as Vader, TextBlob, and SentiWordNet. With exploring different methods and techniques of both topic modeling and sentiment analysis, our team was able to choose methods that would work with our dataset as well as compare our chosen model.

In this report, we propose to develop a BERTopic Modeling and Sentiment Analysis platform to extract patients' views on cancer immunotherapy and provide insights to medical professionals. The platform includes pre-processing 4.9 million text posts, generating topics using BERTopic, and computing the distribution of sentiment in each topic. Our experiments demonstrate that BERTopic outperforms other widely used topic modeling methods. We also provide a dynamic web-based dashboard to support medical professionals in analyzing the data.

Chapter 2: Literature Review - This chapter includes a detailed discussion of different topic modeling and sentiment analysis methods, including their advantages and disadvantages. We also discuss how sentiment analysis scores are computed.

Chapter 3: BERTopic Methodology and Details - This chapter focuses on the original BERTopic and how it is applied to search for topics, compute sentiment analysis, and demonstrate them on our web platform. It includes a block diagram to describe the entire process in detail.

Chapter 4: Experimental Results and Discussion - This chapter includes a comparison of BERTopic with other methods and the topics we were able to extract from the 3.6 million messages, along with their sentiment analysis results and the topics with the sentiment analysis results in the half-year basis changed over time. We also discuss the web platform, which we designed to present our results such that medical

professionals can extract meaningful insight about the immunotherapy patient experience.

Chapter 5: Conclusion and Recommendation - The report concludes with recommendations for future research and highlights the potential benefits of the BERTopic Modeling and Sentiment Analysis platform in providing insights to medical professionals.

Acknowledgments

We extend our heartfelt gratitude to Professor Chun-Kit (Ben) Ngan for initiating this project and providing us with invaluable guidance throughout its development.

We would also like to express our sincere appreciation to Taylor Spangler and Sam Jenkins, data scientists at Microsoft Corporation, for generously sharing their time and expertise with us. Their feedback and insights have been instrumental in shaping our project and overcoming challenges.

Furthermore, we would like to thank Professor Yin-Ting Cheung from the Faculty of Medicine at the Chinese University of Hong Kong for providing us with the initial dataset and for offering valuable feedback on our generated topics.

Finally, we are immensely grateful to our families and friends for their unwavering support and encouragement, which have kept us motivated and sane throughout this project.

Table of Contents

List of Figures

List of Tables

List of Formulas

Introduction

Cancer is one of the greatest public health concerns globally, as the leading cause of death worldwide as of 2023 (World Health Organization). From the variety of currently available cancer treatments, immunotherapy has emerged as one of the most promising approaches that leverages and trains the immune system to fight the cancer cells (Sharma, P., 2022). This approach has shown tremendous potential in improving patient outcomes, and several immunotherapeutic agents have been approved for the treatment of various types of cancer (Sharma, P., 2022; Cancer Research Institute, 2023).

Understanding patients' perspectives and experiences is crucial to help medical professionals provide the best outcomes for their patients (TechTarget, 2019). To this end, in this project we use topic modeling and sentiment analysis techniques to learn how patients feel about the various aspects of the cancer immunotherapy treatment experience.

In this report, we propose a BERTopic Modeling and Sentiment Analysis platform to analyze a corpus of 4.9 million text posts collected from diverse cancer social media and health forums. Our platform aims to provide medical professionals with valuable insights into patient perspectives on cancer immunotherapy, and our report aims to provide a jumping-off point for potential future projects to continue this important work.

Background and Literature Review

1.1 Immunotherapy and research purpose

Immunotherapy is one of the most promising new cancer treatments (Sharma, P., 2022). Improving every year, it is of significant interest to both patients and medical professionals (Wyant, T., 2023; Sharma, P., 2022). The purpose of this research is to give medical professionals insights into the opinions and views of patients on immunotherapy, including their general attitude toward it (positive/negative) as well as their views about the topics they felt were important. For this purpose, our team has inquired multiple Natural Language Processing techniques to gather topics within those patient opinions, sentiment analysis libraries that are able to distinguish the opinions from positive, neutral, and negative documents, as well as techniques to visualize the information for medical professionals. This information will allow these members of the industry to better understand patients so that doctors can come to the best and most informed conclusions about their cancer treatments.

1.1.1 What is Immunotherapy?

Immunotherapy is one of the most common cancer treatments currently, along with surgery and chemotherapy (Gersten, T., & Zieve, D., 2021). Many consider immunotherapy to be the future of cancer treatments because of its new approach to cancer care (Sharma, P., 2022).

Both surgery and chemotherapy have very significant side effects. Having to go through invasive surgery creates side effects from the surgery itself, and also does not guarantee that all of the cancer is removed (Mayo Foundation, 2022). Surgery is also ineffective for later-stage cancer, where the tumor has spread past its initial location.

Chemotherapy also has very harmful side effects. In chemotherapy, drugs are used to attack fast-growing cells in the body, and since cancer cells are fast-growing, they are targeted. However, this treatment does not only target cancer cells, normal cells are attacked as well, causing significant side effects such as nausea, vomiting, hair loss, and fatigue (Mayo Foundation, 2022).

Unlike these two treatments, immunotherapy targets cancer cells specifically by enabling the patient’s immune system to directly identify and target them (Wyant, T., 2023; Sharma, P., 2022). Frequently given using an IV, immunotherapy typically causes only normal flu symptoms (Gersten, T., & Zieve, D., 2021). Additionally, because it helps the patient themselves attack cancer, some medical professionals are now looking at it as a potential path to curing cancer (Sharma, P., 2022). In a recent study in 2022, four people treated for rectal cancer with immunotherapy saw their tumors completely disappear, which as of the writing of the article, had not reappeared after two years (Sunday, J., 2022).

1.1.2 Data and Purpose of Research

In this study, our team performed Natural Language Processing on 3.6 million text posts about immunotherapy cancer treatment on a variety of online forums. These posts were collected by a separate team led by Professor Cheung Yin Ting using a web scraper. We discuss this in depth in sections 3.2 and 3.3.

The purpose of this research is to inform medical professionals about patient views on immunotherapy. This will allow them to understand how to best inform patients about immunotherapy, clear up misunderstandings, and give them better ways to cope from the treatment. Additionally, research shows that increased patient involvement in treatment leads to improved treatment outcomes, and understanding patients' views would be a step towards involving them more in their cancer treatment decisions (Say, R. E., 2003).

Ultimately, we aim to gain insights as to the topics that users discuss related to immunotherapy treatments, and how they feel about these topics (negative, positive, or neutral). We aim to have our NLP pipeline be usable for other future data to gather new insights on patients' views and to present these insights on an interactive website so that we can share them with medical professionals.

2.1 Natural Language Processing

Natural language processing, or NLP, is a domain that focuses on giving computers the ability to decipher input such as text and spoken words (IBM, 2023). Common in everyday life, it can be found in speech-to-text software, digital assistants, and other similar technologies. It combines rule-based models of language with machine and deep learning, often using them together to decipher meaning from human-generated text and speech (IBM, 2023). NLP is broken down into a number of tasks, as shown in Figure 2.1.

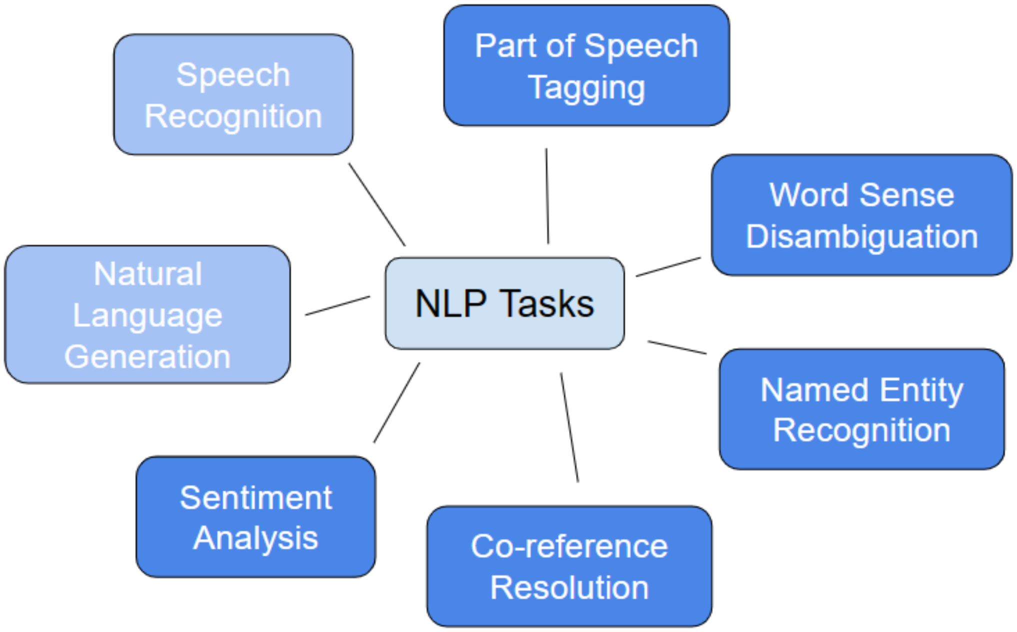


Figure 2.1: NLP Tasks

Out of the tasks listed in Figure 2.1, speech recognition and natural language generation are not included in the scope of this project (What is Natural Language Processing, 2023). Those tasks, which refer to converting voice to text data and creating language out of data, do not apply as our data is already in text form and our results will be reported in topics and sentiment data rather than verbally reported insights.

The tasks we will be focusing on are part of speech tagging, word sense disambiguation, coreference resolution, and sentiment analysis. Part of speech tagging refers to identifying the part of speech a word is based on the context of a sentence (What is Natural Language Processing, 2023). Word sense disambiguation refers to selecting the meaning of a word based on context, such as “ran for office” compared to “ran a marathon” (What is Natural Language Processing, 2023). Coreference resolution refers to identifying when multiple words refer to the same entity, such as determining if a “he” refers to “Steve” mentioned earlier in a paragraph (What is Natural Language Processing, 2023). All three of these tasks are relevant to the first of our two goals in this research, topic generation. The last task is sentiment analysis, which refers to extracting emotion, attitudes, and other subjective qualities from the text (What is Natural Language Processing, 2023).

2.1.1 Text Cleaning

Before our natural language processing of the data could begin, we needed to clean the text. An example figure of this process can be seen in Figure 2.2.

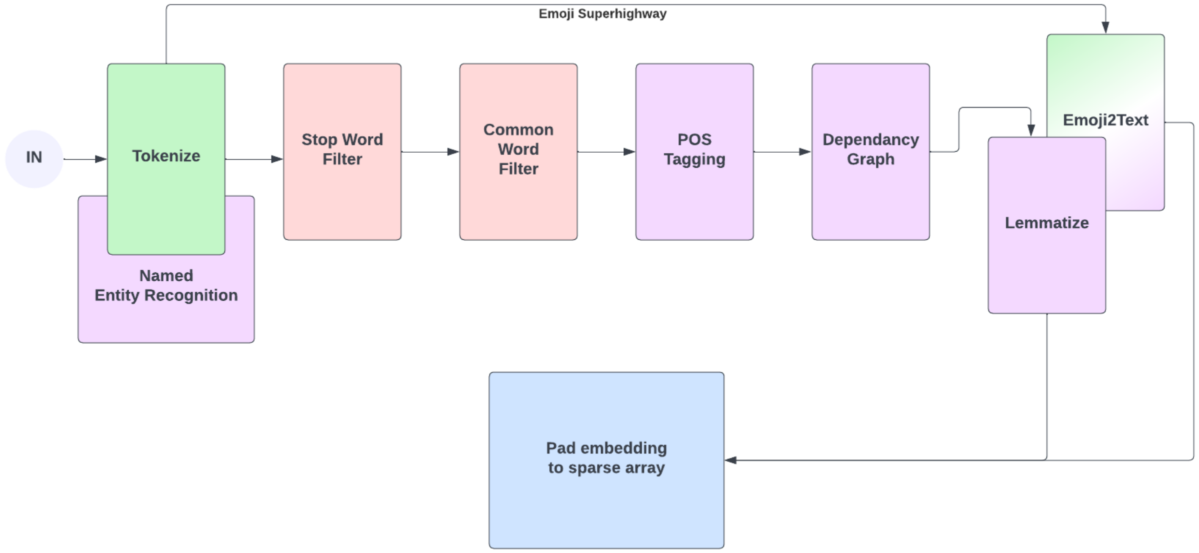


Figure 2.2: Text Cleaning in Natural Language Processing

In Figure 2.2, one possible pipeline of text cleaning we considered upon inspection of the highly unstructured documents, although it is important to note that there are many different libraries and steps that researchers can utilize to clean text, depending on their particular needs.

Tokenization is frequently one of the first steps in text cleaning, breaking a stream of textual data into smaller sections such as words, terms, sentences, or symbols (Menzli, A., 2023). This turns an unstructured document into a set of tokens for machine learning to process. Emoji2Text is similar, removing emojis from the text and later converting them into textual equivalents for the algorithms.

Stop Word and Common Word Filters remove words that are relatively common but also provide little value in determining meaning in text, such as “and,” “are,” “its,” and “of” (Kim, S., 2022; Manning, C. D., et. al., 2019). These words can lead to overfitting of the model, as well as the model creating topics based on irrelevant words (Manning, C. D., et. al., 2019).

Part of Speech tagging, or POS tagging, is a process that involves labeling words in text with their parts of speech, such as noun or preposition (Harsha, A., 2022). This is used to understand the grammatical structure of a sentence, as well as to disambiguate words with more than one meeting, such as “run” down the street versus a bank “run.” POS tagging also helps improve nlp tasks such as named entity recognition (Harsha, A., 2019).

Dependency parsing is the process of examining relationships between phrases in a text to determine its grammatical structure (Sharma, P., 2021). Dependency graphs are often used in this process, an example of which is shown in Figure 2.3.

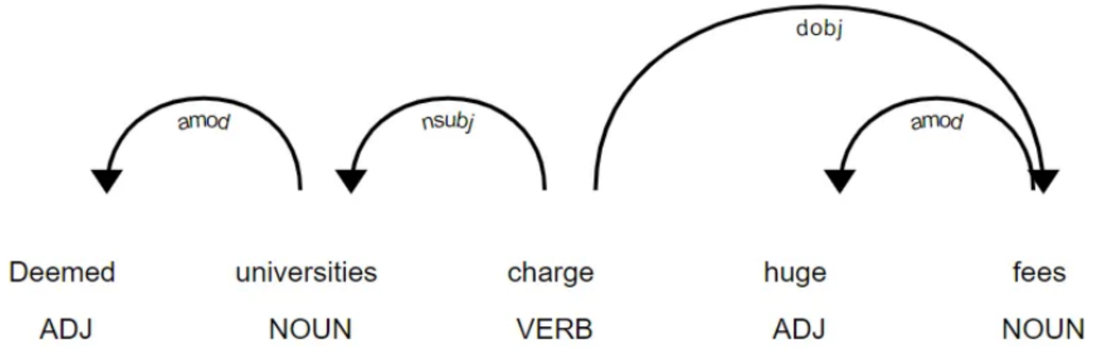


Figure 2.3: Example of a Dependency Graph (Jaiswal, S., 2021)

The dependencies in the text also influence the meaning of the words. For example, in Figure 2.3, “fees” is modified by “huge,” changing the meaning of the noun (Jaiswal, S., 2021).

Lemmatization is the process of reducing words to their base root (Engati, 2021). For example, “leaves” would be lemmatized to “leaf,” and “sleeping” would be lemmatized to “sleep.” Stemming is similar to this, although rather than reducing the word to its base word, it removes common prefixes and suffixes, such as “pre-” or “-ing” (Engati, 2021). Both are used to simplify text input to make it easier for programs to understand words’ meanings. The difference between the two is that while lemmatization is far more precise than stemming, it is also much slower (Engati, 2021).

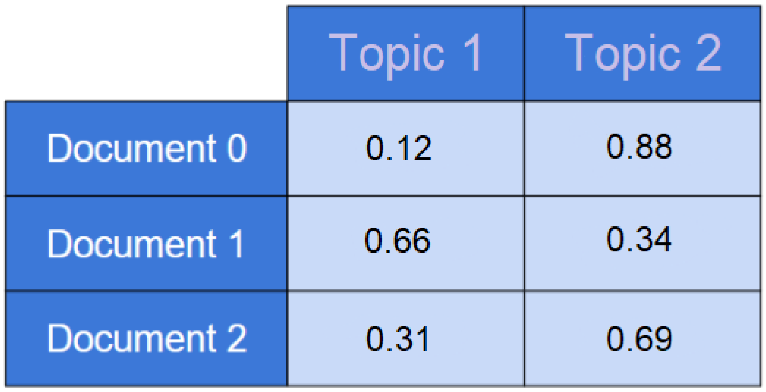
Finally, although they are not listed in Figure 2.2, there are many other steps that are frequently used in text processing. These include but are not limited to, removing URLs and HTML tags, removing links, making all characters lowercase, and removing accents (Kim, S., 2022). Depending on the nature of the text being processed, some of these steps might be necessary for the computer to properly process the input, and some may not. For example, HTML tags would not have to be removed from transcriptions of personal letters, but might have to be removed from text scraped from websites.

2.1.2 Topic Modeling

Our first goal in this research is topic modeling. Topic modeling refers to an unsupervised machine-learning problem with the goal of extracting distinct topics from a set of documents. For the construction of our pipeline, we explored 11 baseline methods for use in topic modeling. In sections 2.1.2.1 to 2.1.2.11, we go into more detail on these baselines.

2.1.2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, or LDA, is a topic modeling method used to assign documents to topics (Kulshrestha, R., 2019). It works by determining which words in the documents belong to which topic, and uses that to assign each document a distribution of possible topics. A visual representation of LDA can be seen in Figure 2.4.



	Topic 1	Topic 2
Document 0	0.12	0.88
Document 1	0.66	0.34
Document 2	0.31	0.69

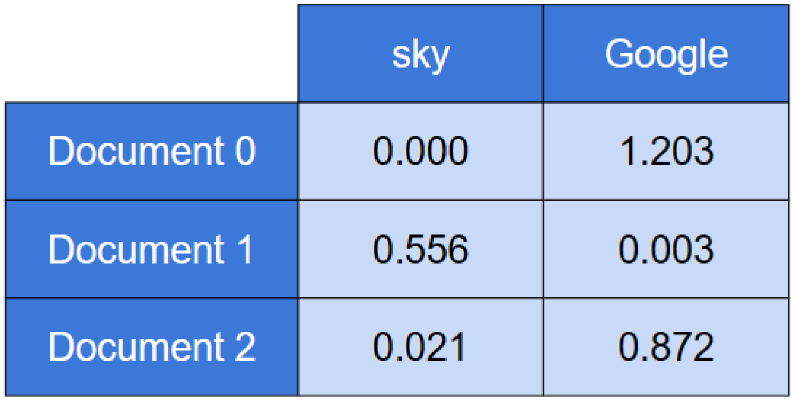
Figure 2.4: Visual Representation of LDA

In Figure 2.4, LDA is run on three documents. Two of them have words that are related to nature, and the other two have words that are related to computers. However, although it's easy for a human to determine that based on looking at the contents of the document, it's much more difficult for a computer to decide what the documents are related to, or assign them to topics. LDA does this by randomly assigning words to *n* topics, where *n* is chosen by the user as a hyperparameter (Qadir, A., 2021). LDA then iteratively changes the probability of which each word is assigned to each topic, so that it is able to most closely re-generate the documents given only the probabilities (Qadir, A., 2021).

As LDA is one of the older approaches to topic modeling, it has historically been popular even though it has many downsides (Kulshrestha, R., 2019). It is still somewhat popular because it is easy to train and very common, although one of its largest disadvantages is that it only considers documents as bags of words, ignoring semantic information (Nathan, J., 2020). It also creates a fixed number of topics defined by the user, and does not take into account evolution of topics over time (Nathan, J., 2020).

2.1.2.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis, or LSA, is an unsupervised topic modeling and dimensionality reduction method (Ioana, 2020). LSA reformulates text data in terms of latent, or hidden, features. One can think of this by creating a document-term matrix, an example of which is shown in Figure 2.5.



	sky	Google
Document 0	0.000	1.203
Document 1	0.556	0.003
Document 2	0.021	0.872

Figure 2.5: LSA document-term matrix

In a document-term matrix, each row is a document, and each column is a word that appears at least once in the set of documents. The intersection between the row and the column reflects how important the word is to that document; for example, if the value is 0, the word does not appear in the document.

LSA also compares documents and latent features, creating a document-latent feature matrix similar to Figure 2.5, as well as a term-latent feature matrix. Through exploring the relationships between the words, latent features (or topics), and documents, LSA can assign each document to a topic. Additionally, by creating less topics than there are documents, LSA also functions as a dimensionality-reduction technique. For example, if there are 100 documents, but LSA decides they are either about the topics technology, politics, and nature, LSA has greatly reduced the dimensionality of the data from 100 to 3.

LSA is a very popular dimensionality reduction algorithm mostly because it is very fast and easy to implement (Topic modeling using LSA, 2021). However, it involves singular value decomposition, which is very computationally intensive and hard to update, and requires a large set of documents to get accurate results (Topic modeling using LSA, 2021). As the dataset for this project contains millions of documents, this is not much of an issue, but LSA's embeddings are also not easily interpretable (meaning topics are not easily discernible from the results), creating challenges for gathering insights and passing along information to medical researchers (Topic modeling using LSA, 2021).

2.1.2.3 Non-negative Matrix Factorization (NMF)

Non-negative matrix factorization, or NMF, is a topic modeling and dimensionality reduction method. It takes in a term-document matrix, as well as the number of k topics that it will create (Goyal, C., 2021). It outputs two non-negative matrices, one of the input terms' relationships with the identified k topics, and one of the documents' relationships with the identified k topics (Goyal, C., 2021). In simple terms, NMF is linear algebra applied to natural language processing. Figure 2.6 shows a visual representation of this process.

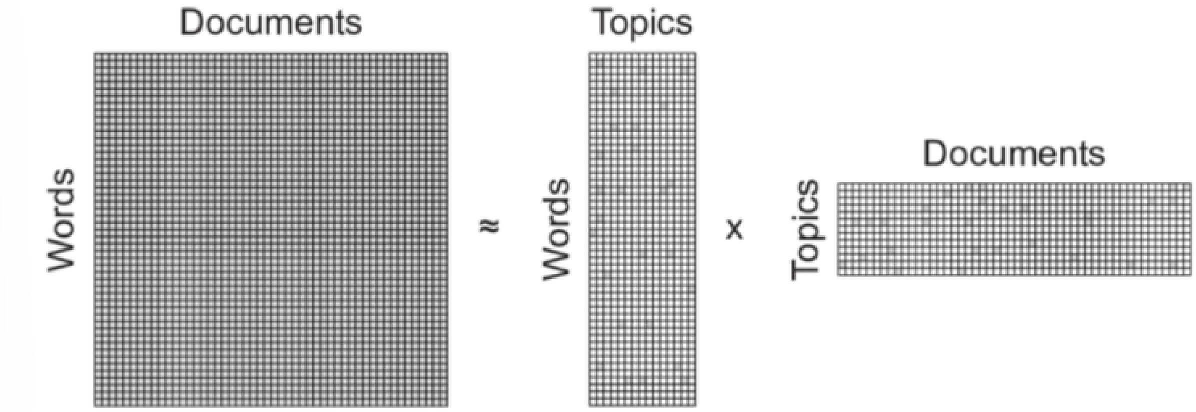


Figure 2.6: Relationship between input and results in NMF (Goyal, C., 2021)

The biggest advantages of NMF are storage, as it works in matrices, and interpretability (Albright, R., et. al., 2005). The matrix generated by NMF contains only positive values. By keeping its results positive, the values in each column can easily be compared to determine the relative probabilities the model associates with each topic (Albright, R., et. al. 2005).

2.1.2.4 Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a dimensionality reduction algorithm commonly used for feature extraction (Jain, Y., 2022). Its main purpose is to decrease model complexity while still maintaining meaning. In brief, it works by standardizing the data and then reducing the dimensionality using the covariance matrix and eigenvalues (Jain, Y., 2022).

One of the biggest advantages of principal component analysis is that it counteracts many of the issues of a high-dimensional data set, such as overfitting, and makes it far easier to create correlations of features (Statistics Globe, 2022). This also improves visualization abilities of the data, as the lower dimensionality makes high dimensional data much easier to visualize. However, there are a few downsides to this algorithm, including losing some information through the dimensionality reduction, especially if the wrong number of principal components is chosen. Additionally, the most significant principal components that the algorithm extracts can be difficult to extract from all of the components (Statistics Globe, 2022). Finally, processing very high-dimensional data is very computationally expensive for PCA compared to other dimensionality reduction techniques like Random Projection, which will be talked about in detail in section 2.1.3.5 (Bingham, E., et. al., 2001).

2.1.2.5 Random Projection

Random projection is a dimensionality reduction tool, and an alternative to PCA that helps with some of PCA's shortcomings (Bingham, E., et. al., 2001). Specifically, as mentioned in section 2.1.3.4, PCA is very computationally expensive for high-dimension data. Random projection is significantly less expensive, as shown in their computational complexity formulas in Formula 2.1 and Formula 2.2, where
d x N is the data matrix that is projected onto k dimensions (Bingham, E., et. al., 2001).

Formula 2.1: Computational Complexity of PCA

$$O\ (d^{\,2}N) + O\ (d^{\,3})$$

Formula 2.2: Computational Complexity of Random Projection

$$O\ (dkN)$$

Although random projection is much more efficient than PCA when processing higher-dimension data, it has a significant cost in accuracy (Bingham, E., et. al., 2001). However, the dramatic decrease in computing time can compensate in some cases with extremely large datasets.

2.1.2.6 K-Means Clustering

K-means is one of the simplest and most popular clustering algorithms currently available (Medium, 2018). An unsupervised machine learning algorithm, K-means identifies *k* centroids, and then assigns every data point to its nearest cluster (javatpoint, 2023). An example of this clustering algorithm used on an example set of points can be seen in Figure 2.7.

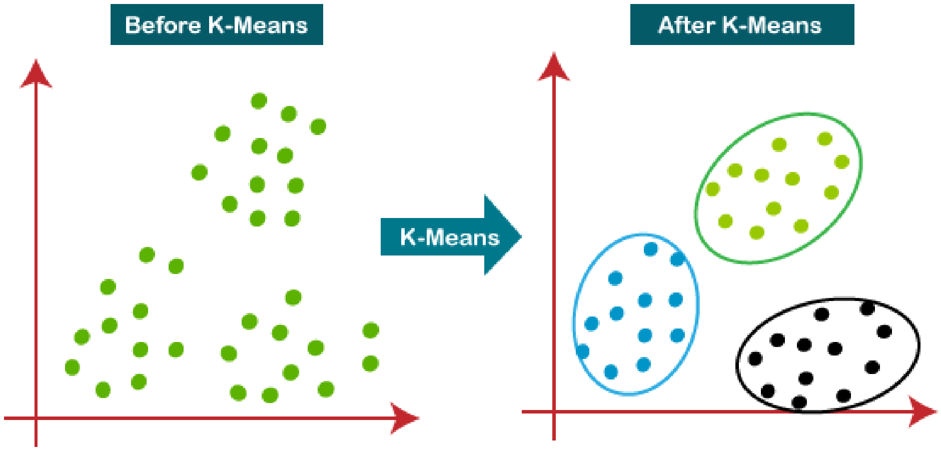


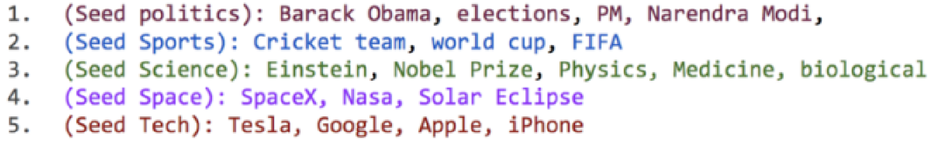
Figure 2.7: K-Means Clustering (javatpoint, 2023)

On the left side of Figure 2.7, a set of unclustered points can be seen, which K-means splits into three clusters as shown on the right side of the figure. The clusters are created based on how close each node is to a selection of K center points, a simple solution that creates very intuitive clusters.

K-means is a very popular choice for clustering because of its simplicity, but also because it also works with large datasets and easily adapts to new examples (K-means advantages and disadvantages, 2023). It also can create clusters of different shapes and sizes. However, there are also a number of downsides to this clustering method, including dependence on a pre-selected *k* number of clusters and problems with outliers dragging or creating new clusters (K-means advantages and disadvantages, 2023).

2.1.2.7 Guided Latent Dirichlet Allocation (Guided LDA)

Guided Latent Dirichlet Allocation, or Guided LDA, is very similar to LDA in that it assigns words and documents to topics. However, while LDA and most topic modeling algorithms are unsupervised, Guided LDA is a semi-supervised learning algorithm (Singh, V., 2017). When using Guided LDA, data scientists provide seed words, or priors, for each topic they want the model to create (Singh, V., 2017). An example of this can be seen in Figure 2.8.



```
1.  (Seed politics): Barack Obama, elections, PM, Narendra Modi,
2.  (Seed Sports): Cricket team, world cup, FIFA
3.  (Seed Science): Einstein, Nobel Prize, Physics, Medicine, biological
4.  (Seed Space): SpaceX, Nasa, Solar Eclipse
5.  (Seed Tech): Tesla, Google, Apple, iPhone
```

Figure 2.8: Guided LDA Seed Words (Singh, V., 2017)

Assigning these seed words to establish the descriptions of the topics helps Guided LDA assign classify words and documents to the intended topics, and can create results that are easier to interpret in some contexts (Singh, V., 2017).

Guided LDA helps to improve the quality of topics discovered by LDA, especially in situations where the underlying corpus is noisy or the topics are difficult to extract. Users can specify a set of relevant keywords or phrases, which can help to extract more meaningful topics that are relevant to their domain of interest (Sande, S., 2020). Finally, the use of guided keywords makes the results more interpretable, as it allows users to better understand how the model arrived at certain topics (Sande, S., 2020).

However, the quality of the results are highly dependent on the relevance and quality of the specified keywords (Sande, S., 2020). It is also very time-consuming, as the user must manually identify keywords, and may overfit if the keywords are overspecific, leading to generalization and inaccurate results (Sande, S., 2020).

In our methodology, Guided LDA was ultimately not used as one of our baseline methods as the quality of the results depends on the relevance and quality of the specific keywords. Additionally, the manual identification of keywords is time consuming and may lead to overfitting and inaccurate results. Overall, Guided LDA can be a useful technique for improving the accuracy and interpretability of topic models, especially in situations where domain-specific knowledge is available (Sande, S., 2020).

2.1.2.8 Top2Vec

Top2vec is one of the newest unsupervised topic modeling algorithms, proposed in 2020 (Angelov, D). Top2vec works by taking a collection of input texts in and directly converting each word into a vector in a semantic space. It then identifies and creates topic vectors (Angelov, D., 2020).

Top2vec works by converting every individual word and document into an embedding vector, which represents the meaning of the word. The idea behind this is that the embedding vectors of similar words and similar documents are similar (Mavuduru, A., 2021). This also allows each document to be treated as an individual point in space so that dimensionality reduction and clustering can then be performed. A visualization of clustering these vectors can be seen in Figure 2.9.

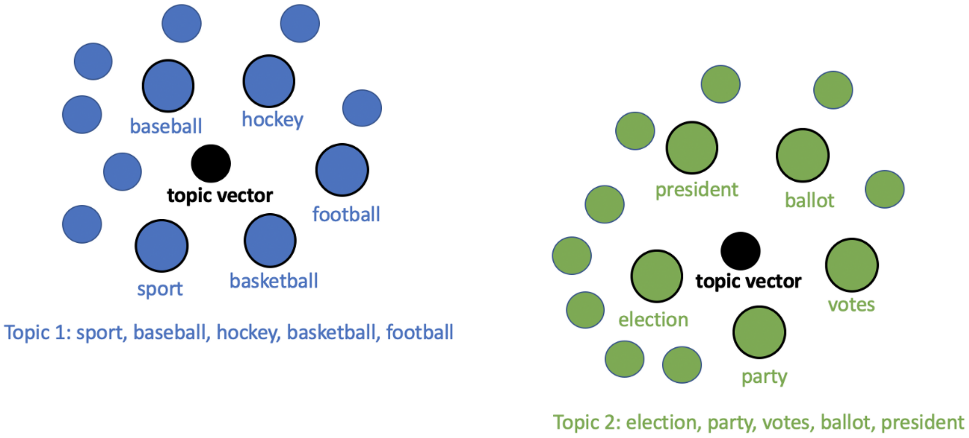

baseball
hockey
topic vector
football
sport
basketball
Topic 1: sport, baseball, hockey, basketball, football
president
ballot
topic vector
votes
election
party
Topic 2: election, party, votes, ballot, president

Figure 2.9: Topic clusters and keywords (Mavuduru, A., 2021)

In Figure 2.9, two topic clusters can be seen, surrounded by “points” representing topic vectors subjected to dimensionality reduction. Because the vectors are created based on how often words appear together in documents, vectors near each other represent similar topics, allowing clustering to be used to classify documents.

Although top2vec has many benefits that rival the industry standard methods of LDA and LSA. While other topic modeling algorithms often require pre-processing such as stemming and lemmatization, as well as ignore ordering and semantics of words (known as bag-of-words representation of documents), top2vec takes in entire document and takes into account word semantic embedding (Angelov, D., 2020). Top2vec also doesn’t require the number of topics to be already known. Although less common than some other topic modeling methods, these benefits make it able to find significantly more informative and representative topics than other models (Angelov, D., 2020).

2.1.2.9 Bidirectional Encoder Representation from Transformers (BERT)

Bidirectional Encoder Representation from Transformers, otherwise known as BERT, is a deep learning model used for NLP tasks that was created and trained by

Google researchers in 2018 (Devlin, J., et al.) It was a leap forward in natural language processing because it took bidirectional training of transformers and applied it to language modeling, which resulted in a deeper sense of language context than other, single-direction language models (Horev, R., 2018).

This deep sense of language context is one of BERT's greatest strengths in the context of NLP when compared to other embedding or modeling techniques; because it incorporates a stronger understanding of contextual meaning of each word in its document representations, its topics are more accurate than traditional, statistical models (Horev, R., 2018). BERT is used to convert textual input into numerical embeddings that capture the contextual importance of each word, and by combining the BERT embeddings from each word in a document (usually by taking the average), a document-wide embedding can be established. Additionally, BERT was trained and is constantly updated on an incredibly large dataset by Google researchers, constantly improving it and maintaining its relevance (Wei, J., 2020). Although BERT is slower than traditional machine learning models, its incredible increase in descriptive power more than makes up for this downside (Horev, R., 2018).

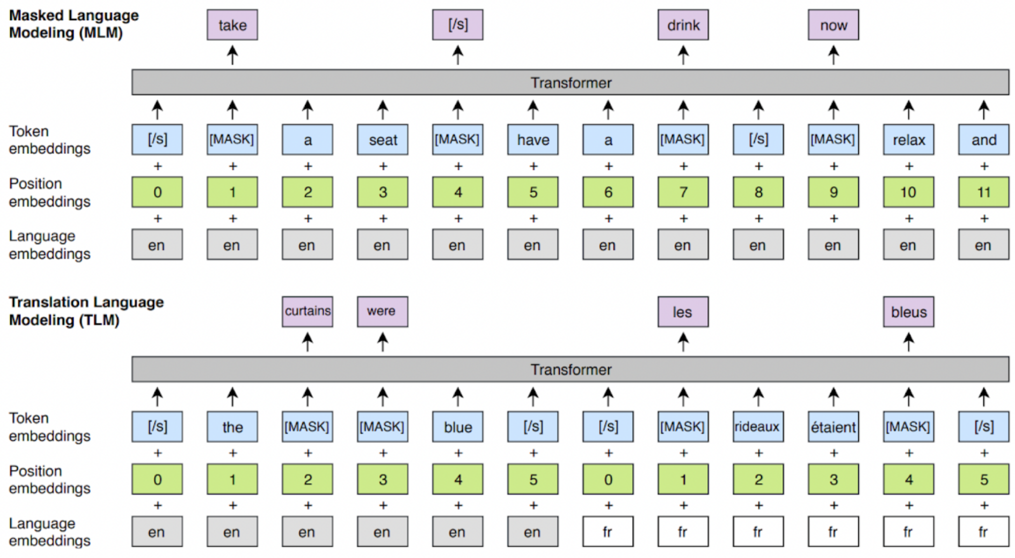


Figure 2.10: BERT Model (Koroteev, 2021)

When applied to the topic modeling context, BERT can be combined with a clustering algorithm to generate topics that contain documents with similar semantic meaning. One such clustering technique is the K-Means clustering algorithm, which serves as a baseline for BERT-based topic modeling and is investigated in the *Methods* section. To further improve the amount of information contained in BERT embeddings generated for a given set of documents (and therefore improve the quality of topics it can extract), another approach is to combine each BERT document vector with the document-topic vector that an LDA model would generate given the same input data. This allows the final embedding to describe both the semantic and statistical content of each document, but these vectors have high dimensionality which can make clustering either difficult or meaningless. This problem is ignored when passing BERT embeddings straight to K-Means, but one technique that can be used to facilitate better clustering (which means better topic modeling) is dimension reduction with an autoencoder before clustering.

2.1.2.10 BERTopic

BERTopic is a topic modeling library created in 2020 by Maarten Grootensorst (Grootendorst, M., 2020; Keita, Z., 2022) that takes advantage of the vectors created by BERT models. The first step of the process is to pass every document through a BERT model to produce embeddings. Once each document has been associated with a vector, the dimensionality of these embedding vectors are reduced with UMAP. UMAP essentially converts large vectors to smaller vectors, but it does so in a way that retains as much of the information from the large vectors as possible. This is done because the next step of the process is clustering the document vectors with HDBSCAN, and small dimensionality is essentially necessary for meaningful clustering. HDBSCAN is a clustering algorithm that automatically determines the number of clusters, which can be used to find the number of topics present in a corpus.

Once the clusters are generated, descriptions of the clusters are created by running a cTFIDF algorithm on the documents. cTFIDF computes the TFIDF score each word would have if all of the documents within each cluster were concatenated, which means that for each cluster, the words with the highest cTFIDF scores give the most meaningful description of that cluster. The final step is to merge the most semantically similar clusters until you're left with the desired number of topics. This can either be a number chosen as an educated guess, provided by a domain expert, or determined by testing the topic quality for different cluster counts

2.1.3 Sentiment Analysis

In our research, our second objective is to perform sentiment analysis on the text. Sentiment analysis is the process of extracting opinions and emotions from a set of the text documents. The sentiment of a given document can often be represented as a numerical score ranging from -1 to 1, where -1 indicates a negative opinion, 1 indicates a positive opinion, and 0 indicates a neutral opinion (Dremali, A., 2020).

There are various methods for conducting sentiment analysis, including rule-based approaches, machine learning algorithms, and deep learning techniques. The selection of the appropriate method depends on the dataset being analyzed, the research question, as well as the computational resources and expertise available (Pak, A., & Paroubek, P., 2010).

A common approach to sentiment analysis is to use a lexicon-based method, where each word in the text is assigned a numerical score based on its sentiment polarity (positive or negative). The scores are then combined to produce an overall sentiment score for the text. Alternatively, a machine learning algorithm such as a Naive Bayes classifier or Support Vector Machine can be utilized to predict the sentiment of new texts by learning patterns in the language used. It is essential to validate the accuracy and reliability of sentiment analysis results since factors such as context, sarcasm, and cultural differences in language use can influence sentiment analysis results (Pang, B., & Lee, L., 2008).

2.2 Technology

In order to test and run our code, we used both Google Colaboratory and Worcester Polytechnic Institute's Turing cluster. We also used Apache Spark's Spark NLP for text processing and other natural language processing methods.

2.2.1 Google Colaboratory

Google Colaboratory, or Google Colab, is a Jupyter Notebook service that can run through browsers, using cloud computing resources to run python code (Google, 2023). This service allows multiple Google account owners to access and simultaneously edit the code base. One of its advantages is that it provides access to a wide range of pre-installed libraries, such as TensorFlow and PyTorch, as well as allows easy integration with other Google services such as Google Drive (Google, 2023). A significant drawback, however, is that without paying for additional resources, there are limits on time and resources available for computations.

In this project, we used Google Colab to run and test our baselines on smaller batches of our data of 10,000 documents.

2.2.2 Turing Cluster

Worcester Polytechnic Institute’s Turing cluster is a research cluster hosted by WPI for students and faculty to do research (Worcester Polytechnic Institute, 2023). It has a total computing power of 1326 CPUs, 9.2 TB RAM, and 64 GPUs (Worcester Polytechnic Institute, 2023). As our dataset has 4.9 million documents, running our pipeline on a powerful distributed system was necessary to quickly and successfully process our data.

Unlike Google Colaboratory, WPI’s turing cluster allows for much more time and computing power to be allocated to python scripts. It does, however, require requesting and reserving resources in advance, as well as some knowledge about managing complex computing resources.

For these reasons, we used Turing to run and test our final pipeline on our complete dataset.

2.2.3 Spark NLP

Spark NLP is an Apache Spark natural language processing library in python, created by John Snow Labs, a company focused on developing AI technologies for healthcare and life sciences. It is the most commonly used open-source natural language processing library currently, used by such companies as Amazon and Google (John Snow Labs, 2023).

In this project, we used Spark NLP to process our initial text data, doing named entity recognition, stop word removal, and other text-processing steps using the library. One of Spark NLP’s main advantages is its scalability, as it can quickly and efficiently process large volumes of text data using distributed computing techniques. Although it can require some knowledge of distributed computing concepts to apply, its user-friendly interface and extensive documentation make it relatively simple to learn.

2.3 Web Visualization

The primary deliverable that we intend to create for this project is a website with visualizations for both topic modeling and sentiment analysis results. Web visualizations are interactive visual representations of data that are created and viewed through web browsers (Heer, J., & Shneiderman, B., 2012). They can include a variety of elements such as graphs, charts, maps, and animations that allow users to explore and understand complex data in a more intuitive and engaging way. These visualizations can be used for a variety of purposes such as data journalism, scientific research, and business intelligence, and can be embedded in web pages, shared on social media platforms, and accessed from mobile devices, making them a versatile and accessible tool for data communication (Heer, J., & Shneiderman, B., 2012). One example of a popular visualization tool, D3.js, is shown below in Figure 2.11. This tool provides significant benefits because it provides a variety of different visualization tools for many possible datasets.

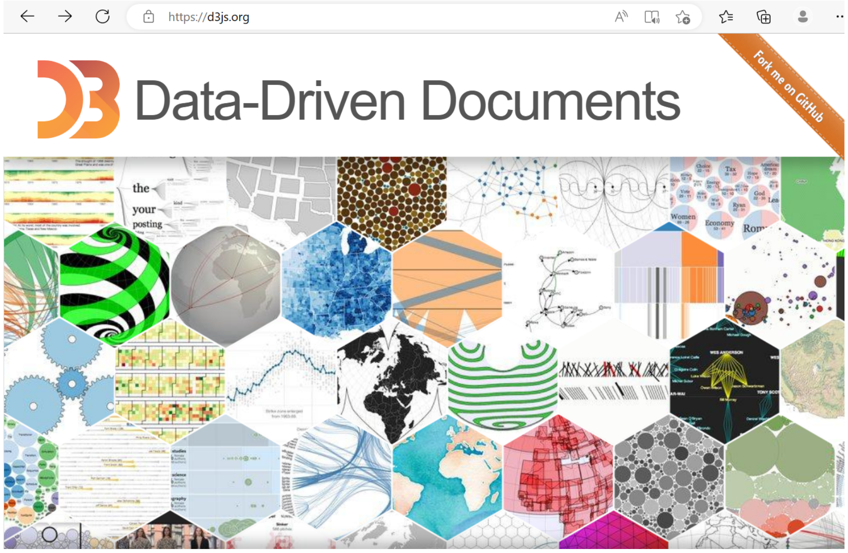


Figure 2.11: D3.js

Some popular tools and frameworks for creating web visualizations include D3.js, Plotly, Bokeh, and Tableau. These tools offer a range of features such as data manipulation, interactivity, and customization options, allowing users to create high-quality and dynamic visualizations with relatively little coding knowledge.

We aimed to use these or similar visualizations to communicate the results of our topic modeling and sentiment analysis clearly and concisely to medical researchers in an easily accessible way.

Methodology

3.1 Initial Plan and Schedule

This project started with the goals of analyzing text posts given to us by Professor Ngan. From this, we extracted a list of tasks, including cleaning the data, testing topic modeling, testing sentiment analysis, and visualizing the results.

In September of 2022, we created our initial schedule for this project. This schedule is shown in Appendix A, but we summarize it briefly below.

In the first term we planned to focus on data cleaning and testing topic modeling with existing methods. We also focused on research in this term, researching natural language processing, topic modeling, and sentiment analysis. In the second term we planned to improve upon and test existing topic modeling methods, and through the second and third term we planned to do the same with sentiment analysis. Finally, we planned to focus on creating a report and visualizations of the results of our research in the last term.

3.2 Initial Data

The posts that we collected were gathered from a variety of cancer chat forums by Professor Cheung Yin Ting in 2021. In total, 4.9 million posts were collected, but due to some duplication and incompatibilities, after our final pipeline of text-processing, we only used approximately 3.6 million posts worth of data.

Below, in Tables 3.1-3.3, simulated data is included as an example of the features data that we worked with. In this example data, a column “Key” has been added that does not exist in the original data so that readers can better understand which columns line up between datasets.

Key	Forum	Keyword	Title	Link	Author	Link_ID
0	CancerChat	Bevacizumab	carbo/taxol/avastin advice	[link]	sundaygirl	0
1	CancerChat	Opdivo	Lung Recurrence. Positive Stories Needed	[link]	ashleybee	1
2	JCCT	Cancer vaccine	It’s that time again…	[link]	Trudes07	2
3	JCCT	Avastin	Avastin	[link]	lorraine1	3

					9	
4	HU	Avastin	Chemotherapy treatment results	[link]	juju16v	4

Table 3.1. Example Data: Indexing Data

Key	Forum	Trimmed_link	Author_type	Author	Subforum	Title	Content	Datetime	TrimmedLink_ID
0	CancerChat	[link]	0	sundaygirl	0	carbo/taxol/avastin advice	Who has ever taken avastin before? My mother …	2019-01-27T13:40:45.000Z	119342
1	CancerChat	[link]	0	ashleybee	0	Lung Recurrence. Positive Stories Needed	i have a biopsy scheduled for monday…	2021-08-02T23:43:05.000Z	207783
2	JCCT	[link]	0	Trudes07	0	It's that time again…	Hi everyone, hope your all doing well. …	2004-11-11T18:20:16.000Z	120123
3	JCCT	[link]	0	lorraine19	0	Avastin	what is the side effect that …	2008-01-13T15:02:55.000Z	241864
4	HU	[link]	0	juju16v	0	Chemotherapy treatment results	My last session ended yesterday…	2009-02-03T08:27:33.000Z	154108

Table 3.2. Example Data: Content Data

Key	Link	Link_ID	Trimmed_link	TrimmedLink_I

				D
0	[link]	0	[link]	119342
1	[link]	1	[link]	207783
2	[link]	2	[link]	120123
3	[link]	3	[link]	241864
4	[link]	4	[link]	154108

Table 3.3. Example Data: Link Data

As mentioned previously, this data was collected by Professor Cheung Yin Ting using a web scraper. Unfortunately, we were not able to obtain any details about or code for the web scraper, which would have provided more insights about how the data was collected. From what we have been able to gather, however, a single web scraper was created and then run on 15 different websites. Unfortunately, this resulted in a significant amount of the final collected data having incorrect dates or missing columns.

3.3 Data Cleaning

In the first and second quarters of this project, we worked on creating a text cleaning pipeline for the text posts. Our final text cleaning pipeline is shown in Figure 3.1.

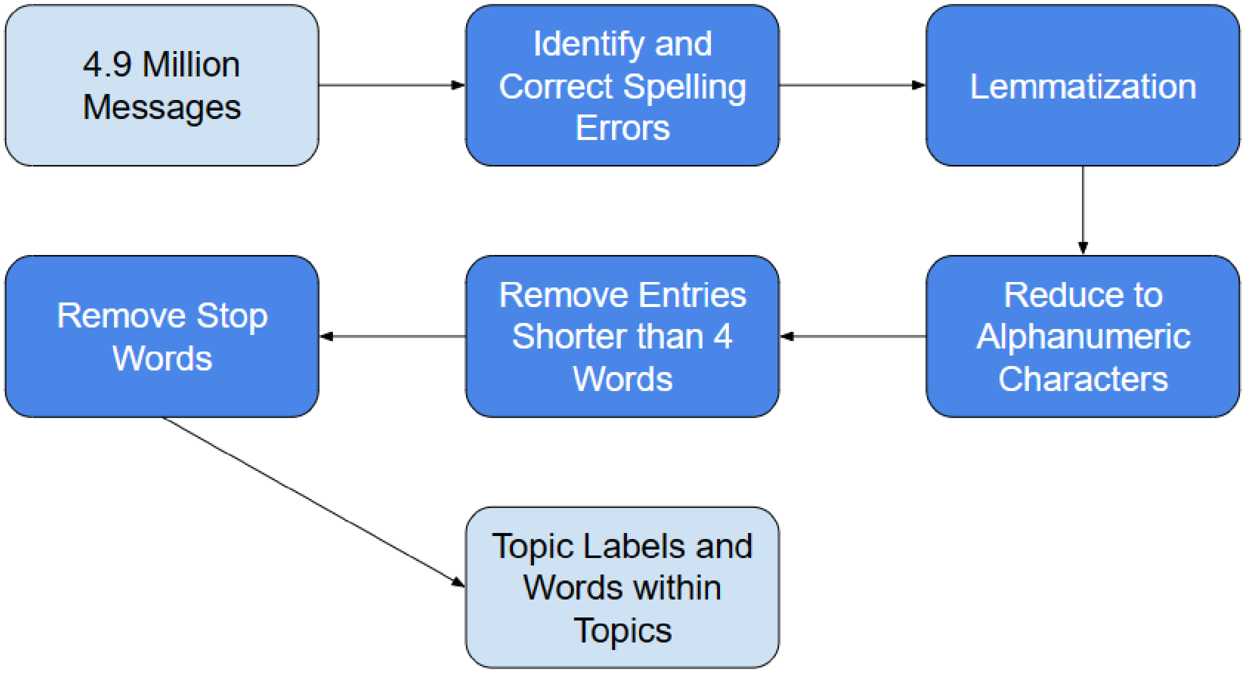

4.9 Million Messages
Identify and Correct Spelling Errors
Lemmatization
Reduce to Alphanumeric Characters
Remove Entries Shorter than 4 Words
Remove Stop Words
Topic Labels and Words within Topics

Figure 3.1. Final Text Processing Pipeline

Our text processing pipeline utilizes SparkNLP, an Apache Spark text processing library. We filter URLs, identify and correct common spelling mistakes, reduce words to their lemma (contextual stem), reduce words to alphanumeric characters, and then filter out stop words.

We also attempted to use Spark NLP to identify ‘named entities,’ or sets of words identifying a single object, location, or person. Some algorithms can use these tokens separately from other words to determine token frequency and relationship. However, not all the text transformers used in the pipeline were capable of processing these tokens, so this was not part of the final pipeline.

After the removal of all documents that were too short to fall under any topic, we were left with 3.6 million documents in a significantly reduced format.

3.4 Baselines

In the process of testing possible topic modeling methods to use on the processed text posts, we ultimately coded and tested 10 baselines, common or well-regarded methods for topic modeling that had the potential to gather quality topics from the data. The baselines we tested are as follows:

Latent Dirichlet Allocation (LDA)
Latent Semantic Analysis (LSA)
Non-negative Matrix Factorization (NMF)
Principal Component Analysis (PCA)
Random Projection
K-Means Clustering
Top2Vec
BERTopic
BERT
BERT + LDA

More details about the baseline methods can be found in section 4.1.

The output of each of our pipelines was represented by a table relating each document in the dataset to the topic number most highly associated with that document.

3.5 Initial Pipeline

At the beginning of this project, we started with an initial set of steps given to us by Professor Ngan as a starting point for us to find a final baseline. This initial baseline contained the following steps:

1. Obtain seedwords from the original 3.6 million messages
2. Sent the seedwords to the medical team so that they can isolate the important topics
3. Optimize number of encoder layers within BERT Encoders for our data
4. Choose an embedding method
5. Find the optimal combination for UMAP reduction
6. Pass clusters to cTFIDF to extract topics

3.5.1 Roadblocks

Although the set of steps we started working with for this project was a great starting point, we ran into a significant number of roadblocks while attempting to implement it.

The first steps in this process were to obtain seedwords from the original set of messages and send them to Professor Cheung Yin Ting and her team so they could isolate the important topics. However, this was complicated by the fact that there was no correlation between the metadata and the content of the documents. To remedy this, we decided to first perform unsupervised topic modeling on the data to isolate keywords, and then send those lists to the medical team to determine the topics as they apply to the immunotherapy domain.

The third through fifth steps involved optimizing BERT and its embeddings for optimal UMAP reduction performance on our dataset. Although theoretically this would allow us to optimize the topic modeling on our dataset, this involved opening up the “black box” that is the interior of BERT and changing its variables. As BERT, as a neural network, is pre-trained and incredibly complex, this would have to involve retraining and recreating BERT, an unreasonably large task for a project of this size.

Finding the best combination of BERT and its embeddings for UMAP reduction was also a challenge, due to both time necessary to process the data, and lack of results to compare it to to determine accuracy.

3.6 Sentiment Analysis Pipeline

For sentiment analysis, we used both VADER and TextBlob. To analyze the sentiment of the text data, we employed two popular sentiment analysis libraries, VADER and TextBlob. VADER is a rule-based library that uses a lexicon of sentiment-related words to analyze text, while TextBlob is a machine learning-based library that uses a Naive Bayes classifier. We opted to use both libraries to increase the accuracy of our results and avoid biases that may be present in a single library. To generate a final sentiment score for each document, we averaged the scores generated by VADER and TextBlob and performed multiple iterations to ensure the stability of our results. The sentiment analysis results are presented in a visual format in Figure 3.2.

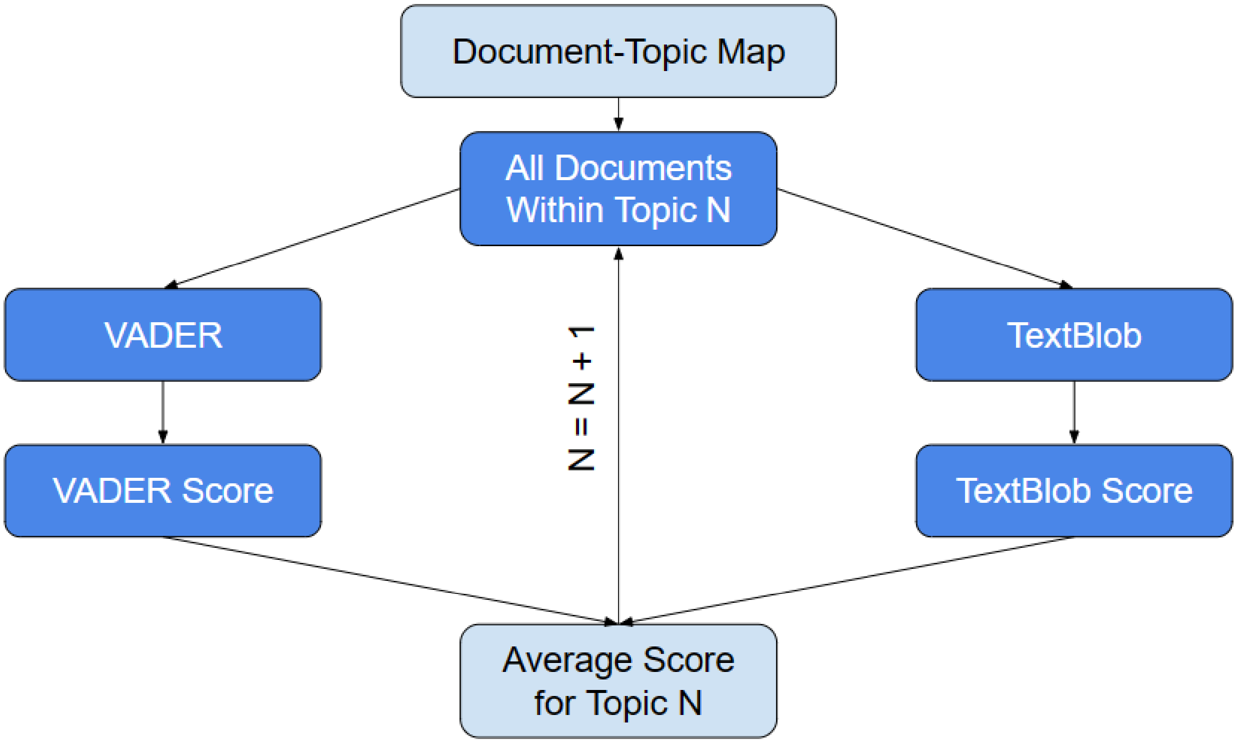


Figure 3.2: Final Sentiment Analysis Pipeline

3.7 Experimental Clustering

One of our attempts at creating our final pipeline was creating a novel clustering and reduction method based on BERTopic but with parallelization for large datasets in mind.

We used precomputed transformer embeddings similar to BERTopic, however chose to replace UMAP and HDBSCAN with simple PCA and our own modified bisecting 'kmeans'. We used the variance information from PCA to progressively divide our document vectors into clusters of roughly equal size and variance until either k partitions have been made or the silhouette score of our clusters converged to an acceptable minima. We did this by finding the mean value of the top n principal components and choosing some combinations of them to partition clusters while maximizing the margin (not unlike SVM). We then recursed this on each of the newly created partitions in parallel.

We tried to test the validity of such a method on the 20 NewsGroup dataset. This dataset has roughly 20 categories, though additional latent topics do exist bringing the 'true' K value only slightly above 20 (California University, 2013). Various iterations of this method were competitive with LDA in terms of topic Coherence but varied in terms of topic Diversity. We also found issues when applying it to our own dataset as there were unknown K values and an overabundant number of latent topics according to HDBSCAN.

3.8 Final Pipeline

In our final pipeline, we used BERTopic along with cTF-IDF to cluster our data. A visualization of this pipeline can be seen in Figure 3.3, below.

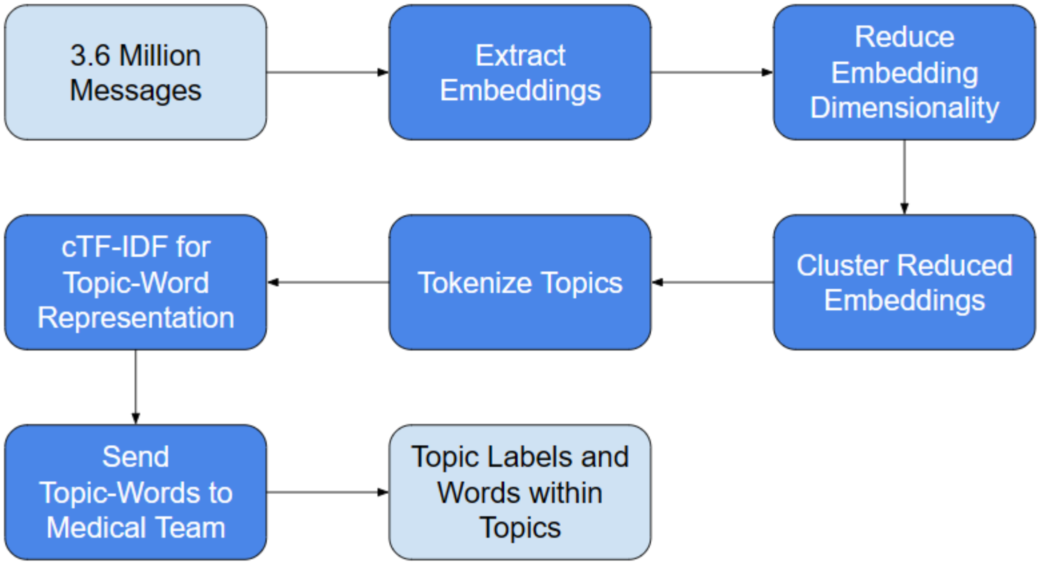

3.6 Million Messages
Extract Embeddings
Reduce Embedding Dimensionality
Cluster Reduced Embeddings
Tokenize Topics
cTF-IDF for Topic-Word Representation
Send Topic-Words to Medical Team
Topic Labels and Words within Topics

Figure 3.3: Final topic modeling pipeline

Although we initially ran 4.9 million messages through our pipeline, our final results were gathered from the pipeline run on 3.6 million messages. When running this method, our team tried looking at different topic counts from 5 topics to 25 topics if metrics would be better depending on topic count. The best results of topic count for this specific dataset and modeling method will be discussed in section 4.2. After clustering our data, we extracted the most frequent and meaningful words from each topic. The topic-word representations included the weight of each word in relation to the other important words, which can be interpreted as its relative importance within that topic. After getting all the topics with the different words and weights within the topics, we sent it to the medical group for assessment. The medical group would give labels to the topics as well as group topics together when topics are closely aligned. With this information, we included visuals of the topics and labels generated both by our optimal baseline and as aggregated according to the professionals on our final website.

Analysis and Results

4.1 Baseline Results

The first steps of this project, after initial research, were to do baseline tests on a select set of our data. These tests, detailed in sections 2.1.3.1 to 2.1.3.10, were run using a selection of the most popular and best-performing topic modeling frameworks currently available. Table 4.1 shows baseline data for ten baselines we tested.

In Table 4.1, each of the rows contains the data for one of the baselines we tested. Sample size and input count are the percentage and number of text posts, respectively, that the baseline was tested on. Topic count is the number of topics the method is classifying between. Finally, time to fit is the amount of time it took to process the data using the method.

Topic diversity, coherence, and quality are all measurements of how good a job the method did isolating topics from the text posts. Topic diversity is a measurement of how well the model is able to capture how different each of the generated topics are from each other (Heka.ai, 2022). Topic coherence is a measurement of how often the documents within each cluster contain multiple of the descriptive words of the topic, providing insight as to how similar documents within each cluster are to each other (Heka.ai, 2022).

Topic quality is topic diversity multiplied by topic coherence. It is a way of combining both measurements into one cohesive number (Heka.ai., 2022).

Method	Sample Size	Input Count	Topic Count	Time to Fit	Topic Diversity	Topic Coherence	Topic Quality
LDA	100%	3,596,380	10	15m	25.00%	81.61%	20.40%
LSA	100%	3,596,380	10	1h 23m	35.00%	83.37%	29.18%
NMF	100%	3,596,380	10	25m	40.00%	80.86%	32.34%
PCA	100%	3,596,380	10	3m	60.00%	81.57%	48.94%
Random Project	100%	3,596,380	10	2m	17.00%	80.66%	13.71%
K-Means	100%	3,596,380	10	36m	37.00%	80.36%	29.73%
Top2Vec	100%	3,596,380	10	2h 9m	76.00%	73.15%	55.59%
BERT + K-Means	100%	3,596,380	10	7h 29m	33.00%	80.71%	26.63%
LDA + BERT	100%	3,596,380	10	22h 1m	36.00%	80.51%	28.98%
BERTopic	100%	3,596,380	10	7h 43m	81.11%	80.48%	65.28%

Table 4.1: Baseline Test Results

$$\textbf{Topic Diversity} = (2/(n*(n - 1)) * \Sigma_{i<j}[(|T_i \cap T_j|) / (|T_i \cup T_j|)]$$

Formula 4.1: Topic Diversity (Tran, N. K., et. al., 2013)

$$\text{Coherence } (x, y) = \frac{\underline{max\{log\,f(x),\, log\,f(y)\} - log\,f(x, y)}}{logM - min\{log\,f(x),\, log\,f(y)\}}$$

$$\text{Average Coherence } (T) = (1/n)*\text{Coherence}(T_j)$$

$$\textbf{Topic Coherence } (T) = 2/(m * (m - 1)) * \Sigma_{wi,\, wj\, \in w}\text{Coherence } (w_i, w_j)$$

Formula 4.2: Topic Coherence (Tran, N. K., et. al., 2013)

$$\textbf{Topic Quality} = \text{Topic Coherence * Topic Diversity}$$

Formula 4.3: Topic Quality (Tran, N. K., et. al., 2013)

As seen in Table 4.1, all of the methods were run on the entirety of the text posts. The data was also fit on 10 topics for each of the methods. While 10 topics was unlikely to be the optimal number of topics, keeping the topic number constant for our methods allowed us to compare them easier.

Time to fit varied wildly across the baselines, with times ranging from a few minutes to nearly 24 hours. Topic diversity, coherence, and quality also varied significantly. However, the baseline with the highest scores by far was BERTopic, with an 81.11% topic diversity score, an 80.48% topic coherence score, and a 65.28% topic quality score.

4.2 Pipeline Results

Since BERTopic was found to be the best performing baseline, we used this in our pipeline and performed additional tests to analyze its performance on different numbers of topics.

When BERTopic was performed without being given a specific number of topics, it extracted 5,548 topics. However, this would have been far too many topics to easily gain insights on, and many of the topics it extracted were not related to cancer treatment, rather things like “nail polish” and “running.” Therefore, we ran BERTopic on a set of different numbers of topics to determine which it performed best on, from 5 to 25 topics. The results of these tests are shown in Table 4.2.

Topic Count	Time to Fit	Topic Diversity	Topic Coherence	Topic Quality
5	14h 42m	87.50%	76.74%	67.15%
10	7h 43m	81.11%	80.48%	65.28%
12	6h 10m	84.55%	80.19%	67.80%
13	6h 20m	82.50%	80.20%	66.17%
14	9h 54m	82.31%	79.57%	65.49%
15	9h 27m	87.86%	80.21%	70.47%
16	8h 53m	70.00%	82.84%	57.99%
17	10h 12m	77.50%	81.76%	63.36%
18	13h 43m	78.24%	80.61%	63.07%
20	7h 54m	75.26%	79.09%	59.52%
25	9h 49m	76.67%	78.34%	60.06%

Table 4.2. BERTopic Tests

From these tests, we determined that BERTopic performed the best with 15 topics overall, with a topic diversity score of 87.86%, a topic coherence score of 80.21%, and a topic quality of 70.47%. For these tests, we chose to test numbers of topics in the range of 5-25 to maintain a small number of topics that we hoped would be general enough to not include irrelevant topics but big enough to gather new insights.

4.2.1 Final Pipeline Results

After generating the final topic modeling results, we sent the topics and the words they contained to Cheung Yin Ting, who sent us back an analysis that we discuss in detail in section 4.2.2.

Our results from our final pipeline, when run with 15 topics, are shown in Figure 4.1. The full list of the words for each of these topics can be found in Appendix B. It is worth noting that there are only 14 topics listed below. This is because the 15th topic is "other," composed of text posts that did not cleanly fit into other topics.

1. hair, good, time, day, hope, year
2. acupuncture, pain, treatment, feel, session
3. yeast, fight, hope, stage, cancer, news
4. gallbladder, liver, ultrasound, bile, issue
5. cows, hope, day, bovine, love, morning, huge
6. saint, faith, family, love, god, strength, miracle
7. ptsd, feel, therapist, anxiety, traumatic
8. antioxidant, supplement, radiation
9. simulation, rad, treatment, appointment, session
10. keto, monsanto, diet, gmo, farmer, pesticide
11. binder, scrapbook, make, post
12. hyperthermia, cell, fever, burn, kill, therapy
13. blog, risk, patient, study, benefit, favorable
14. pain, symptom, mayoclinic, hope, painful, scan

Figure 4.1: Final topic results

When we gathered our results, we also used Topic Coherence, Diversity, and Quality to determine how well our methods created clear, discrete topics. A visualization of our results is shown below in Figure 4.2.

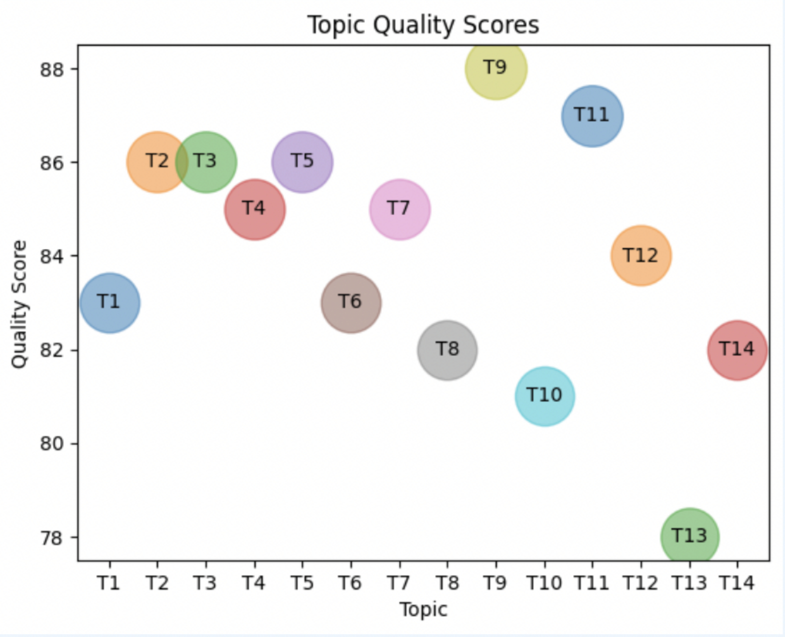


Figure 4.2: Topic Modeling Quality

As can be seen in the above figure, the topic qualities of our 14 topics were found to be between around 78% and 88%, with most falling around 84%. Most of the topics have similar topic qualities, but we can see that specifically topic nine has the highest topic quality and topic thirteen has the lowest topic quality. This shows that some topics may be less coherent than others, but the distribution of scores is very high given the nature of the highly unstructured data.

4.2.2 Final Results Analysis

After the topic results were gathered from the final pipeline, we sent the topics and their contents to Professor Cheung Yin Ting, who analyzed them for their meanings, gathering insights and clustering topics that gave similar insights.

Professor Yin Ting clustered topics 1, 3, 5, and 6 as signifying feelings and positivity and hope among patients being treated with immunotherapy. This may be because this offers the possibility of improved treatment, especially for late-stage cancers. Topics 4 and 13 were clustered as indicating that patients being treated with immunotherapy may have interest in favorable evidence suggesting immunotherapy is especially effective for treating cancers that are difficult to treat with conventional therapies. Topic 8 was identified as indicating that patients believe that the combination of immunotherapy and conventional therapies, such as chemotherapy or surgery, have a higher efficacy against cancer.

The significant number of topics indicating feelings of positivity and hope, along with adjacent topics indicating that patients are interested in favorable studies and believe that immunotherapy in combination with other therapies, is an indicator that patients have generally very positive views of immunotherapy.

Topic 7 was identified as indicating psychological distress associated with immunotherapy. Topics 12 and 14 were clustered as physical side effects and symptoms associated with immunotherapy. Topic 9 seems to indicate that immunotherapy treatment causes significant disruption to daily life. These topics indicate that patients face significant distress and disruption to their lives when treated by immunotherapy, but notably, even these topics include words such as “hope,” and do not seem to indicate any regret or lack of confidence in the treatment.

Topics 2 and 10 seem to indicate that patients treated with immunotherapy are also likely to explore special diets to help prevent cancer recurrence, or use alternative treatments to address physical side effects of immunotherapy. These insights could be useful for medical professionals, so that they could both research other treatments that would help alongside immunotherapy, or advise patients on which alternative treatments or diets they should avoid.

Finally, topic 11 was identified as coping strategies that patients use to address side effects and negative feelings associated with being treated with immunotherapy. These strategies seem to include such things as scrapbooking and crafts. This could provide value to medical professionals because it could give them strategies to suggest to their patients to help them cope with their side effects and emotions.

Through these insights, the 14 topics initially gathered are narrowed down to 8 topics. Descriptions, word clouds, and figures for these topics are shown in Appendices C1-C8.

4.3 Sentiment Analysis Pipeline Results

In addition to our topic clustering pipeline, we ran our sentiment analysis pipeline on the entirety of our text data, using VADER and TextBlob, averaging the results and iterating through our documents within the topics multiple times to get our results. A diagram of these results is shown below, in Figure 4.3.

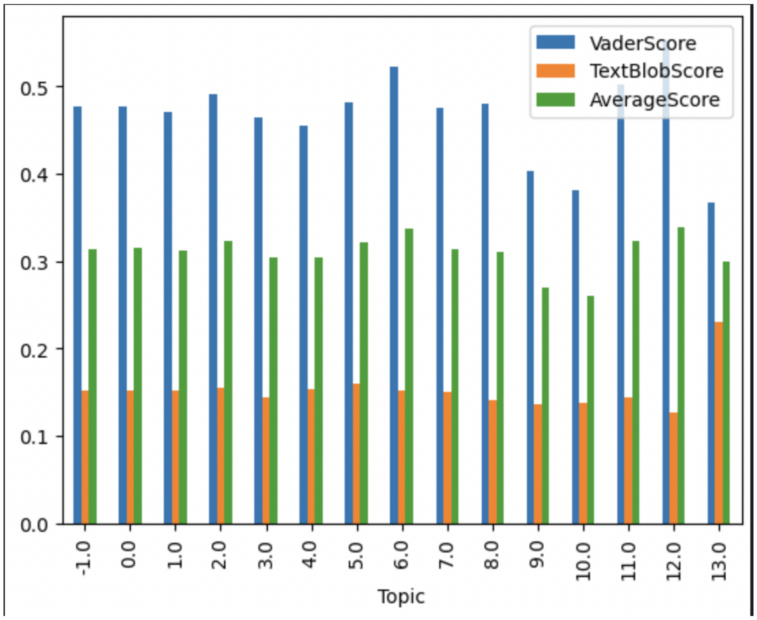
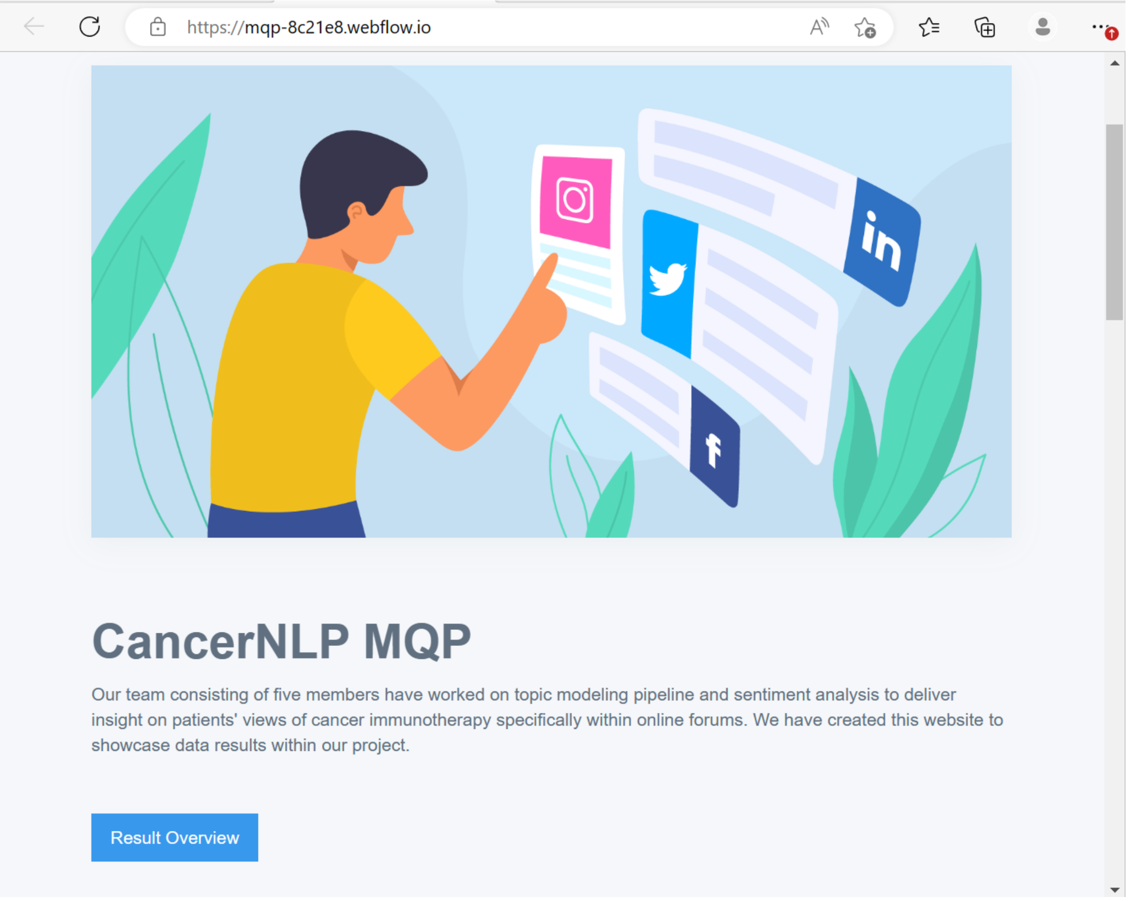


Figure 4.3: Sentiment Analysis Results

In general, the Vader sentiment analysis scores were much higher than the TextBlob sentiment analysis scores, with the average falling around 0.3. As sentiment analysis scores are between -1 and 1, with -1 being very negative and 1 being very positive, these results can be interpreted as generally positive outlooks on the treatment. However, we leave further analysis to the medical researchers, who have the context to better analyze these results. The sentiment analysis visuals of the individual topics are shown in Appendices C1-C8. The grouped topic labels with the average sentiment scores given by Professor Cheung are shown in Appendices D1-D4.

4.4 Website

The final objective of this project was to create a website for the use of medical professionals so that they could better understand their patients' views of immunotherapy. To this end, we have created a website that includes statistics on the sentiment analysis and topic modeling of our data, as well as information on each topic and their associated sentiments. An image of the front page of this website can be seen in Figure 4.4.



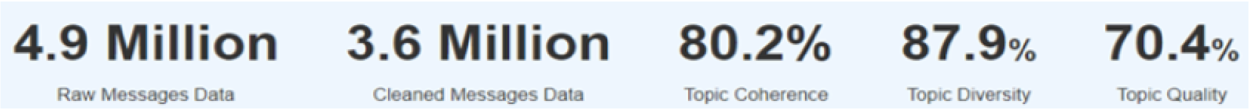
CancerNLP MQP

Our team consisting of five members have worked on topic modeling pipeline and sentiment analysis to deliver insight on patients' views of cancer immunotherapy specifically within online forums. We have created this website to showcase data results within our project.

Figure 4.4: Main page of Cancer NLP Website

4.4.1 Graphs and Statistics

One of the main sections we included in this website was data on the accuracy of our topic modeling and sentiment analysis, including graphs and descriptions. Images of these graphs and our data visualization can be seen in Appendices C1-C8. In addition, we have the grouped topics per topic label given by the medical professionals. Those data visualizations can be seen in Appendices D1-D4. We also included the visualization below, showing basic statistics about the input and cleaned data, as well as topic coherence, diversity, and quality as generated by our final pipeline.



4.9 Million	3.6 Million	80.2%	87.9%	70.4%
Raw Messages Data	Cleaned Messages Data	Topic Coherence	Topic Diversity	Topic Quality

Figure 4.5: NLP Statistics Website Visualization

4.4.2 Topic Word Clouds

Perhaps the most valuable section of our website was the topic word clouds and sentiment analysis data we included. Figures of these word clouds can be seen in Figure 4.6.

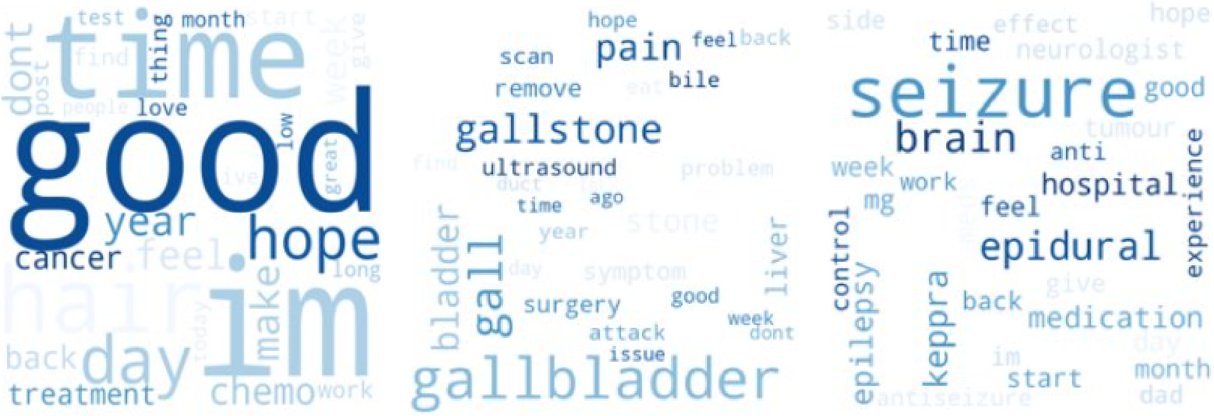


Figure 4.6: Topic Word Clouds

As can be seen above, these word clouds emphasize the more frequent words using bolder colors and larger fonts. This allows medical professionals to gather an understanding of the focus of each of the topics, and gather insights from that.

4.4.3 Word Frequencies

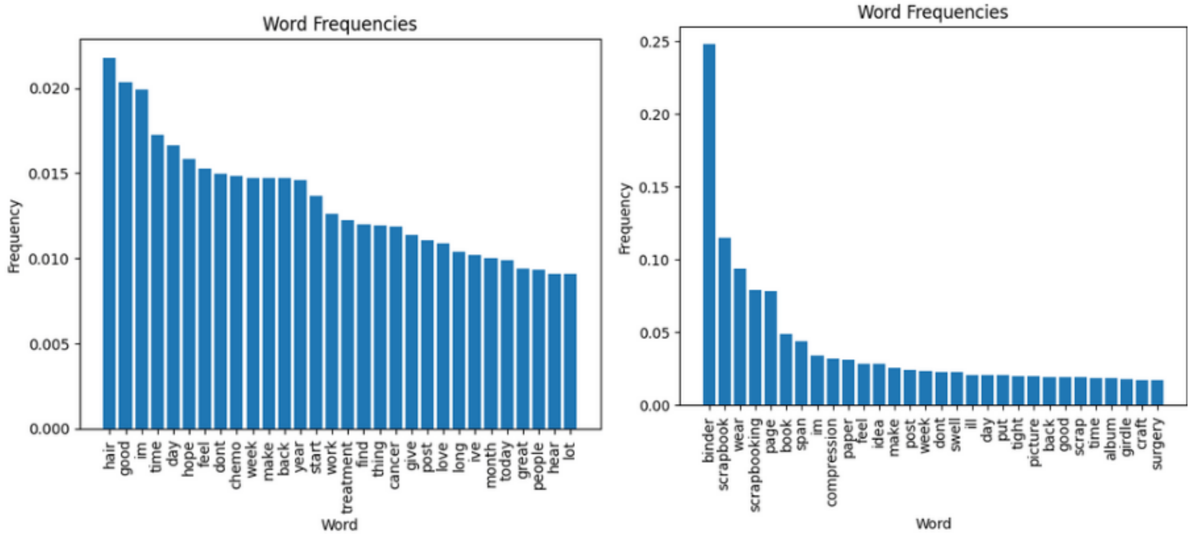


Figure 4.7: Word Frequencies

These graphs provide a numerical summary of the word clouds and give a clear picture of the relative importance of words within each topic.

4.4.4 Sentiment Distributions

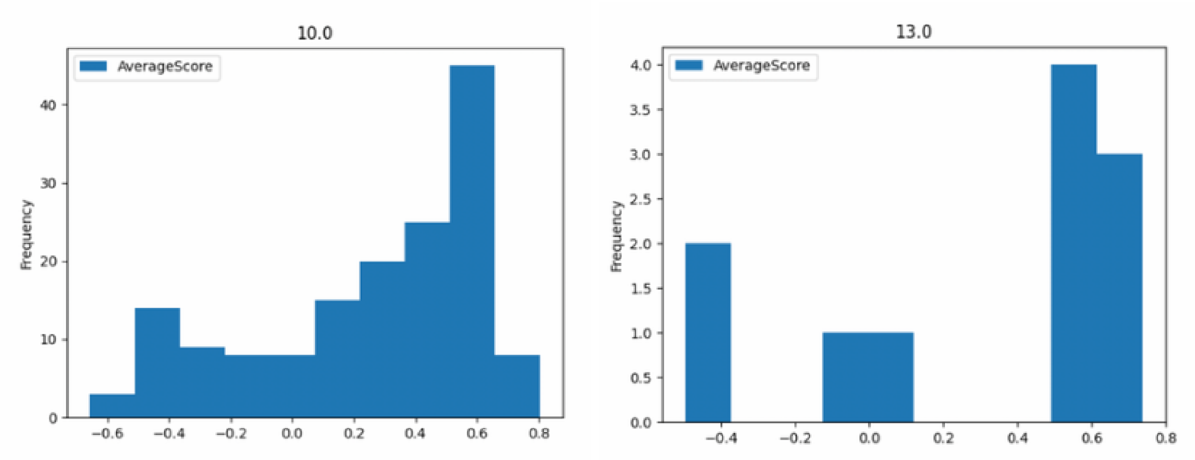


Figure 4.8: Sentiment Distributions

These visuals highlight the distribution of the sentiment of documents within each topic. This can be used to assess public opinion about the topic indicated by the label of the associated topic.

4.4.3 Interactivity

An initial goal of this project was to make the website interactive, through either real-time processing of the data, querying of the data, or other methods. For the website, our team wanted the medical professionals to be able to have user input when looking at the data visuals. With our current data, we didn’t have much flexibility when trying to have user input as most of the data gathered by results were static. With this data not having much flexibility it limited our ability with making it interactive through user inputs. Another major factor was time and resources. When gathering our final results, we didn’t have enough time to get an interactive component to our website as that would have needed more variety in our data results as well as resources such as programs to get interactivity.

Conclusion and Recommendations

In conclusion, through a combination of modern NLP methods, we were able to: extract topics from a database of almost 5 million text documents, analyze the sentiment of each of these forum posts, determine the distribution of sentiment for each topic, and present our results through a web-visualization platform. Our hope is that these results can be used by medical professionals in the immunotherapy domain to determine which aspects of the experience must be prioritized in order to best improve the patient experience. Although we achieved our goals, there are quite a few threads left to pull in terms of the potential for this project in the future:

In this project, we used BERTopic for our topic clustering. Although this gave us good results with high topic quality, there is potential for improvement here. One possible direction for future research would be to enhance BERT for embedding extraction, as well as optimizing dimension reduction and the clustering method used. This would improve the topic modeling on this data, and allow future groups to obtain even more accurate results.

Another possible direction for a future project is to test more sentiment analysis libraries or implement a novel technique. This could be compared with the current pipeline to analyze comparative accuracy, as the current pipeline uses two very common sentiment analysis techniques.

Finally, a future project could link the output from our analysis pipeline directly to a website for a more interactive experience. The website we created for this project allows users and medical professionals to view visualizations we created to present examples of words from each topic cluster and the corresponding sentiment analysis scores. This is a valuable tool, but presents potential for more interactive features such as querying data based on date range, sentiment, targeted topics, etc.

This project was a very successful MQP in that we achieved the main objectives we set out to accomplish. The visualizations we created have the potential to provide value to current and future medical professionals working with immunotherapy patients around the world. We hope that future MQPs will continue this research.

References

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, *167*, 114155. https://doi.org/10.1016/j.eswa.2020.114155

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00042

Albright, R., et. al. (2005). *Algorithms, Initiations, and Convergence for the Nonnegative Matrix Factorization* [Unpublished paper]. Department of Mathematics, College of Charleston.

Angelov, D. (2020, August 19). Top2Vec: Distributed Representations of Topics. Retrieved February 12, 2023, from https://arxiv.org/abs/2008.09470.

Basmatkar, P., & Maurya, M. (2022). An overview of contextual topic modeling using bidirectional encoder representations from Transformers. *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*, 489–504. https://doi.org/10.1007/978-981-16-8862-1_32

Berkeley Institute International. (2021, March 20). *Immunotherapy vs. Chemotherapy: What's the difference?: Berkeley Institute International*. Berkeley Institute International. Retrieved February 10, 2023, from https://berkeley-institute.com/immunotherapy-vs-chemotherapy-whats-the-difference/.

Bingham, E., et. al. (2001). Random projection in dimensionality reduction: Applications to image and text data. *ACM*. Retrieved March 15, 2023, from https://cs-people.bu.edu/evimaria/cs565/kdd-rp.pdf.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and Trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and Topic modeling study. *JMIR Public Health and Surveillance*, *6*(4). https://doi.org/10.2196/21978

Cai, G., Sun, F., & Sha, Y. (2018). Interactive Visualization for Topic Model Curation. *CEUR Workshop Proceedings, 2068.*

Cancer Research Institute. (2023, February 2). *FDA approval timeline of active immunotherapies*. Cancer Research Institute. Retrieved April 7, 2023, from https://www.cancerresearch.org/fda-approval-timeline-of-active-immunotherapies

Chae, B., & Park, E. (2018). Corporate Social Responsibility (CSR): A survey of topics and trends using Twitter data and topic modeling. *Sustainability*, *10*(7), 2231. https://doi.org/10.3390/su10072231

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. Retrieved February 20, 2023, from https://arxiv.org/pdf/1810.04805.pdf.

Dremali, A. (2020, November 4). *What is sentiment analysis in NLP?* AndPlus. Retrieved February 12, 2023, from https://www.andplus.com/blog/what-is-sentiment-analysis-in-nlp-

Engati. (2021). *Lemmatization*. Engati. Retrieved March 18, 2023, from https://www.engati.com/glossary/lemmatization#:~:text=REQUEST%20A%20DEMO-,What%20is%20Lemmatization%20in%20NLP%3F,form%2C%20having%20the%20same%20meaning.

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. *Applied Sciences*, *11*(23), 11091. https://doi.org/10.3390/app112311091

Gersten, T., & Zieve, D. (2021, October 28). Cancer Treatments. Retrieved February 10, 2023, from https://medlineplus.gov/ency/patientinstructions/000901.htm#:~:text=The%20most%20common%20treatments%20are,cancer%20and%20how%20they%20work.&text=Surgery%20is%20a%20common%20treatment%20for%20many%20types%20of%20cancer.

Ghasiya, P., & Okamura, K. (2021). Investigating covid-19 news across Four nations: A topic modeling and sentiment analysis approach. *IEEE Access*, *9*, 36645–36656. https://doi.org/10.1109/access.2021.3062875

Google. (2023). *Colaboratory*. Google . Retrieved February 12, 2023, from https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory%2C%20or%20%E2%80%9CColab%E2%80%9D%20for,learning%2C%20data%20analysis%20and%20education.

Google. (2023). *K-means advantages and disadvantages*. Google. Retrieved March 15, 2023, from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

Goyal, C. (2021, June 26). *Topic modelling using LSA*. Analytics Vidhya. Retrieved March 15, 2023, from https://www.analyticsvidhya.com/blog/2021/06/part-16-step-by-step-guide-to-master-nlp-topic-modelling-using-lsa/

Goyal, C. (2021, June 26). Topic Modelling using NMF [web log]. Retrieved February 15, 2023, from https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/.

Grootendorst, M. (2020). *BERTopic*. GitHub. Retrieved February 21, 2023, from https://github.com/MaartenGr/BERTopic

Grootendorst, M. (2021, January 6). Interactive Topic Modeling with BERTopic. *Towards Data Science*. Retrieved February 21, 2023, from https://towardsdatascience.com/interactive-topic-modeling-with-bertopic-1ea55e7d73d8.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

arXiv. doi.org/10.48550/arXiv.2203.05794

Harsha, A. (2022, December 20). *Understanding Part-of-Speech Tagging in NLP: Techniques and Applications*. Naukri Learning. Retrieved March 18, 2023, from https://www.naukri.com/learning/articles/pos-tagging-in-nlp/

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Communications of the ACM, 55*(4), 45-54. doi: 10.1145/2133806.2133822

Heka.ai. (2022, November 10). *An end-to-end process for semi-automatic topic modeling from a huge corpus of short texts*. Medium. Retrieved April 6, 2023, from https://heka-ai.medium.com/topic-modeling-an-end-to-end-process-for-semi-automatic-topic-modeling-from-a-huge-corpus-of-bfa905d8c2bf

Hoang, M., et. al. (2019). Aspect-Based Sentiment Analysis using BERT. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.

Horev, R. (2018, November 10). Bert Explained: State of the art language model for Nlp. *Medium*. Retrieved February 20, 2023, from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

IBM. (2023). *What is natural language processing?* IBM. Retrieved February 10, 2023, from https://www.ibm.com/topics/natural-language-processing

Ioana. (2020, May 10). Latent Semantic Analysis: intuition, math, implementation. *Towards Data Science*. Retrieved February 15, 2023, from https://towardsdatascience.com/latent-semantic-analysis-intuition-math-implementation-a194aff870f8.

Jain, Y. (2022, May 8). Principal Component Analysis (PCA). *Medium*. Retrieved February 15, 2023, from https://medium.com/@yashj302/principal-component-analysis-pca-nlp-python-ce9caa58bd7a.

Jaiswal, S. (2021, August 1). *Natural language processing‐dependency parsing*. Medium. Retrieved March 18, 2023, from https://towardsdatascience.com/natural-language-processing-dependency-parsing-cf094bbbe3f7

javatpoint. (2023). *K-means clustering algorithm*. javatpoint. Retrieved February 12, 2023, from https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

John Snow Labs. (2023). *John Snow Labs - State of the art NLP in python*. John Snow Labs. Retrieved April 2, 2023, from https://nlp.johnsnowlabs.com/?utm_term=sparknlp&utm_campaign=Search%2B%7C%2BSpark%2BNLP&utm_source=adwords&utm_medium=ppc&hsa_acc=7272492311&hsa_cam=12543136013&hsa_grp=121056973604&hsa_ad=605485254464&hsa_src=g&hsa_tgt=kwd-1243265465686&hsa_kw=sparknlp&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwz6ShBhCMARIsAH9A0qXZ6-DkASWTjXL-hj3s4YonS5RmTeCw-0x22s8dka84B1QdUxBB_DUaAqdjEALw_wcB

Keita, Z. (2022, November 21). *Meet BERTopic- BERT's cousin for Advanced topic modeling*. Medium. Retrieved February 23, 2023, from https://towardsdatascience.com/meet-bertopic-berts-cousin-for-advanced-topic-modeling-ea5bf0b7faa3

Kim, S. (2022, September 5). *Primer on Cleaning Text Data*. Medium. Retrieved March 15, 2023, from https://towardsdatascience.com/primer-to-cleaning-text-data-7e856d6e5791#:~:text=Text%20cleaning%20here%20refers%20to,reducing%20noise%20in%20text%20data.

Koroteev, M. V. (2021, March 22). *Bert: A review of applications in Natural Language Processing and understanding*. arXiv.org. Retrieved April 26, 2023, from https://arxiv.org/abs/2103.11943

Kucher, K., Paradis, C., & Kerren, A. (2017). The state of the art in sentiment visualization. *Computer Graphics Forum*, *37*(1), 71–96. https://doi.org/10.1111/cgf.13217

Kulshrestha, R. (2019, July 19). A Beginner's Guide to Latent Dirichlet Allocation(Lda). Towards Data Science. Retrieved February 14, 2023, from https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2.

Kwon, H.-J., Ban, H.-J., Jun, J.-K., & Kim, H.-S. (2021). Topic modeling and sentiment analysis of online review for Airlines. *Information*, *12*(2), 78. https://doi.org/10.3390/info12020078

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167. [This book provides a comprehensive introduction to sentiment analysis and opinion mining, covering various aspects such as text preprocessing, feature extraction, sentiment classification, and evaluation. It discusses different approaches to sentiment analysis, including lexicon-based methods, machine learning algorithms, and deep learning techniques, and provides examples of their use in different applications. It also highlights some of the challenges and opportunities of sentiment analysis, such as the need for multilingual and multimodal analysis, and the potential for personalized and context-aware analysis.]

Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A survey of sentiment analysis based on Transfer learning. *IEEE Access*, *7*, 85401–85412. https://doi.org/10.1109/access.2019.2925059

Mahadzir, N. H., Omar, M. F., & Nawi, M. N. (2018). A sentiment analysis visualization system for the property industry. *International Journal of Technology*, *9*(8), 1609–1617. https://doi.org/10.14716/ijtech.v9i8.2753

Manning, C. D., Raghavan, P., & Schütze Hinrich. (2019). *Introduction to information retrieval*. Cambridge University Press.

Mavuduru, A. (2021, November 17). How to perform topic modeling with Top2Vec. *Medium*. Retrieved February 12, 2023, from https://towardsdatascience.com/how-to-perform-topic-modeling-with-top2vec-1ae9bb4e89dc.

Mayo Foundation for Medical Education and Research. (2022, August 25). *What to know about surgery for cancer*. Mayo Clinic. Retrieved February 10, 2023, from https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-surgery/art-20044171

Mayo Foundation for Medical Education and Research. (2022, March 22). *Chemotherapy*. Mayo Clinic. Retrieved February 10, 2023, from https://www.mayoclinic.org/tests-procedures/chemotherapy/about/pac-20385033

Medium. (2018, September 12). *Medium*. Retrieved February 12, 2023, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

Menzli, A. (2023, January 25). *Tokenization in NLP: Types, challenges, examples, tools*. neptune.ai. Retrieved March 15, 2023, from https://neptune.ai/blog/tokenization-in-nlp#:~:text=Among%20these%2C%20the%20most%20important,to%20perform%20the%20tokenization%20process.

Mike Bostock. (2023). *Data-driven documents*. D3.js. Retrieved April 7, 2023, from https://d3js.org/

Mishra, R. K., Urolagin, S., Jothi, J. A., Neogi, A. S., & Nawaz, N. (2021). Deep learning-based sentiment analysis and topic modeling on tourism during covid-19 pandemic. *Frontiers in Computer Science*, *3*. https://doi.org/10.3389/fcomp.2021.775368

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Applied Sciences*, *11*(18), 8438. https://doi.org/10.3390/app11188438

Nathan, J., & Adyatama, A. (2020, October 26). *Topic modelling with Latent Dirichlet allocation*. Algoritma Technical Blog. Retrieved March 15, 2023, from https://algotech.netlify.app/blog/topic-modeling-lda/

Padmanee Sharma, M. D. (2022, March 24). *What is the future of immunotherapy?* MD Anderson Cancer Center. Retrieved February 10, 2023, from https://www.mdanderson.org/cancerwise/what-is-the-future-of-immunotherapy.h00-159538167.html#:~:text=Immunotherapy%20represents%20a%20new%20paradigm%20in%20cancer%20care.,tumor%20cells%3B%20instead%2C%20we%27re%20targeting%20the%20immune%20system.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 1320-1326). European Language Resources Association (ELRA). [This paper describes different approaches to sentiment analysis, including lexicon-based methods and machine learning algorithms. It also discusses some of the challenges and limitations of sentiment analysis, such as the need for domain-specific resources and the difficulty of handling irony and sarcasm.]

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. [This survey article provides an overview of the field of sentiment analysis, including its history, applications, and techniques. It discusses both lexicon-based methods and machine learning algorithms, and provides examples of their use in different domains.]

Priya, B. (2022, March 16). *Advanced topic modeling tutorial: How to use SVD & NMF in python*. HackerNoon. Retrieved February 12, 2023, from https://hackernoon.com/advanced-topic-modeling-tutorial-how-to-use-svd-and-nmf-in-python-to-find-topics-in-text

Qadir, A. (2021, April 1). *NLP with LDA (latent Dirichlet allocation) and text clustering to improve classification*. Medium. Retrieved February 10, 2023, from https://towardsdatascience.com/nlp-with-lda-latent-dirichlet-allocation-and-text-clustering-to-improve-classification-97688c23d98

Sande, S. (2020, November 11). *Pros and cons of popular supervised learning algorithms*. Medium. Retrieved April 11, 2023, from https://medium.com/analytics-vidhya/pros-and-cons-of-popular-supervised-learning-algorithms-d5b3b75d9218

Say, R. E. (2003). The importance of patient preferences in treatment decisions--challenges for doctors. *BMJ*, *327*(7414), 542–545. https://doi.org/10.1136/bmj.327.7414.542

Shankar, S. (2021, July 22). *An introduction to topic modeling with Latent Dirichlet Allocation (LDA) - natural language processing (NLP).* LinkedIn. Retrieved February 10, 2023, from https://www.linkedin.com/pulse/introduction-topic-modeling-latent-dirichlet-lda-natural-shankar

Sharma, P. (2021, December 3). *Dependency parsing in natural language processing*. Analytics Vidhya. Retrieved March 18, 2023, from https://www.analyticsvidhya.com/blog/2021/12/dependency-parsing-in-natural-language-processing-with-examples/

Sharma, P. (2023, February 9). *K-means clustering algorithm in python - the ultimate guide*. Analytics Vidhya. Retrieved February 12, 2023, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Shi, H., & Wang, C. (2020). Self-supervised Document Clustering Based on BERT with Data Augment. *ArXiv*. Retrieved February 12, 2023, from https://www.semanticscholar.org/paper/Self-supervised-Document-Clustering-Based-on-BERT-Shi-Wang/1a2894a9fef2b4b668e19306f98594490f2bde05.

Singh, V. (2017). *GuidedLDA Documentation*. Welcome to GuidedLDA's documentation! - GuidedLDA 1.0 documentation. Retrieved January 16, 2023, from https://guidedlda.readthedocs.io/en/latest/

Singh, V. (2017, October 6). *How we changed unsupervised LDA to semi-supervised GuidedLDA*. freeCodeCamp.org. Retrieved February 19, 2023, from https://www.freecodecamp.org/news/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164/

Statistics Globe. (2022, November 24). *Advantages & Disadvantages of PCA: Pros & Cons explained*. Retrieved March 15, 2023, from https://statisticsglobe.com/advantages-disadvantages-pca

Sunday, J. 5. (2022, June 5). *Rectal cancer disappears after experimental use of immunotherapy*. Memorial Sloan Kettering Cancer Center. Retrieved February 10, 2023, from https://www.mskcc.org/news/rectal-cancer-disappears-after-experimental-use-immunotherapy

TechTarget. (2019, March 1). *Patient engagement strategies for improving patient activation*. PatientEngagementHIT. Retrieved April 7, 2023, from https://patientengagementhit.com/features/patient-engagement-strategies-for-improving-patient-activation#:~:text=Patients%20must%20be%20engaged%20and,a%20key%20component%20of%20treatment.

Tran, N. K., Zerr, S., Bischoff, K., Niederée, C., & Krestel, R. (2013). Topic cropping: Leveraging latent topics for the analysis of small corpora. *Research and Advanced Technology for Digital Libraries*, 297–308. https://doi.org/10.1007/978-3-642-40501-3_30

University of California, Irvine. (2023). *Twenty newsgroups data set*. UCI Machine Learning Repository. Retrieved April 25, 2023, from https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

Wei, J. (2020, September 2). *Bert: Why it's been revolutionizing NLP*. Medium. Retrieved March 19, 2023, from https://towardsdatascience.com/bert-why-its-been-revolutionizing-nlp-5d1bcae76a13

Worcester Polytechnic Institute. (2023). *High performance computing*. WPI Academic & Research Computing. Retrieved February 12, 2023, from https://arc.wpi.edu/computing/hpc-clusters/#:~:text=Housed%20in%20the%20Gateway%20Park%20server%20room%2C%20Turing,serving%2073%20different%20faculty%20members%20across%2014%20departments.

World Health Organization. (2023). *Cancer*. World Health Organization. Retrieved April 7, 2023, from https://www.who.int/news-room/fact-sheets/detail/cancer

Wyant, T. (2023). *Treating cancer with immunotherapy: Types of immunotherapy*. Treating Cancer with Immunotherapy | Types of Immunotherapy. Retrieved February 10, 2023, from https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/immunotherapy/what-is-immunotherapy.html#:~:text=The%20main%20types%20of%20immunotherapy%20now%20being%20used,which%20helps%20it%20recognize%20and%20attack%20cancer%20cells.

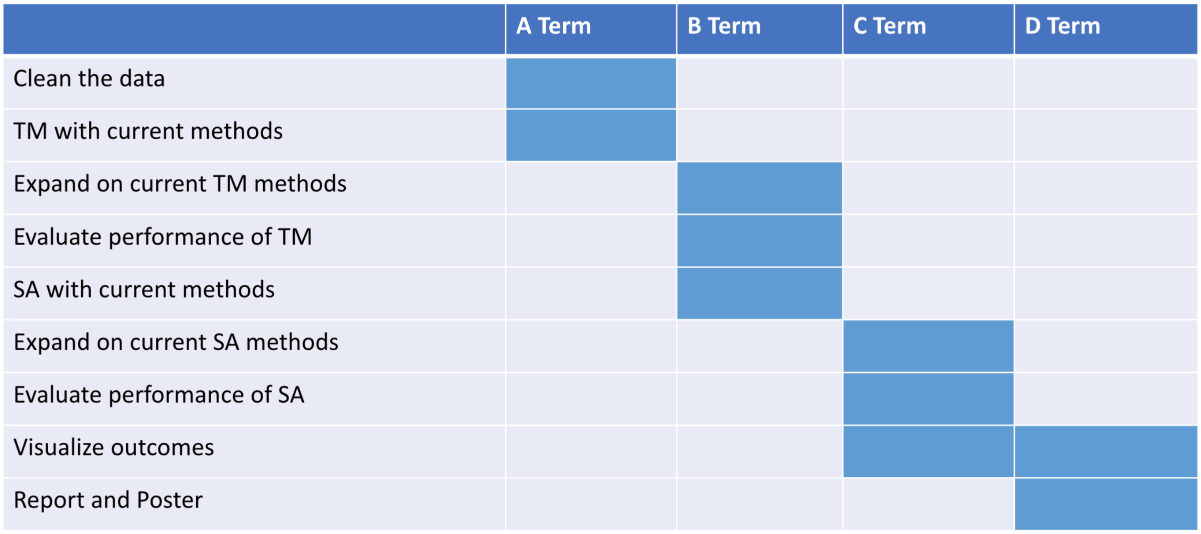
Yadav, A., & Vishwakarma, D. K. (2019). Sentiment analysis using Deep Learning Architectures: A Review. *Artificial Intelligence Review*, *53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yin, H., Song, X., Yang, S., & Li, J. (2022). Sentiment analysis and topic modeling for COVID-19 vaccine discussions. *World Wide Web*, *25*(3), 1067–1083. https://doi.org/10.1007/s11280-022-01029-y

Appendices

Appendix A. Initial Schedule

Our MQP (Major Qualifying Project) is divided into four terms, each lasting approximately 8 weeks. During the first term, the focus was on cleaning the data and using current topic modeling (TM) methods. The second term involved expanding on current TM methods, evaluating their performance, and also working on sentiment analysis (SA) with current methods. In the third term, the team focused on expanding on current SA methods, evaluating their performance, and visualizing the outcomes. The final term, or D term, was dedicated to finishing up the visualization of the outcomes, completing the report, and preparing for the presentation.



	A Term	B Term	C Term	D Term
Clean the data	■			
TM with current methods	■			
Expand on current TM methods		■		
Evaluate performance of TM		■		
SA with current methods		■		
Expand on current SA methods			■	
Evaluate performance of SA			■	
Visualize outcomes			■	■
Report and Poster				■

Appendix B. Topics

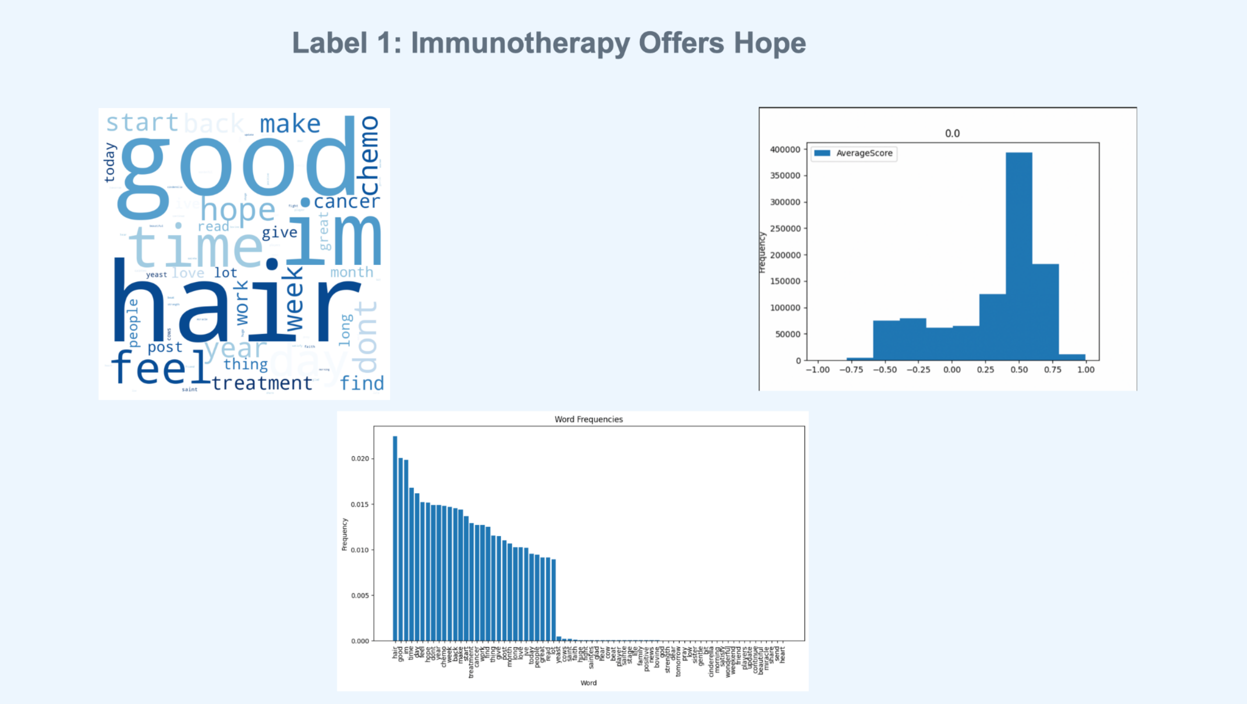
The 14 full topics gathered using BERTopic.

1. [hair, good, im, time, day, feel, hope, dont, year, chemo, week, back, make, start, treatment, work, cancer, find, thing, give, post, month, long, ive, love, today, people, read, great, lot]
2. [acupuncture, pain, hot, im, work, treatment, good, feel, week, session, hope, needle, day, flash, effect, chemo, massage, find, time, side, start, back, ive, flush, give, neuropathy, make, today, hear, thing]
3. [yeast, treatment, good, fight, im, hope, year, chemo, time, back, beat, post, dont, feel, cancer, hear, give, great, love, huge, stage, life, day, find, make, thing, positive, start, news, glad]
4. [gallbladder, gall, pain, gallstone, bladder, stone, liver, symptom, ultrasound, scan, surgery, remove, bile, back, problem, year, duct, im, eat, attack, good, issue, feel, week, hope, time, pancreas, ago, dont, test]
5. [cows, hope, cow, good, today, im, day, bovine, feel, time, back, thing, week, tomorrow, glad, low, great, dont, ive, work, love, hear, post, bit, morning, satisfy, weekend, huge, year, make]
6. [saint, faith, saintes, player, sainte, family, huge, love, god, strength, dear, pray, sister, gentle, cinderella, wonderful, post, friend, hope, players, good, glad, update, continue, beautiful, day, miracle, share, send, heart]
7. [ptsd, feel, therapist, anxiety, traumatic, treatment, experience, im, time, dont, diagnosis, bc, depression, year, cancer, work, trauma, thing, find, life, ive, symptom, back, good, stress, post, people, diagnose, make, disorder]
8. [antioxidant, resveratrol, cell, supplement, cancer, study, radiation, vitamin, protect, chemo, antioxidants, treatment, effect, research, radical, rapamycin, article, kill, rad, dose, interfere, high, free, read, dont, good, herb, grape, avoid, find]
9. [simulation, rad, start, week, ro, im, treatment, monday, today, appointment, day, rat, schedule, ct, hope, tuesday, ill, thursday, map, time, tomorrow, plan, rads, good, forward, tech, chemo, hour, finish, session]
10. [keto, monsanto, diet, gmo, seed, ketosis, carb, food, gmos, genetically, eat, fat, organic, weight, crop, ketone, farmer, mitogenic, corn, im, pesticide, lose, sugar, low, cancer, dont, body, company, year, modify]
11. [binder, scrapbook, wear, scrapbooking, page, book, span, im, compression, paper, feel, idea, make, post, week, dont, swell, ill, day, put, tight, picture, back, good, scrap, time, album, girdle, craft, surgery]
12. [hyperthermia, heat, cell, tumor, cancer, hot, temperature, treatment, degree, immune, chemo, body, clinic, trial, burn, fever, system, damage,

kill, high, therapy, tissue, tnbc, area, low, water, infrared, healthy, skin, protein]

13. [blog, risk, branch, sert, patient, adt, toliver, question, salvage, study, high, ecological, hdr, data, treat, monotherapy, benefit, situation, link, favorable, pelvic, exhaustively, journal, comprehensively, evidence, unpunished, discussion, combining, effect, article]
14. [costochondritis, pain, symptom, costochondritisan, sweep, mayoclinic, teresa, somereassurance, hope, arthritis, mention, constant, okso, give, esmerelda, painful, scan, response, hear, leia, haven, mimic, cathi, issue, colitis, thing, bc, identical, experience, tomorrow]

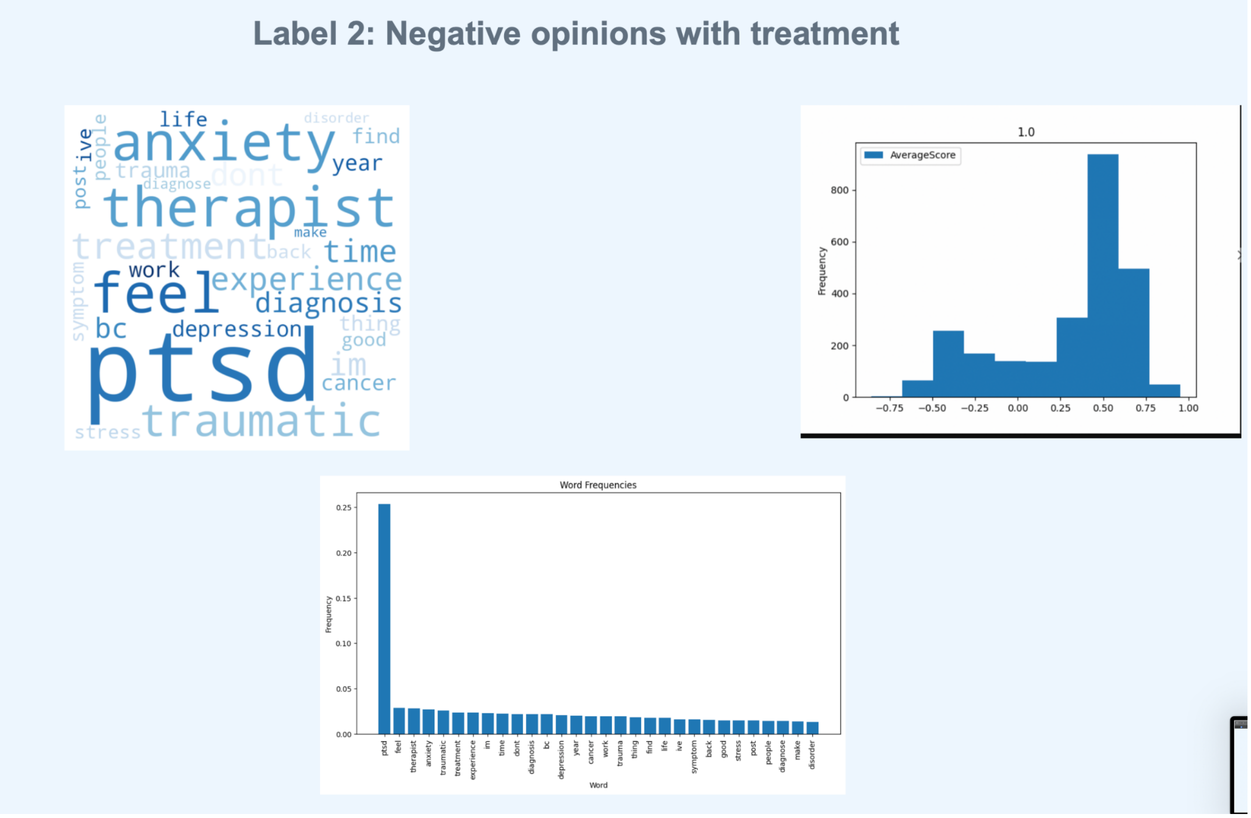
Appendix C.1. Topic 1



Label 1: Immunotherapy Offers Hope

This topic primarily talked about how getting immunotherapy treatment has given hope to most patients as well as the majority feeling positively towards getting immunotherapy treatment.

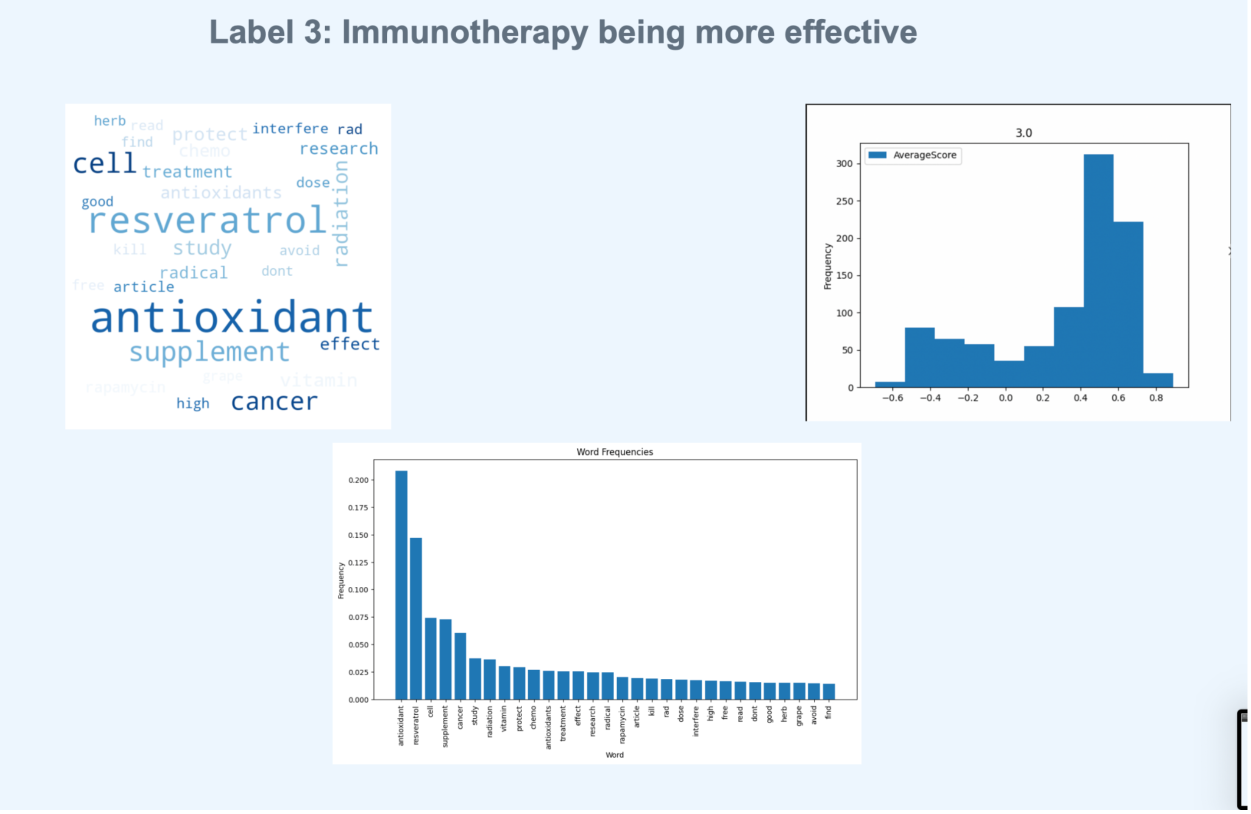
Appendix C.2. Topic 2



Label 2: Negative opinions with treatment

Topic 2 expressed the bad or negative experiences of patients when undergoing the treatment. The treatment had given them negative effects personally. Even though they are talking about bad experiences the overall sentiments were still around neutral with some peaks that are positive.

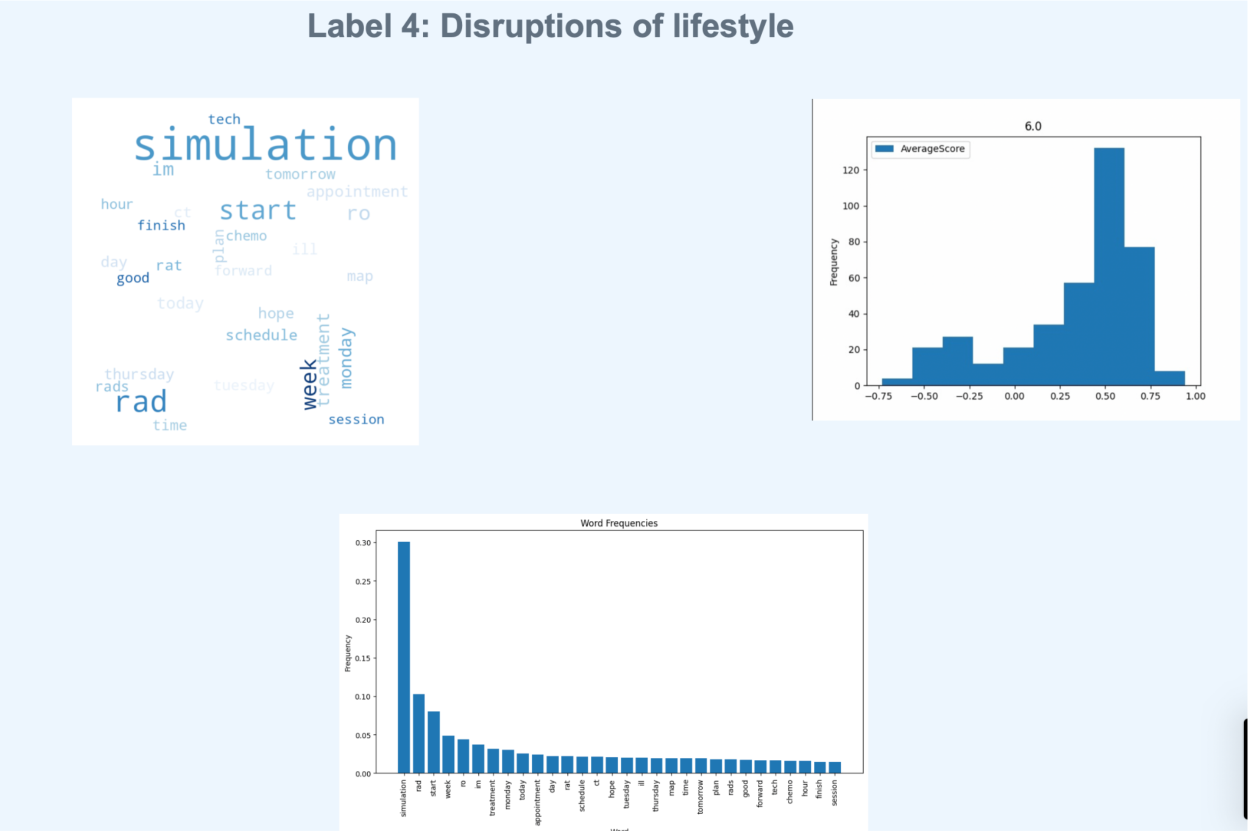
Appendix C.3. Topic 3



Label 3: Immunotherapy being more effective

Topic 3 showcases how patients believe that the immunotherapy treatment has been more effective, whereas there are more levels of sentiments in the positive and negative.

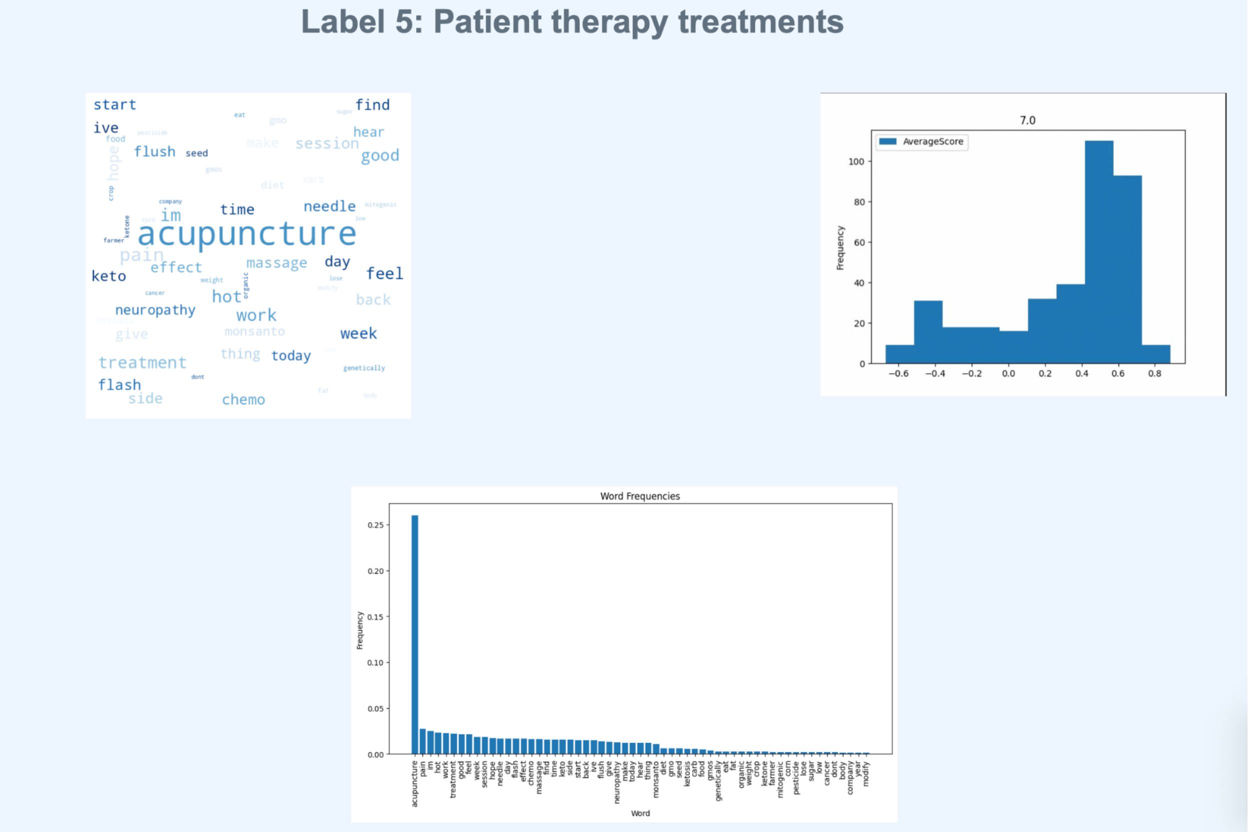
Appendix C.4. Topic 4



Label 4: Disruptions of lifestyle

Topic 4 showcases patients' experiences with the treatment and how the treatment is causing lifestyle problems and disruptions. For example, the patients' treatments cause disruption as they have to go get treatments weekly. Again, there are still more positive sentiments than negative.

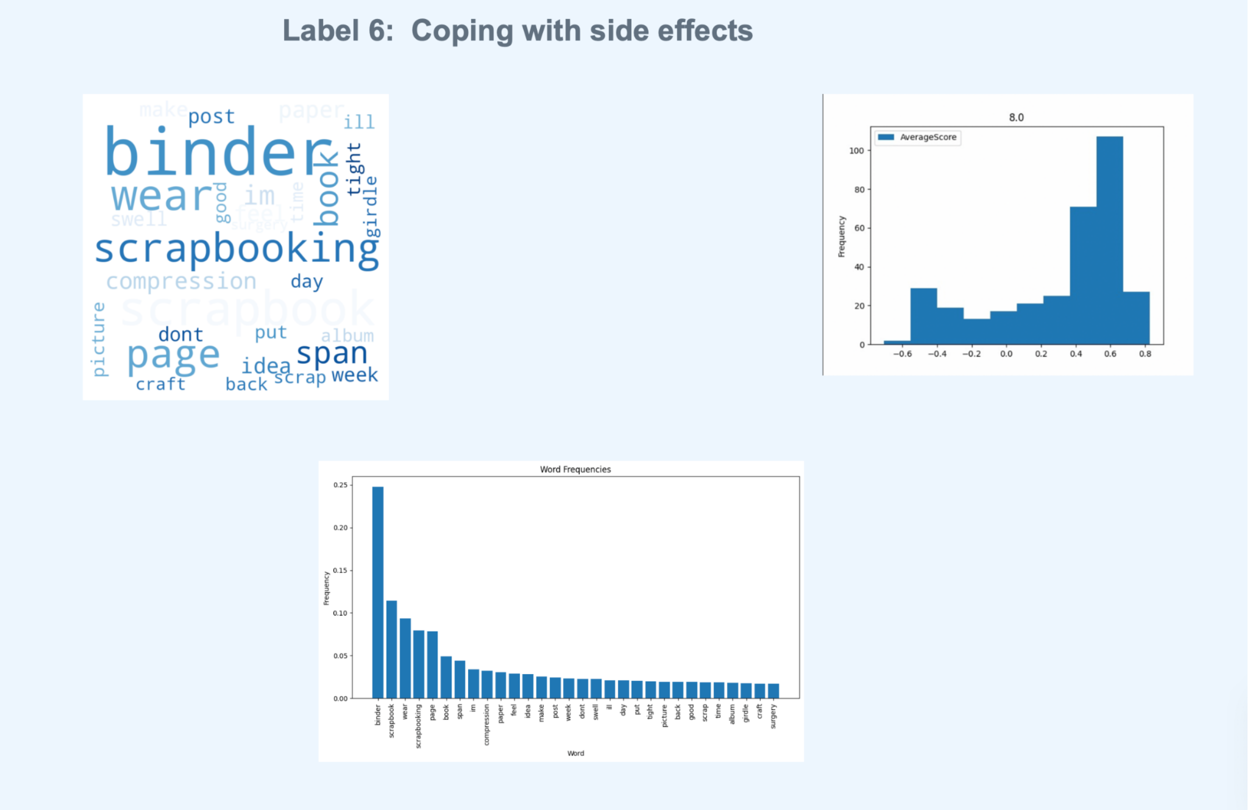
.Appendix C.5. Topic 5



Label 5: Patient therapy treatments

Topic 5 shows how patients have been dealing with the treatments. Most patients have side effects with treatments in general. Many of them reserve other therapies to help with the treatments. Acupuncture is one of the main therapies when dealing with the treatments. Once again, the sentiments have neutral to positive sentiments.

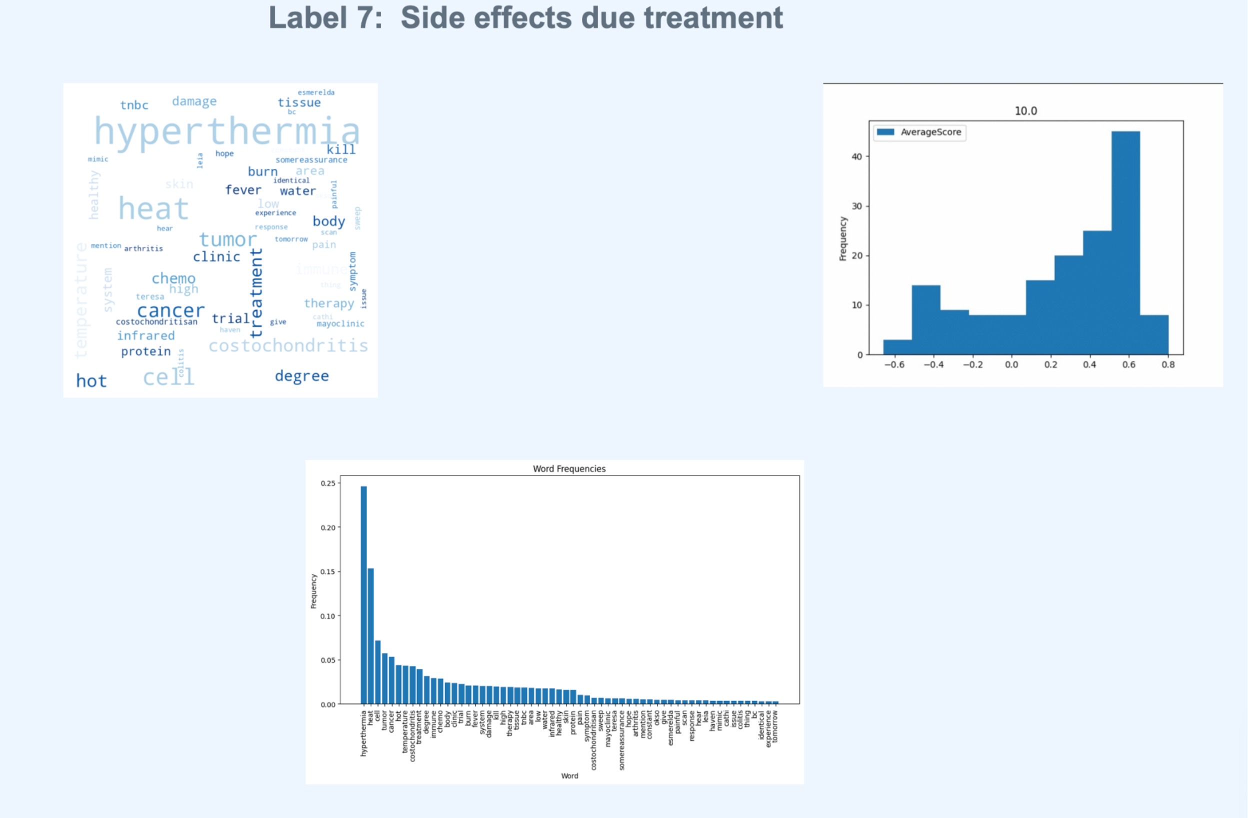
Appendix C.6. Topic 6



Label 6: Coping with side effects

Topic 6 deals with coping with the side effects of the treatment. In this case many patients have done scrapbooking as one of their coping mechanisms when dealing with side effects. The average sentiments still has negative and mostly positive peaks within this topic.

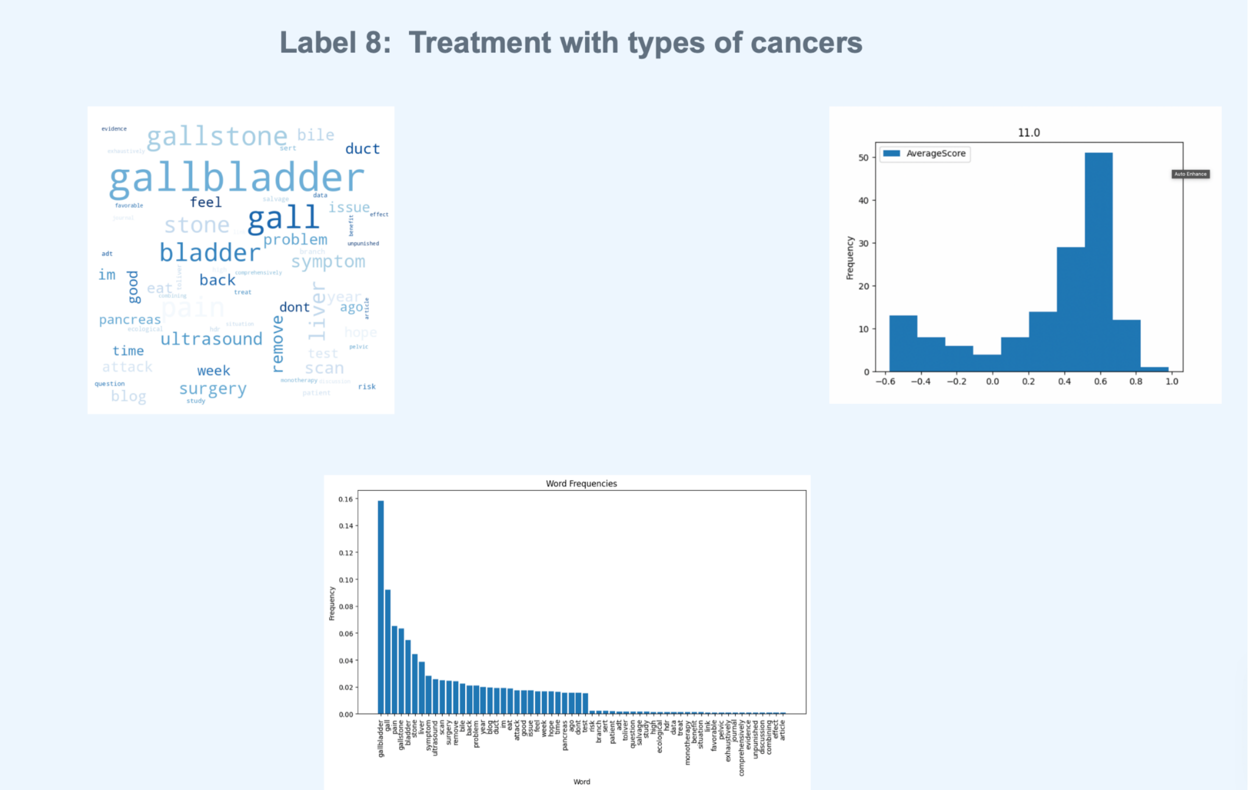
Appendix C.7. Topic 7



Label 7: Side effects due treatment

Topic 7 deals with more side effects due to treatment. Specifically, we can see side effects such as hyperthermia, fever, feeling hot, and other effects. The sentiments overall have a range of negative and more neutral sentiments.

Appendix C.8. Topic 8



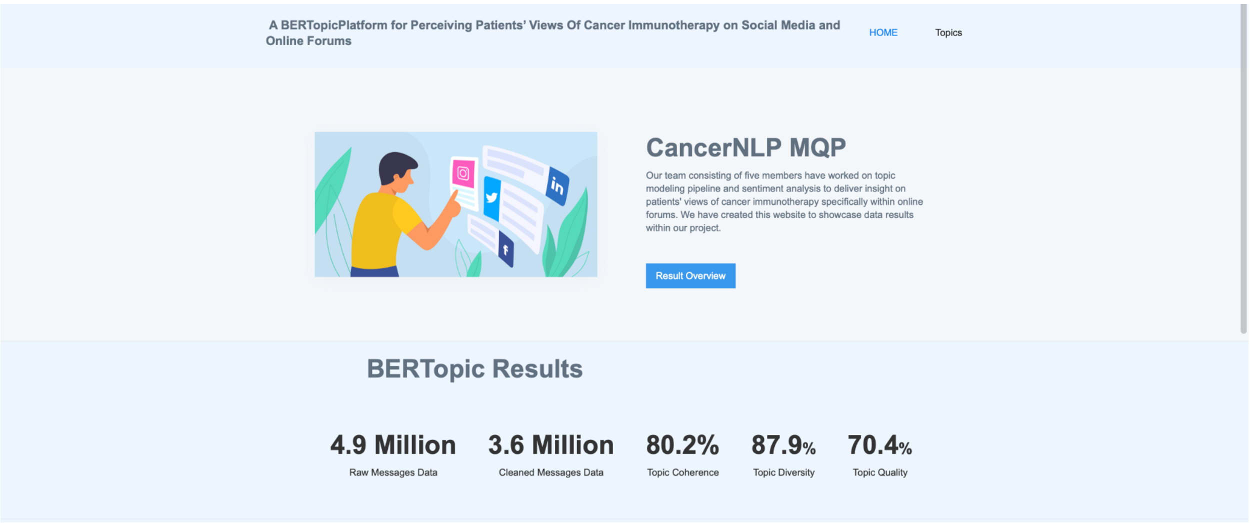
Label 8: Treatment with types of cancers

Topic 8 deals with the different types of cancers being treated within patients. Specifically, bladder cancer was a very popular topic within the forums. Overall, the sentiments show a range of negative and positive sentiments. The topic does have more positive sentiments, but the negative sentiment is more severe when it does occur.

Appendix D. Website

Our website can be found at https://mqp-8c21e8.webflow.io Images of the main pages of the website are shown below.

Appendix D.1. Main Page



A BERTopicPlatform for Perceiving Patients' Views Of Cancer Immunotherapy on Social Media and Online Forums

CancerNLP MQP

Our team consisting of five members have worked on topic modeling pipeline and sentiment analysis to deliver insight on patients' views of cancer immunotherapy specifically within online forums. We have created this website to showcase data results within our project.

BERTopic Results

4.9 Million
Raw Messages Data

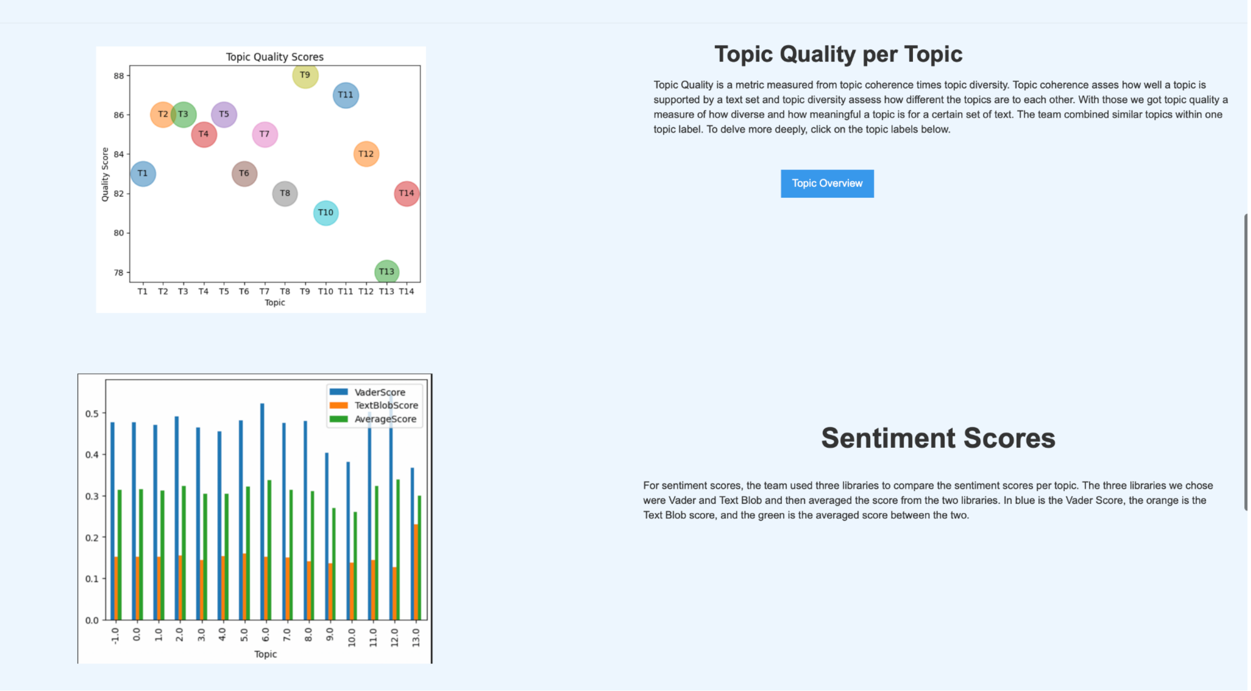
3.6 Million
Cleaned Messages Data

80.2%
Topic Coherence

87.9%
Topic Diversity

70.4%
Topic Quality

Appendix D.2. Topic Quality and Sentiment Scores



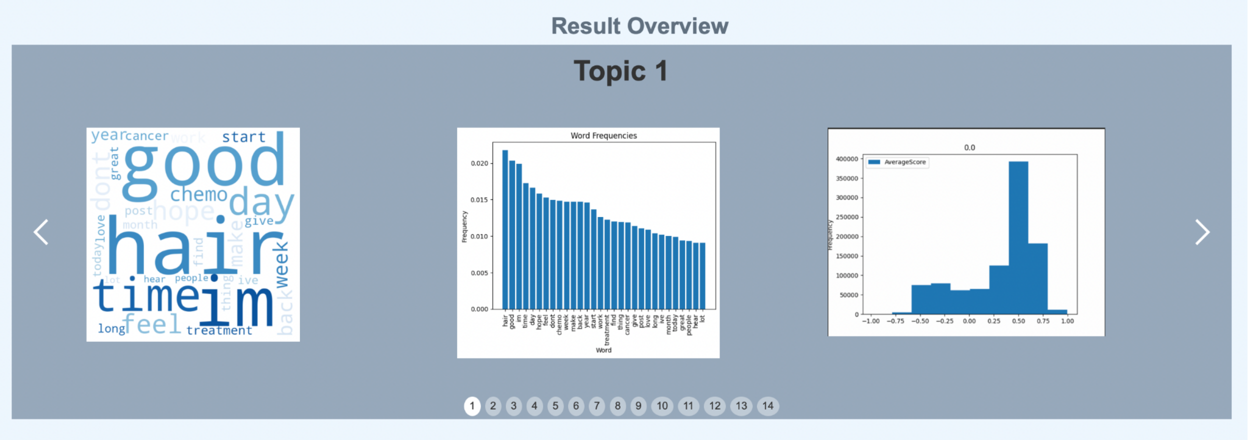
Topic Quality per Topic

Topic Quality is a metric measured from topic coherence times topic diversity. Topic coherence asses how well a topic is supported by a text set and topic diversity assess how different the topics are to each other. With those we got topic quality a measure of how diverse and how meaningful a topic is for a certain set of text. The team combined similar topics within one topic label. To delve more deeply, click on the topic labels below.

Sentiment Scores

For sentiment scores, the team used three libraries to compare the sentiment scores per topic. The three libraries we chose were Vader and Text Blob and then averaged the score from the two libraries. In blue is the Vader Score, the orange is the Text Blob score, and the green is the averaged score between the two.

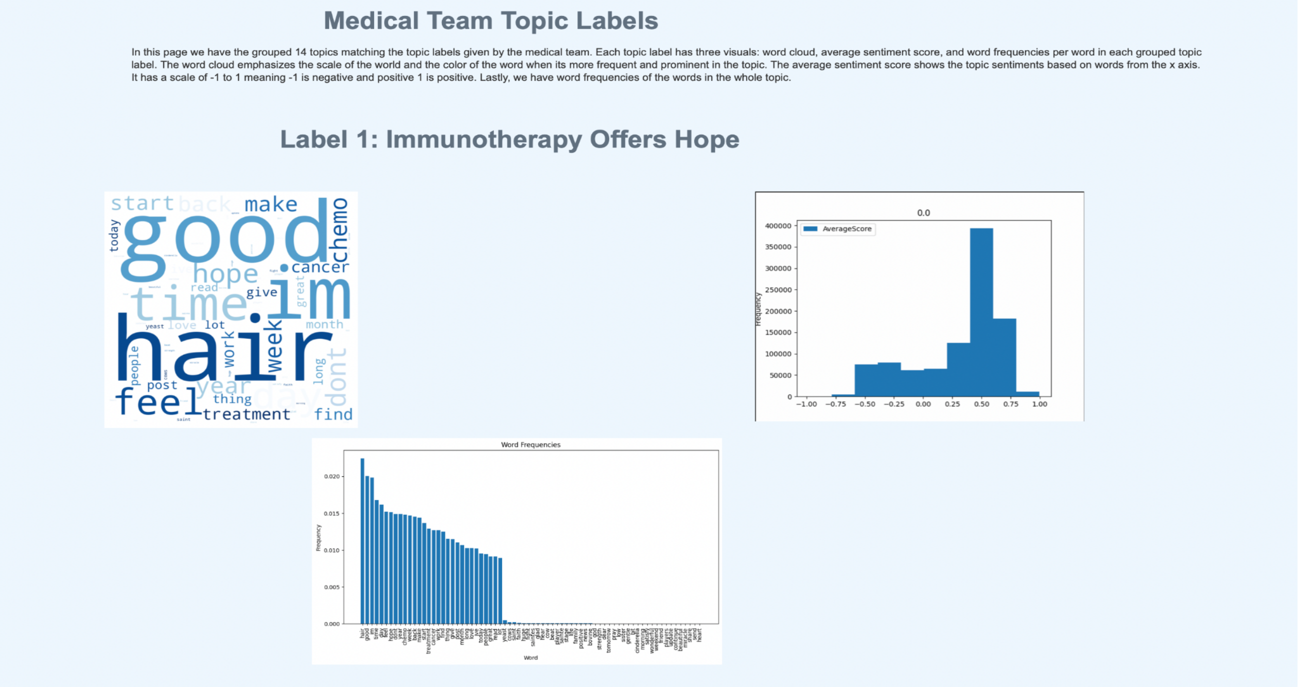
Appendix D.3. 14 BERTopic Visuals



Result Overview

Topic 1

Appendix D.4. Topic Label Groups



Medical Team Topic Labels

In this page we have the grouped 14 topics matching the topic labels given by the medical team. Each topic label has three visuals: word cloud, average sentiment score, and word frequencies per word in each grouped topic label. The word cloud emphasizes the scale of the world and the color of the word when its more frequent and prominent in the topic. The average sentiment score shows the topic sentiments based on words from the x axis. It has a scale of -1 to 1 meaning -1 is negative and positive 1 is positive. Lastly, we have word frequencies of the words in the whole topic.

Label 1: Immunotherapy Offers Hope