University of British Columbia
Data Science Institute
Data Science for Social Good Fellowship

Understanding Online Mentoring Relationships: Using NLP Techniques for Analysis and Program Improvement

Authors: Makafui Amouzouvi, Tiffany Chu, and Jonah Curl

August
2024

University of British Columbia
Data Science Institute
Data Science for Social Good Fellowship

Understanding Online Mentoring Relationships: Using NLP Techniques for Analysis and Program Improvement

Authors: Makafui Amouzouvi, Tiffany Chu, and Jonah Curl

Abstract:

This report results from the research of the 2024 UBC Data Science for Social Good (DSSG) Fellows on an eMentoring project done in partnership with Rural eMentoring British Columbia (ReMBC).

ReMBC is an online initiative that offers mentorship programs for youths from rural, Indigenous and undeserved communities in BC.[11] The goal is to prepare these youths for life after high school and transition to post-secondary education and the workforce. This program uses an online platform called MentorCity for all Mentor-Mentee conversation exchanges. A format which allows Rural eMentoring BC (ReMBC) to evaluate the effectiveness of their programs by having direct access to these conversations along with collecting surveys and reviews from the users (mentors and mentees). However, human reviews have limitations in terms of manpower and time. Therefore, this project aimed to analyze the conversations using machine learning techniques, with the goal of contributing to social good by identifying trends and patterns that the ReMBC team can leverage to enhance future relationships.

This research project involved analyzing text conversations exchanged by over 500 Mentors and over 1,000 Mentees. Specifically, the team explored five major natural language processing techniques: topic modeling, sentiment analysis, mood detection, abstractive summarization, and Mentor-Mentee matching (recommendation systems) to find insights in these converstations.

The result of this exploration is the deployment of a user-friendly web application, which will enable the ReMBC initiative to continuously benefit from the relevant findings and tools of this research, making it easy for even non-technical experts to use NLP tools and obtain clear, interpretable insights.

Although the primary beneficiary of this research is ReMBC, it also serves a broader social good goal by helping to better support students from Indigenous and rural communities. As a result, we hope to see more students from Indigenous and rural areas pursue post-secondary careers in the coming years.

August
2024

Contents

1 Background

In British Columbia (BC) 17% of the population lives in Rural and Indigenous communities.[13] Rural and Indigenous Communities have increased need for healthcare due to their aging population, increased mortality rate, and increased rate of chronic conditions compared to urban regions.[4] However, due to increased distances to hospitals, increased costs to access healthcare, and reduced number of physicians, these regions have more difficulty to access health services.[4] Research has shown that students from Rural and Indigenous communities are more likely to consider pursuing careers in their home communities[6], leading the University of British Columbia (UBC) to develop programs to help students access healthcare education.[9] However, many students still experience barriers to access and are often unaware of these opportunities to help them succeed.

As an initiative of the Rural Education Action Plan and UBC, Rural eMentoring BC (ReMBC) was developed to promote and inspire new rural and Indigenous healthcare practitioners in BC.[11] ReMBC offers online mentorship programs designed to help students through career exploration and development, starting with junior high and high school where students begin exploring their career options and guiding them through their transition to post-secondary.[11] Their programs ensure they have the resources and skills necessary to succeed in various careers including healthcare and other post-secondary programs. Their high school program pairs students with Mentors in professional programs such as Medicine, Law, and Dentistry, who are trained to help engage students through their career exploration. ReMBC further supports relationship development through their specially designed curriculum which includes 13 units students can choose from including Transitioning from Rural to Urban Environment's, Developing Good Study Habits, and "Adulting".

ReMBC uses Mentor City, an online platform designed to manage eMentoring programs.[8] Through Mentor City students can access the ReMBC curriculum, complete assignments, and receive guidance from their Mentor. Mentor City allows ReMBC to maintain a record of conversations between Mentors and Mentees, allowing continuous and retrospective analysis of what students are engaging with. ReMBC has used this data to assess the success of their program, but with over 1000 mentorship relationships to date, it makes it more time consuming to parse every relationship and draw conclusions about all the mentorship relationships. Through Natural Language Processing (NLP) techniques including Topic Modeling, Mood Detection,

and Abstractive Summaries, analysis of mentorship relationships can be done at a larger scale.

2 Project

2.1 Problem Statement

What are successful Mentor-Mentee pairs doing to sustain their relationships that the ReMBC could further leverage in future relationships?

2.2 Project Objectives

The problem statement broken down into quantifiable tasks means addressing the following:

- What factors correlate with success?
- What curriculum units do Mentees find most engaging?
- Are there any off topic subjects that Mentees also find engaging ?
- How does engagement change over time among Mentees?

2.3 Deliverables

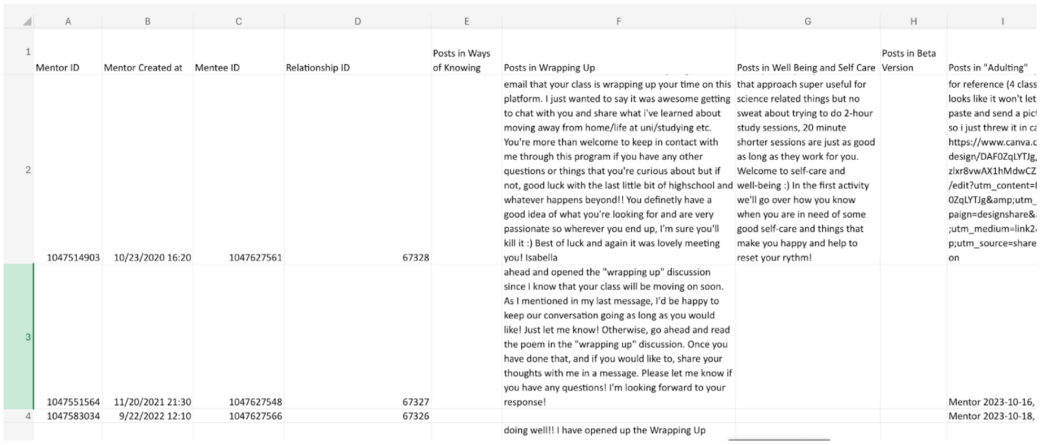
- **A Topic Modeling Pipeline** to reveal the themes of interest among Mentees
- **An Abstractive Summary tool** for researcher to quickly get relevant information and summaries
- **A Mood Detection tool** to point out the attitude/engagement of Mentees over time
- **Mentor-Mentee Matching Recommender System** to aid in forming relationship pairings
- **A Web Application** that integrates NLP tools, providing researchers with streamlined and accessible resources for their analyses.

2.4 Dataset

The original dataset comprises all the conversations between Mentors and Mentees in the ReMBC program. The pairs discuss in units. Each unit has background info, like an article, video or text, and an activity or discussion prompts that encourages discussions among the pairs. The dataset contains discussions among over 1000 Mentees and over 500 Mentors. It is usually one-on-one relationships with the exception of some rare cases of a Mentor managing two or more Mentees.

Most of the conversations dataset was provided to us as excel spreadsheets containing 17 columns (4 identification columns and the 13 curriculum categories):

- **Mentor ID:** unique identifier for the Mentor of the relationship
- **Mentor Created at:** date-time the Mentor was created
- **Mentee ID:** unique identifier for the Mentee of the relationship
- **Relationship ID:** unique identifier for the relationship
- **Posts in ways of Knowing:** conversations between the Mentor and Mentee in the given category
- **Posts in Wrapping up:** conversations between the Mentor and Mentee in the given category
- ...
- ...(the remaining categories)



	A	B	C	D	E	F	G	H	I
1	Mentor ID	Mentor Created at	Mentee ID	Relationship ID	Posts in Ways of Knowing	Posts in Wrapping Up	Posts in Well Being and Self Care	Posts in Beta Version	Posts in "Adulting"
2	1047514903	10/23/2020 16:20	1047627561	67328		email that your class is wrapping up your time on this platform. I just wanted to say it was awesome getting to chat with you and share what i've learned about moving away from home/life at uni/studying etc. You're more than welcome to keep in contact with me through this program if you have any other questions or things that you're curious about but if not, good luck with the last little bit of highschool and whatever happens beyond!! You definetly have a good idea of what you're looking for and are very passionate so wherever you end up, I'm sure you'll kill it :) Best of luck and again it was lovely meeting you! Isabella	that approach super useful for science related things but no sweat about trying to do 2-hour study sessions, 20 minute shorter sessions are just as good as long as they work for you. Welcome to self-care and well-being :) In the first activity we'll go over how you know when you are in need of some good self-care and things that make you happy and help to reset your rythm!		for reference (4 class looks like it won't let paste and send a pic so i just threw it in c https://www.canva.c design/DAF0ZqLYTJg, zlxr8vwAX1hMdwCZ /edit?utm_content=I 0ZqLYTJg&utm_ paign=designshare& ;utm_medium=link2 p;utm_source=share on
3	1047551564	11/20/2021 21:30	1047627548	67327		ahead and opened the "wrapping up" discussion since I know that your class will be moving on soon. As I mentioned in my last message, I'd be happy to keep our conversation going as long as you would like! Just let me know! Otherwise, go ahead and read the poem in the "wrapping up" discussion. Once you have done that, and if you would like to, share your thoughts with me in a message. Please let me know if you have any questions! I'm looking forward to your response!			Mentor 2023-10-16,
4	1047583034	9/22/2022 12:10	1047627566	67326		doing well!! I have opened up the Wrapping Up			Mentor 2023-10-18,

Figure 1: Table of Example Data From Excel Spreadsheets

Each row represented the entire conversation history between a Mentor and a Mentee. An example of this can be seen in **figure 1**.

A small piece of the dataset provided to us was Word files containing conversations of individual relationships, specifically conversations that overflowed the capacity of the spreadsheets cells. An example can be seen in **figure 2**.

Mentee commented at 10:25AM January 6

Hello! Happy New Year!

I think the first floor of the library is our go-to study spot because we're able to talk but also get work done, and it's kinda central. I spent most of last year studying on the library's second floor, and I haven't studied there at all this year.

My finals went well, and I'm happy with my class grades, which is good. The interesting thing I've found with the second year is that we're all just better at school. Classes are more difficult, but we know how to study, we're more relaxed for exams because we know how they work, and we better understand what we need to do to succeed.

Okay, I feel the same way about if January 2023 me talked to December 2023, so much happened last year (most of it good) but ultimately it was a good year for personal growth!

How did the December (mid-year) performance review go? Did you get a raise?

MCAT studying sounds exciting and stressful! How's it going?

I appreciate that we're in the getting hired section because I'm about to start applying for a summer job! I'm going for the same approach as last year, getting a job in a lab at UBC. So I've updated my resume and getting my references in order. Wish me luck!

I've got to get back to studying

Your mentee,

Mentee

Mentor commented at 9:03PM November 29

Hi Mentee!

WELCOME BACK TO MENTORING TO THE BOTH OF US! Goodness, 2 months pass just like that and now look at all these upgrades ReMBC has for us... I'm not even sure what half of these buttons mean, so let's have fun trying them out as we get back into the swing of things. Like are those... Emojis?! (I think we can stick with using our Apple ones for the variety though so shhh). My heart felt so warm reading all your life updates, because I always feel like we have this telepathic understanding

Figure 2: Example from .docx files

3 Methods

3.1 Data Cleaning

During data cleaning, we separated the conversations into individual responses. In our cleaned dataset each row represents a single message sent by either a Mentor or a Mentee, recorded at a specific time with the original unit it was in. This approach allows us to compare the content of responses both between different mentorship relationships and within the same relationship over time. An example transformed data set can be seen in **table 1**.

Response dt	Relationship ID	Mentor	Response	Category
2024-01-01 10:03	1584937	Mentor	Hi, Hope you are well	Ways of Knowing
2024-01-01 12:00	1584937	Mentee	I am well, thank you	Ways of Knowing

Table 1: Mentor-Mentee Conversation Data example including: datetime of the response, relationship ID, the role (Mentor/Mentee), their response message, and the post category

Using Regular Expressions (REGEX) we split the conversations into individual rows and split each response into three pieces:

- **Response Datetime:** The date and time the response was sent
- **Mentor:** categorical variable representing the sender (Mentor or Mentee)
- **Response:** A string representing the response itself.

Using Figures 1 and 2 as a reference, we see that conversations had two different formats, they both had the same order being Mentor/Mentee, Response Datetime, and Response. Some significant differences were that they had different strings between the different components and different formats for the Response Datetime, with one format excluding the year of the response. These formats were used to develop a RegEx required to split the conversations into individual responses and maintain the data for each response. This splitting was done for each of the "Posts In" columns and the category for each response was set to its original column name to maintain the category data. This was done for all the excel and document files and they were combined to produce the format seen in table 1.

3.2 Exploratory Data Analysis

Table 2 summarizes the descriptive statistics for a dataset that contains information on the number of responses, relationship duration, and word

count. The mean, maximum, and standard deviation values are provided for each variable.

	Number of Responses	**Relationship duration (days)**	**Word Count**
Mean	20	101	121
Max	70	189	2095
SD	17	43	159

Table 2: Summary of Descriptive Statistics from the 'Master' dataset containing conversations from 2022-2024

3.3 Natural Language Processing Tools

We employ Natural Language Processing (NLP) techniques on conversational text data to analyze mentorship interactions. We focus on methodologies: Emotion Detection, Topic Modeling, and Abstractive Summarization. These techniques allowed us to uncover hidden insights from vast amounts of conversations, giving us a glimpse into how these relationships can evolve.

Text Vectorization is a major aspect of natural language processing, this method was used in all our models to represent the textual information in a numerical form (vectors), we use BERT (Bidirectional Encoder Representations from Transformers) to generate embeddings. This produces contextualized word embeddings based on the context in which words appear, giving the numbers meaning.

Four main Natural Processing techniques were explored in this project: Topic Modelling, Mood detection, Abstractive Summaries and Recommender System. While other techniques were also explored throughout this project, their results are not sufficient enough to be taken into consideration.

3.3.1 Topic Modeling

Topic Modelling is the process of clustering documents together based on similar themes and words contained in the documents. Topic labels can also be extracted from these clusters using the common words within them. We explored 2 different models, Latent Dirichlet Allocation (LDA)[7] and BERTopic[3], but decided to focus on BERTopic since it had minimal preprocessing compared to LDA and was more modular allowing for more customization and optimization.

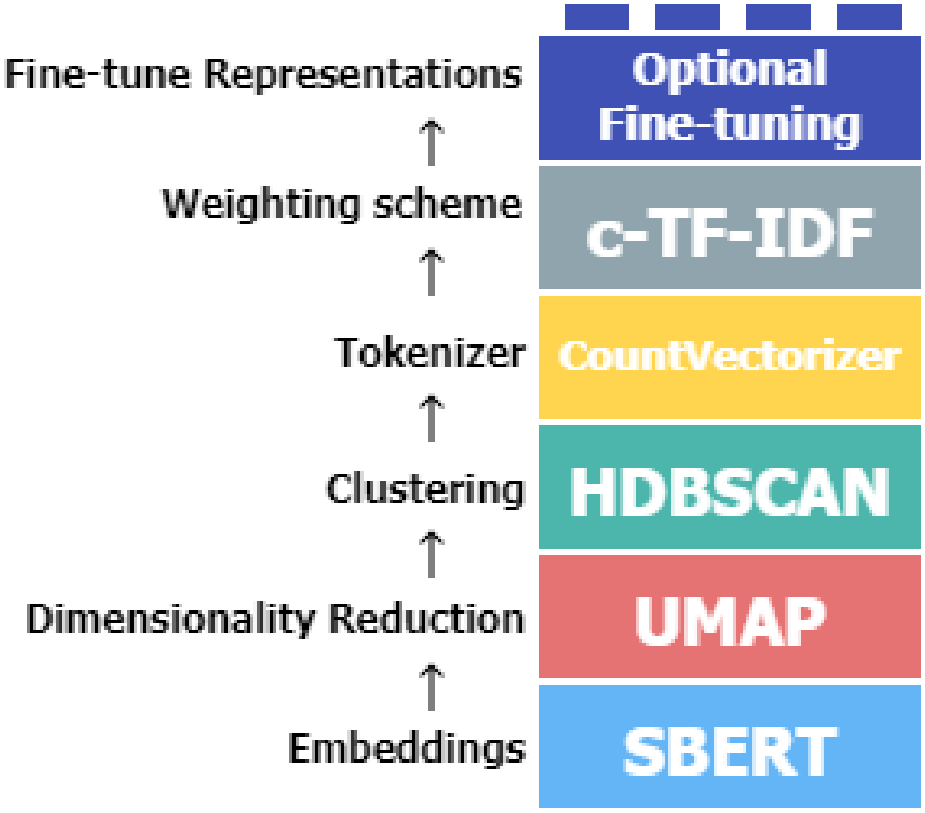


Figure 3: Bertopic Steps for topic modelling[3]

Since BERTopic uses Bidirectional Encoder Representation from Transformers (BERT) embeddings, stop words should not be removed from the documents before performing topic analysis with BERTopic. BERTopic has very little preprocessing, allowing you to feed your documents directly into the model, allowing it to produce embeddings, tokenize and cluster all within the fitting stage of the model.

Each stage of the BERTopic pipeline can be modified with different models that produce the same class of results (eg. K-means instead of HDBSCAN for clustering) and tuned individually to improve the output models. Since the main focus of this project was to determine Mentee engagement, the topic models were produced from Mentee responses.

3.3.2 Mood Detection

Mood detection is an NLP task that involves identifying the emotional tone or sentiment expressed in a piece of text. It aims to classify text into different emotional categories, such as happiness, sadness, anger, fear, etc. This is often a more nuanced version of sentiment analysis, which typically categorizes text as positive, negative, or neutral. This was used in an attempt to understand the Mentees' emotional attitude towards the program, and overtime.

For this task, we used the zero-shot classification approach, a technique that allows us to make predictions without needing to train the model on specific examples. This was particularly useful because our data was unlabeled, meaning we didn't have predefined categories to guide the model. The model we used is "facebook/bart-large-mnli" [1] from Hugging Face.[5]

For prediction, the model takes the text dataset and potential mood categories we define, such as "happy," "nervous," "frustrated," or "motivated" etc, and assess the likelihood each text fits each of these categories. For each piece of text, the model assigns a score to each mood category, indicating how closely the text matches that emotion. These scores reflect the model's confidence in each prediction.

The model's predictions were then aggregated to analyze emotions both for individual Mentees over time and across all Mentees collectively.

3.3.3 Abstractive Summaries

Abstractive Summarization models can take in huge volumes of text, and then condense and rephrase it to generate concise key points about the text. We use the Pegasus model, an advanced neural network architecture specifically designed for abstractive summarization tasks. In the pre-training stage of this model, it uses Gap Sentences Generation (GSG) [10], essentially, it masks 10-15 percent of words and predicts these masked works, resulting in 'self-supervised' learning as opposed to typical reinforcement learning. To shorten the time researchers spend reaching through text conversations, this summarizer allows them to select specific rows to analyze at a time. They can also paste an entire paragraph for it to output a summarized version.

As well, they can upload paragraphs of unlimited length into the site. The model, which is pre-trained on google/multi-news which is a dataset of

56k pairs of new articles and their human-written summaries which are taken from the site newser.com.

The pre-training corpus is for the model to understand and learn a diverse range of language patterns and contexts Pegasus was pre-trained on C4, which is a cleaned version of Common Crawl – text from 350 million webpages, as well as HugeNews – a dataset of 1.5 billion news/ news-like articles.

3.3.4 Mentor-Mentee Matching

To automate the process of matching Mentors and Mentees based on their backgrounds and profiles, we developed a recommendation system that uses K-Nearest Neighbors (KNN) algorithm to match similar Mentors and Mentees. This automatically pairs Mentees with Mentors with similar goals and backgrounds. First we used BERT[2] to vectorize the Mentor and Mentee profiles to generate numerical embeddings of the text, which also captures its semantic meaning. Before vectorizing, we got rid of numerical and irrelevant columns that created conflicting results, we then added more weight to specific columns such as "Career Interests" and "Hobbies" which would amplify the students' answers to these when calculating similarity. We also normalized the vectors to reduce them to unit length, however we should note that this does not change its direction, only the magnitude, and this transforms the values to a similar scale. Then we calculate the cosine similarity between the vectorized profiles, giving us a value between -1 and 1, representing its similarity. The scores then allow us to use KNN to find the most similar Mentors for each Mentee. The KNN algorithm identifies the top 5 Mentors whose vectorized profiles are closest to the Mentee's vector. We selected our k to be 5. The output is an downloaded into a csv file with all the original profile ID on the left column, and its 5 matches plus their IDs on the right. After generating the initial list of recommended Mentor-Mentee pairs, Juliet the program coordinator will review these recommendations. This human oversight ensures that the recommendations align with her qualitative input.

3.4 Application Deployment

A goal of ours was to make a pipeline so that ReMBC could clean and process their data in the future, and this would require us to make a simple application that could do this all for them. Building a web application required free deployment and hosting, Streamlit was a great candidate as it was easy to use and was free.[12]

One struggle was the maintenance of the site, so we've decided to employ the use of containers. Containerizing the app allows it to perform consistently, regardless of the computer system it runs on. This makes sure that all dependencies are the same as they are all in one packaged unit within our container. This required us to make a devcontainer.json file which tells VSCode how the container will be accessed, instructing how specific tools, libraries, or runtimes are used for working with the codebase. In this file, it instructs the container to refer to the requirements.txt file, which is a file we made with all the libraries and modules this application uses.

4 Results

In this section, we will present the results of our research. Given our unsupervised learning approaches, evaluation metrics are pretty subjective or conducted on a small data points sample.

4.1 Topic Modeling

4.1.1 Topics from Mentees

Topic Number	Top 3 Representative Words	Interpretation
0	careers, college, university	Career and Post-Secondary Exploration
1	rural, towns, community	Rural to Urban Living
2	health, selfcare, stress	Wellness and Self-care
3	studying, memorize, practice	Developing Good Study Habits
4	mentoring, talk, helping n	mentorship Conversations
5	spring, holidays, break	Free Time
6	inspiration, passion, creation	Finding Inspiration
7	hobbies, sports, skiing	Hobbies
8	login, messages, app	Online Platform
9	careers, pursuing, profession	Career Exploration

Table 3: Top 10 Mentee Topics Produced by BERTopic

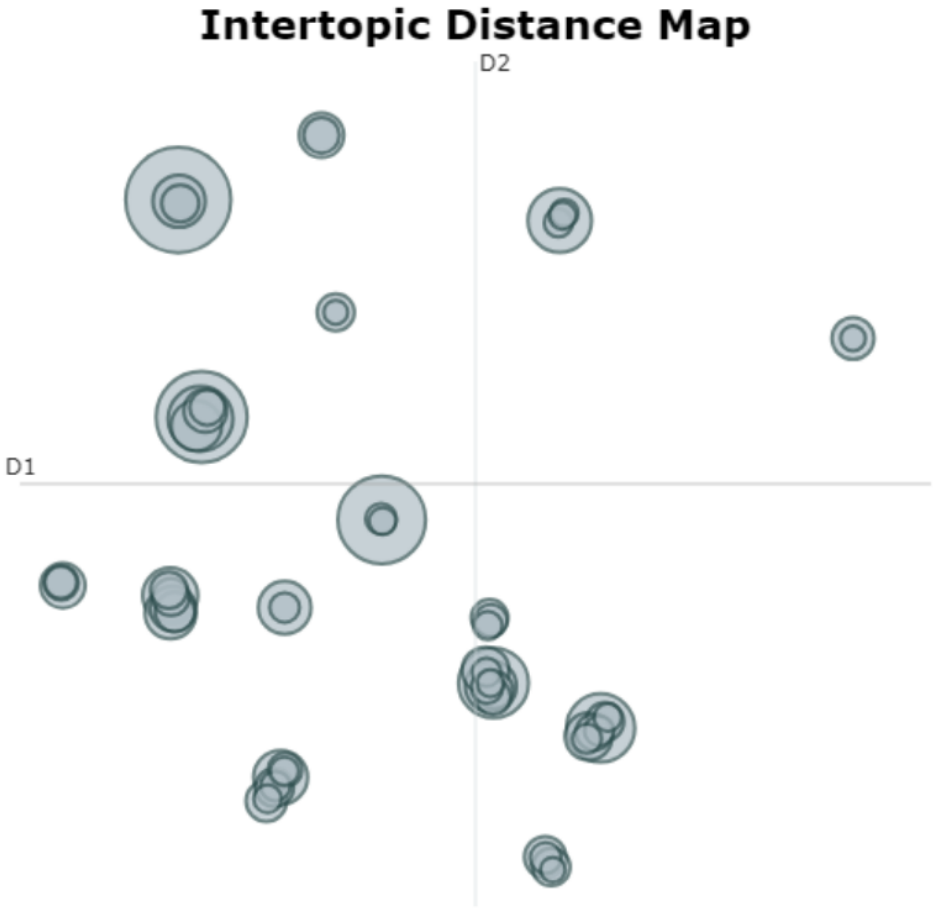


Figure 4: Topic clusters from Mentee responses:
each dot represents a topic projected onto a 2D plane. Larger circles represent topic with more diverse word representations, while smaller circles represent topics with more concise topic word representations.

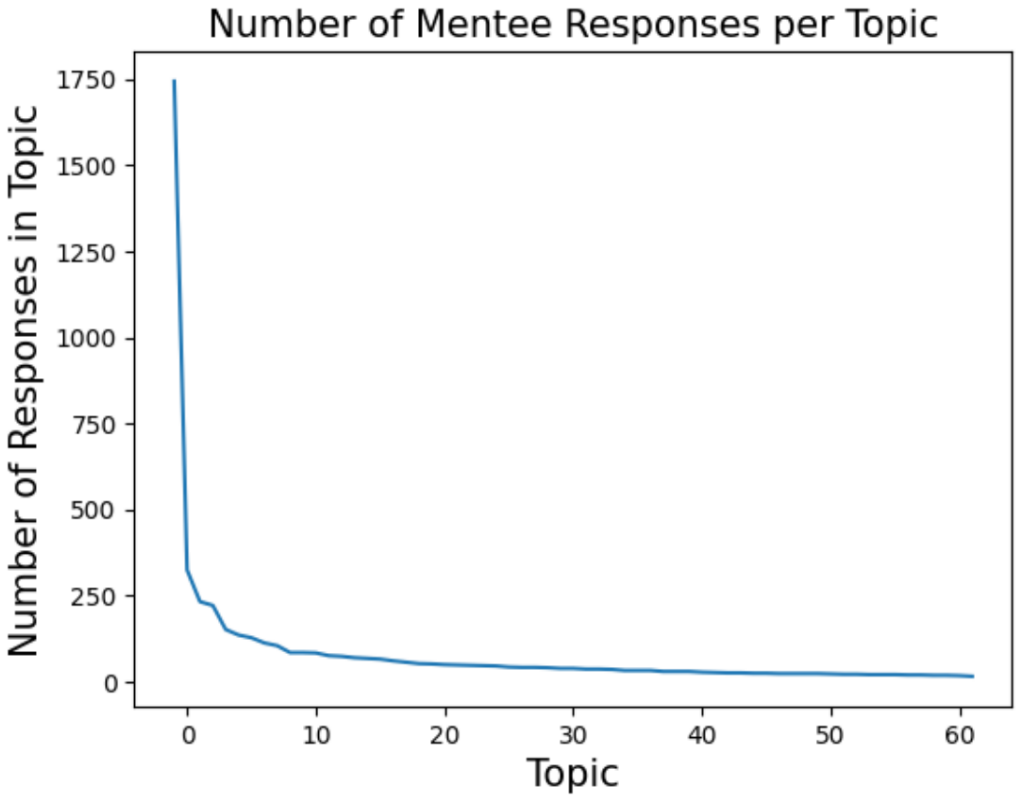


Figure 5: Topic Size vs Topic Number

Through BERTopic we were able to determine the top 10 topics from Mentee Responses (**figure 3**). The topic labels are human made, while the representative words are assigned by the model based on the probability of finding those words in documents assigned to that topic.

Six of the ten topics come directly from the curriculum units, namely Wellness and Self Care, Post-secondary Exploration, Career Exploration, Developing Good Study Habits and Finding Inspiration. We do expect most of the topics to be related to the curriculum units since these are used to stimulate conversations between Mentor-Mentee pairs. These topics give an indication that our model is accurate, since we expect many of the topics to be related to the curriculum material. Although students are prompted by Mentors and the curriculum to talk about these topics, they are not required to spend a specific amount of time on them, suggesting that these top 6 subjects are the units that Mentees are engaging with the most and can be used as a guideline for other units that ReMBC could use to develop additional curriculum units.

Two of the topics were related to online mentorship, one being about Mentee gratitude towards their Mentors and the mentorship relationship and the other being about the eMentoring platform and potential Mentee comments. We expect some of the conversations to relate to the nature of the relationship themselves and these conversations could provide insight into what Mentors do to engage Mentees in successful relationships and how the mentoring platform could be improved.

The last general category for these topics are the off-topic subjects that are less predictable. Two of these topics are on holiday plans and hobbies. The representative words suggest that these topics have much more personalized answers and that the Mentees are opening up to their Mentors, indicating positive relationship development and could provide signs of a successful relationship developing. These topics could provide a suggestion for new curriculum units for ReMBC to develop to help improve Mentee engagement earlier in their relationship.

4.1.2 Topics Within Categories

By splitting the responses according to their original categories, we were able to produce topics from within the categories. This did produce some challenges due to class imbalances, many categories had few or no responses in them, which makes it difficult to produce meaningful topics. Many Mentors

and Mentees would have conversations in the "Start Here" category, instead of specific subject chats, resulting in some really small topics, essentially having only one or no response per relationship. Generally the top topic from each subject had representative words that were expected, such as studying containing memorize, studying, school, but one notable exception was career and post secondary exploration, which had a topic containing words such as holiday and travel as the first topic.

4.1.3 Using Predefined Labels to Cluster Responses

BERTopic allows you to give your topics human readable labels, this is done by either feeding BERTopic predefined topic labels into which the model will try to cluster the documents into, or feeding the representative word list into a LLM to generate human readable labels.

When given predefined topic labels, the size of those topics were orders of magnitude smaller than the topics without the predefined labels, making it difficult to draw meaningful connections. It is possible that better engineering of the labels could produce better clustering and is a future direction for this project.

By changing the representative model used to fit the model (last step in **Figure 3**) you can use AI models such as ChatGPT to generate topic labels from the representitive words. However, this proved difficult due to some topics having very small cluster sizes, making it difficult for the LLM to create labels. Further tuning of our model to only include large clusters could allow this procedure to be more useful in the future.

4.2 Mood Detection

Figure 6 shows the Mood Detection Tool's analysis of eight successful mentorship relationships over time, represented by the average emotion scores of among all Mentees throughout the program. Key points include:

- Excitement (green line) and contentment (orange line) remain approximately constant, suggesting consistent engagement.
- Boredom (red line) is consistently low, indicating that Mentees are generally stimulated throughout the program.
- Notably, nervousness (blue line) remains continuously high, which may prompt the ReMBC team to explore the reasons behind this ongoing anxiety among Mentees

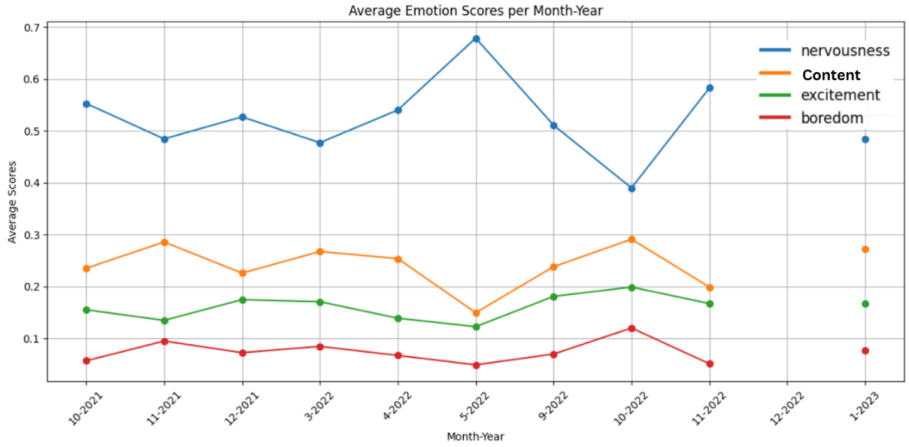


Figure 6: Mood detection results on all Mentees across the years 2021-2022

We can also analyze individual Mentee experiences. For example, **Figures 7 and 8** compare two Mentees over the duration of their respective mentorships. In **Figure 7**, the first Mentee's excitement (green line) decreases slightly, while contentment (orange line) shows a modest increase. In contrast, **Figure 8** shows that although the second Mentee's contentment (orange line) rises significantly over time, their excitement (green line) drops sharply.

These observations suggest that Mentee 2 may require additional support. Mood detection could offer valuable insights into how Mentor-Mentee relationships evolve and how ReMBC can better facilitate this development.

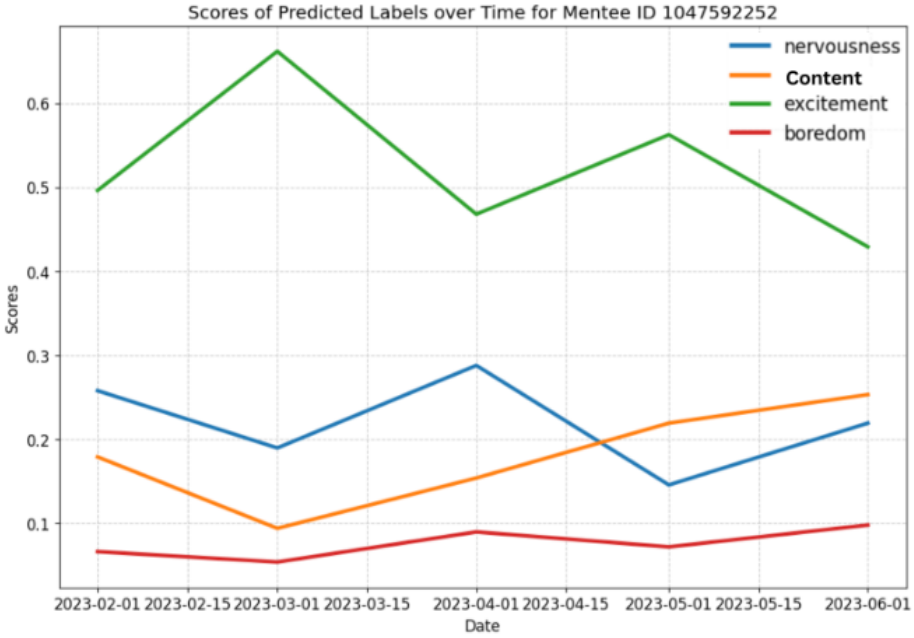


Figure 7: Enter Caption

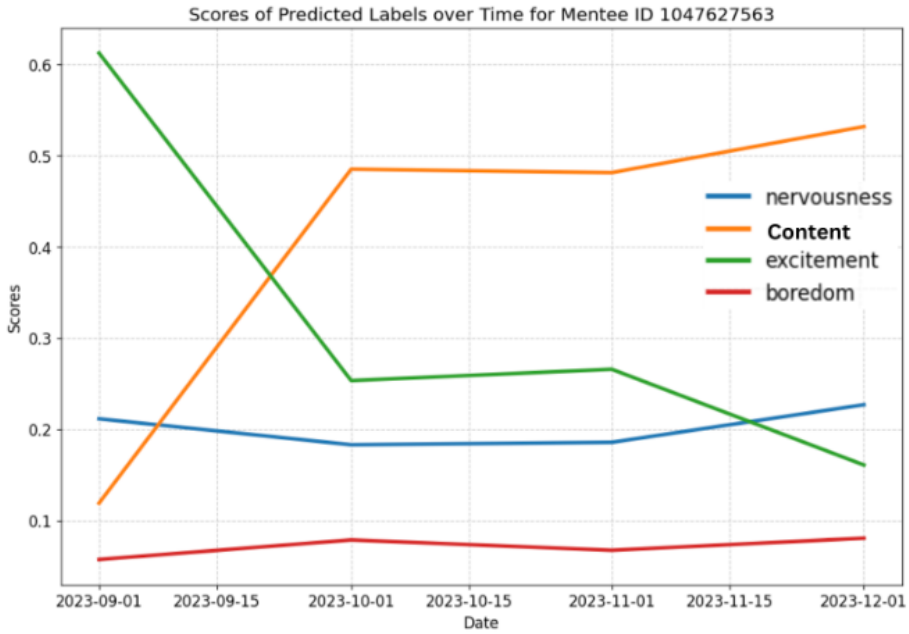


Figure 8: Enter Caption

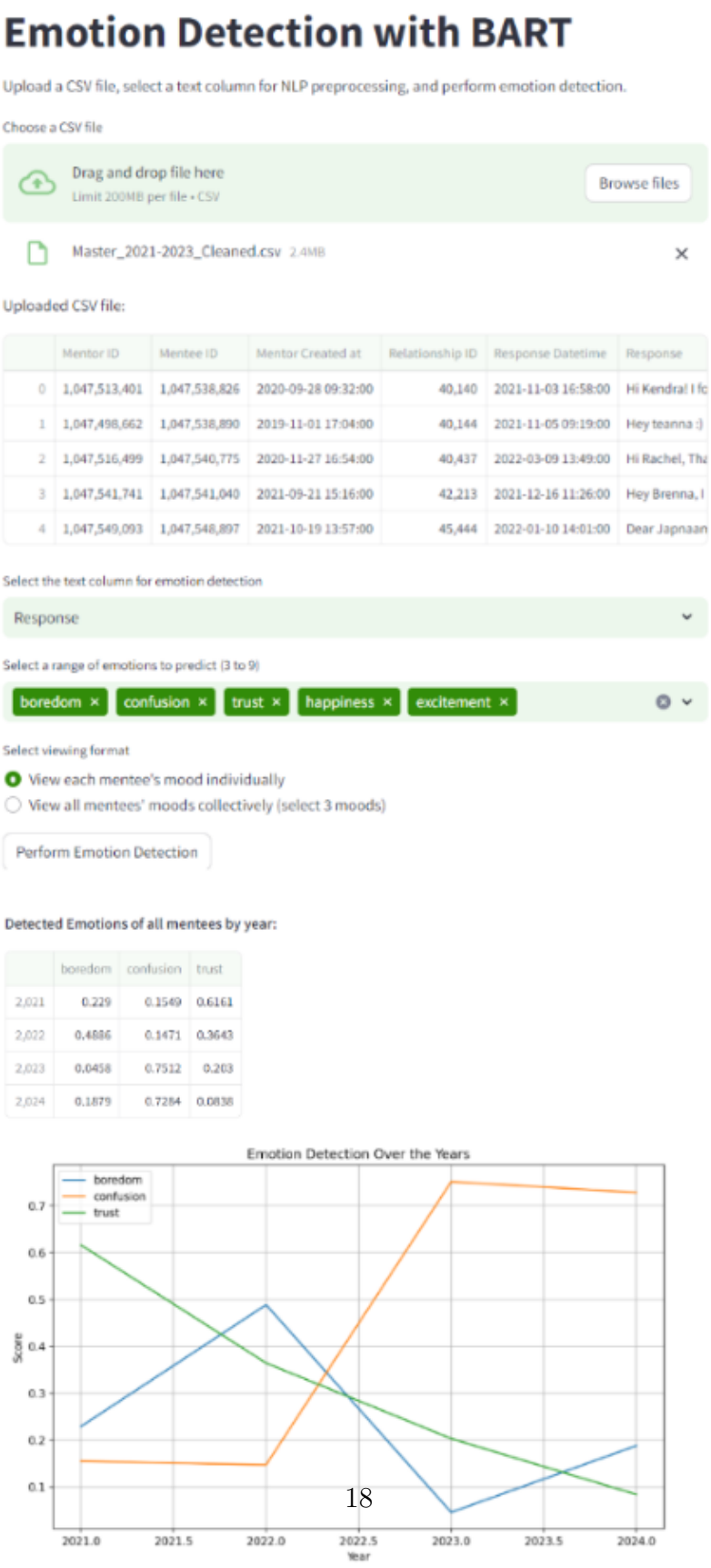


Figure 9: table and graph of the moods of all users. Combines the text of that entire year and then performs mood detection on the collective text.

	Mentee ID	Relationship ID	Month	Year	Scores	Predicted Labels
0	1,047,538,812	40,139	9	2,021	0.9917650818824768 0.007824769243597984 0.0004101244849152863	boredom confusion trust
1	1,047,538,814	40,516	9	2,021	0.6917086839675903 0.29578399658203125 0.012507263571023941	confusion boredom trust
3	1,047,538,816	40,138	9	2,021	0.7880495190620422 0.1418982297182083 0.07005226612091064	confusion boredom trust
5	1,047,538,822	40,133	9	2,021	0.8401774168014526 0.14668583869934082 0.013136738911271095	confusion boredom trust
6	1,047,538,823	40,135	9	2,021	0.9586277008056641 0.03478942811489105 0.00658293766900897	confusion boredom trust
8	1,047,538,824	40,143	9	2,021	0.9951383471488953 0.0037733460776507854 0.001088273362256586	confusion trust boredom
10	1,047,538,825	40,141	9	2,021	0.9947504997253418 0.0038456108886748552 0.001403880771249532	confusion trust boredom
11	1,047,538,826	40,140	9	2,021	0.9551961421966553 0.025120733305811882 0.019683126360177994	confusion boredom trust
14	1,047,538,827	40,142	9	2,021	0.9025152325630188 0.09122481942176819 0.006259890738874674	confusion boredom trust
18	1,047,538,891	40,137	9	2,021	0.8709954023361206 0.10864584147930145 0.020358772948384285	confusion boredom trust
19	1,047,540,772	40,168	9	2,021	0.8103339672088623 0.18076880276203156 0.008897243067622185	confusion boredom trust
24	1,047,540,777	40,136	9	2,021	0.9941662549972534 0.004713636822998524 0.001120130647905171	confusion boredom trust
25	1,047,540,957	40,440	9	2,021	0.9123660326004028 0.08588754385709763 0.0017463797703385353	boredom confusion trust
27	1,047,540,958	40,350	9	2,021	0.9948634505271912 0.003355069551616907 0.0017815037863329053	confusion trust boredom
30	1,047,540,960	40,359	9	2,021	0.6778873801231384 0.31871503591537476 0.0033975925762206316	confusion boredom trust
33	1,047,540,961	40,345	9	2,021	0.8577352166175842 0.12426720559597015 0.01799757406115532	confusion trust boredom

Figure 10: Top emotions per individual eMentoring user

4.3 Abstractive Summaries

In **figure 11** is an example output from our app. We have copy and pasted a long description of the Solarized Theme from VSCode extension, which is a multiparagraph description of how the theme works and how it was made. The summary output is around 1/4th the length of the input, and sounds relatively human.

4.4 Mentor-Mentee Matching

A screenshot from the output of our app, **figure 12**, saved and opened as a .csv file. We are currently working to filter the nearest neighbours so that Mentors will only be matched with Mentees, and vice versa. The 'relationship role' column displays the role of ID. The 'nearest neighbours' column displays the 5 nearest neighbours to ID in the format [('ID', Similarity Score, 'Relationship Role of Neighbour') . . .] as shown in **figure 13**.

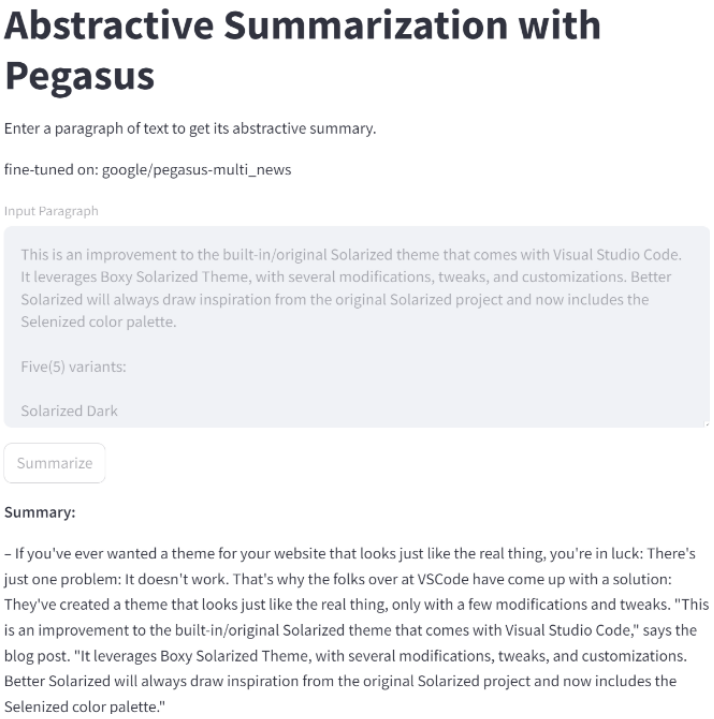


Figure 11: Screenshot of Abstractive Summarization page from our app

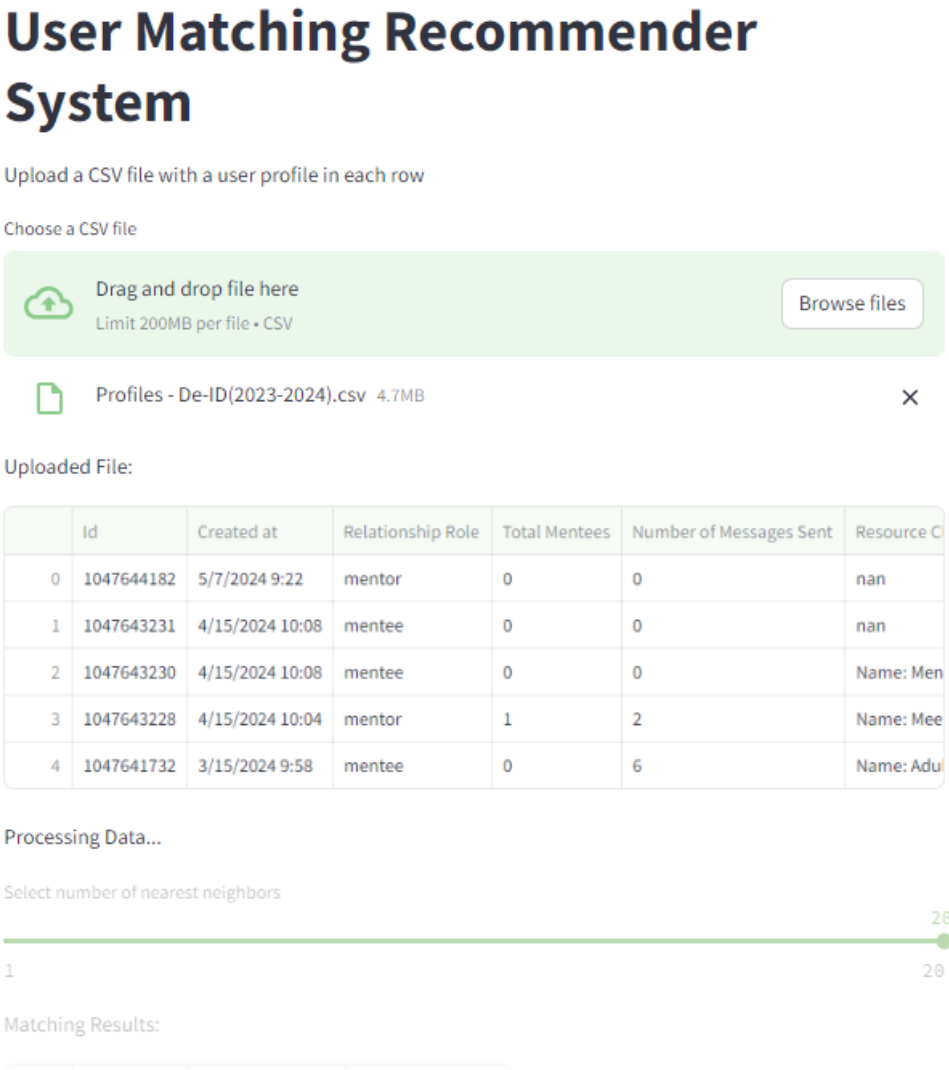


Figure 12: Mentor Mentee Matching from our app

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Id	Relationsh	Nearest Neighbors										
2	1.05E+09	mentor	[('1047632515', 0.84, 'mentee'), ('1047640865', 0.83, 'mentee'), ('1047639743', 0.82, 'mentee'), ('1047593961', 0.82, 'mentee'										
3	1.05E+09	mentor	[('1047627550', 0.8, 'mentee'), ('1047592265', 0.77, 'mentee'), ('1047595006', 0.77, 'mentee')]										
4	1.05E+09	mentee	[('1047585105', 0.77, 'mentor'), ('1047596809', 0.63, 'mentor'), ('1047631785', 0.59, 'mentor'), ('1047587030', 0.58, 'mentor')]										
5	1.05E+09	mentor	[('1047627550', 0.84, 'mentee'), ('1047594497', 0.83, 'mentee'), ('1047585791', 0.82, 'mentee'), ('1047634155', 0.82, 'mentee'										
6	1.05E+09	mentor	[('1047640494', 0.83, 'mentee'), ('1047592066', 0.83, 'mentee'), ('1047585094', 0.82, 'mentee'), ('1047594497', 0.8, 'mentee'),										
7	1.05E+09	mentee	[('1047633458', 0.6, 'mentor'), ('1047629229', 0.59, 'mentor'), ('1047630611', 0.58, 'mentor')]										
8	1.05E+09	mentor	[('1047593958', 0.98, 'mentee'), ('1047594026', 0.96, 'mentee'), ('1047592604', 0.95, 'mentee'), ('1047584583', 0.95, 'mentee'										

Figure 13: Mentor Mentee matching results when downloaded

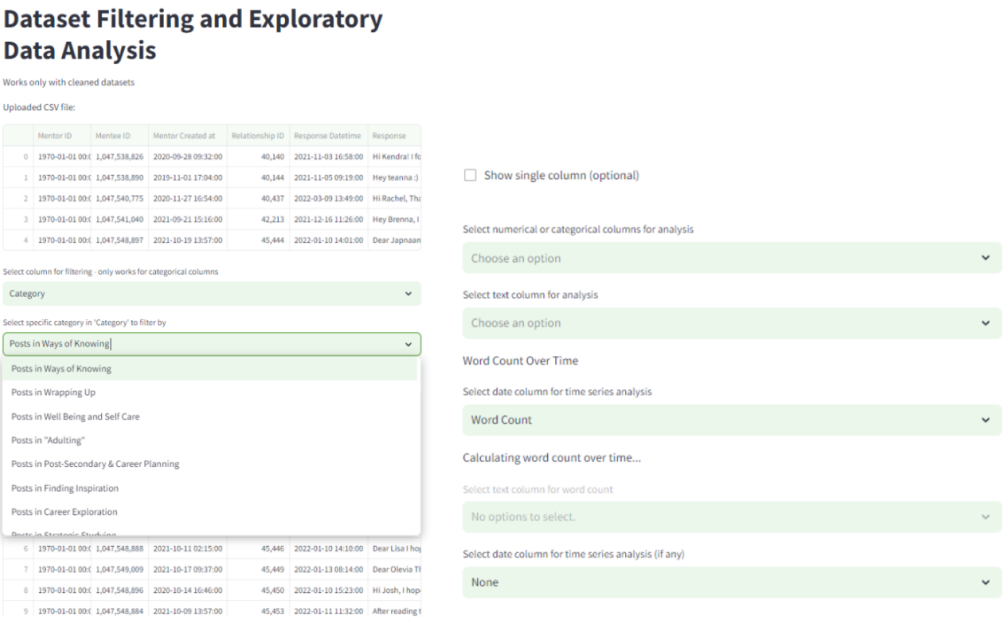


Figure 14: A snapshot of our filtering and Analysis page

4.5 Cleaning and Filtering

Figure 14 shows the filtering page, allowing users to filter through their dataset to view specific columns, or (sub)categories within columns. Each time it is filtered, the new filtered dataset can be downloaded, providing a much smaller file. Other smaller tools include searching for response rows with over x (input) number of words. There are visualization functions included, options include a bar chart, heatmap, boxplot, histogram, and scatter matrix. An example for the analysis is – visualizing the word count over time in a bar or line graph. The cleaning page allows ReMBC to upload uncleaned datasets in different file types, and converts them into cleaned CSVs that can then be used in other other pages. For example, datasets with many responses in one cell will be outputted as a CSV where each response is in a different row.

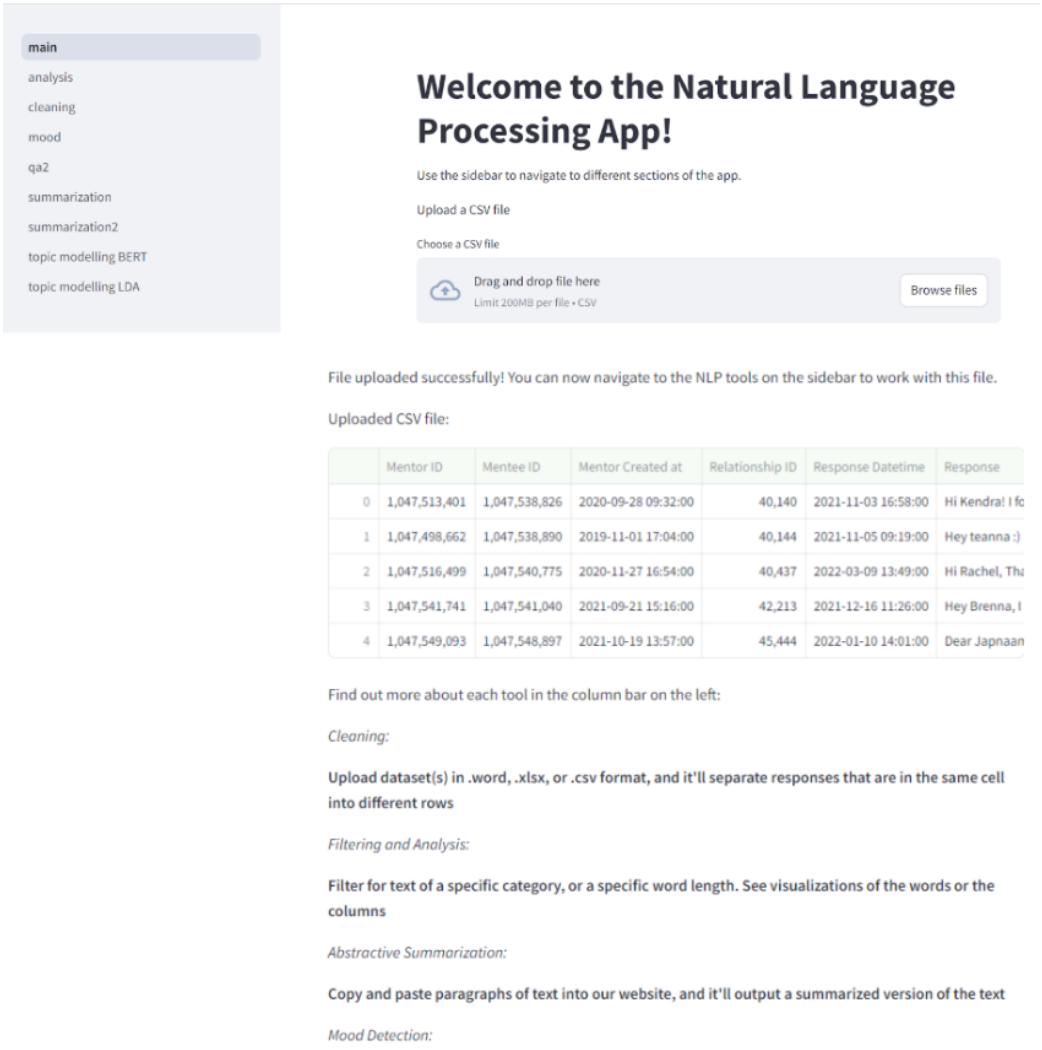


Figure 15: Snapshot of the app interface

4.6 Application Development

Finally, we have developed a website with the models and tools we described above, these models are located in different pages on the side bar to the left. A screenshot of our app interface is in **figure 15**. This site can also be used by researchers in non-tech domains, so they can experiment with natural language processing tools for text analysis. It is deliberately easy to use and user-friendly, with instructions on how each tool can be used. It requires users to upload a CSV file, one which contains a column(s) of text, then they must select the text column for further natural language processing. Our app can be accessed at https://nlp-tool.streamlit.app/.

5 Discussion

5.1 Contributions to Social Good

As a project of the 2024 Data Science for Social Good Fellowship, the primary goal of our work is to contribute to social good. Specifically to this project, the overarching objective is to better support students from Indigenous and rural communities. Natural Processing Language (NLP) models can understand the intricacies of text data, thus providing meaningful key insights.

The web Application that will be launched as a result of our project will predominantly benefit Rural eMentoring BC, our partner organization, allowing them to make use of NLP tools without the required technical expertise.

5.2 Topic Modeling

Using topic modeling we see that most of the topics produced are from within the curriculum, which are expected and provide a qualitative metric for the success of the model. Even more significantly, some Mentees are engaging with more personal topics that are outside of the curriculum. The presence of these topics can be used as an indication that Mentees are becoming comfortable with their Mentors. Delving further into the conversations of the relationships with lots of topics outside of the curriculum - particularly more personal ones - could allow us to see what those Mentors are doing to develop relationships and inspire Mentees to open up to them. These relationships could provide guidance into ways to improve Mentor training.

5.3 Mood Detection

Mood detection in Mentor-Mentee conversations reveals emotional trends that can prove beneficial to ReMBC by enabling them to monitor the Mentees' attitude towards the program and offer personalized support when necessary. Our analysis over months and years provides key insights into the development of the Mentor-Mentee relationships, allowing ReMBC to adjust its practices to better support these students.

5.4 Abstractive Summaries

This technology can help researchers and program leaders obtain quick summaries while maintaining the original text's meaning, making it easier to monitor and evaluate Mentor-Mentee interactions. Users are able to quickly grasp the essence of Mentor-Mentee relationships without reading their entire conversations.

5.5 Mentor-Mentee Matching System

This automated Mentor-Mentee matching system greatly enhances the efficiency of the matching process. It becomes quicker and more accurate, leading to better-aligned Mentor-Mentee pairs. It also opens the door for the program leaders to focus on the quality of those recommended pairs and get some insights into what features or criteria to focus on while matching.

6 Limitations

In our study, we identified two primary categories of limitations: those related to the models used and those related to the dataset.

6.1 Modeling Limitations

One of the key challenges we encountered with mood detection was evaluating the model's performance. Since our approach relied on zero-shot classification applied to unlabeled data, we faced difficulties in accurately assessing the model's effectiveness. To approximate evaluation, we manually labeled a small sample of 50 data points and achieved an 80 percent accuracy rate. However, this method is subject to potential biases, as the definition and interpretation of emotions can vary significantly between and even within individual labellers.

Similarly, this applied to our other models as well, topic modelling, KNN recommender systems, and summarization tasks – are all unsupervised machine learning methods. Limitations such as evaluation challenges, as performance metrics were difficult to automate with a script when the data is unlabelled. Oftentimes, metrics such as F1, accuracy, and word error rate, are used to fine-tune the models, however, we'd have to manually label each conversation with its success, mood, or the outcome we want to predict in order to get these measurements. Using qualitative metrics by continuously testing our model with different datasets was our alternative, as well as confirming our model findings with our partner, Juliet.

Using pre-trained models, such as pegasus on multi-news, provided some assurance that the model has been initially trained before being used by us. As well, with unsupervised methods, the risk of overfitting is always present, it may identify patterns in noise as significant, as it does not have a feedback mechanism to account for this risk.

6.2 Dataset Limitations

A major limitation in our research was the use of unlabeled data. Our objective was to identify the strategies employed by successful Mentor-Mentee pairs to sustain their relationships, with the intention of using these insights to enhance future Mentor-Mentee interactions in the ReMBC program. However, the dataset we initially worked with did not include labels indicating which relationships were successful and which were not. While we eventually obtained examples of successful pairs, the number of these examples was too limited to allow for conclusive analysis of the key characteristics of successful relationships.

A priority of ours was keep costs minimal, so the maintenance and use of this application wouldn't incur fees over time. We ended up using only open source tools, including the deployment of this app. Hosting on Streamlit is free, but the consequence is major memory and computational limitations. Often, the app would encounter an error on the deployed site, but would work fine when run locally. This error messages usually resolves itself and disappears after a few hours. We assume this is a memory leak issue or a constraint of using Streamlit.

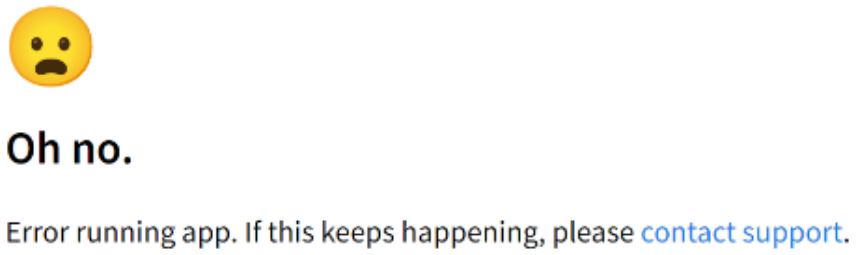


Figure 16: Error Message

7 Recommendations and Future Directions

To further enhance the insights derived from our study and to improve the applicability of the results, we propose several potential directions for future research.

- **Labeling and Classifying Relationships:** A promising area for future work can evolve from developing a spectrum for classifying Mentor-Mentee relationships. Instead of a binary classification (e.g., successful vs. unsuccessful), researchers could explore a broader spectrum that includes categories such as "great engagement," "good engagement," and "average engagement." This expanded classification would lead to the application of more machine learning techniques, particularly in the realm of supervised learning.

- **Exploring Tone Detection Among Mentors:** Another fruitful direction for future research can be found within the tone and writing style in Mentor communications. Specifically, investigating how the tone and writing style of Mentors affect Mentee engagement could provide relevant insights into what makes successful relationships. This information could be instrumental in improving the training of Mentors, equipping them with the tools and insights needed to foster more effective and successful Mentor-Mentee relationships.

- **Combining Mood Detection With Topic Modeling:** by determining the moods and topics that each response is labeled as, we could determine how Mentees are responding to different topics from an emotional perspective. ReMBC's goal is to improve Mentee's sense of comfort with their Mentors, so looking at topics that improve this

would help them introduce improve existing subjects that Mentees are responding negatively to and find new subjects that Mentees are responding positively to.

8 Conclusion

In conclusion, this project has laid a foundational framework for Rural eMentoring BC (ReMBC) and other e-mentoring programs aiming to better support students from underprivileged and rural communities. Through the application of advanced Natural Language Processing (NLP) techniques, we have successfully developed tools that provide valuable insights into Mentor-Mentee interactions. These tools, including topic modeling, mood detection, and a Mentor-Mentee matching system, are now accessible through a user-friendly web application, making sophisticated data analysis available to all users.

These tools are designed to complement and supplement, rather than replace the human oversight essential in managing ReMBC, offering additional perspectives that can enhance the quality of support provided. Looking ahead, our hope is that the results and insights from this project will not only benefit ReMBC but also inspire similar initiatives to adopt data-driven approaches in supporting students from rural and Indigenous communities. Possibly in 2 or 3 years from now, we can see more students from Indigenous and rural areas venture more into post-secondary careers.

9 References

References

[1] *BART Transformer*. URL: https://huggingface.co/docs/transformers/en/model_doc/bart. (accessed: 07.01.2024).

[2] *Bert Transformer*. URL: https://huggingface.co/docs/transformers/en/model_doc/bert. (accessed: 06.01.2024.

[3] *BERTopic: The Algorithm*. URL: https://maartengr.github.io/BERTopic/algorithm/algorithm.html. (accessed: 05.06.2024).

[4] Halseth G. Hanlon N. "The greying of resource communities in northern British Columbia: Implications for health care delivery in already-underserviced communities." In: *Candian Geographer* 49.1 (2005), pp. 1–24.

[5] *Hugging Face*. URL: https://huggingface.co/models?other=emotion-detection. (accessed: 05.25.2024).

[6] *Indigenous Health Improves But Health Status Gap With Other British Columbians Widens*. URL: https://www.fnha.ca/about/news-and-events/news/indigenous-health-improves-but-health-status-gap-with-other-british-columbians-widens. (accessed: 07.10.2024).

[7] *Latent Dirichlet Allocation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html. (accessed: 05.25.2024).

[8] *Mentor City*. URL: https://www.mentorcity.com/. (accessed: 08.08.2024).

[9] *Pathways to Medicine*. URL: https://mdprogram.med.ubc.ca/admissions/before-you-apply/pathways-to-medicine/. (accessed: 07.10.2024).

[10] *Pegasus Gap Sentneces Generation*. URL: https://huggingface.co/docs/transformers/en/model_doc/pegasus. (accessed: 08.01.2024).

[11] *Rural eMentoring BC*. URL: https://ps-ementoringbc-2023.sites.olt.ubc.ca/our-mission/. (accessed: 06.15.2024).

[12] *Streamlit*. URL: https://docs.streamlit.io/. (accessed: 07.03.2024).

[13] *The Geography of BC*. URL: https://pressbooks.bccampus.ca/ccedarrproject/chapter/the-geography-of-british-columbia/#:~:text=As%20of%202021%2C%20about%20half,population%20resides%20in%20rural%20areas.&text=A%20significant%20portion%20of%20rural,BC%20for%20over%2010%2C000%20years.. (accessed: 07.09.2024).