Natural Language Processing of Applications of Pipeline Projects

Xinya Gao, Zhengzhi Liang, Rachel Lobbay, Sidney Lu

Supervisor: Raymond Ng

UBC Data Science Institute

The University of British Columbia

Canada Energy Regulator

Abstract

From 2002 to present, approximately 11,000 letters have been submitted to the Canada Energy Regulator (CER) from parties and individuals who expressed their comments and concerns regarding large pipeline applications. These letters of comment are important because they are how people share their views on the pipeline projects to the institution that adjudicates the applications for these projects. Since reviewing and determining the content of letters is done manually at present, in this project we developed and applied Natural Language Processing (NLP) techniques to help to streamline the letter analysis through employing automation. Our goals were to utilize NLP techniques to extract what people said and how they felt about three major pipeline applications. To obtain what people said, we utilized topic modelling to obtain major topics and subtopics that were discussed in the letters. Text summarization was also implemented to obtain the key points from each letter. Sentiment analysis and emotion analysis techniques were utilized to ascertain how the people felt and the primary emotions present in their letters. Our work on this project resulted in visualization dashboards for topic modelling and for sentiment analysis. These powerful tools not only help to communicate our results, but they also facilitate collaboration on the research and development of tools to better extract the core of what people say and how they feel from their letters of comment.

Acknowledgements

We would like to thank the public who submitted Letter of Comment for pipeline applications. It is important for the project to have insights and diversities from all communities. We hope our analysis and tool can help Canada Energy Regular to have better abilities to understand public's opinions and better engage with the communities in the future.

We would like to give our special thank to Canada Energy Regulator (CER), who is our partner organization and the University of British Columbia (UBC), Data Science Institute's (DSI) Data Science for Social Good (DSSG) program to offer us the opportunity to work on the project.

Specifically, we would like to extend our appreciation to Sousan Yazdi, who is the data scientist from CER for her assistance and help on this project. We are grateful for the guidance and support provided by Dr. Raymond Ng and Dr. Kevin Lin at the UBC Data Science Institute, especially during the special circumstance of the COVID - 19 situation.

Contents

List of Figures

1 Introduction

The Canada Energy Regulator (CER) is a quasi-judicial institution that adjudicates applications for pipelines that cross inter-provincial or international borders. The CER regulates pipelines, power lines, energy development, and trade in the Canadian public interest and makes decisions on whether pipeline applications get approved or rejected.

While an application is being adjudicated, other parties and individuals can submit letter of comments to voice their interest and concerns to the CER. Examples of parties that have commented on past applications are shippers, other energy companies, Indigenous peoples, landowners, and non-governmental organizations. There are approximately 11,000 letters, from 2002 to the present, for large pipeline applications (length greater than 40 km; e.g., Trans Mountain Expansion).

Due to a change in legislation allowing broader participation from interested parties, the CER expect to receive a higher volume of letters of comment moving forward. Currently, reviewing and determining the content of the letters is done manually. The expanded mandate to receive comments from any interested parties may make this manual process harder. Therefore, we are going to explore ways to automate the processing of letters of comment in this report. Our outcome can support the CER in processing a larger volume of letters.

In this project, we used Natural Language Processing (NLP) tools. We have used some latest development in Natural Language Processing including: Latent Dirichlet Allocation Topic Modelling, Sentiment Analysis, Emotion Analysis and Text Summarizing by TextRank Algorithm. The project is hoping to answer the following questions:

1. What did people say in essence and details in the past letter of comment?
2. How did people feel for when they wrote the letter?
3. How can Canada Energy Regular better engage with the public and submitters in the future?

2 Description of dataset

The dataset consists of 11,000 letters of comment from 2002-2020. Each letter of comment consists of the personal information of the submitter, submission date, the company or institution of the submitter, the pipeline project that the submitter is referring to and the submitter's comment for the specific pipeline project. These metadata provided are useful for us to extraction information and to group the letters of comment by the pipeline project.

Among all the letters we received, there were certain kinds of letters of comment that were not easily processed, and we decided to take these out due to time restrictions.

The unprocessable letters of comments include 1972 scanned PDFs, approximately 19.5% handwritten letters, letters contains tables or figures and approximately 12% of letters are rotated. There are other formats of letters which we decided to take out including a small amount of emails, brochures and website screenshots.

After the preliminary analysis of the dataset, we found there are three types of structured letter of comments which we can easily extract the information within the time restriction. Therefore, we decided to conduct our analysis based on these three structured types of letters of comment and continue analyzing the unstructured letter of comment for future development.

All the structured letter of comments consist of three big pipeline applications: Enbridge Northern Gateway Project, Trans Mountain Pipeline Expansion Project, Brunwick Pipeline Project. The total number of these three types of structured letters of comment is 4,877. So, our analysis is based on the 4,877 letters.

3 Topic Modelling and LDA

To address the question of "What did people say?", we performed topic modelling on past letters. Topic modelling is an unsupervised machine learning technique that clusters words and phrases together to form a topic, and returns the importance of each word/phrase to the topic. One major limitation of topic modelling is that the number of topics in the texts (which we call the hyperparameter k) cannot be automatically determined, and must be specified by the user prior to running the algorithm. The hyperparameter k is of high importance to the performance of the final model, and should be selected with great care. There are many topic modelling algorithms out there, including Latent Semantic Analysis (LSA) and Negative Matrix Factorization (NMF), but we chose to use Latent Dirichlet Allocation (LDA) for this project, as it seems to be the most common and effective approach for topic modelling.

Since we suspected that different projects would have different topics, we performed LDA separately on each of the projects. Prior to applying LDA, we did some preprocessing on our data. All text was converted to lowercase, punctuation and digits were removed, stop words and other common, unimportant pieces were removed, and all words were lemmatized. This data was then fed into the LDA algorithm.

As a first attempt, to choose the number of topics k, we ran the LDA algorithm with all k from 1 to 50, and recorded their topic coherence score for each model. Topic coherence scores attempt to quantify how much the words within a topic make sense with one another. However, we found that these scores were not particularly effective at doing what they were supposed to. In fact, the two most commonly used topic coherence scores, UMASS and CV, were contradictory to one another within our data. Due to this, we abandoned using topic coherence scores for our model evaluations. We then tried manually going through the words within the topics we got, and making subjective evaluations as to whether they made sense with one another. After reading through many of the letters, we chose a much narrower range of k to evaluate, and tried using this manual method to evaluate the models. However, we found this difficult to do with individual words within the topics, as a lot of context was missing.

To try and improve the situation, we tried using gensims Phrases API to create ngrams. An individual word is a unigram, while a phrase of three words is a trigram. We applied this tool to our data to generate ngrams, and then ran it through LDA once more. This was an improvement, however we found that very few bigrams and above actually appeared within the top important words to a topic, and so evaluating models was still the same. We then decided to filter out all unigrams, and stick with only bigrams and above.

To help us with evaluating the models, we used a handy visualization tool called LDAvis. An example of LDAvis is shown in Figure 3.1. Within the figure, the 3 bubbles that we see on the left represent topics, and the further they are from one another, the more they are dissimilar. Each bar on the right represents a word, and within those bars is a blue bar and a red bar. The blue bars represent the overall frequency of the terms within all the documents. When a topic is selected, the red bars represent the frequency of that term within the topic. Setting the relevance metric to 0 on the top right, the visualization will then show the top words that are the most exclusive to that topic. Using these tools made evaluating the models much easier.

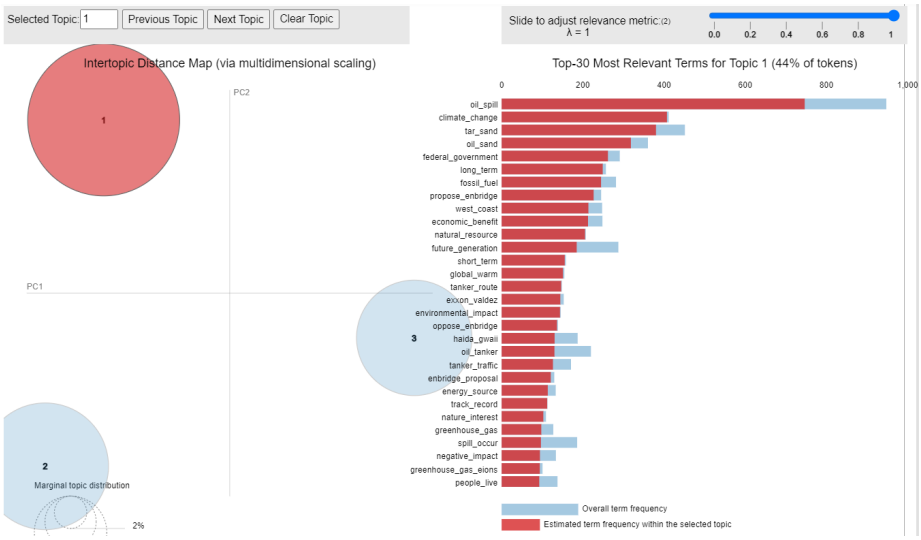

Figure 3.1: LDAvis with topic 1 selected

With the help of LDAvis and our most updated methodology, we were able to find a model that we were happy with. Using this

model, we were able to identify 3 main topics within the Enbridge project, 3 main topics within the Transmountain project, and 2 main topics within the Brunswick project.

The topics we were able to extract were accurate, but very general. We wanted much more specific topics. To accomplish this, we used our LDA model to assign every document to a main topic, and then performed LDA once more on each separate main topic. By doing this, we were able to extract subtopics from the letters. Using this approach, we identified 6 subtopics within Enbridge, and 8 subtopics within Transmountain. Within the Brunswick project, we didn't find this hierarchical approach to improve results. We think that this is mainly because Brunswick had significantly less letters than the other two projects. The topics and subtopics for all the projects are shown in Figure 3.2, 3.3 and 3.3 below.

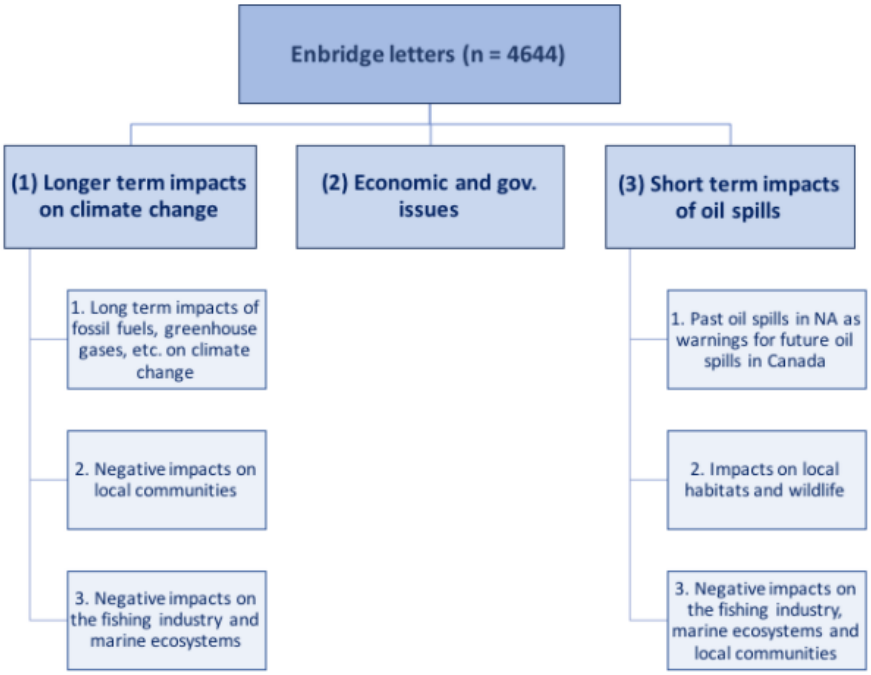

Figure 3.2: Enbridge Topics and Subtopics

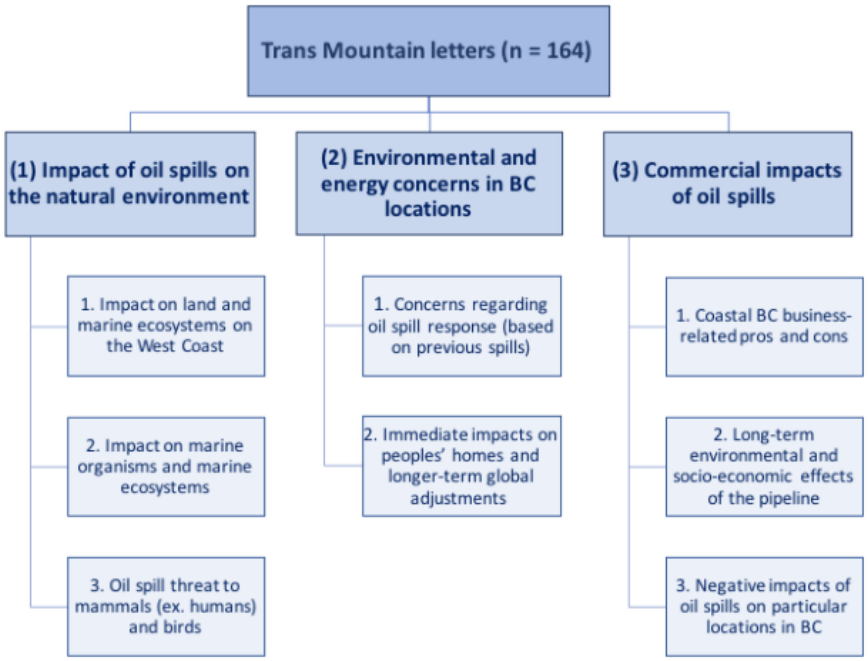

Figure 3.3: Transmountain Topics and Subtopics

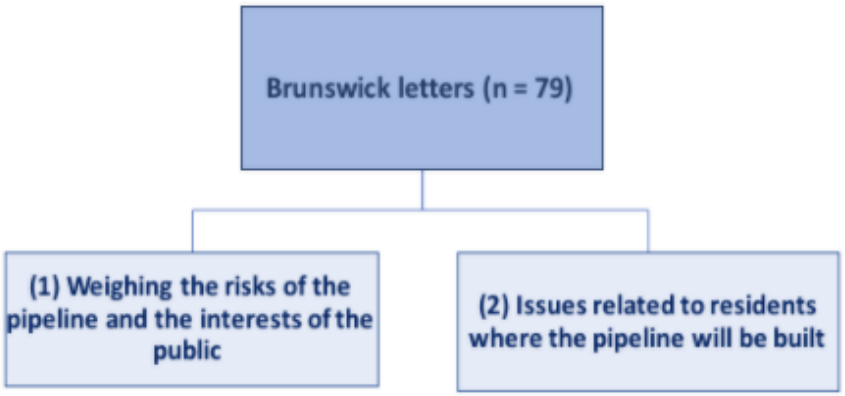

Figure 3.4: Brunswick Topics and Subtopics

4 Emotion Analysis Using NRC Emotion Lexicon

4.1 Goals and Tasks

Our goal was to provide the CER with interesting, interactive visualizations that would allow them to explore and learn about what people said and how people felt about those pipeline projects. To better understand how people felt from multiple dimensions, we went beyond the sentiment not just positive or negative by using The NRC Emotion Lexicon(aka EmoLex) in document-level emotion analysis. The NRC Emotion Lexicon is a list of English words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Since our training data is limited and unlabelled, this lexicon is especially useful in this unsupervised setting. First, we begin by providing the emotion words cloud from various pipeline projects within subtopics obtained from the LDA results to users. We want users to be able to compare the emotion words from various pipeline projects within subtopics and to observe the distribution of emotion words for that project in isolation from the others. Afterwards, we want users to explore the connections between topic modelling and emotion analysis. We want them to obtain information about people's major concerns and emotions regarding subtopics such as oil tanker traffic. We also want users to learn more about the top emotions and major concerns within the document that has the subtopics they are interested in, to find which pipeline project has raised the most concerns. Finally, we want users to gain insight into people's concerns about those various pipeline projects and their emotions towards specific projects. They can observe the distribution of different emotions and compare them between different projects or subtopics.

4.2 Data Visualization

We started off with an overview of how people felt differently among three pipeline projects and what people were mainly concerned about within that project. We utilized pie charts to show the part-to-whole relationship of the main topics gained from the LDA results to the overall particular project such as the Enbridge pipeline project. To compare the emotion trends among three pipeline projects, we used

a stacked bar chart with a categorical color scheme as the emotion words are broken down into 8 categories. Our main focus here is the comparison among different projects, so we aligned it on the bottom axis so that viewers are still able to easily read the values of the particular emotion contributed on its own using the y-axis. Users can view a detailed topic-level breakdown of the top 2 negative and positive emotions across different main topics by selecting the pipeline project they want to view. By selecting the emotion on top of the stacked bar chart, users would be able to observe the trend of that emotion across different pipeline projects.

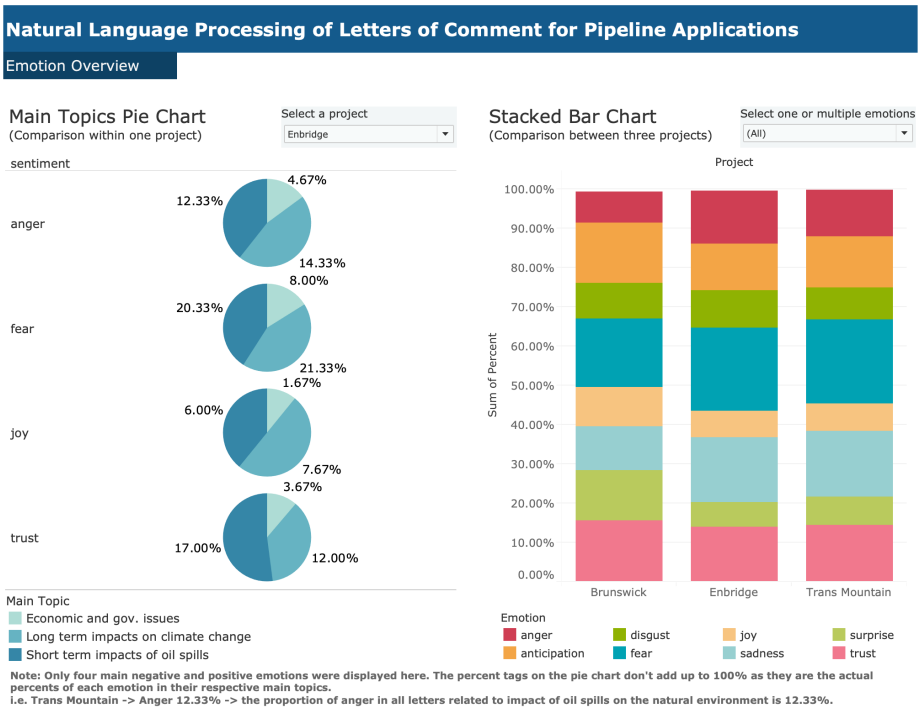

Figure 4.1: Emotion Overview

From going into detail about the specific emotions that people had of the projects and their related frequencies, we decided to take a step back and provide another detailed overview of all documents. Users can select the specific document they are interested in and check the percentage of the top 3 negative emotions as well as the words corresponding to each emotion of that document. Viewers

can also search the documents with a high concern rate by setting the emotion percent bar to the value they want. Hovering over the light bulb in the upper right legend, users would be able to see the instructions step by step in full detail.

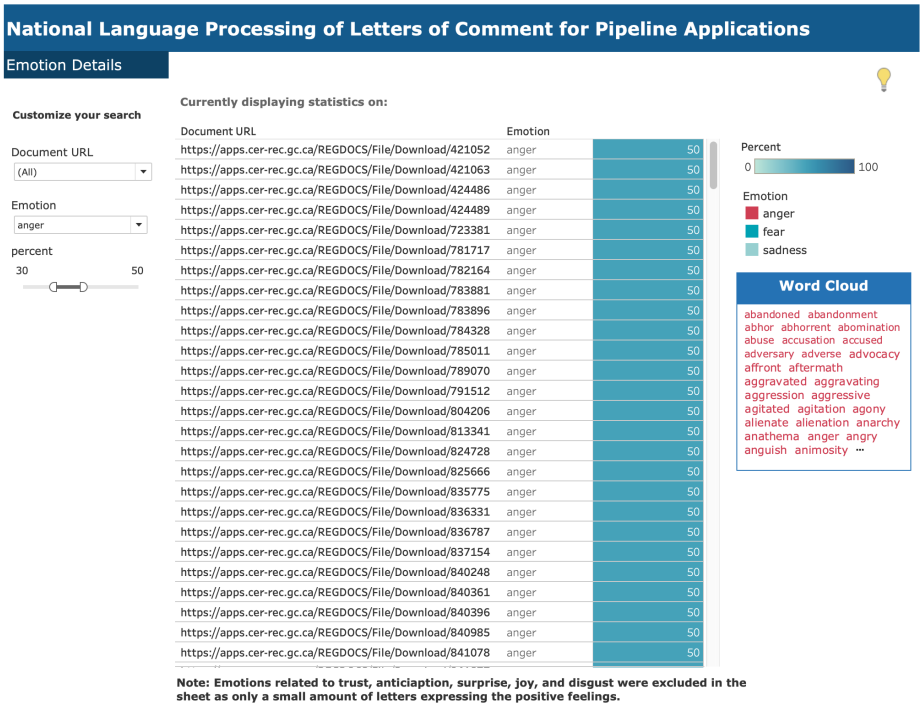

Figure 4.2: Emotion Details

5 Automatic Letter Summarization Using the TextRank Algorithm

To better understand what people said, we summarized the letters of comment using the TextRank Algorithm, which is based on the famous Google PageRank Algorithm used to rank webpages. The TextRank Algorithm is an extractive text summarization approach which means that it only extracts sentences from the original letter without rephrasing or changing it.

To start, we compared the performance of a couple different extractive summarization approaches including word frequency and the TextRank Algorithm. From our preliminary tests, the TextRank Algorithm slightly outperformed the other approaches in terms of returning the sentences that best summarized the letters. However, it is important to acknowledge the subjectivity as to whether or not the chosen sentences best captured the meaning of the letters. In addition, our selection of the TextRank Algorithm was, in part, due to its incorporation of how often words appeared next to each other (word co-occurences) in the scoring of sentences. Many of the other text summarization approaches that we explored did not take into account context and instead dealt with words as entirely separate entities. The other notable benefits of the TextRank Algorithm are its speed, simplicity, and reliability. In our experience, the time to summarize many documents was fast as it only took a few hours to obtain all letter summarizations. Secondly, the approach is relatively easy to understand compared to other more complicated summarization approaches. Thirdly, there was no danger of losing the original meaning of the sentences when summarization was performed. This is a major concern in abstractive text summarization, where new sentences are generated from the original text (potentially using words and phrases that were not include in the original text). A limitation of the algorithm is that it looks for similarities between sentences. Therefore, the focus of the summarization may be on a single topic and not on all of the topics within a letter. In addition, it may be difficult to pinpoint how many sentences best summarize each document. The exact number likely varies from document to document and is subject to personal preference.

The TextRank Algorithm was applied to letters after some data cleaning and processing to remove the words that did not contribute

greatly to the meaning of each sentence (for example, “a”, “an”, “the”). Any of the cleaned and processed letters could then be summarized. So, the first step in our implementation of the TextRank Algorithm was to separate each letter into individual sentences. Then, we used pre-trained word vectors to obtain vectors for our sentences. More precisely, we obtained the Global Vectors for Word Representation (GloVe) word vectors from Stanford NLP (which were trained on a large Wikipedia text corpus) for all of the important words in each sentence. We took the average of those vectors to arrive at one vector for each sentence. We next scored the similarity between each pair of sentences by using the cosine similarity approach (where the cosine angle between two sentence vectors was calculated to see if they were in a similar location in the vector space or not) and stored them in a matrix. Then, the Google PageRank Algorithm was applied to obtain the sentence rankings. It is important to note that the sentences with greater overlap (higher similarity scores) were given higher ranks. Finally, we extracted the top-ranked sentences to summarize each letter.

Although the number of sentences that best summarizes a letter is subjective and depends on the letter of interest, we thought that a reasonable summary is about 20% of the entire word count of the letter. Hence, for any letter containing 200 words or more, our TextRank Algorithm function returned 20% of the letter or less depending on the word count. For example, if a letter has 1000 words, the function identified that 200 words is the max word count of the output summary. It went through the top sentences, adding the sentence to the summary if the number of words in the sentence does not make the output exceed 200 words. For letters less than 200 words, which were considered to be short letters, we did not perform text summarization. That is, the output of any such letter was the entire original letter.

6 Conclusion

The topics that were modelled using LDA are indicative of what many individuals said for each the three major pipeline projects that we explored. They bring major public concerns to the forefront that should be taken into consideration in future decision-making processes. That said, our primary expectation is that our LDA modelling will provide a framework for analyzing future letters of comment for pipeline applications. A general workflow for a project is to first use our LDA Overview tool to label the main topics according to the top terms that are shown by bar charts and word clouds. This gives the CER team the opportunity to label the topics as their expertise may enable them to more precisely label them. Then, the CER team can utilize the Concern Details dashboard to filter for the letters according to the topics that they are most interested in. In addition, they can incorporate additional criteria into the letter extraction such as the author's organization and the year of when the letter was filed. After showing people's top concerns in the Concern Details dashboard, the CER team should be able to explore the emotion related to individual pipeline project and certain topics that they are interested in in the Emotion Details dashboard.

Since our letter summarizations are based on sentence similarity, they provide snapshots of what people are saying and highlight some of their top concerns. However, even though the summarizations call attention to some of the key points of what people are saying and how they are feeling, they should not be used as substitutes for reading the letters in full. Rather, the text summarizations should be utilized in addition to other NLP tools to gain a more comprehensive understanding of what people are saying.

We can easily incorporate text summarization into the LDA workflow. After the Concern Overview tool is utilized to extract the letters of interest, the summarizations of the letters can be referred to get a quick rundown of what people are saying before going through and reading the letters more thoroughly.

7 Future Work

We highly recommend the continuous improvement of our visualization dashboards as those tools provide simple ways to inspect and assess the results. We recommend improving the LDA visualization dashboard by incorporating text summarization. In the Concern Details dashboard, the text summarization for each letter may be included as a separate column so that the user can get a sense of what the letter contains before moving on to reading the full letter. Next, we recommend performing text summarization within the LDA results to find the top sentences within each topic and subtopic to give more informative descriptions of what they mean. Then, the LDA Overview dashboard can easily be improved by including the summarizations for each topic and subtopic.

To improve upon our LDA, it would be informative to obtain the percentage of each LDA topic (and subtopic) within each document as some of the letters may discuss multiple topics (and/or subtopics). A potential issue with this is that a labelled (training) set of letters is likely that has sufficiently large numbers of positive and negative letters is likely required. It may be difficult to construct the labelled set of letters because the dataset we used has very limited positive feedback.

To improve upon our sentiment analysis, we recommend using a binary classifier to say whether each author has positive or negative feelings about the pipeline project. A potential issue with this is that a labelled (training) set of letters is likely that has sufficiently large numbers of positive and negative letters is likely required. It may be difficult to construct the labelled set of letters because the current dataset has very limited positive feedback. If more positive letters are available, this classification project would be a good research direction to pursue.