MyLivingCity

Capstone Project Plan

## Introduction

The rise of social media has dramatically transformed how communities engage in decision-making processes. However, this transformation is not without challenges. Current platforms often fail to effectively harness community dialogues for constructive outcomes, with discussions frequently degenerating into repetitive, unorganized, and unproductive exchanges. MyLivingCity, a nonprofit organization based in Victoria, BC, aims to address these shortcomings through innovative computational solutions that refine how community feedback is gathered and analyzed. This initiative presents a fascinating problem in the realm of computational linguistics. The project’s focus on developing functionalities like the Comment Funnel Function and advanced filtering systems underscores a significant step toward more meaningful and structured community interactions. This approach not only promises to elevate the quality of community dialogues but also enriches the decision-making process by effectively summarizing and prioritizing community sentiments.

## Data

To address the limited scope of development data provided by MyLivingCity, our team plans to enrich this dataset by sourcing additional open-source text data from popular social platforms like Twitter and YouTube. This data will be meticulously selected to reflect diverse community dialogues relevant to the MyLivingCity user profile. Once collected, the data will be formatted into CSV files for uniformity and ease of processing. It will undergo rigorous cleaning to remove any irrelevant or sensitive content, ensuring that it adheres to the standards of non-sensitive, open-source data.

Moreover, to enhance the data's utility for training computational models, it will be annotated to categorize key aspects of community feedback such as sentiment and relevance. This annotated data will then be securely stored and shared via the Google Drive folder “UBC” created by MyLivingCity.

## Deliverables

The goals for this project, in order of priority, would be the following:

**Comment Funnel Function(Core):** The primary deliverable will be the development and implementation of the Comment Funnel Function within the MyLivingCity platform. This function will automatically suggest similar existing comments to users during the comment submission process, thereby reducing duplicate content and enhancing the clarity of discussions. The system will prompt users to agree with existing comments rather than posting duplicates, streamlining the conversation flow and emphasizing community consensus.

**Input Filtering System(Core):** Another critical deliverable is the creation of a robust input filtering system designed to detect and prevent the submission of inappropriate comments. This system will use advanced text analysis techniques to identify offensive language, off-topic remarks, and other unwanted content. A warning message will prompt users to modify their inputs accordingly, ensuring that the dialogue remains constructive and relevant to the community's interests.

**Gauge Statistics and Display Dashboard:** We will develop a comprehensive dashboard to display various metrics and statistics derived from community interactions on the platform. This dashboard will provide insights into the level of agreement or disagreement among different user segments (e.g., residents, students, business workers) and will be instrumental for both users and administrators to understand the community dynamics better.

**Summation Comment Display:** An exploratory deliverable will be the enhancement of the current comment display interface. We aim to present comments in a more structured and visually engaging manner, possibly integrating techniques like word clouds or fishbone diagrams to better represent community sentiment and feedback.

## Methods

### Comment Funnel Function

We plan to utilize a natural language processing approach to develop a system that groups similar comments together to reduce duplicates. By leveraging pre-trained models(LLMs) for semantic similarity, comments will be encoded into vectors and clustered via algorithms such as DBSCAN. This will enable the platform to suggest existing comments to users, encouraging them to endorse these instead of adding redundant entries.

### Input Filtering System

We will integrate a dual-layer approach to maintain comment quality. The first layer will consist of a predefined dictionary of inappropriate terms to filter out common offensive content directly. In the second layer, we'll deploy a machine learning model, potentially using a fine-tuned version of an NLP model(could be BERT), to identify subtler forms of toxicity and off-topic comments that a static dictionary might miss.

We would also consider if users can manually report some of the illegal words or contents, which could be added to our dictionary for filtering. The effectiveness of this filtering system will be validated by monitoring reductions in flagged content and improvements in user-reported satisfaction.

### Gauge Statistics and Display Dashboard

We are going to focus on creating a straightforward and functional interface to visualize key metrics from community interactions. Using Python's libraries such as Matplotlib and Seaborn for basic plotting, along with Dash for web-based dashboard integration, we will design simple yet informative displays. The dashboard will show statistics like agreement levels, comment frequency, and sentiment analysis results, segmented by user groups such as residents and business workers.

### Summation Comment Display

LLM is the best way to synthesize and summarize community feedback effectively. The system will process extensive comment threads and generate concise summaries that capture the predominant sentiments and key points raised by the community. By fine-tuning the LLM, we can show how the effectiveness of the Summation Comment Display is evaluated based on user engagement metrics and feedback, assessing whether the summaries and visualizations are helpful in understanding the overall community sentiment.

## Schedule

| Week | Dates | Detail |
| --- | --- | --- |
| Week 1 | May 6 - May 12 | * Data Collection and Preprocessing; * Finalize project plan; * Comment Funnel Function development start |
| Week 2 | May 13 - May 19 | * Continue Comment Funnel Function development; * Input Filtering System development start |
| Week 3 | May 20 - May 26 | * Continue Input Filtering System development |
| Week 4 | May 27 - June 2 | * Gauge Statistics and Display Dashboard Development |
| Week 5 | June 2 - June 8 | * Summation Comment Display Development |
| Week 6 | June 9 - June 15 | * Cleaning and standardizing existing codes; * Handover document preparation |
| Week 7 | June 16 - June 22 | * Final project report preparation; * Presentation preparation |