Capstone Project Review:
Call Type X99

Hayden Chiu, Jingyi Liao,
Jarrett MacFarlane, Yuxi Wang

Our Capstone Team

Master of Data Science in Computational Linguistics at UBC

Team members:

Hayden Chiu

Jarrett MacFarlane

Jingyi Liao

Yuxi Wang

The Problem: X99 Miscellaneous Calls

The right calls need to go into the right event types!

- The most common event type for Computer Assisted Dispatch (CAD) calls made by Calgary citizens in 2023 was **X99 - Miscellaneous**
- Goal: **Analyze text data** in order to understand why this shift has occurred
- Impact:
  - Calls can be reclassified accurately
  - More effective resource allocation
  - Better-informed response strategies

The Data

- Received data for **3 months of CAD call logs**, which were collected between January 1, 2023 and March 30, 2023
  - These calls had personal data redacted by CPS
- 72034 CAD entries, which detail **9752 unique events**
  - All event type X99 Miscellaneous
- **Event Remarks Text**: Description of the event, scene, callers, etc.

Our Methods

Data Preprocessing:

- Cleaned, normalized, and compiled text for analysis

Analysis:

- Discover “clusters” of calls by **grouping similar events together**

Tools and libraries used:

- Python, BERTopic, LaBSE, Mistral 7B, Llama-3, scikit-learn

Our Methods - Techniques for Analysis

- **Topic Modeling with LDA**:
  - Traditional method to identify hidden topics

- **Sentence Embedding & K-Means Clustering**:
  - Used `LaBSE` and `Mistral 7B` to get numerical representations of sentences (“embeddings”)
  - Used K-Means to cluster embeddings into topic groupings

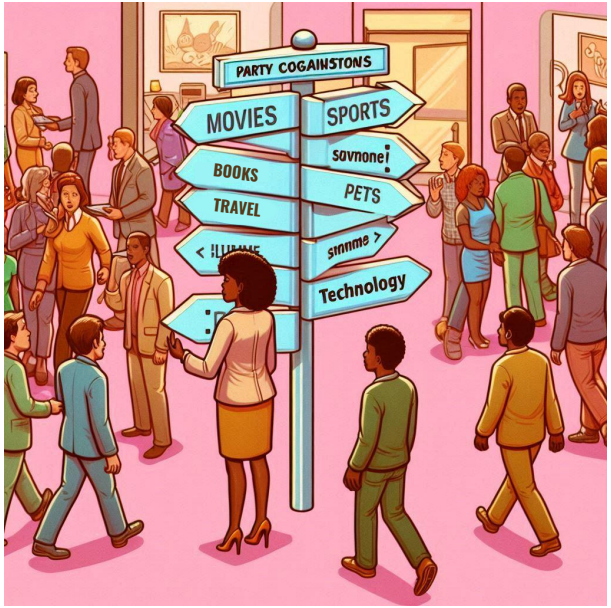
- **Zero-shot Topic Modeling with BERTopic**:
  - Leveraged `gte-large` embeddings
  - Used BERTopic's modular approach with LLM-based representation models for advanced topic labeling

*Images generated by Microsoft Designer

BERTopic - A High-level Overview

Embeddings

Clustering



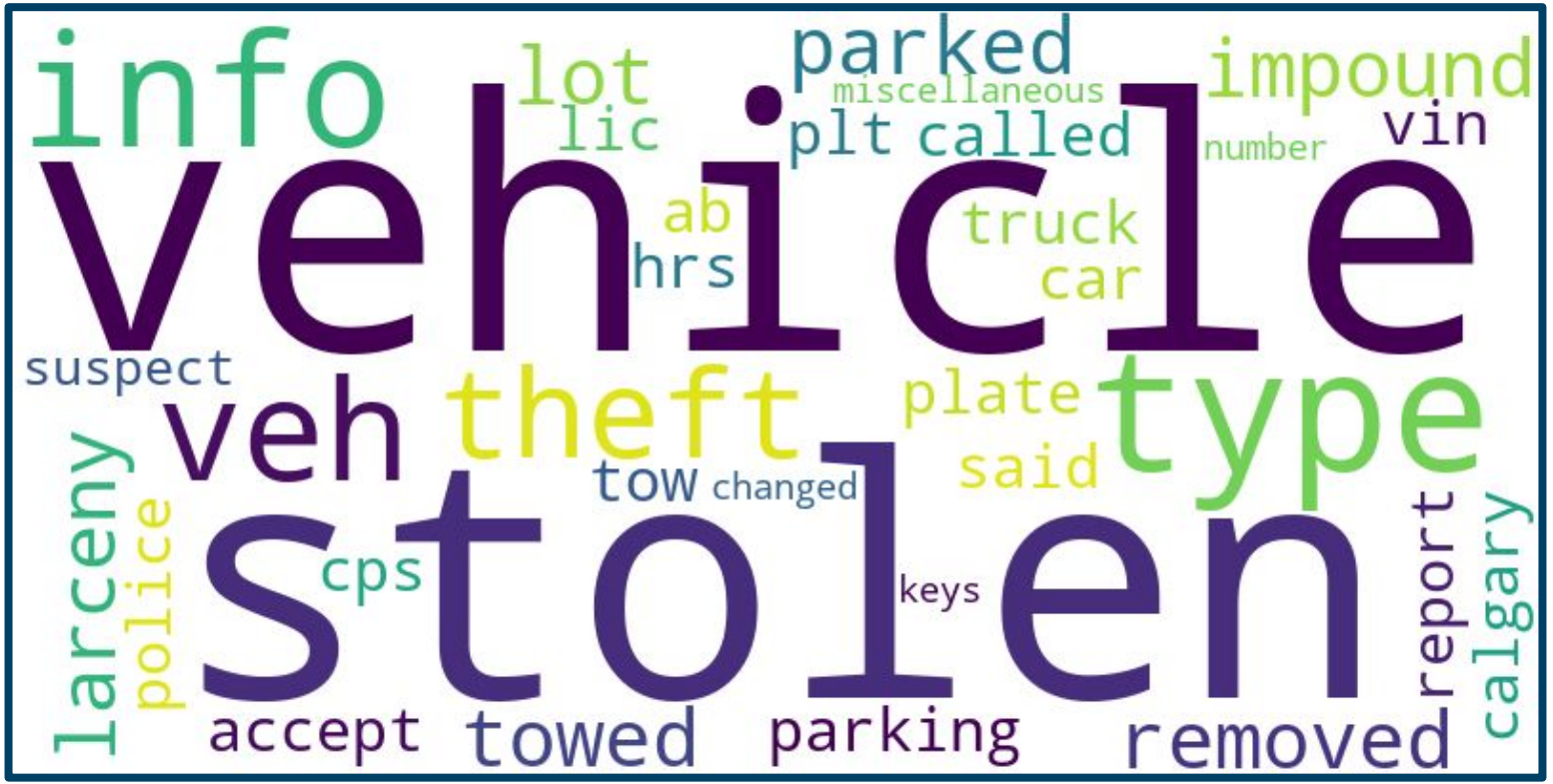
Summarizing
and
Fine-Tuning

Our Results

- Used BERTopic to identify clusters of similar calls, with at least **25 events per cluster**
  - Resulted in 76 total clusters
- **3580** out of 9752 events “actually” miscellaneous (36.7%)
- Successfully created groupings of calls which can offer insight as to what types of calls are being classified as X99
- Automatically labeled clusters using language model Llama-3

Topic	Count
MISC	3580
1	588
2	242
3	188
4	175
5	174
6	171
7	169
8	161
9	136
10	136

Group labeled “Vehicle Theft”- 588 events

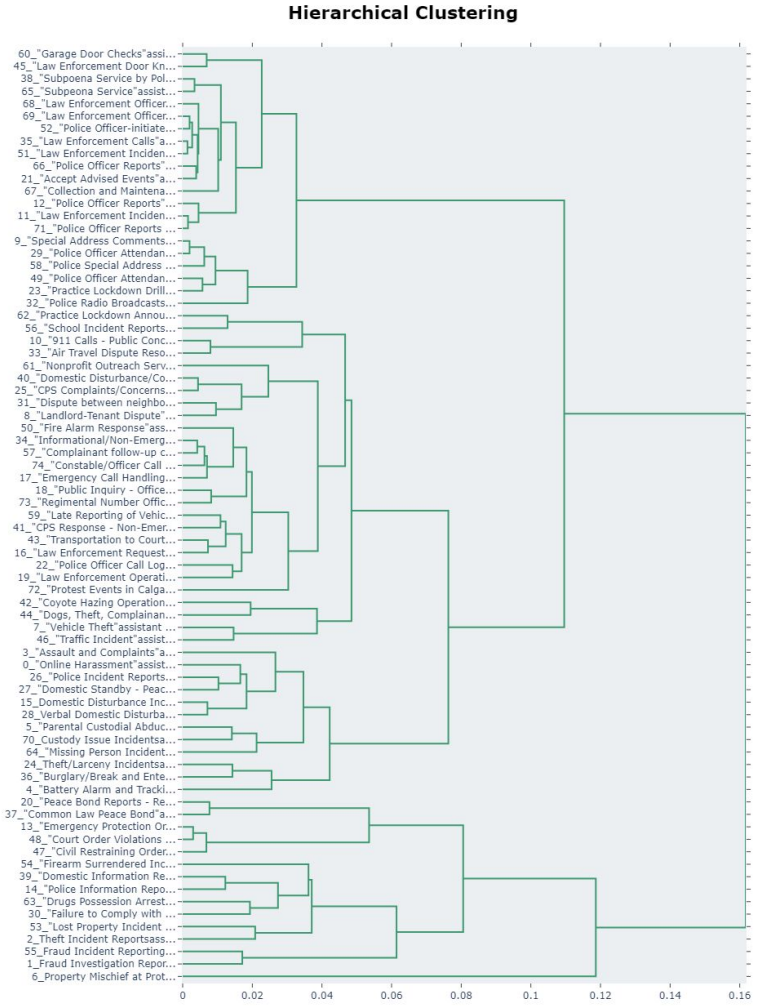


Group labeled “Emergency Protection Orders (EPOs) in Restricted Access Areas”- 242 events

Group labeled “Law Enforcement Incident Reports - Officer Attendance” - 174 events

Our Results

- Hierarchical clustering
- Exploring structures among the clusters
- Form clusters of lower granularity



Insights

Technical Insights:

- Benefits of modular approaches in NLP pipelines.
- LLMs significantly enhance topic interpretability and labeling.

Challenges Faced:

- Difficulty in model quality evaluation (relied on manual reviews)

Future work

Continuous improvement:

- Regularly update models with new data
- Incorporate feedback from officers and analysts

Automated monitoring:

- Implement automated systems for 911 call analysis
- Real-time alerts based on identified topics and trends in CAD Event Remarks Text

So, to wrap things up...

Summary:

- Successfully implemented advanced NLP techniques for CAD call data analysis
- Achieved meaningful insights and high-quality topic identification

Impact:

- Improved resource allocation and response strategies
- Enhanced situational awareness and decision-making for CPS

Thank you for listening!

Any Questions?