Canadian 2025 Election Reddit Sentiment Analysis

Ethan Do, Marjan Hassanzadehjahromi, Tianjing Gan, Zhiran (Andy) Tong

CMPT 732 Fall 2024

1 Problem Definition

Analyzing social media to understand the popularity of different parties is helpful in predicting the election results. This project uses Reddit comments to analyze the public sentiment toward the Liberal and Conservative parties to gain insights into the potential outcome of the 2025 Canadian federal election. Historical Reddit comment data was collected and assessed to get a deeper understanding of the relationship between social media presence and the 2021 election outcome. Additionally, daily Reddit comments are collected that analyze current sentiment trends and track public opinions toward the different parties.

2 Methodologies

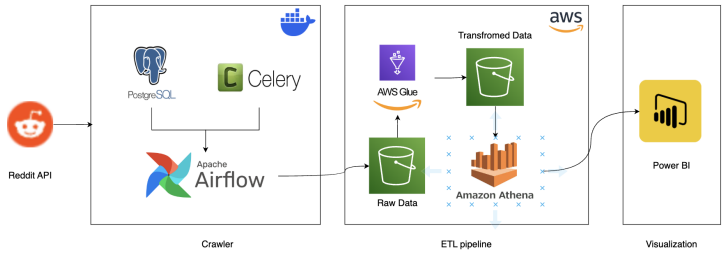

Figure 1: Project Architecture Overview

2.1 Data Extraction

The 3-year historical Reddit data from the year 2019 to 2021 was large, comprising 2.4 terabytes of comment data. Seeing as this project focuses on poli-

tics, a subset of political subreddits: (*'CanadaPolitics', 'Canada', 'Canadian', 'OnGuardForThee', 'CanadianConservative', and 'canadaleft'*) was handpicked to reduce the data size. Nevertheless, since our dataset was still large, we used big data tools to extract data from SFU's cluster data lake. Extraction for each year took around 60 minutes on 33 cores, submitted using a GNU screen command.

To address Reddit's API limit of 1,000 recent comments, an optimal daily crawling schedule was implemented using Airflow, scheduling jobs to extract recent Reddit comments every 3 hours from selected subreddits. The data pipeline operates automatically in a Docker container on an SFU virtual machine.

At the end of each day, raw data is converted to Parquet files for efficient storage and querying, then securely transferred to **AWS S3** for scalable, durable storage.

2.2 Data Transformation

The entire process was implemented for the historical data on the cluster. A similar approach was used with **AWS Glue** for automated daily execution.

Data Cleaning

Empty or deleted comments were omitted, and special characters, newlines, urls were removed from the text of the comments in order to improve data quality.

Party Labelling

Spark was used to perform an extensive keyword search on the comments and label each as one of four possible categories: *Liberal, Conservative, Both, or Neither*. The liberal keywords used were: *'trudeau', 'justin trudeau', 'liberals', 'liberal party', 'libparty', 'justintrudeau'*. Conservative keywords were: *'conservatives', 'conservative party', 'scheer', 'andrew scheer', 'o'toole', 'erin o'toole'*. Comments labeled as *Neither* were discarded and those labeled as *Both* were split into segments (according to punctuation) on which the same keyword search method was reapplied.

Sentiment Analysis Tasks

We leveraged pre-trained Large Language Models (LLMs) from Hugging Face Transformers, integrated with Apache Spark, to analyze Reddit comments at scale. Sentiment classification was performed using DistilBERT, which identified comments as positive or negative and computed class-based signed sentiment scores, while DistilRoBERTa was used for emotion classification across seven categories such as anger, joy, and sadness. For hate speech detection, we employed a BERT model trained on the HateXplain dataset to categorize comments as hate speech, offensive, or normal. This approach eliminated the need for manual labeling, enabling efficient and scalable Natural Language Processing (NLP) analysis.

After calculating the scores for each comment, **Apache Spark** was used to aggregate historical data. For recent datasets, **AWS Lambda** was configured to trigger SQL queries in **AWS Athena** whenever transformed data was uploaded.

Visualization

Finally, the processed and analyzed data is visualized using **PowerBI**, transforming complex sentiment data into intuitive and interactive dashboards. Regarding the recent Reddit data, while the pipeline updates the data daily, **PowerBI** limits the automatic refreshing to paid **PowerBI** Pro subscription which was out of our budget for the project.

2.3 Problems

Size of the dataset

This project faced multiple problems, mainly the large dataset size. Methods like breaking data into years and using big data tools such as Spark were employed during extraction, cleaning, and transformation, but the size still posed significant challenges for NLP techniques.

Spark NLP, designed for large-scale text processing based on parallel implementations of Hugging Face Transformers, approaches were implemented and tested on a small scale. But regardless of the models or the methods used, the same technical difficulties would arise whenever the models were used on the yearly dataset on the SFU cluster.

Therefore, Spark NLP was abandoned and **Hugging Face Transformers** were mainly used for the NLP tasks. Even so, we were still faced with memory, computation, and time limitations and had to use other methods—such as using a smaller subsequence of each comment for the analysis and tolerating the fault that comes with it, manually increasing the allocated memory of executor and driver nodes of Spark, and more—to overcome these challenges.

Variety

At the time of writing, Hugging Face offers 1,422 sentiment analysis, 6,183 emotion-related, and 608 hate-speech detection models. Selecting models for each task was challenging, with considerations like performance versus speed, figuring out non-standard formatting, and lack of documentation. In the end, several models were tested and trade-offs were made to pick simpler (less accurate) but more efficient models that could handle such a large dataset in an acceptable time frame—a few hours for each task for each year of data!

Reddit's Submissions

While we initially considered using both posts and comments from the subreddits, we only analyzed comments because upon deeper inspection we noticed that the submissions mainly consisted of simple links and questions but didn't

offer any sentiment towards the parties. Furthermore, we tried joining submissions with comments but that decreased the labelling and sentiment analysis performance. We did—and have continued to—extract, label, and store the submissions so we can figure out how to utilize the relationship between comments and submissions to improve the labeling and analysis of the comments in the future.

2.4 Results

Between 2019 and 2021, we collected over 600,000 comments from Reddit, primarily from four subreddits: *'canada' (57.08%), 'CanadaPolitics' (26.22%), 'onguardforthee' (12.33%)*, with other subreddits contributing smaller percentages. Liberal Party generally received more comment counts than the Conservative Party, except in the *'CanadianConservative'* subreddit during 2019 and 2020. In the dominant *'canada'* subreddit, the Liberal Party garnered 67.82% of the sentiment scores from 2019 to 2020, and most subreddits reflected over 50% Liberal Party sentiment.

The Liberal Party consistently received higher sentiment scores, which means positive sentiment, overall. The Liberals' sentiment was particularly high from August to October in 2019, and another high from July to September in 2021; these were the timing of the federal elections in both years. These findings align with the 2021 election results, in which the Liberal Party won the election, further confirming the positive sentiment expressed toward the party during this time.

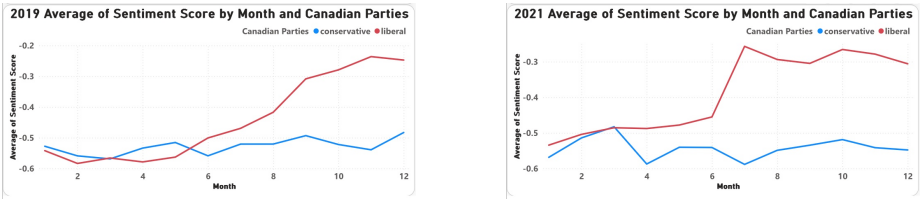

Figure 2: Sentiment Scores Over Time for 2019 and 2021

The analysis also included an evaluation of emotional tones and hate speech within the comments. The top three emotional tones for comments related to both parties were: neutral, anger, and surprise. The Conservatives had slightly fewer anger-related posts compared to the Liberals. Additionally, according to the upvotes, which was used to measure popularity, the Conservative Party scored higher at most of the time, indicating higher popularity towards the conservative related comments.

Further analysis was drawn from the analysis of emoji usage in comments. The most commonly used emojis in most subreddits were 😂 and 🙄. Interestingly, the third most popular emoji provided a deeper understanding of the comments. For the Conservatives, positive emojis, such as: 😉 and 😄, consistently ranked third from 2019 to 2021. However, 😠 was the third most popular

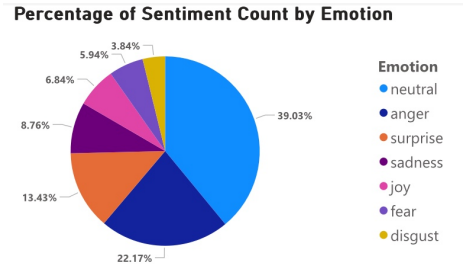

Figure 3: Percentage of Sentiment Count by Emotion

emoji in 2021 for the Liberal Party.

We analyzed data from November 2024 onwards to gain insight into the current sentiment for the upcoming 2025 Canadian Federal Election. The average sentiment score for the Liberal Party was -0.18, while the score for the Conservative Party was -0.23. Compared to 2020, which was the year before the previous federal election, both parties showed improvements. However, when compared to sentiment scores from November in election years, the Conservative Party showed a larger improvement, and narrowed the sentiment gap with the Liberal Party. An analysis of emotion and hate speech in November 2024 didn't show a big difference to the previous years. The distribution of emotional tones and hate speech followed the historical trends.

Project Summary

In this project, we developed an automated ETL pipeline to crawl Reddit comments and apply multiple NLP models from Hugging Face Transformers, gaining insights into Canadian public sentiment towards the Liberals and Conservatives. Big data tools such as **Apache Spark** and cloud storage **AWS S3** were used to manage data extraction and aggregation computations on a large dataset. Additionally, technologies like **Airflow**, **PostgreSQL**, **Docker**, **AWS Glue**, **AWS Lambda**, and **AWS Athena** were leveraged to automate and streamline the pipeline. The complete code is available at: Github.

- Getting data: 3
- Algorithms: 3
- Parallelization: 2
- ETL: 2
- UI: 2
- Visualization: 2
- Problem: 2
- Technology: 3