UNIVERSITY OF CALGARY
FACULTY OF GRADUATE STUDIES

DATA 608

Developing Big Data Applications

Final Project:

Sentiment Analysis of Reddit Posts

Kane Smith

Jordan Keelan

Rodrigo Rosales Alvarez

03/14/2023

Index

1) INTRODUCTION

1.1) Background and Context

The internet has fundamentally changed the way we communicate, learn, socialize, and conduct business. It has enabled connection in a way that is unprecedented in human history; distance and time are no longer factors in one's ability to share opinions, discuss topics with friends or strangers, and learn from a variety of sources of information, misinformation, and disinformation. One of the most popular online communities today is Reddit, which calls itself "the front page of the internet." It is the world most popular internet forum, allowing users to share content, engage in discussions, and connect with others who share similar interests or perspectives.

Reddit provides a unique window through which to view the discourse of many communities both geographic and subject based. As the majority of the content generated by the users of reddit is text based, and is already grouped by subject matter, we are able to explore the thoughts, ideas, and ideations of both public and anonymous voices.

1.2) Objectives of the Study

It is the goal of this project to allow this group to explore the potential and limitations of natural language processing (NLP) algorithms, to learn to interact with and scrape data via a website API, to extract meaningful insights from our analysis, and to share those insights in an interesting way.

NLP enables computers to understand, interpret, and generate human language. It involves developing algorithms and models that can analyze and understand natural language text, speech, and other forms of communication. NLP has a wide range of applications, including sentiment analysis, language translation, chatbots, speech recognition, and more. The goal of NLP is to bridge the gap between human communication and computer processing, making it easier for people to interact with technology in a more natural way (Lutkevich & Burns, 2023).

Reddit lends itself to being the optimal means of accomplishing the goals of this project. It has a robust API for interacting with and extracting data from the website. It contains massive amounts text data on various subjects moderated and validated by the subreddits that host it.

Our project consists of four main components.

1. Conduct sentiment analysis on both those post titles and content, and the comments contained under those trending topic posts. With the goal of analyzing how differences and similarities of how topics are described and discussed across Canada's cities.
2. Compare the time series results of our NLP analysis against weather data, to test the hypothesis that weather can influence online behavior and tone.
3. Conduct topic modelling on post titles to identify trending topics within each Canadian city subreddit to identify what are the most common topics of discussion within each, and to identify common topics shared by various cities.
4. Visualize and communicate the findings of our analysis in an interesting and interactive way.

We created an Extract-Transform-Load (ETL) data pipeline. This pipeline collects data from various Canadian city Subreddits and load it into local CSVs. The data pipeline runs automatically one per week, every Monday at 08:00 GMT-6.

2) PROJECT DATASET

2.1) Reddit

Reddit is composed of various communities/subforums, known as subreddits. Subreddits seem unlimited in the scope of topics that they cover, there are certainly less places and topics that do not have subreddits than do.

The focus of our exploration of Reddit will be on the subreddits of major Canadian cities. Each city in Canada has a user generated and moderated subreddit on which its citizens, and anyone else who would like to, may create posts and comment on the posts of others. We are interested in the distribution of topics that Canadians discuss, and the sentiment behind those topics.

Here below are the mentioned cities and links to go directly to their specific subreddit:

- Edmonton: https://www.reddit.com/r/Edmonton/
- Calgary: https://www.reddit.com/r/Calgary/
- Vancouver: https://www.reddit.com/r/vancouver/
- Toronto: https://www.reddit.com/r/toronto/
- Ottawa: https://www.reddit.com/r/ottawa/
- Winnipeg: https://www.reddit.com/r/Winnipeg/
- Saskatoon: https://www.reddit.com/r/saskatoon/
- Victoria: https://www.reddit.com/r/VictoriaBC/
- Hamilton: https://www.reddit.com/r/Hamilton/
- Halifax: https://www.reddit.com/r/halifax/

2.2) Open-Mateo

Provides current weather data, forecasts, and historical weather data for locations worldwide that can be consulted by their own API. We were able to retrieve weather information, including temperature, humidity, wind speed and direction, precipitation, and more. We downloaded historical data from 2010 all the way into mid-March of 2023.

2.3) Data Collection Process

We will use the Reddit´s API to extract data from the main subreddit in the most important Canadian cities where English is the first language.

The data will be extracted using the Python library “PRAW” (Python Reddit API Wrapper). Wrappers are modifications to functions or classes which change their behavior in some way. They are called wrappers because they “wrap” around the existing code to modify it.

PRAW and the Reddit API extract the data in JSON form, but with functions easily transform it into a Pandas Data Frame.

The data collection process involves extracting data in JSON format, converting it to a Pandas Data Frame, and then storing it locally as CSV files. The data to be collected includes post titles, post content, comment content, post and comment timestamps, and user metadata.

An initial data download of 1000 posts for each subreddit and 10 comments of each post was subtracted, which gave us posts up to 2012 for some subreddits; then an automatic ETL pipeline that runs in python scheduled by Task Manager in windows will scrape 50 posts with 10 comments per city every Monday at 08:00 GMT-6.

Posts and comments are retrieved depending on the relevance of it, Reddit has a complex mechanism to show some posts and comments first depending on how popular the post or comment is, using that we can always extract the most important posts and comments of each subreddit.

Lastly the ETL pipeline will check for duplicates and delete the first appearance of the post or comment, so we will always have the most updated information.

We decided to have two CSV files in which we are storing the data, the first one is for the posts and the second one for the comments. Some of the columns were autogenerated and needed for the sentiment analysis and topic modelling stage of the project.

Posts Table

A post is a piece of content submitted by a user to a specific subreddit. A post can take different forms, such as a text-based discussion, a link to an article, an image, a video, or a combination of these; for this project we are only considering text-based posts. Each post can be upvoted or downvoted by other users, which determines its visibility on the subreddit's front page or within search results.

- Title: tittle of the post, it is used to catch the attention of the users
- Content: if needed, the content space can be used to write things related to the title of the post.
- ID: unique identifier for the post.
- Date: datetime in which the post was created.
- City: name of the subreddit.
- Text: concatenation of title and content
- Date_8d: date in which the post was created.
- Cleaned_text: formatted text suitable for NLP.
- Neg: indicates the degree of negativity in the text, from 0 to 1. (1 is extremely negative sentiment)
- Pos: indicates the degree of positivity in the text, from 0 to 1. (1 is extremely positive sentiment)
- Neu: measure indicates the degree of neutrality in the text, from 0 to 1. (1 is extremely neutral sentiment)
- Compound: measure represents an overall score of sentiment in the text, from -1 to 1. (-1 indicates extremely negative sentiment, 1 indicates extremely positive sentiment and 0 indicates neutral sentiment)

Comments Table

A comment is a response or feedback left by a user in relation to a post or another comment on a subreddit. Users can leave comments on posts to express their thoughts or opinions related to the content of the post. Additionally, users can reply to other users' comments, creating a threaded conversation under the original post. Comments can be upvoted or downvoted by other users, which determines their visibility on the subreddit's comment section.

- Post_id: unique identifier for the post.
- Body: content of the comment.
- Comment_id: unique identifier of the post.
- Comment_score: Score of the comment, based on upvotes and downvotes.
- Date: datetime in which the comment was posted.
- Date_8d: date in which the comment was posted.

- Cleaned_text: formatted text suitable for NLP.
- Neg: indicates the degree of negativity in the text, from 0 to 1. (1 is extremely negative sentiment)
- Pos: indicates the degree of positivity in the text, from 0 to 1. (1 is extremely positive sentiment)
- Neu: measure indicates the degree of neutrality in the text, from 0 to 1. (1 is extremely neutral sentiment)
- Compound: measure represents an overall score of sentiment in the text, from -1 to 1. (-1 indicates extremely negative sentiment, 1 indicates extremely positive sentiment and 0 indicates neutral sentiment)

2.4) Data Pre-processing for Sentiment Analysis

Cleaning text is a crucial step in Natural Language Processing because it helps to improve accuracy of the NLP algorithms. The main reasons to clean text are:

- Removing irrelevant information: this step includes the removal of special characters.
- Standardizing text: text can come in different ways, like slang, uppercases, misspelling, etc. By removing those, NLP algorithms will give us better results.
- Reducing noise: examples of noise include typos, extra whitespace, and grammatical errors.
- Improving efficiency: reducing the dataset can help to improve efficiency.

For our project we used methods provided in the NLTK library to clean our text. The steps followed were:

- Tokenization: breaking down the text into smaller units, in this case by each word.
- Lowercasing: transform all text in lower cases.
- Stop word removal: getting read of words like “a”, “the”, “and”, etc.
- Stemming: removing suffixes in words.
- Lemmatization: converting words into their base form.

2.5) Data Pre-processing for Topic Modeling

The saved 'posts.csv' file, which contains sentiment analysis scores, is loaded into the post_df DataFrame. New stop words are added to the existing set of stop words for each stemmed city name, as they appear too frequently and disrupt the algorithm. These stop words are then removed from the 'cleaned_text' column. A BERTopic instance is created, and the model is fit to the previously cleaned and stemmed data in the 'cleaned_text' column. Topic numbers and probabilities are generated, and the topic_model, containing all data on representative topic words, is saved. The generated topics are then appended to the post_df.

To address the issue that sentiment derived only from posts is likely to read neutral, as posts are often very short strings of text, the comment sentiment is used to inform topic sentiment. To achieve this, the comments.csv is opened and cleaned, and is grouped by post_id, averaging the compound score. This is then merged to the post_df with ‘post_id’ as the key.

As a result, we have a post_df containing cleaned post content, sentiment scores for both the post and its comments, and assigned topic numbers, enabling us to create visualizations and draw meaningful conclusions.

3) DATA ANALYSIS

3.1) Sentiment Analysis

Sentiment analysis is a subfield of NLP that involves computational tools to automatically determine the emotional tone texts. There are many models that can perform sentiment analysis tasks, the one that we used for our project was VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon.

We decided to use VADER as it is designed to handle social media text, which is usually short and informal and non-standard language. Moreover, the model uses a lexicon approach, meaning that it has a pre-defined dictionary of words and phrases with pre-assigned scores for their sentiment polarity, making it more accurate than other methods that rely solely on machine learning algorithms.

3.2) Topic Modeling

Topic modeling is a natural language processing (NLP) technique used to identify and analyze recurring themes or topics within large collections of textual data. By examining word co-occurrences or clusters, topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), group related terms with the aim of generating coherent multi-word representations of topics present in the input text. Topic modeling leverages semantic structures in text to make sense of unstructured, unlabeled data (Sheridan).

Initially, I chose LDA as our algorithm, but after several attempts to fine-tune its parameters, the model failed to yield meaningful results.

BERTopic is a newer, more advanced algorithm that utilizes Bidirectional Encoder Representations from Transformers, or BERT. In broad terms, BERTopic embeds data as vectors in a multidimensional space, performs dimensionality reduction, and clusters the data into topics. It generates as many topics as necessary to describe the text, assigning a probability for each text segment to belong to a specific topic. Topics are ranked from -1 (noise) upwards, with 1 being the most common, 2 the second most common, and so on (Grootendorst, 2021).

3.2.1) Topic Modeling Methodology

Topic modeling is applied to posts rather than comments, as posts generally introduce the main topic for discussion, while comments elaborate on that topic. Including comments in the analysis could significantly obscure the results.

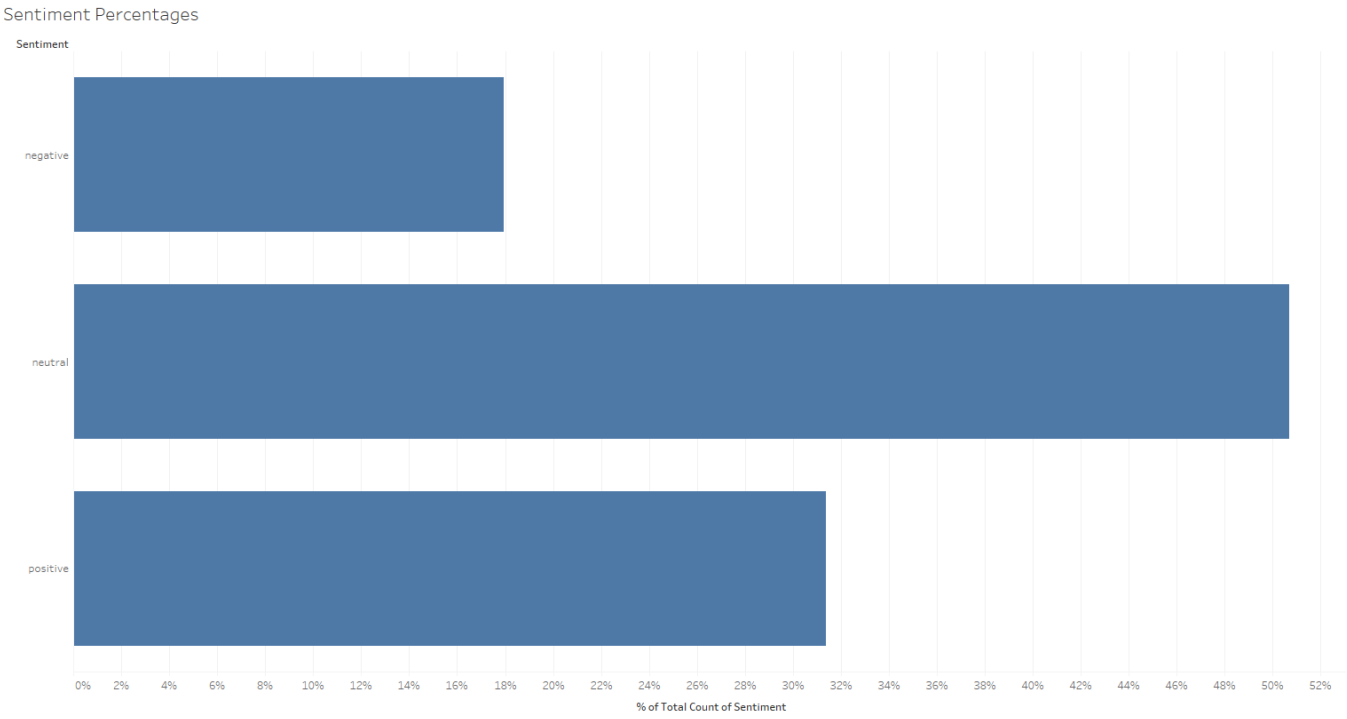
However, comment sentiment is a better metric for analyzing

4) RESULTS

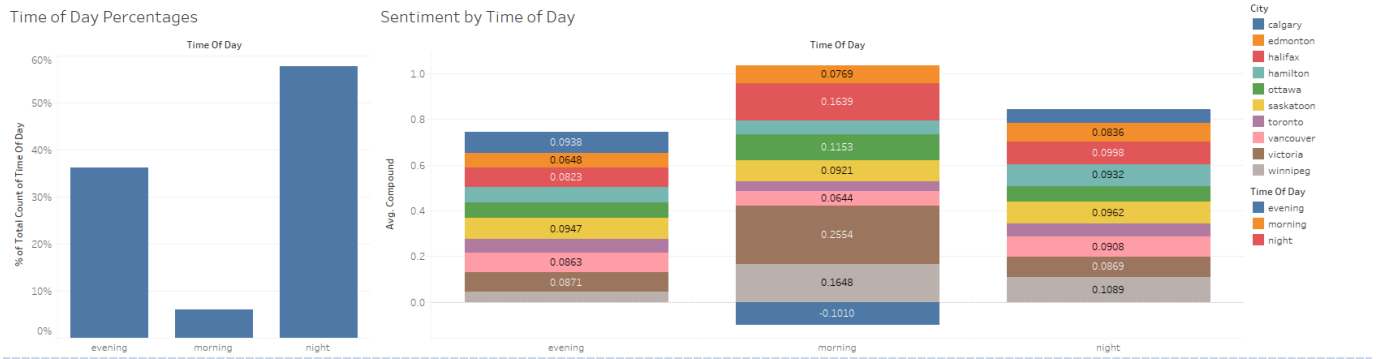
4.1) Sentiment Analysis Results

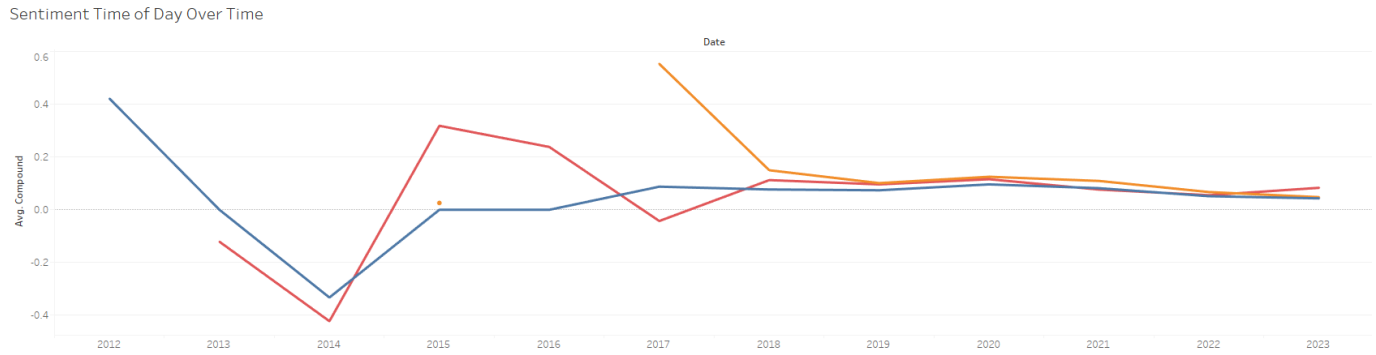
4.1.1) General

The overall sentiment analysis of the posts and comments found that VADER classified approximately 50% of the posts as neutral, 30% as positive and 20% as negative. We then broke our analysis into 3 different categories: the time of day of the post/comment, the season of the post/comment, and the day of the week of the post/comment.



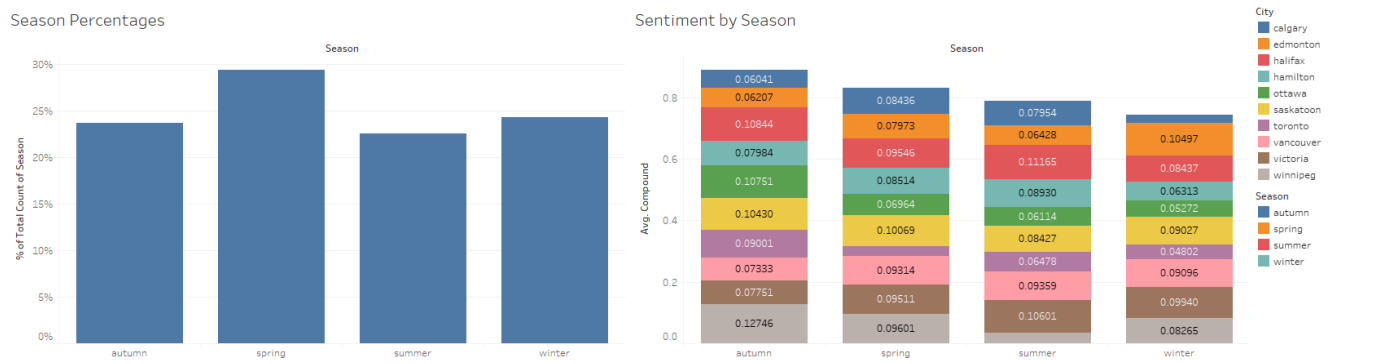
Time of Day

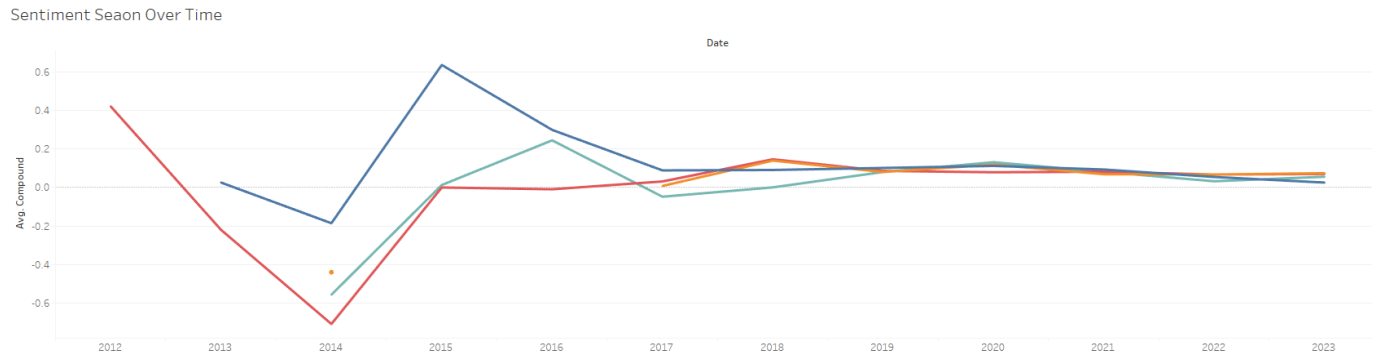




We can see that posting frequency during the mornings is extremely low compared to the other time of the day, with people posting most frequently at nighttime. However, the mornings have the highest average compound (higher is more positive) with Calgary having the only negative compound.

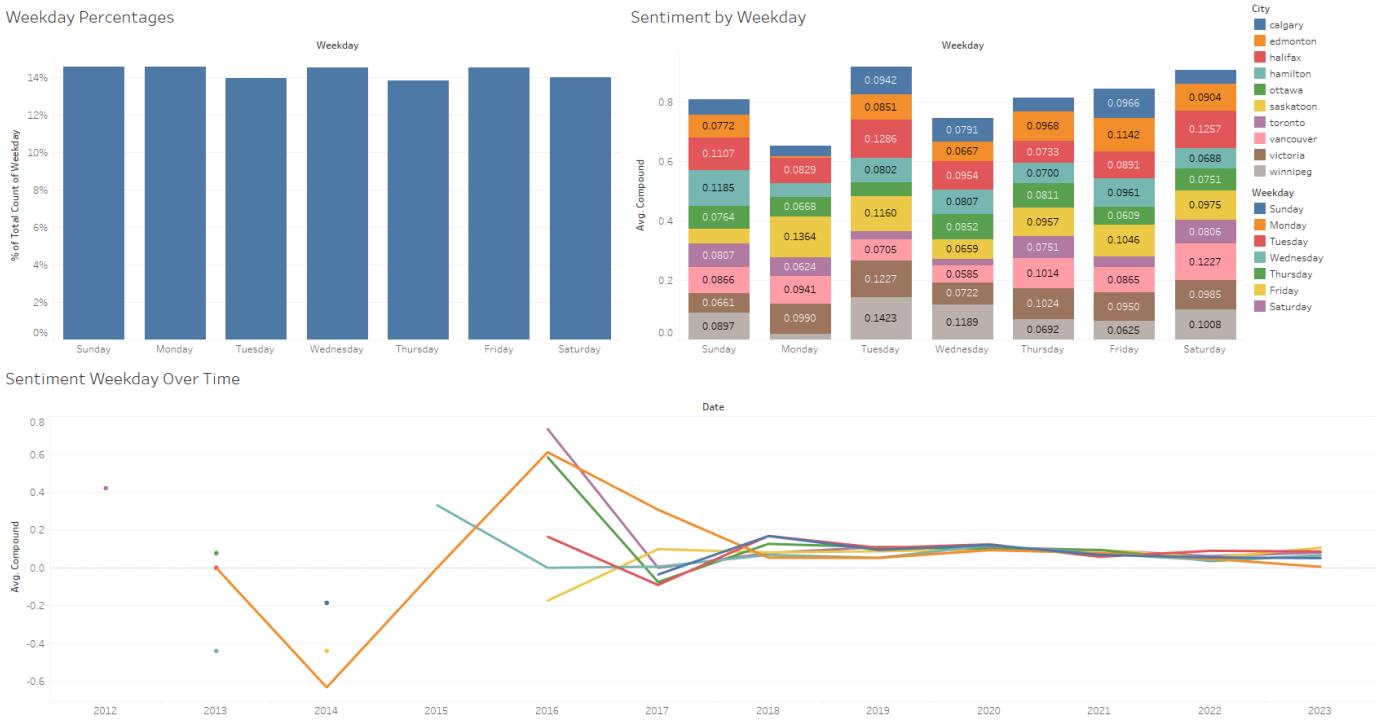
Season





Above we can see that during the Summer, there is a small drop in posting frequency compared to the other three seasons. The difference in average compound is the highest in the Summer.

Weekday



The posting frequency is essentially the same for every day of the week, but we see that Tuesday has the highest average compound, followed by Saturday.

4.1.2) Weather

We also wanted to see if we could find any exogenous variables that may correlate with the sentiment of the Reddit posts over time. We decided to look at different weather variables including the temperature, cloud cover and snow fall. We wanted to look at the top and bottom 5 days of their respective categories, but obviously different cities in Canada experience different weather so it would not make sense to pick an arbitrary cutoff to define the top or bottom 5. To get around this, we looked at each city separately and looked at the top 5 that way. Below are our findings:

Overall, we did not find any significant interesting results, which is why we did not include it in our presentation. We see some change in sentiment during the days with the most extreme weather conditions, but it is not obvious whether this is random or potentially a product of how we defined extreme weather.

4.2) Topic Modelling Results

The BERTopic model identified 176 distinct topics and one noise topic. Since this is an unsupervised machine learning algorithm, there is no direct way to calculate accuracy or error scores. The success of the algorithm must be assessed qualitatively, using the following criteria:

1) Coherency – Do the representative words for each topic make sense together? Is the generated topic easily identifiable?
2) Reproducibility – Given that the ETL pipeline is designed to scrape new posts weekly, adding them to a larger body of previously scraped posts and re-running the algorithm on the entire dataset, the results should not be static but should also not vary drastically with each iteration.
3) Pervasiveness – For our analysis, it is essential that topics appear in more than one city subreddit. Although different cities may discuss different topics, it is unlikely that common themes are not present across all subreddits. For example, if the model generated 8,000 topics for 10,000 posts, we would not consider this successful.

Our ETL pipeline generates five outputs:

1) topic_hierarchy_tree.html – the decision tree in topic space generated by the BERTopic model fitting.
2) Intertopic_distance.html - The dimension reduced plot of the topic vectors in 2D space with clustering performed based on intertopic distance.
3) topic_info.csv - A list of all topics, their count within the data, and a name containing the representative words for that topic.
4) topic_word_clouds.html – Word clouds generated displaying the most common words used to define each topic.
5) topics_by_city_staked_bar.html - Staked bar chart for each city subreddit showing the top ten topics.
6) city_topic_sentiment_boxAndWhisker.html - Box and whisker plots representing the sentiment associated with each topic for each city.

View All Interactive Plotly HTML Visuals Here

Topic Hierarchy Tree

The topic hierarchy tree provides a visualization of the organization, relationships, and clustering of topics, as well as the keywords associated with each topic. The top-level comments represent broad topics, and more granularity is revealed as the depth increases. While I attempted to trim this tree, the topics generated by the algorithm were superior to those resulting from trimming. Additionally, trimming significantly increased the complexity of incorporating the changes to the rest of the model and visualizations (Grootendorst).

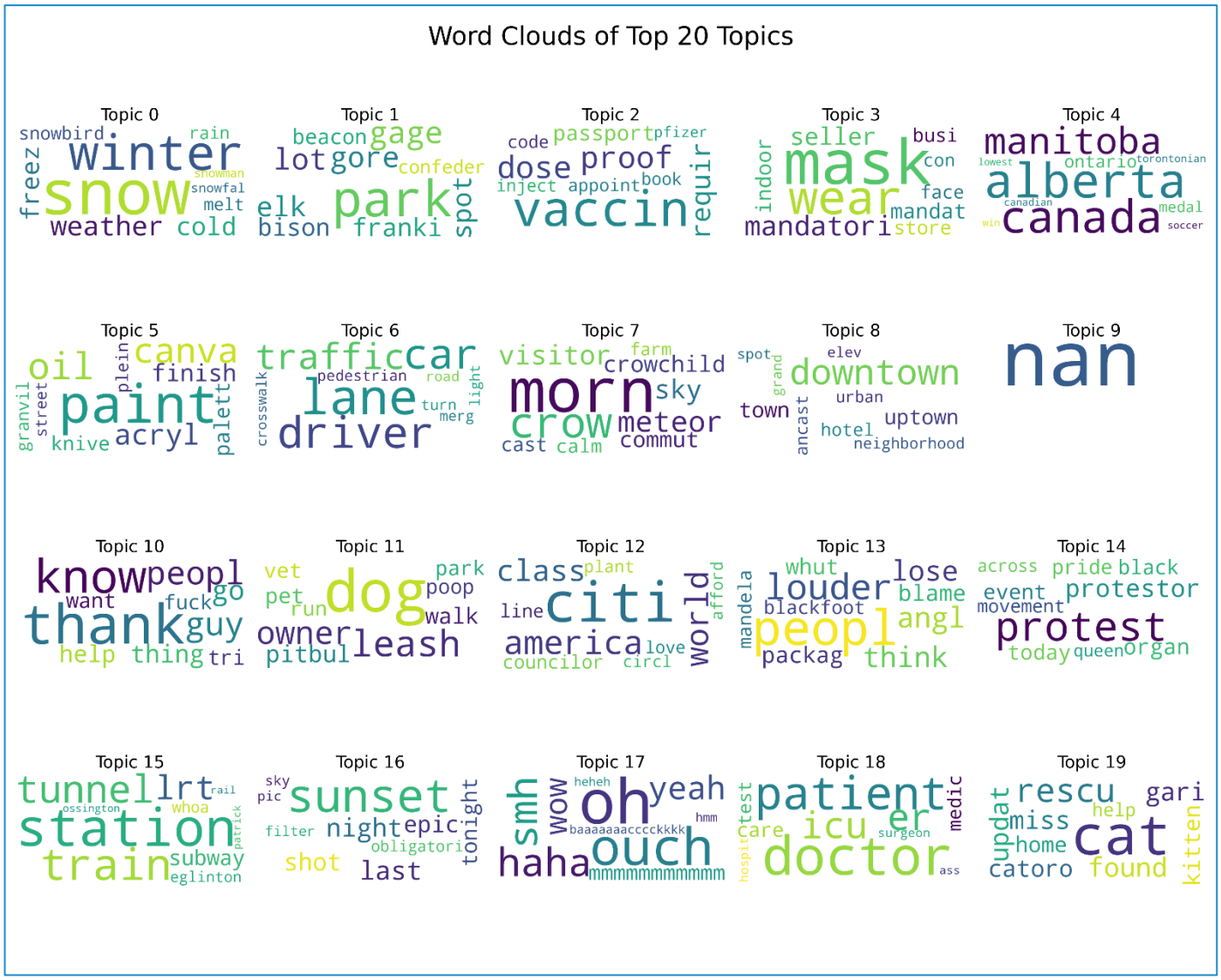
Intertopic Distance Map

This visualization employs dimensionality reduction to display high-dimensional embedded topic vectors in 2D space. Ideally, this presents the generated topics as clusters of semantically related items, providing insight into how BERTopic constructs the topic hierarchy tree (Grootendorst).

By using the slider, you can select a topic, which will then be highlighted in red. Hovering over a topic reveals general information about it, including the topic's size and its associated words (Grootendorst).

Initially, I had hoped to use these larger clusters and their intertopic distances to create broader clusters. However, upon examining the semantically similar clusters, it becomes apparent that deriving a single overarching topic to encompass all topics within the cluster is not feasible.

Topic Word Clouds



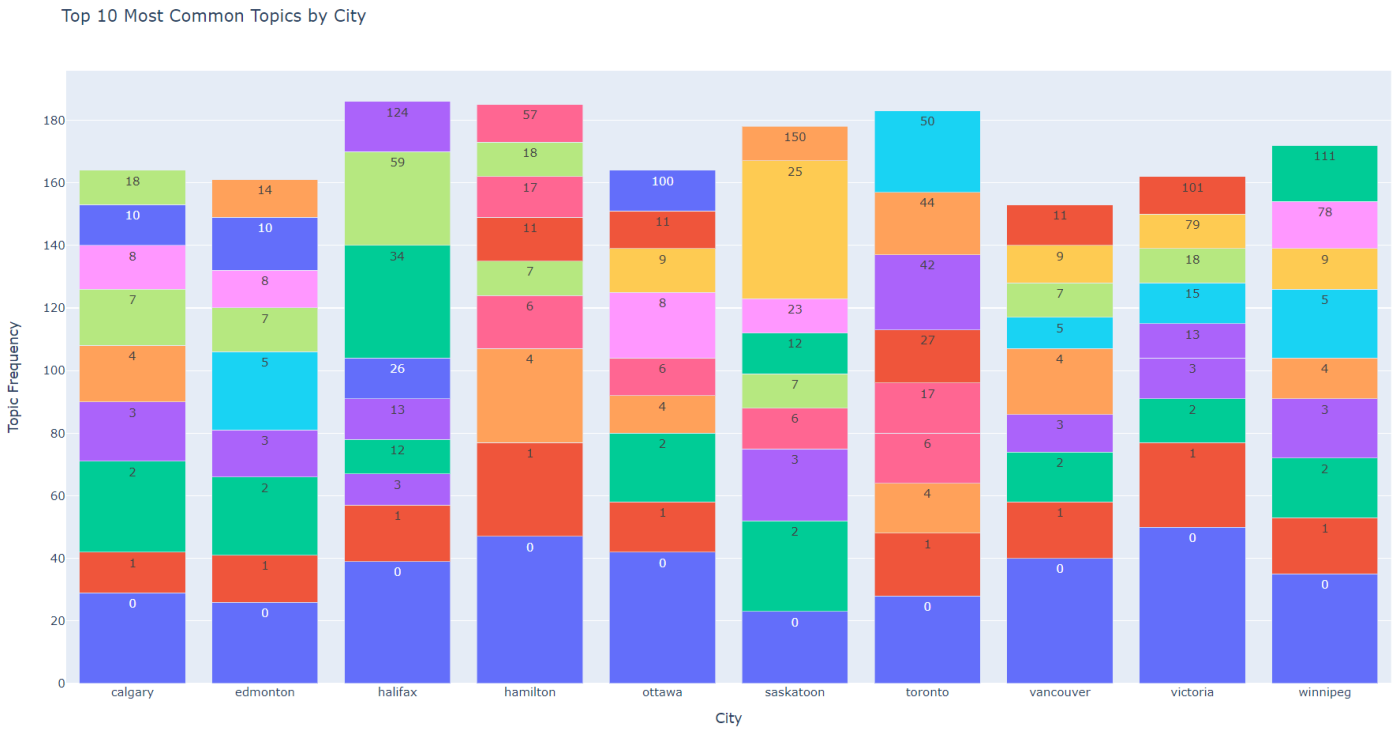
To showcase the topics generated by BERTopic, we employed word cloud plots that highlight the most prominent words within each topic. Note that these words are still stemmed from our preprocessing. We attempted to un-stem them using various algorithms, but those were more likely to cause confusion than our brains un-stemming them.

Some of these topics exhibit excellent coherence and align with the types of topics we aimed to generate for comparing sentiments across cities. These topics include: topic 0 (winter weather), 2 (vaccine proof), 3 (mask mandates), 4 (provinces), 6 (drivers/traffic), 11 (dogs/leashes/pitbulls), 15 (light-rail), and 14 (protests).

Others, such as "17_oh_ouch_smh_haha," require an understanding of typical Reddit/internet discourse to grasp their potential meaning. Words like 'ouch,' 'smh,' or 'haha' usually accompany content and serve as the poster's reaction to it. For example, a person struggling to park their car might say 'smh' (shaking my head).

Some topics are more challenging to comprehend, like topic 1, which we interpret to encompass words like 'gorgeous' or 'gorge,' 'confederation,' 'elk,' and 'bison,' potentially relating to topics such as city parks, national parks, and beautiful spots.

City Topic Stacked Bar Chart



This stacked bar chart is useful for evaluating the persistence of topics across subreddits and gaining insights into the relative prevalence of topics discussed in each subreddit. It's easy to trace the more common topics across multiple cities, as most cities have topics in the top 10. However, it's also evident that each city has its own unique set of topics, or topics that are more common to geographically similar cities. For example, consider the similarities between Calgary and Edmonton, Hamilton and Toronto, or Vancouver and Victoria.

Topic Sentiment by City

As our pipeline codes each post with the average sentiment of its top comments, we are enabled to plot the distribution of sentiment for each topic across each city subreddit. As explained in the sentiment analysis, the y-axis of these box and whisker plots has 1 as the most positive, -1 as the most negative, and 0 as neutral. We originally tried this with just sentiment derived from the post title and body but noticed that with such a small amount of text, the majority of post sentiments were neutral. It is highly likely that due to the brevity of posts, most valuable sentiment would be contained within the comment section and not well represented in the post title and body. Adapting the sentiment towards the comment sentiment has made for much more interesting results.

Some interesting findings:

- Sentiment for the top 20 topics is largely positive, with nearly all median sentiment scores above 0.
- The most negative topic across all cities is Topic 14: Protests.

- Halifax speaks the most negatively of their drivers, Topic 6. Though there is quite a lot of spread on this one for all cities.
- Unsurprisingly, topics with lower coherency see larger spreads in city sentiment. Topic 10 and 13, for examples
- Winnipeg, Hamilton, and Calgary speak the highest of their downtowns, Toronto speaks the least (Topic 8).

5) CONCLUSIONS

VADER sentiment analysis classified approximately 50% of the posts as neutral, 30% as positive and 20% as negative. While mornings are the least active time of day for top-level Reddit comments, they yield the most positive sentiment. Season of the year doesn't appear to impact Reddit sentiment across the Canadian cities. And aside from Tuesday having the most positive comments, there is likely no statistically significant difference between the days of the week.

From our topic modeling we can draw several conclusions about the discussion trends in various city subreddits. First, it is evident that some topics, such as winter weather, vaccine proof, mask mandates, and protests, are consistently prevalent across multiple cities, indicating common concerns and interests among residents. And validating the pervasiveness of our BERTopic model. Also, geographically similar cities share similar topic distributions, suggesting regional factors play a role in shaping the discussion themes.

Sentiment associated with topics were mostly positive across all cities and topics. With very few instances of median sentiment dipping below zero, and when it does, it's not by much. The most negative topic was regarding protests, and most negatively in Calgary.

In conclusion, the topic modeling analysis provided valuable insights into the commonalities and differences among city subreddits, highlighting the potential significance of regional factors in shaping online discussions. Additionally, the sentiment analysis demonstrated the varying emotional responses to different topics. Further exploration such as pulling more comments per post and spending more time preprocessing to ensure high quality data, could help deepen our understanding of these discussions and enable more targeted comparisons of sentiment across cities.

6) FUTURE WORK

While the BERTopic model successfully identified some coherent topics, it also generated topics that were more difficult to interpret. Future research could explore alternative topic modeling approaches or fine-tune the model to improve topic coherence and interpretability. Additionally, the current study only focused on English language posts, potentially overlooking important discussions in other languages, especially French in the Canadian context.

Analyzing the sentiment of posts over different time periods (seasons of the year, times of the day, days of the week) provided interesting insight into the different posting patterns on Reddit. We also see the contrasts between each Canadian city's subreddit. For our current analysis, we only collected 1000 posts across all 10 subreddits; a future consideration would be to scrape more posts dating further back to get a more complete picture of the sentiment of the subreddits. To do this, we could run multiple scripts at once to get around the request limit of the PRAW API.

To improve the efficiency and stability of the ETL pipeline, future work could involve migrating it to AWS cloud services. By leveraging AWS services like AWS Glue, Amazon S3, and Amazon Redshift, all stages of extraction, transformation, analysis, and visualization could be enabled, reducing the computational burden on local devices and increasing scalability as the dataset grows. The integration of AWS Glue would facilitate the automation and orchestration of the ETL process, while storing raw data in an S3 bucket would allow for increased storage of extracted data (AWS). The utilization of Amazon Redshift for data warehousing would enable querying and analysis (AWS). Furthermore, adopting AWS cloud services presents the potential to integrate advanced machine learning and natural language processing tools, such as Amazon Comprehend, to enhance the topic modeling and text analysis capabilities.

7) REFERENCES

AWS. (n.d.). AWS Glue. Amazon Web Services, Inc. Retrieved April 10, 2023, from https://aws.amazon.com/glue/

AWS. (n.d.). Amazon Simple Storage Service (Amazon S3). Amazon Web Services, Inc. Retrieved April 10, 2023, from https://aws.amazon.com/s3/

AWS. (n.d.). Amazon Redshift. Amazon Web Services, Inc. Retrieved April 10, 2023, from https://aws.amazon.com/redshift/

AWS. (n.d.). Amazon Comprehend. Amazon Web Services, Inc. Retrieved April 10, 2023, from https://aws.amazon.com/comprehend/

Grootendorst, M. (2021, January 12). Interactive topic modeling with Bertopic. Medium. Retrieved April 10, 2023, from https://towardsdatascience.com/interactive-topic-modeling-with-bertopic-1ea55e7d73d8

Grootendorst, M. P. (n.d.). Hierarchical topic modeling. BERTopic. Retrieved April 10, 2023, from https://maartengr.github.io/BERTopic/getting_started/hierarchicaltopics/hierarchicaltopics.html#linkage-functions

Lutkevich, B., & Burns, E. (2023, January 20). What is natural language processing? an introduction to NLP. Enterprise AI. Retrieved March 12, 2023, from https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

NLTK: Natural Language Toolkit. Retrieved March 6, 2023. NLTK: Natural Language Toolkit

PRAW: The Python Reddit API Wrapper. Bryce Bowe. Retrieved March 6, 2023, from PRAW: The Python Reddit API Wrapper — PRAW 7.7.0 documentation .

Sheridan, S. (n.d.). What is topic modeling? A beginner's guide. RSS. Retrieved April 10, 2023, from https://levity.ai/blog/what-is-topic-modeling

Top 100: The most visited websites in the US [2022 top websites edition]. SEMrush Blog. (n.d.). Retrieved March 6, 2023, from https://www.semrush.com/blog/most-visited-websites/.

8) Appendix

8.1) Extract Transform Load

Please review our code by clicking on the following link.

8.2) Video Discussion

To review our findings and project description click on the following link.