ProposalDomain Description and motivation**What is the Data domain:**

Political Tweets in Canada specifically splitting it amount the different political ridings

The data includes the tweets contents (description), username, time, location, hashtags as well as other information about the user(Tweet streams, csv files , xlsx, JSON, Opinion Polls)

What is the goal of your project?Goal:

/*In this project, we want to do a compound analysis to examine Taxi and Uber datasets. This analysis aims to extract patterns between the customers of Taxi and Uber services and find the correlations between them as well as the differences. The results of this analysis can be then used to further improve the services that Taxis and Uber provide and help them better understand their customers.*/

Conduct a sentiment analysis over Canadian twitter data relating to politics. Throughout our analysis, we wish to understand political sentiment across Canadian ridings and use the sentiment to identify what political party that specific region will be. We hope that our predictive model can be used as a platform by politicians..etc. Regions that have a high percentage of a poltical shift in the future. (regions that are susceptible to change)

We also aim to identify across the different ridings collectively what each riding values ex. Decrease taxes or make housing more affordable

What is the motivation for rigorous data analytics?

/*The need arises from the demand for a better analysis platform which can extract as much insight as possible about the behavior of customers in New York City. Furthermore, rigorous data analysis allows us to make predictions that can help the service providers improve and personalize their services. */

The motivation for our data analysis stems from helping politicians better understand what regions to focus their campaign on. We aim to provide as much insight as possible about the people's sentiment in Canada narrowed down by region/riding.

Through twitter sentiment analysis, (predicition)**Questions to answer:**

1. What is the political sentiment of a geo-location in Canada?
2. Which riding/city/province have a deep sentiment towards a political party?
3. What do people value in each riding/swing state (e.g. taxes, housing, education..etc.)
4. What's the percentage that a specific riding will have a political shift in the upcoming election
5. Ratio of each political party per geo location?

---

6. What percentage of users are politically active?

Why is the analysis important?

-Understand and gain insights from our data analytics to help political parties understand people's sentiments and values across Canada (in each riding).

----What factors effect political sentiment. e.g To what extent does housing rate effect political sentiment?For every factor we should be able to have **What are a few potential applications?**

Provide a platform for politicians improve their political campaigns and make data driven decisions to wiin people's hearts (political marketing), Create a platform to allow politicians to understand what drives sentiment.

Provide real-time analysis on the political status of ridings**The architecture of the Proposed Solution**Use Hadoop Distributed File System with a Spark distributed cluster to retain

scalability, high fault tolerance, and efficiency

- Use Spark's columnar storage layout to improve performance
- Perform MapReduce algorithms for aggregation using the Spark
- visualize data using Python's MatPlotLib and other data visualization modules

Description of the data collection/ingestion process, data storage, data processing, data serving and data visualization.

Data Ingestion:

- Stream tweets using Twitter's API Tweepy and Twitter Intiligence Tool (TWINT)
- Preprocess and clean the data on Spark through MapReduce algorithms

Data Storage:

- Spark's columnar storage
- RDMS

Data Processing:

- python modules dataframes pandas good stuff numpy
- create a predictive model using

Spark's MLlib

Data Serving/visualization:

- Provide geographic maps / graphs to serve data using Python's matplotlib and other modules

Overall architecture and data flow in the

System

Draw chart make sure to include streaming and batch together

Limitations and difficulties with the chosen approach

Only a small number of tweets have geolocation enabled which could provide incomplete data

Twitter's API Tweepy limitation on the number of tweets collected in certain time frame

Corelating Twitter's geolocation to the actual ridings

Tweets can be written in various languages, such as french, which can make it difficult to process (somehow). Solution: use google translate API to translate tweets and stuff! Cool! wooho

System evaluation and Data Analysis

How will you evaluate your system and architecture?

Evaluation will be done based on the following criterion.

1. Data cleaning and preprocessing will be evaluated by removing special characters, translating emojis, making sure tweet has no unncessary information (stop words)

2. Optimize Spark to deal with stragglers using different RDD actions/algorithms

3. Ensure our system and architecture is scalable and is able to process massive amount of data

4. Ensure our system can be generalized and re-used on different countries/dataset domains

5. ensure efficiency of the system

What results do you plan to obtain?Interactive map of political sentiment across regions in Canada.

Model to predict political shift in the upcoming election

Obtain the most important factors affecting people's sentiment (most frequent keywords)**What type of data analysis will you perform?**

Natural language processing, Sentiment analysis, machine learning classification on batch data, and time series graph analysis on streaming tweets

How this type of analysis is adequate for the data, problem and the issues posed?

Twitter data consists of content written in English which can be analyzed through natural language processing algorithms. Using machine learning algorithm to corelate sentiment to voter outcome of future election

What other datasets can be used?

Other social media platforms which provide intensive APIs for querying user feeds (e.g. facebook/instagram).

Our platform can be generalized to fit other countries' data by accomodating different variables such as language and riding locations

What are the steps you need to take to scale your solution?

Our solution can be scaled both vertically and horizentally

Vertical scaling:

Increasing computational power through better CPUs and GPUs

Horizental scaling: by adding more nodse to the spark cluster.