**3. System evaluation and Data Analysis**

**3.1 How will you evaluate your system and architecture?**

Data Quality: The data, being compromised mostly of tweets, is unstructured data. This requires the handling of special characters, emojis, and finding a balance in removing “stopwords” from these tweets to find only relevant “clean” information.

Performance Tuning: Optimize Spark to deal with stragglers by using RDD and DataFrame APIs.

Scalability: Ensure the tools being used can preform on a larger scale by running analysis on massive datasets.

Generalization: Ensure the system can be generalized to analyze different countries’ political sentiment as well as different data domains. Ensure the code is reusable, outlined and documented.

Overall Efficiency: Run stress tests to evaluate the ability of the system’s performance as well as identify programmatic bottle necks and attempt to resolve them.

**3.2 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 change

**3.3 What type of data analysis will you perform?**

Natural language processing, sentiment analysis, and machine learning classification as well as time series graph analysis will be preformed on batch data and streaming tweets.

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

The sheer amount of data would not be feasible for one machine and therefore Spark is used to distribute the data over a cluster. Twitter data consists of content written in English, which can be analyzed through natural language processing algorithms to determine sentiment towards a party or ideology. The use of machine learning algorithms can be used to create a prediction model that classifies different region’s voting outcomes based on sentiment and can therefore become a tool for political campaigns.

**3.5 What other datasets can be used?**

Other social media platforms which provide intensive APIs for querying user feeds (e.g. Facebook/Instagram) can be used to follow sentiment both in the past, and live.

The platform can be generalized to fit other countries’ data by accommodating different parameters such as language and riding locations.

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

Our solution can be scaled both vertically and horizontally.

Vertical scaling can be done by increasing computational power through better CPUs and GPUs, whereas horizontal scaling can be achieved by adding more nodes to the spark cluster.