EECS 4415 Project Proposal

Sentiment Analysis of Political Canadian Tweets

Matthew Longphee  
 York University

Raymond Barakat  
 York University

Sami  
 York University

1 **Domain Description and Motivation**

1.1 **What is the data domain?**

The data domain for the project consists of political tweets in Canada specifically splitting tweets among the different political ridings of Canada. The data set includes tweet contents (what the user has tweeted), username, user ID, the time and date the user posted the tweet, the location (latitude and longitude and/or city) and hashtags used.

1.2 **What is the goal of your project?**

Our goal over the course of this project is to conduct a sentiment analysis over Canadian twitter data relating to politics. Throughout our analysis, we aim 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 sentiment analysis model can be used as a platform by politicians to better understand what people value in Canada and to use the information to further improve their campaign.

1.3 **What is the motivation for rigorous data analytics?**

The motivation for our data analysis stems from helping politicians better understand the public perception of their campaign and improve public sentiment about their campaign and themselves. A single misquote or negative news about a candidate can be the difference between him/her winning or losing the election. It becomes key to have a platform to guide and direct the candidate’s campaign. As a result, we aim to provide an analysis platform that provides as much insight as possible about people’s sentiment of the Canadian election. The sentiment analysis is narrowed down by province/city/riding and what each of those locations value allowing candidates to improve their campaign and public sentiment.

1.4 **What are the questions that you want to answer?**Conference Short Name :WOODSTOCK’18

* What is the political sentiment distribution of a major geo-locations in Canada?
* Which riding/city/province have a positive/negative sentiment towards a political party?
* What do people value in each riding/swing state? (e.g. taxes, housing, education, etc.)
* What’s the percentage that a specific riding will have a political shift in the upcoming election?
* Time Analysis: How does the public sentiment about a candidate change over time?

1.5 **Why is the analysis important?**Conference Short Name :WOODSTOCK’18

The analysis is important because our data analytics help politicians understand and gain insights of people's sentiment and values across Canada. As a result, this helps politicians improve their campaign and public sentiment by guiding their campaign to prioritizing specific campaign activities.

1.6 **What are a few potential applications?** Conference Short Name :WOODSTOCK’18

* Provide a platform for politicians that analysis’s their campaign and public sentiment to help them better understand how people view their political campaign and allows them to make data driven decisions to further improve their public sentiment and campaign.
* Train different deep learning models to predict future election results using the data from our sentiment analysis as well as including other information such as past twitter sentiment analysis and voting behavior
* Creating a generalized twitter sentiment analysis model that can be used in different countries during their electionsConference Location:El Paso, Texas USA

2 **Domain Description and Motivation**ISBN:978-1-4503-0000-0/18/06

2.1 **Data analytics architecture**

***Stream Layer:***

* Real-time input of twitter streams using Twitter’s API (Tweepy)
* Use Spark Streaming, part of Apache’s language-integrated API, to provide scalable and fault-tolerant cleaning and processing of real-time tweet streams.
* Apply ML Algorithms using Spark’s MLib’s RDD-based and/or Dataframe-based APIs

***Batch Layer:***

* Store batches of raw data in Spark’s columnar storage layout.
* Store and process structured data in a relational database model using Spark SQL for efficient querying
* Clean and process stored data using MapReduce algorithms in Spark
* Create a predictive machine learning model with logistic regression and classification using Spark’s MLib

***Data Visualization:***

* Visualize data using D3, MatPlotLib, and other Python modules

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

***Data Collection/Ingestion:***

* **Streaming Data:** Receive tweet streams using Twitter’s API (Tweepy) and Twitter Intelligence Tool (TWINT)
* **Batch Data:** Receive massive amounts of batch data (e.g. old tweets, economic data...etc.) from Kaggle and web-scraped tweets.

***Data Storage:***

* Store cleaned and processed data using Spark's columnar storage layout
* Store structured tweets in a relational model to be queried fast

***Data Processing:***

* Clean and process data using MapReduce algorithms in Spark
* Run logistic regression on data to create a predictive ML model using Spark's MLib
* Use Pandas and other Python modules to process data efficiently

***Data Serving and Visualization:***

* Serve and visualize data in the browser as an interactive geographic map by integrating D3 and MatPlotLib modules with related Spark API

2.2 **Overall architecture and data flow in the system**

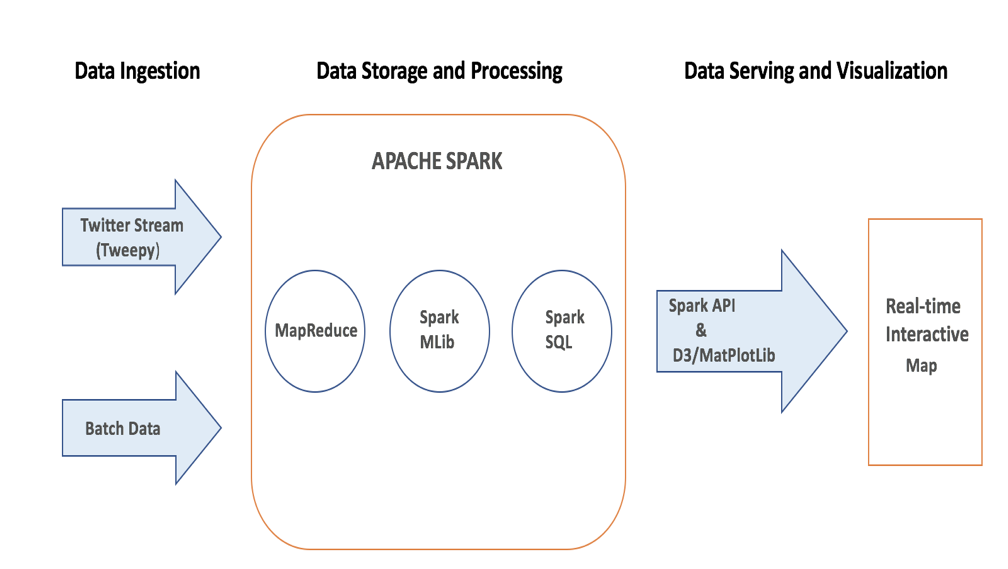
****Year:2018

Figure 1: **The Architecture and the flow of data in the proposed solution**

2.4 **Limitations and difficulties with the chosen approach**

* Streaming tweets with geo-locations can impose a limitation on the accuracy of the data because only a small number of users enable location permissions on Twitter.
* Since Canada is a bilingual country, streaming tweets in French may require additional translation/cleaning algorithms.
* Mapping Twitter's geo-locations with Canada's electoral districts (ridings) maybe a challenging task. However, given the latitude and longitude of all electoral districts, we will be able to map the ridings to the tweet location quite accurately.

3 **System Evaluation and Data Analysis**ISBN:978-1-4503-0000-0/18/06

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.

* 1. **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 performed on batch data and streaming tweets.

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

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.

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00