CSCI 5408, Summer 2018

**Final Project Report**

**Twit-Predict**

Due: August 5, 2018, 11:59PM

Andrea Christians (B00412494) and Navkaran Kumar (B00782012))

Dalhousie University

Link to GitHub: https://github.com/Navkaran0105/Twit\_Predict.git

Table of Contents

[Introduction and Problem Statement 4](#_Toc521249157)

[Business Idea 4](#_Toc521249158)

[Value Proposition 4](#_Toc521249159)

[Elections 5](#_Toc521249160)

[Choosing New Brunswick 5](#_Toc521249161)

[39th New Brunswick General Election 5](#_Toc521249162)

[Role-based distribution of work 5](#_Toc521249163)

[Work Breakdown 5](#_Toc521249164)

[Implementation 6](#_Toc521249165)

[Data Sources 6](#_Toc521249166)

[Data Analytical Tools 7](#_Toc521249167)

[Algorithms Used 7](#_Toc521249168)

[Visualizations 8](#_Toc521249169)

[Limitations of Current Work 9](#_Toc521249170)

[Future Work 10](#_Toc521249171)

[Critical Review 10](#_Toc521249172)

[Conclusions 10](#_Toc521249173)

[References 11](#_Toc521249174)

Table of Figures

Figure 1 4

Figure 2 5

Table 1 6

Figure 3 6

Figure 4 7

Figure 5 8

Figure 6 8

Figure 7 9

# Introduction and Problem Statement

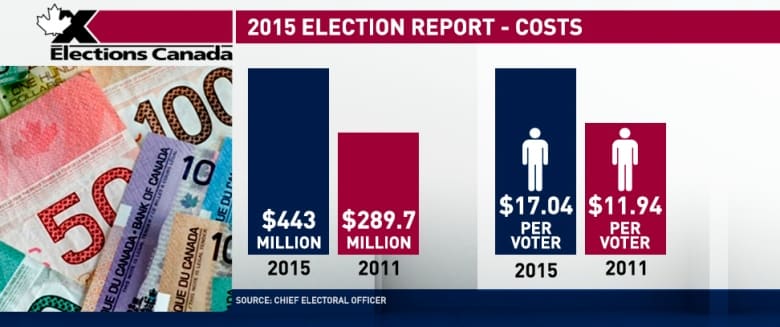
The result of any political campaign is the ability to sway public opinion. Campaigns are complex projects that require significant amounts of manpower, time, resources and money. They also appeal to a wide geographical demographic and thus large amounts of travel in order to engage with the local citizens that they are appealing to vote. Social Media is a powerful tool to extract public opinion in any regard and this can specifically be targeted for the benefit of politics. By studying the sentiments of a tweets in a given location, a consensus on whether a candidate is positively or negatively received in that area can be determined. Political predictions could yield benefits for both citizens and the candidate. Specifically, citizen maybe be more inclined to vote if the party that they oppose is in the lead. For the candidate, they could adjust their campaign to target specific demographic in an area. In this report, we will explain and discuss the workflow of creating Twit-Predict from a case study perspective for the New Brunswick General Election.

# Business Idea

Twit-Predict is a political prediction tool that foresees the success of given candidates in a given location through machine learning algorithms. By using machine learning in conjunction with customized training and test data, Twit-Predict can gauge the public opinion of a given candidate or party in a region. This information can be used to adjust or redirect a campaign.

# Value Proposition

During the 2015 Federal Election in Canada, the political parties spent more than $400 million on their political campaign (Fig. 1). Since private funding is not acceptable in Canada (as it may encourage corruption), these campaign costs are funded from the taxpayers. By using Twit-Predict, a campaign could have a better understanding of public perception. By understanding their voters, they can adapt their campaign to target that demographic to acquire more voters. This would reduce, time, manpower and most importantly, money. For example, by running the model directly after a political rally a candidate would almost have immediate feedback of their reception in that location. A shift up in percentages for example, may indicate that the candidate was well received in that area and they could perhaps move on with their campaign. If, however, there was a dramatic decrease in percentage, the candidate may choose to spend more time in that location rather than moving to a different area that is favourable to him already.



**Figure 1:** The cost of elections to Canadians [1].

For the citizens there is a dual benefit, and that is the decreased cost of the election due selective and careful planning by using the model and if the results are published, they could see whether or not their favoured candidate is winning or losing. If their candidate is losing, they may for example be more inclined to vote.

# Elections

## Choosing New Brunswick

Initially, the idea was to choose an international election because of the large amount of social media that would be available with that election. However, it quickly became clear that it was difficult to find a federal election in the upcoming months for us to currently run our model on. One example that we considered was an election that was taking place in Sweden, however the language barrier would have been a huge issue for our model so we did not use foreign countries. In retrospect, we could have run a model in different language (Swedish, for example), however finding a political lexicon in Swedish would have proven to be very difficult as was already challenging finding one in English. Furthermore, searching for the relevant and trending topics would have been more complex.

The Annual New Brunswick Election will be held on the 28 September 2018 [2]. This election was strategically chosen for several reasons. Firstly, the election is about two or three months shy of our project, so there was a lot of hype on social media about the candidates. Secondly, it was Canadian (local) and also in English to make it easier for our model to recognize the dictionary of words.

## 39th New Brunswick General Election

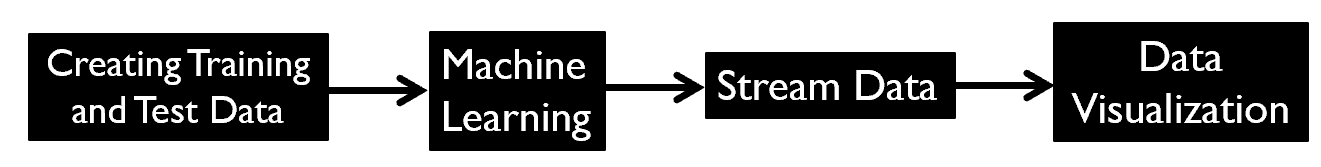
Currently the New Brunswick election will have two main competing parties, the Liberal and Conservative Parties. Although the Liberals won the last election, it was a narrow win. As a result of this, Brian Gallant is the current Premier and he is also running for his second term against Blaine Higgs of the Conservative Party. Currently, the Liberal party holds 27 seats (as the majority) and the Conservative and Green Party hold 21 and 1 seat respectively although this may shift with the upcoming election [2]. Twenty-five seats are required for a majority.

# Role-based distribution of work

Andrea was the Data Engineer and gathered the data from News API and tweets. This entailed creating the training and test data using several python scripts. Nav was the Data Scientist and trained the model using logistic regression with Apache Spark (on a local machine) and then streamed/labelled the tweets with the model. Finally, Andrea did the final visualization in PowerBI.

# Work Breakdown

Due to the one month project deadline, we had four, one week sprints. The initial phases were data collection, followed by training the model (machine learning), streaming the data, and then finally the visualization stage (Fig. 2).



**Figure 2:** Simple depiction of the workflow in this project.

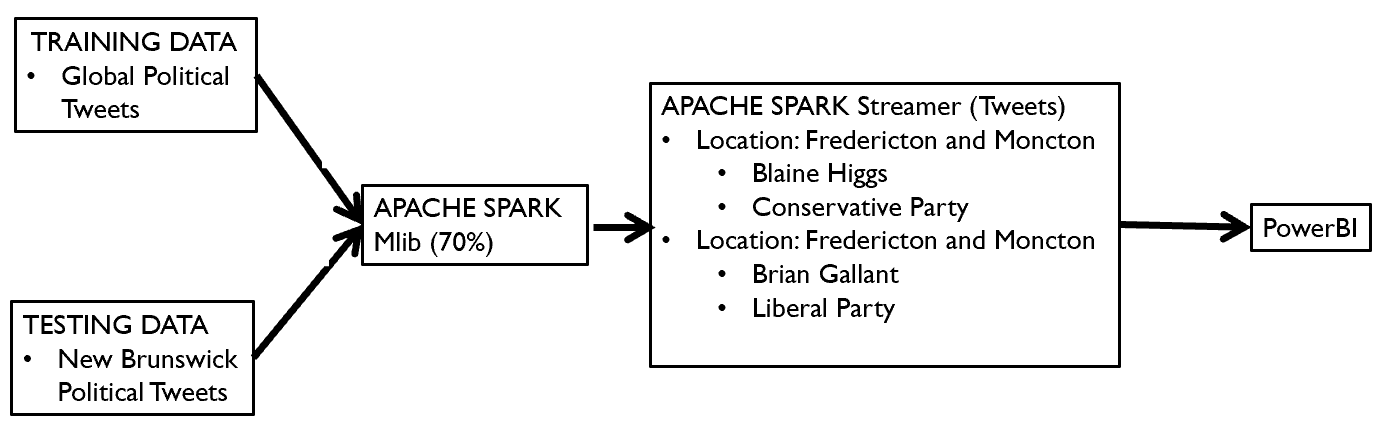
The sprints are listed below in Table 1. Week 1 was finding a political lexicon, finding sources of for the training and test data and then streaming it. Week 2 was more complex and entailed running sentiment analysis on the imported data and training the model. During week 3, we were able to run model with the labelled tweets and import the data into PowerBI. The final sprint entailed final analysis of the PowerBI charts.

|  |  |  |
| --- | --- | --- |
| Sprint # | Date | Tasks |
| 1 | 4-11 July | Find a political lexicon  Find training and test data sources |
| 2 | 12-18 July | Run Sentiment Analysis on training and test data  Train the model |
| 3 | 19-25 July | Run the model with the tweets  Import data into PowerBI |
| 4 | 26-2 August | Final analysis and visualization |

**Table 1:** Details for Sprint’s 1 to 4.

# Implementation

The implementation can be broadly divided into 4 parts namely: creating training and test data, model training, live twitter stream labelling and visualization (Fig. 2). A detailed view of the ETL pipeline is depicted in Figure 3 and will be discussed in more detail in this section. These have been in covered in this section as Data Sources, Data Analytical Tools, Algorithms and Visualization.



**Figure 3:** Overall view of the ETL process in the project.

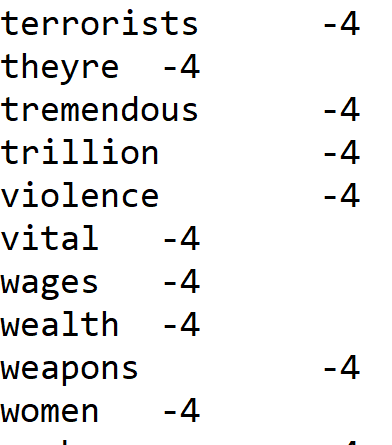
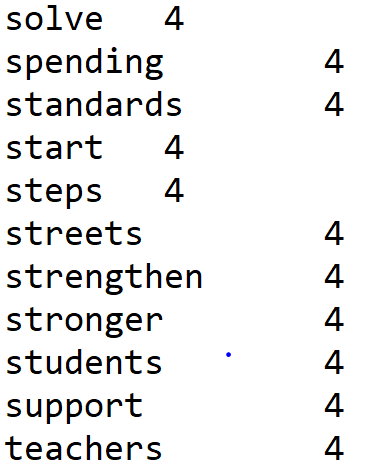
## Data Sources

Sprint 1 entailed searching for sources of data, specifically training and test data and a lexicon. We initially searched for general training and test data with sentiments that we could use with our model but could only find datasets with generic classifications (Fig. 3). Since we wanted to customize our dataset to political scenarios it was important to find a training and test data that was customized for this project and thus we created our own with various python scripts.

Training data consisted of tweets imported with ‘#ABCPolitics’, ‘#FoxPolitics’ which gave us a rich vocabulary of the political keywords. Additionally, by using popular news, a large amount of data was able to be gathered.

Twitter was the main source of raw data and along with the News API, however only gave around 30 headlines in one stream. In contrast, Twitter gave large amounts of data. The data came in the form of json and we had difficulty parsing the file. Due to the limited amounts of entries, it was cleaned manually. Test data was imported from Twitter using ‘#nbpoli’ and this yielded tweets about the New Brunswick elections.

To customize our training and test data, it was important to have a specific set of lexicons which had political keywords and associated weighted emotions i.e. a political lexicon. We were able to locate a blog with a list of political lexicons in English [3]. Unfortunately, there wasn’t a download link in the blog, so we copy-pasted the list into a text file and then ran cleaning python scripts to convert the lexicon text file into a dictionary. Using the lexicon, we were able to run sentiment analysis with the raw data collected. The lexicon emotion ranged from 4 to -4 which allowed us to gauge a range of positive and negative emotions in the raw data (Fig. 3).



**Figure 4:** Sample view of the political lexicons with a range of 4 to -4. It contains words more common to politics such as violence, weapons and terrorists (as a negative context).

## Data Analytical Tools

Data Analytics were conducted using the Machine Learning Library (MLib) of Apache Spark. We chose Logistic Regression for training and final prediction of the data since we did not much variation in the emotions. Our sentiment analysis script used the poltical lexicon and removed neutral interpretations. Thus Binary Classification for labelling was used as we had only two emotions (positive and negative) when live streamed and labelled.

## Algorithms

Sentiment Analysis was run on both training and test data for data preparation. A political lexicon dictionary was used for finding out accumulative sentiment of each line in our raw data. Once this was completed, the data was labelled as ‘positive’ and ‘negative’ for both training and test data. We chose to remove neutral as it is assumed an individual would have polar responses to a candidate, i.e. either negative or positive. Positive meaning they would vote in favour or negative, against the candidate.

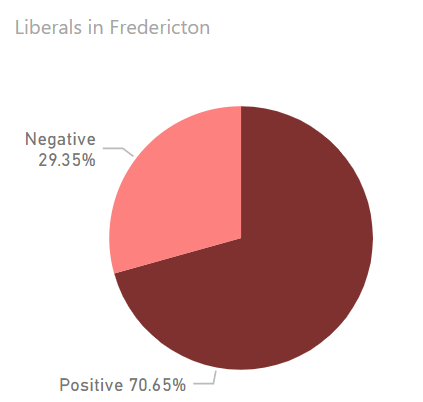
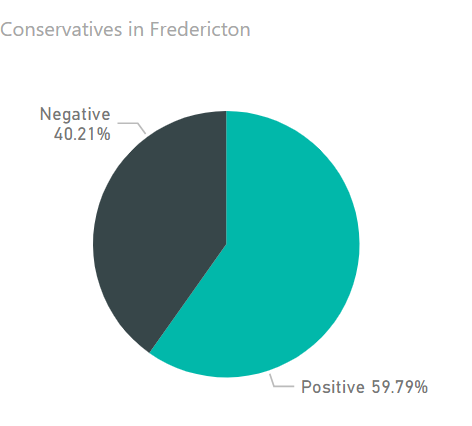
The sentiment was labelled as positive if the total score was greater than 0 and negative if total score was less than 0. After running sentiment analysis using lexicons, the final csv contained two fields with its tweet and sentiment.

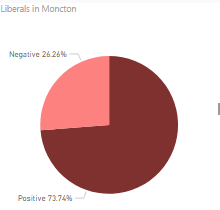
At this point, this data was run with the machine learning model. We chose to go with Logistic Regression’s ‘Binary Classification Evaluator’ because we were concentrating on positive and negative emotions regarding a candidate. The prepared training data was imported into the model and it was then tested for its efficiency in prediction by feeding the prepared test data. It came out to be approximate of 70% when global vocabulary (training data) was tested with the local vocabulary of New Brunswick (test data).

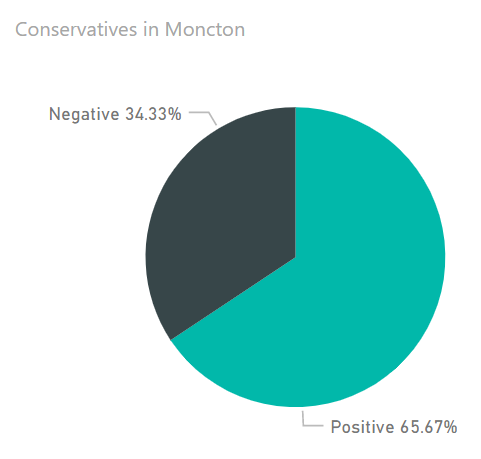
In order to determine the opinion of a candidate in a region the tweets were streamed for a location labelled with our trained model. The results were saved as batches of csv which were later merged for visualisation. The tweets were filtered and streamed for three scenario’s i.e. tweets for a political candidate in Moncton, Fredericton and overall New Brunswick. There was difficulty in merging csv’s because of the naming conventions when a batch is labelled and saved and thus it was merged manually for presenting the tweets in a graphical presentation.

## Visualizations

Visualization has broadly divided into two main scenarios. Firstly, the predictions for both of the candidates in Moncton and Fredericton. Both of the candidates (Blaine Higgs and Brian Gallant) were plotted with the data visualization tool, Power BI using PI charts. The PI charts created a percentage for public responses. This was done for both Conservative and Liberal parties in both regions (Fig.’s 5 and 6).

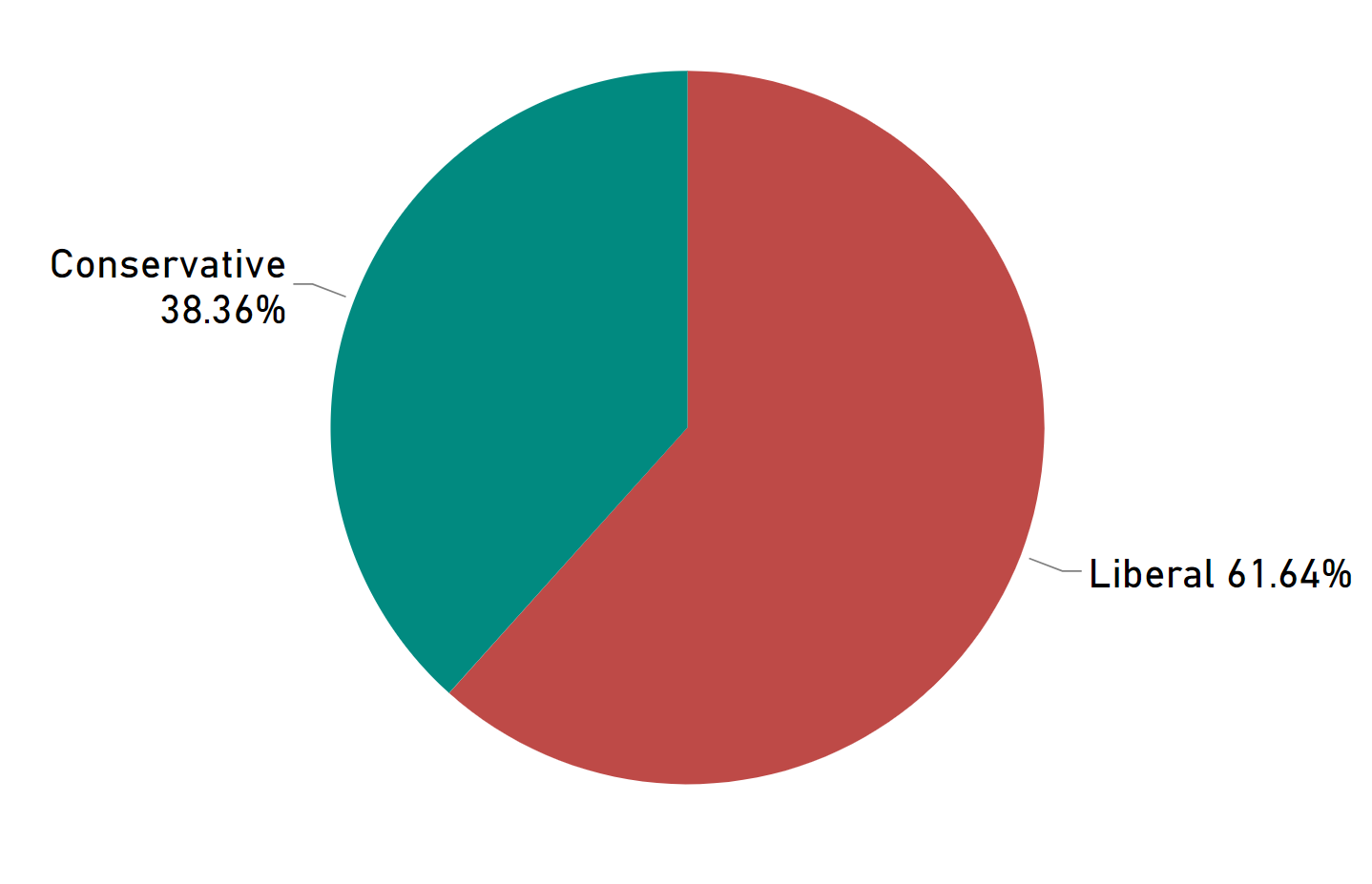


**Figure 5:** Current depictions of public approval for political candidates in Fredericton. The right plot shows the liberal and left shows the conservative ratios.



**Figure 6:** Current depictions of public approval for political candidates in Moncton. The right plot shows the liberal and right shows the conservative ratios.

The final step was to gauge the overall standing for both the parties in New Brunswick (Fig. 7). In order to do that, tweets were streamed for both parties in overall New Brunswick. The batches for both the parties were merged into a single csv respectively for each candidate. Positive tweets from both the csv’s were filtered out and put into a new csv. For each party, their ‘positive’ sentiment was replaced with Liberal or Conservative respectively. The final merged csv was then plotted in Power BI and grouped by party name.



**Figure 7:** Final political predictions in New Brunswick.

# Limitations of Current Work

We had several limitations in our workflow that were quickly apparent. Firstly, there were limitations based on the size of the political lexicon. Sentiments can only be predicted on words matched in the lexicon for our training and test data and there is an underlying assumption that all words are caught and gauged correctly prior to importing into the model. We are aware that this is not the case and thus is it is a limitation in our results.

Tweet language and spelling changed from user to user and the lexicon assumes correct spelling where the model assumes a uniform type of spelling. In reality, not all users are of same age, education and background and this may cause large variations in spelling and use of language. The model and lexicon would not be able to handle these scenarios.

Another issue was with the filtering of our tweets at the final stage of project. Once we had trained our data, we had difficulty filtering the tweets to a location and topic i.e. there were some discrepancies in the data streamed. Specifically, data from topics like American elections headlines, and other world elections were present and these entries are anomalous in our data. Furthermore, if a user chooses to tag #nbpoli and write nothing about a given candidate, it may change the result. We also realize that tweets from an official news page (used for training data) and personal tweets are different in style and this may influence our result as well.

Finally, our application is missing a user-interface and it would be near impossible for an average person to configure it.

# Future Work

Absence of a working front-end is a big limitation of our analysis process. It could help us present our idea better with minimal interaction of a user with the python code which would allow the user to effortlessly get results.

A subset of creating the UI, is figuring out how to merge the csv batches that are labelled and saved by model. This would streamline and automate the ETL process to process the prediction.

# Critical Review

Although we were able to create our own training data, one large issue with our process is the lack of automation. Furthermore, if we had more time, it would have been better to read through the training and test data to ensure that all of it was relevant. It was apparent that tweets were not always political and this would create anomalies in our data.

# Conclusions

In conclusion, we were able to create customized training and test data that improved the results of our final. Results-wise, as anticipated, the Liberals have a lead in the polls when compared to the conservatives and this was to be expected as they won the previous election and appear to be historically favoured. We did our best to customize the results for our project to ensure optimal results.

# References

[1]"New Brunswick election is now just three months away", Global News, 2018. [Online]. Available:

https://globalnews.ca/news/4293738/new-brunswick-election-2018/. [Accessed: 31- Jul- 2018].

[2]"Marathon 78-day election cost taxpayers $443 million says Elections Canada | CBC News", *CBC*, 2018. [Online]. Available: https://www.cbc.ca/news/politics/elections-canada-443-million-1.3436139. [Accessed: 31- Jul- 2018].

[3]C. Vail, "MBA 676: Political Sentiment Lexicon", *Rstudio-pubs-static.s3.amazonaws.com*, 2018. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/338458\_3478e1d95ccf49bf90b30abdb4e3bd40.html. [Accessed: 04- Aug- 2018].