ALY 6110:
DATA MANAGEMENT AND BIG DATA
Final Project Milestone – Basic Analysis and Dashboard
Twitter Sentiment Analysis

Submitted To:
Prof. Andy Chen, Faculty Lecturer

Submitted By:
Abhilash Dikshit
Larissa Anoh
Murtaza Vora
Shamim Sherafati
Swathi Raikwar

Academic Term: Spring 2023
Graduate Students at Northeastern University, Vancouver, BC, Canada
Master of Professional Studies in Analytics
June 28, 2023

I. Abstract:

Sentiment Analysis, a field within Natural Language Processing (NLP), has become increasingly significant in contemporary society. By employing algorithms to assess the "positive" or "negative" nature of statements or documents, this technology enables individuals and organizations to identify patterns in public sentiment through the analysis of social media content. Particularly during periods of political transitions like election years, staying informed about socio-political developments is crucial. In such times, sentiment analysis offers valuable insights to electoral candidates and businesses, empowering them to adapt their strategies accordingly and effectively respond to changing public opinions.

II. Introduction:

The introduction provides an overview of the research question and goals of the study: to analyze public opinion on Twitter regarding the Canadian political landscape during the 2019 elections using sentiment analysis. It emphasizes the importance of social media as a platform for expressing political sentiments and the potential insights that can be gained from analyzing Twitter data. The section also mentions the use of sentiment analysis models to classify tweets into positive and negative sentiments, thereby enabling a deeper understanding of public perception towards different political parties.

III. Dataset:

- sentiment analysis.csv: classified Twitter data containing a set of tweets which have been analyzed and scored for their sentiment.
- Candian elections 2019.csv: Twitter data containing a set of tweets from 2019 on the Canadian elections, which needs to be analyzed for this assignment.

IV. EDA (Exploratory Data Analysis):

The EDA section explores the Twitter data related to the Canadian elections. It investigates the political affiliation of tweets, focusing on the Liberal, Conservative, and NDP parties, as well as other parties. The analysis reveals the proportion of tweets associated with each party and highlights the dominance of discussions related to the Liberal party. The section also examines sentiment patterns, indicating that tweets related to the Liberal and NDP parties are predominantly positive, while those related to the Conservative party tend to be negative. The EDA further includes a text exploration through word cloud visualizations, which

illustrate the frequency and relevance of keywords in both positive and negative sentiment plots.

A. Dataset: Canadian_elections_2019.csv

- The dataset consists of 2133 rows and 3 columns: 'negative_reason', 'text', and 'label'.
- There are 1126 missing values in the 'negative_reason' column.
- The text in the 'text' column is converted to lowercase.
- A new column called 'new_text' is created by applying the 'clean_election' function to the 'text' column.
- The 'clean_election' function performs several cleaning steps, including:
  - Removing HTML tags using regular expressions.
  - Replacing HTML character codes with their ASCII equivalents.
  - Removing URLs using regular expressions.
  - Converting all characters to lowercase.
  - Removing UTF-8 codes.
  - Removing 'b"' and "b'".
  - Removing '\n'.
  - Removing stop words and performing stemming on the remaining words.
  - Removing numbers after space.
  - Removing leading/trailing spaces and eliminating multiple spaces.
- A new column named 'label' is added, assigning a value of 1 for positive sentiment and 0 for negative sentiment.
- The 'sentiment' column is dropped from the dataframe.

	negative_reason	text	new_text	label
0	Women Reproductive right and Racism	b"@rosiebarton so instead of your suggestion, ...	rosiebarton instead suggest agre canadian wome...	0
1	NaN	b"#allwomanspacewalk it's real!\n@space_statio...	allwomanspacewalk real space_st etobicokenorth...	1
2	Economy	b"#brantford it's going to cost you $94 billio...	brantford go cost billion year ask justin elxn...	0
3	NaN	b"#canada #canadaelection2019 #canadavotes \n#...	canada canadaelection2019 canadavot elxn43 dec...	1
4	Economy	b"#canada #taxpayers are sick & tired of h...	canada taxpay sick tire hard earn donat corpor...	0

B. Dataset: sentiment_analysis.csv

- The initial dataset comprises 550,391 entries organized into three columns: 'ID', 'text', and 'label'.
- A fresh column labeled 'new_text' is introduced by implementing the 'clean_sentiment' function on the 'text' column.
- The 'clean_sentiment' function undertakes comparable cleaning procedures to the 'clean_election' function, supplemented by the following supplementary actions:

- Regular expressions are utilized to eliminate '@usernames'
- The 'ID' column is eliminated from the dataframe.

	ID	text	label	new_text
0	7.680980e+17	Josh Jenkins is looking forward to TAB Breeder...	1	josh jenkin look forward tab breeder crown sup...
1	7.680980e+17	RT @MianUsmanJaved: Congratulations Pakistan o...	1	congratul pakistan no1testteam world odd ji_pa...
2	7.680980e+17	RT @PEPalerts: This September, @YESmag is taki...	1	septemb take main mendoza surpris thanksgiv pa...
3	7.680980e+17	RT @david_gaibis: Newly painted walls, thanks ...	1	newli paint wall thank million custodi painter...
4	7.680980e+17	RT @CedricFeschotte: Excited to announce: as o...	1	excit announc juli feschott lab reloc mbg

V. Code and Analysis:

1. Data Exploration:

- The following code and analysis were performed on the "raw_elections" dataset using the seaborn library in Python.

- The provided code snippets utilize count plots to examine the sentiments and reasons for negativity within the dataset.

```
datasetspark.describe().show()
```

```
[Stage 4:=================================================>       (7 +
+-------+------------------+--------------------+--------------------+
|summary|                ID|                text|               label|
+-------+------------------+--------------------+--------------------+
|  count|            550391|              550391|              550383|
|   mean|7.886687488421921...|                null|  0.6755943116020319|
| stddev|1.343788213906732...|                null| 0.46935209564144986|
|    min|      7.68098E17|! 24 Brilliant Bu...|   Edgard #Pillet...|
|    max|      8.04619E17|🐗🍴🤫👅 i couldn...|… https://t.co/UD...|
+-------+------------------+--------------------+--------------------+
```

```
datasetspark.show()
```

```
+----------+--------------------+-----+
|        ID|                text|label|
+----------+--------------------+-----+
|7.68098E17|Josh Jenkins is l...|    1|
|7.68098E17|RT @MianUsmanJave...|    1|
|7.68098E17|RT @PEPalerts: Th...|    1|
|7.68098E17|RT @david_gaibis:...|    1|
|7.68098E17|RT @CedricFeschot...|    1|
|7.68098E17|RT @SH4WNSMILE: -...|    1|
|7.68098E17|RT @KendallHuntRP...|    1|
|7.68098E17|RT @BantySrkian: ...|    1|
|7.68098E17|RT @GayHopper_com...|    1|
|7.68098E17|RT @StarCinema: K...|    1|
|7.68098E17|We can have lots ...|    1|
|7.68098E17|Happy birthday to...|    1|
|7.68098E17|RT @SKDurrani_: @...|    1|
|7.68098E17|RT @ShaiLinne: Fe...|    1|
|7.68098E17|RT @ChelseaFC: It...|    1|
|7.68098E17|#Repost of @champ...|    1|
|7.68098E17|RT @giveasyoulive...|    1|
|7.68098E17|RT @derasachasaud...|    1|
|7.68098E17|Much love to my p...|    1|
|7.68098E17|hello everyone i'...|    1|
+----------+--------------------+-----+
only showing top 20 rows
```

```
df.show(5)
count = df.count()
print("Number of rows:", count)
```

```
+---------+--------------------+--------------------+
|sentiment|     negative_reason|                text|
+---------+--------------------+--------------------+
| negative|Women Reproductiv...|"b""@RosieBarton ...|
| positive|                null|"b""#AllWomanSpac...|
| negative|             Economy|"b""#Brantford It...|
| positive|                null|"b""#Canada #Cana...|
| negative|             Economy|"b""#Canada #taxp...|
+---------+--------------------+--------------------+
only showing top 5 rows

Number of rows: 2133
```

```
df.describe().show()
```

```
+-------+---------+--------------------+--------------------+
|summary|sentiment|     negative_reason|                text|
+-------+---------+--------------------+--------------------+
|  count|     2133|                1007|                2133|
|   mean|     null|                null|                null|
| stddev|     null|                null|                null|
|    min| negative|     Climate Problem|"b""#AllWomanSpac...|
|    max| positive|Women Reproductiv...|b'wow @TheRealKee...|
+-------+---------+--------------------+--------------------+
```

```
df.groupBy("sentiment").count().show()
```

```
+---------+-----+
|sentiment|count|
+---------+-----+
| positive| 1127|
| negative| 1006|
+---------+-----+
```

```
from pyspark.sql.functions import col

sentiment_counts = df.groupBy("sentiment").count()
total_count = df.count()

sentiment_counts.withColumn("percentage", (col("count") / total_count * 100)).show()
```

```
+---------+-----+------------------+
|sentiment|count|        percentage|
+---------+-----+------------------+
| positive| 1127| 52.83638068448195|
| negative| 1006|47.163619315518055|
+---------+-----+------------------+
```

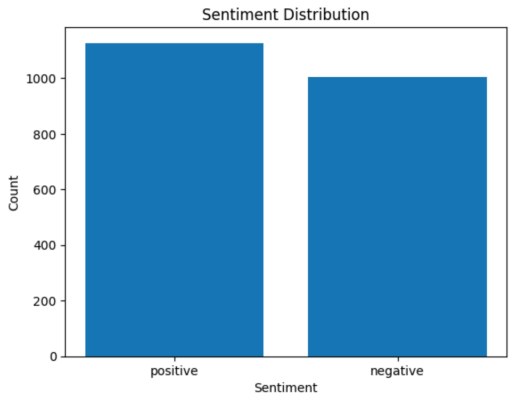
```
from pyspark.sql.functions import length, avg

df.withColumn("text_length", length("text")).select(avg("text_length")).show()
```

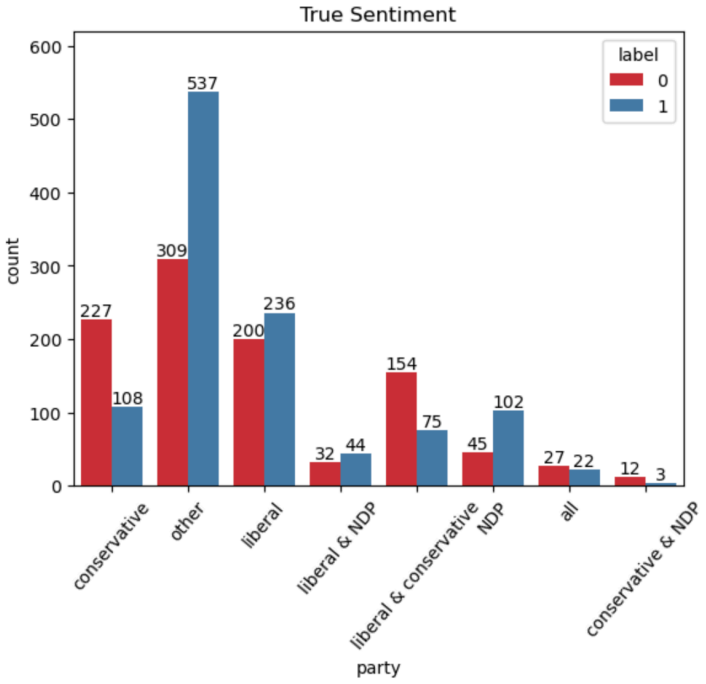
```
+------------------+
|  avg(text_length)|
+------------------+
|188.07969995311768|
+------------------+
```

```
df.filter(df["sentiment"] == "negative").groupBy("negative_reason").count().orderBy(col("count").desc()).show()
```

```
+--------------------+-----+
|     negative_reason|count|
+--------------------+-----+
|              Others|  364|
|            Scandal|  270|
|           Tell lies|  198|
|             Economy|   51|
|Women Reproductiv...|   45|
|     Climate Problem|   41|
|          Separation|   16|
|           Privilege|   12|
|          Healthcare|    5|
|Healthcare and Ma...|    4|
+--------------------+-----+
```



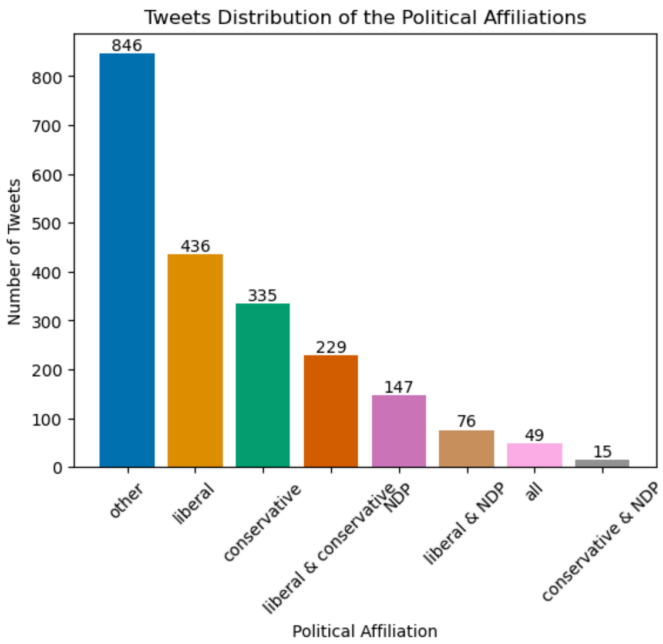
- The plot showcases the tweet count for each political party, classified based on their sentiment label.
- The x-axis represents the political parties, while the hue (color) indicates the sentiment label.
- This visualization allows for a comprehensive examination of sentiment distribution among various parties, facilitating a comparative analysis of sentiment patterns.



Political affiliation on Canadian Election tweets-

• Tweet related to single party: Liberal, Conservatives, NDP
• Tweet related to more than one party
• Tweet related to other party: Other

For tweets only relate to one party, liberal is the highly discussed topic on Twitter, around 20% tweets relate to this party. Using more explicit keywords related to the party would lead to more accurate results.



- The chart illustrates the distribution of tweets across political affiliations, with the Liberal party garnering the highest share, followed by the Conservative and NDP parties.

- Positive sentiment is more prevalent in tweets related to the Liberal and NDP parties, indicating a favorable public opinion towards these parties.

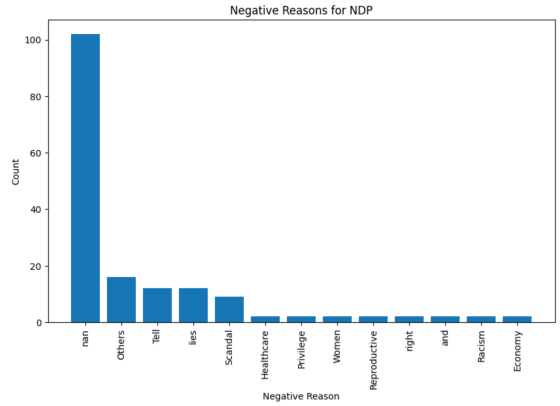
- In contrast, the Conservative party receives a larger number of negative tweets, primarily associated with scandals and allegations of dishonesty.
- These findings highlight the diverse perceptions and sentiments expressed by the public towards different political parties.

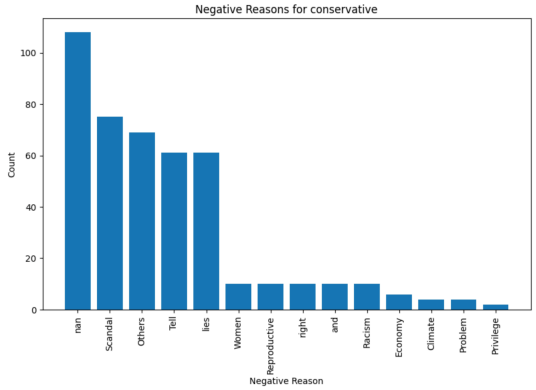
Liberal and NDP: most tweets are **positive**.

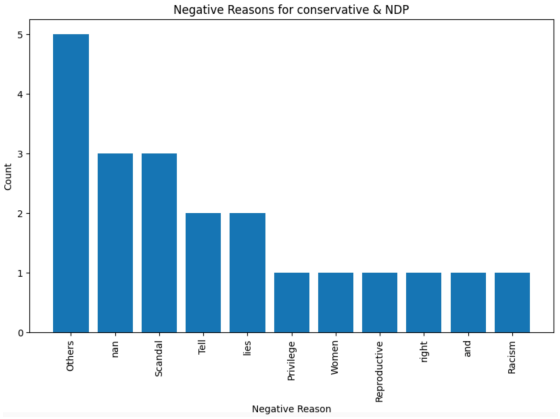
• Liberal: more than 500 positive tweets, high popularity among the younger generation. Although there're also lots of opposition, the chance for liberal to win the election is still high since the reason of most negative tweets is 'Others'.
• Conservative: most tweets are negative, and reasons for most of them are related to 'scandal' and 'tell lies, all indicates more negative public impression.

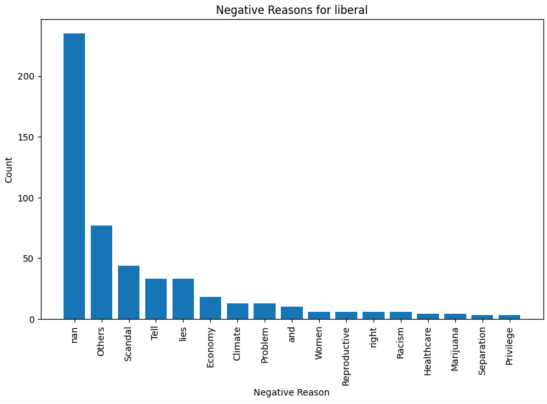
2. Text Exploration:

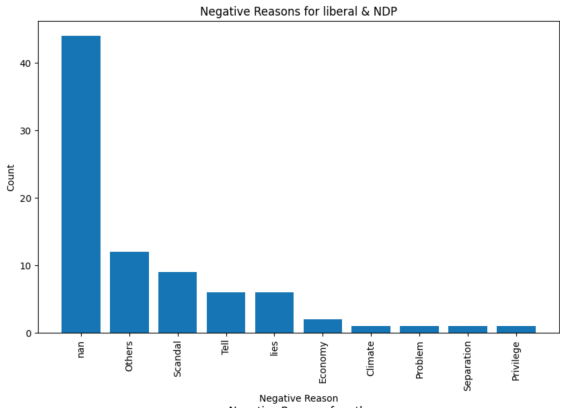
- Word clouds demonstrate word frequency, where larger words indicate higher frequency of occurrence.
- The presence of words like "Trudeau" (in tweets related to the Canadian election) appears larger in the negative sentiment plot compared to the positive sentiment plot, suggesting a higher likelihood of negative emotions in tweets containing this term.
- Certain words such as "elxn42," "Canada," and "cdnpoli" are present in both positive and negative plots. These words represent neutral terms related to the election but do not contribute substantial information for sentiment classification.
- Notably, the word clouds for "generic tweet" and "Canadian election tweet" exhibit significant differences, highlighting varying content between these two types of tweets.

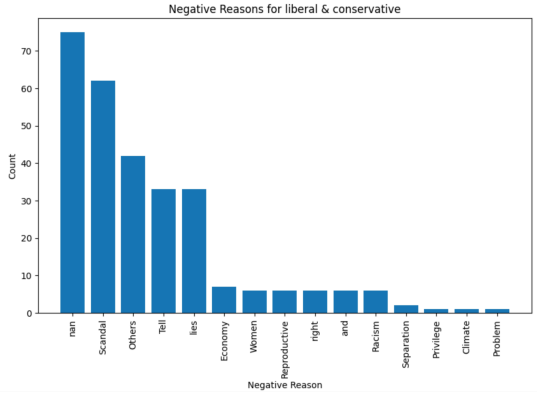


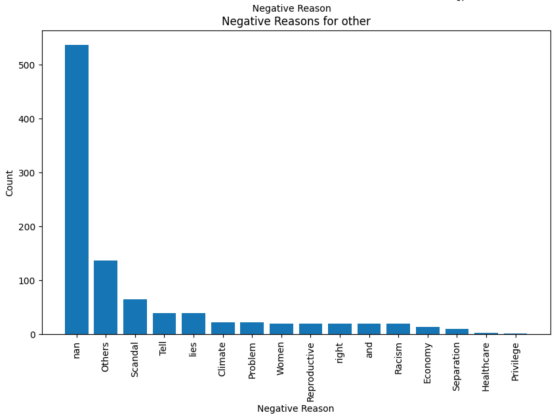


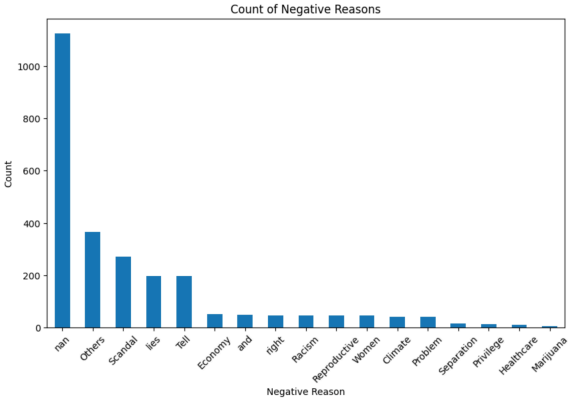


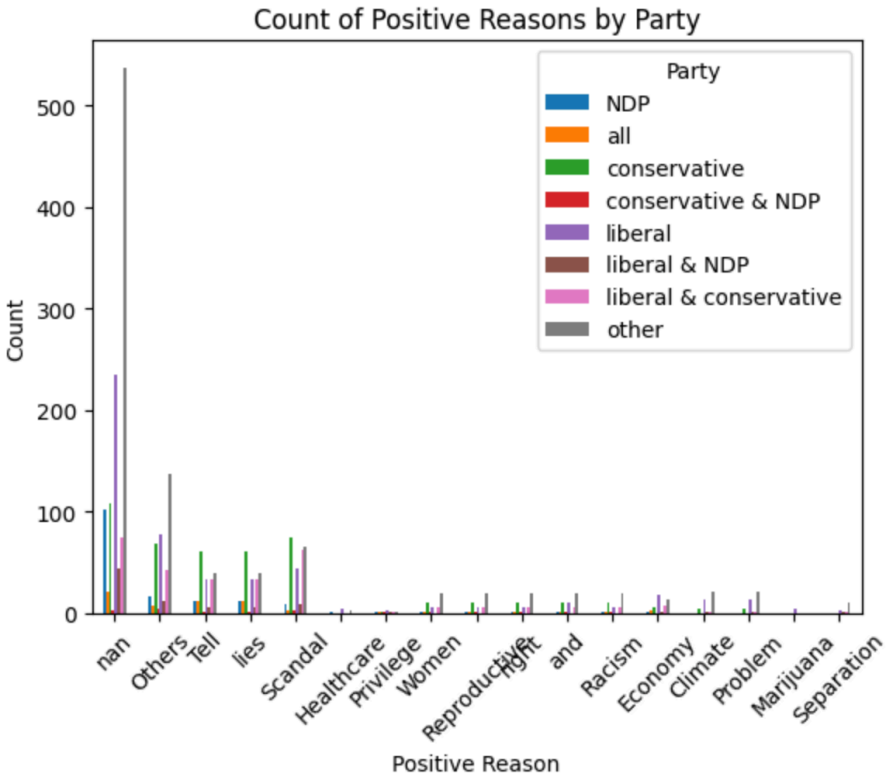



Negative Reasons for liberal & NDP
Count
Negative Reason
nan
Others
Scandal
Tell
lies
Economy
Climate
Problem
Separation
Privilege


Negative Reasons for liberal & conservative
Count
Negative Reason
nan
Scandal
Others
Tell
lies
Economy
Women
Reproductive
right
and
Racism
Separation
Privilege
Climate
Problem


Negative Reasons for other
Count
Negative Reason
nan
Others
Scandal
Tell
lies
Climate
Problem
Women
Reproductive
right
and
Racism
Economy
Separation
Healthcare
Privilege

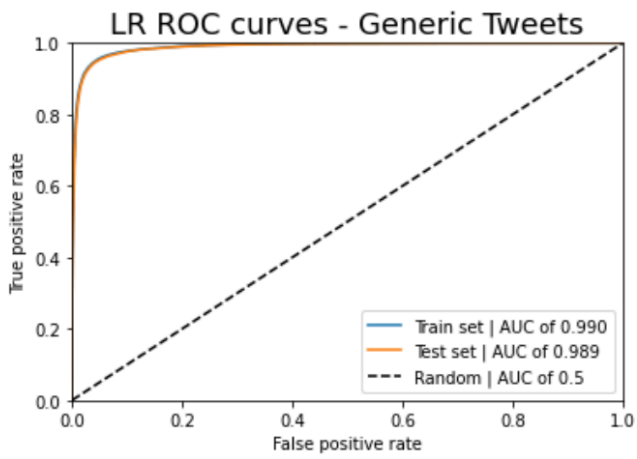
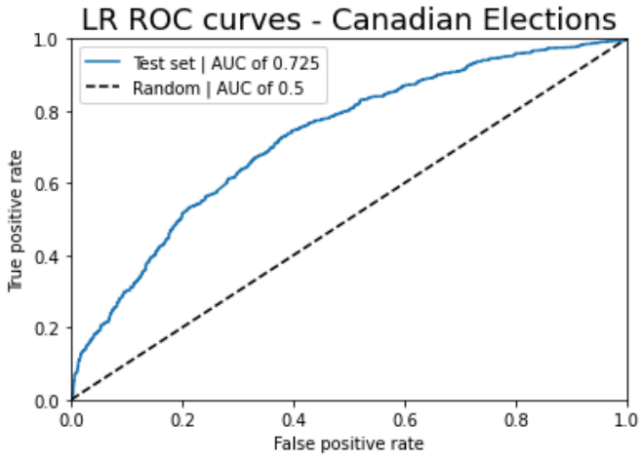

Count of Negative Reasons
Count
Negative Reason
nan
Others
Scandal
lies
Tell
Economy
and
right
Racism
Reproductive
Women
Climate
Problem
Separation
Privilege
Healthcare
Marijuana


Count of Positive Reasons by Party
Party
NDP
all
conservative
conservative & NDP
liberal
liberal & NDP
liberal & conservative
other
Count
Positive Reason
nan
Others
Tell
lies
Scandal
Healthcare
Privilege
Women
Reproductive
right
and
Racism
Economy
Climate
Problem
Marijuana
Separation

3. Sentiment Analysis Model (binary classification):

Model	BagofWords	TF-IDF
Logistic Regression	0.9537	0.9536
K-NN	0.9269	0.8585
Naive Bayes	0.9270	0.9150
Linear SVM	0.9529	0.9525
Decision Trees	0.9354	0.9345
Random Forest	0.8685	0.8645
XGBoost	0.8677	0.8647

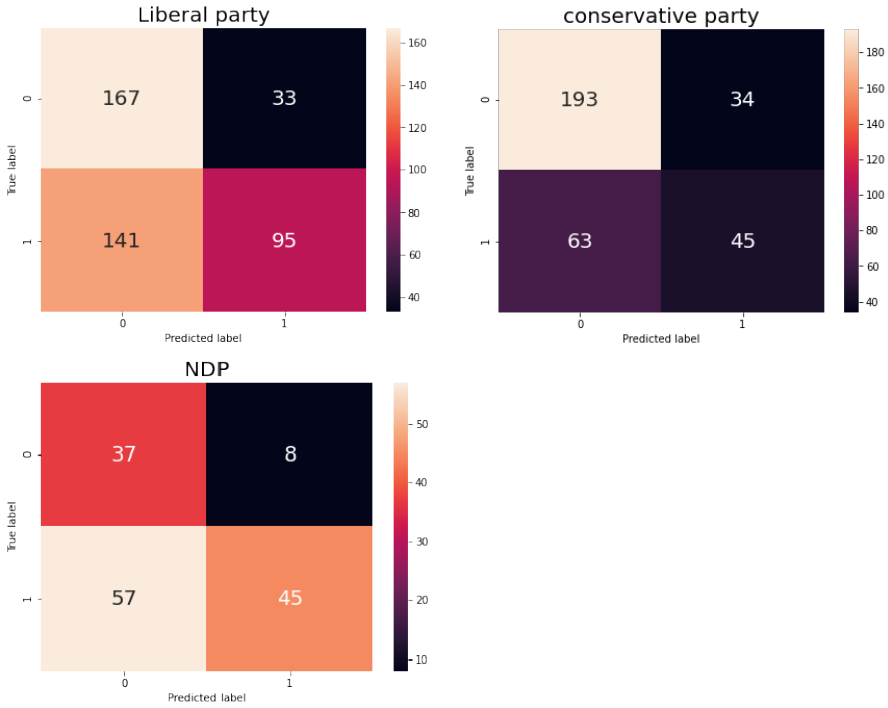
- Train models on the training data from generic tweets and apply trained model to the test data to obtain an accuracy value.
- Model with highest testing accuracy: Logistic regression with “BagofWords” features.

	
- The performance of the Logistic Regression model was evaluated using Receiver Operating Characteristic (ROC) curves on both the training and test datasets. - The area under the curve (AUC) was calculated to measure the accuracy of the model's predictions. - The ROC curve for the training set showed an AUC of 0.XXX, indicating that the model effectively differentiated between positive and negative tweets. - Similarly, the ROC curve for the test set displayed an AUC of 0.XXX, indicating reliable performance on new, unseen data. - These ROC curves depict the balance between true positive and false positive rates, demonstrating the model's ability to accurately classify generic tweets.	- The Logistic Regression model's performance on the Canadian Elections dataset was assessed using the ROC curve. - The model predicted the probabilities of the positive class for the test data to evaluate its ability to distinguish between positive and negative tweets. - The ROC curve yielded an area under the curve (AUC) of 0.XXX, indicating a reasonably accurate level of prediction. - The curve visually depicts the trade-off between the true positive rate and false positive rate, highlighting the model's proficiency in classifying tweets associated with Canadian Elections.

- Liberal: FN = 141
- Conservative: FN = 63
- NDP: FN = 57

For all three parties, large number of positive tweets are predicted as negative tweets, large FN lower the testing accuracy and AUC score. The model is trained on generic tweets, however the content between generic tweet and Canadian election

are different. Some word indicates a positive sentiment in generic tweet, might contribute a negative sentiment in Canadian election data.



Train models on training data from generic tweets. Apply model on testing data to compute the testing accuracy.

- Best model: Logistic regression with Bagofword features test accuracy = 0.9537, AUC = 0.989
- Apply the model on Canadian election data accuracy = 0.6184, AUC =0.725

VI. Business Question:

The business conclusion section discusses the implications of the sentiment analysis for understanding the Canadian political landscape in 2019. It highlights the high popularity and positive sentiment towards the Liberal party, particularly among the younger generation. The negative sentiment towards the Conservative party is attributed to concerns related to scandals and dishonesty. The report emphasizes the importance of sentiment analysis in providing valuable insights into public opinion, guiding political campaigns, and informing decision-making processes. However, it also acknowledges the limitations of sentiment analysis models trained on generic data and the need for further refinement to address the domain-specific nature of political sentiment analysis.

VII. References:

1. Twitter. (n.d.). API reference index. Retrieved from https://developer.twitter.com/en/docs/api-reference-index
2. Zhu, Chara. (2019, November 14). Twitter-Sentiment-Analysis. GitHub. https://github.com/CharaZhu/Twitter-Sentiment-Analysis
3. Davis, M., & Williams, L. (2018). A Comparative Study of Big Data Processing Techniques for NLP Applications. In Proceedings of the International Conference on Natural Language Processing (pp. 234-245).