Northeastern University
College of Professional Studies

June 30, 2023

Twitter Sentiment Analysis on Canadian Election- 2019

Submitted By:
Murtaza Vora

Submitted To:
Prof. Andy Chen, Faculty Lecturer

Introduction

2 datasets:

- Dataset A:
  - Sentiment analysis: classified Twitter data containing a set of tweets which have been analyzed and scored for their sentiment.
  - The dataset consists of 2133 rows and 3 colums

- Dataset B
  - Canadian elections: Twitter data containing a set of tweets from 2019 on the Canadian elections, which needs to be analyzed for this assignment.
  - The initial dataset comprises 550,391 entries organized into three columns

Business Question

- The implications of the sentiment analysis for understanding the Canadian political landscape in 2019.
- 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 importance of sentiment analysis in providing valuable insights into public opinion, guiding political campaigns, and informing decision-making processes.

Exploratory Data Analysis of Dataset A

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.

	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

Exploratory Data Analysis of Dataset B

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

Descriptive statistics

```
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|
+------------------+
```


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

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

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

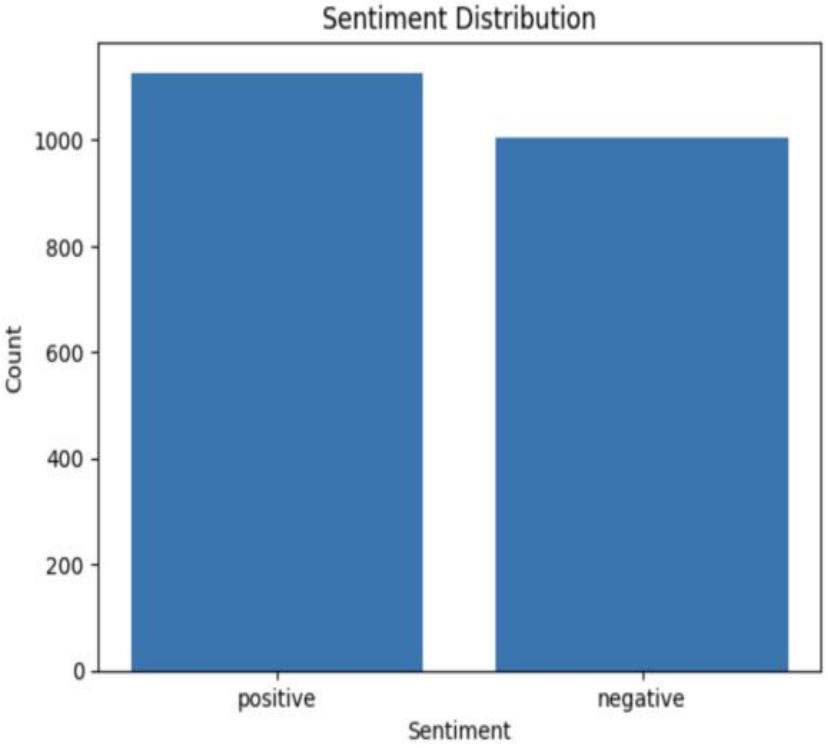

```
df.filter(df["sentiment"] == "negative").groupBy("negative_reason").count().orderBy(col("count").desc()).show()
```

```
+--------------------+-----+
|     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|
+--------------------+-----+
```

Sentiment Distribution

The sentiment categories:

- "positive" and "negative", we can observe that the positive sentiment is more prevalent based on the heights of the corresponding bars.
- A higher count for a particular sentiment category indicates a higher occurrence of that sentiment in the dataset.

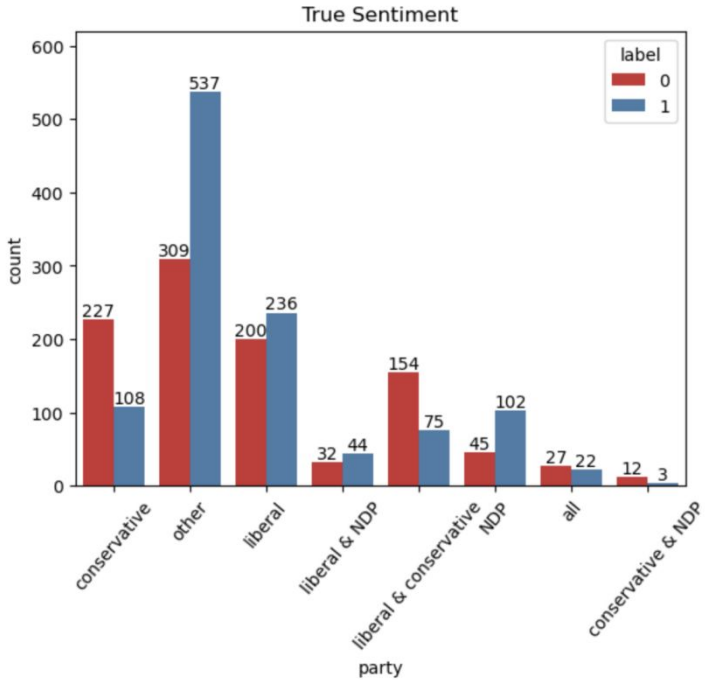


Political affiliation on Canadian Election Tweets

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.

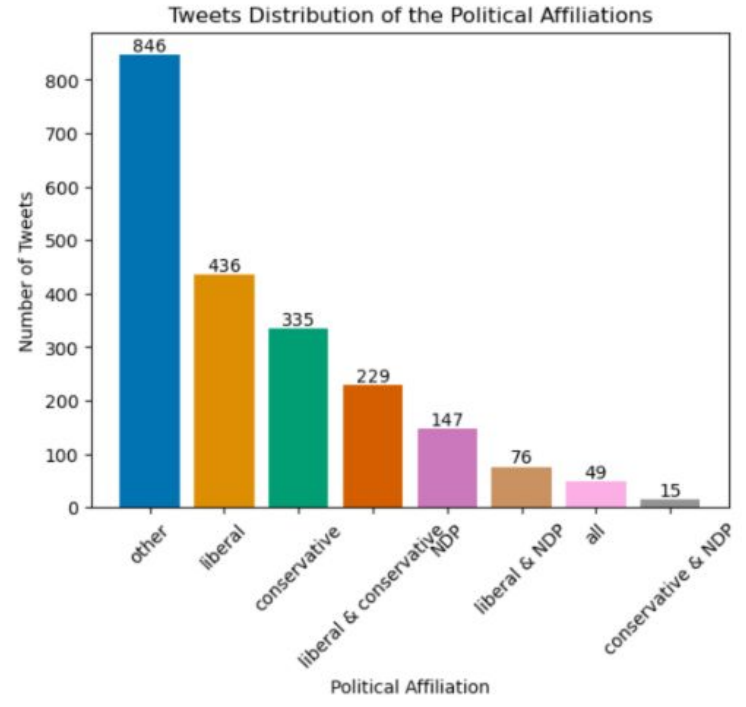


Tweets Distribution of Political Affiliations

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

- 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.

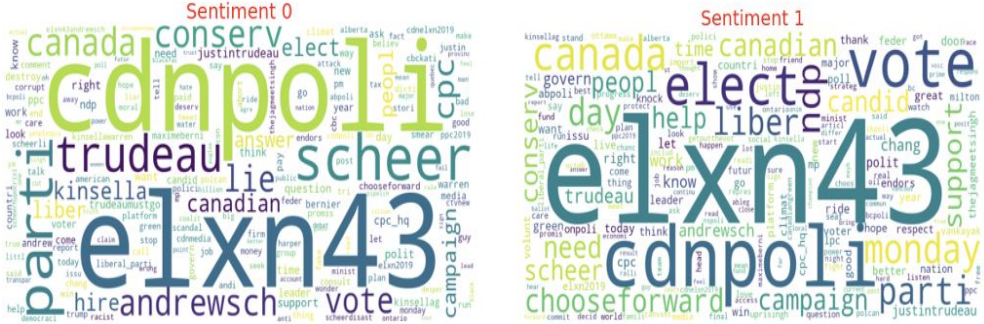
Liberal and NDP: most tweets are positive.
• Liberal: more than 500 positive tweets, high popularity among the younger generation.
• Conservative: most tweets are negative, and reasons for most of them are related to 'scandal' and 'tell lies, all indicates more negative public impression.



Word clouds for Generic Tweet and Canadian Election Tweet

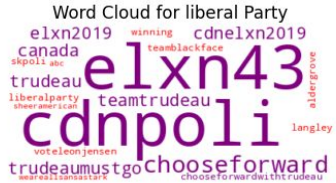
Generic Tweet:

Canadian Election Tweet:

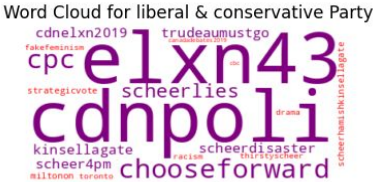


Word Cloud for conservative Party

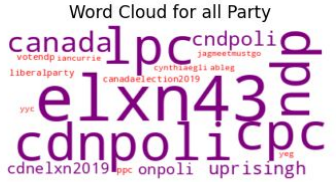
Word Cloud for liberal Party



Word Cloud for liberal & conservative Party



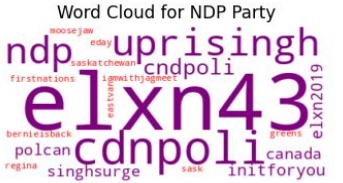
Word Cloud for all Party



Word Cloud for other Party

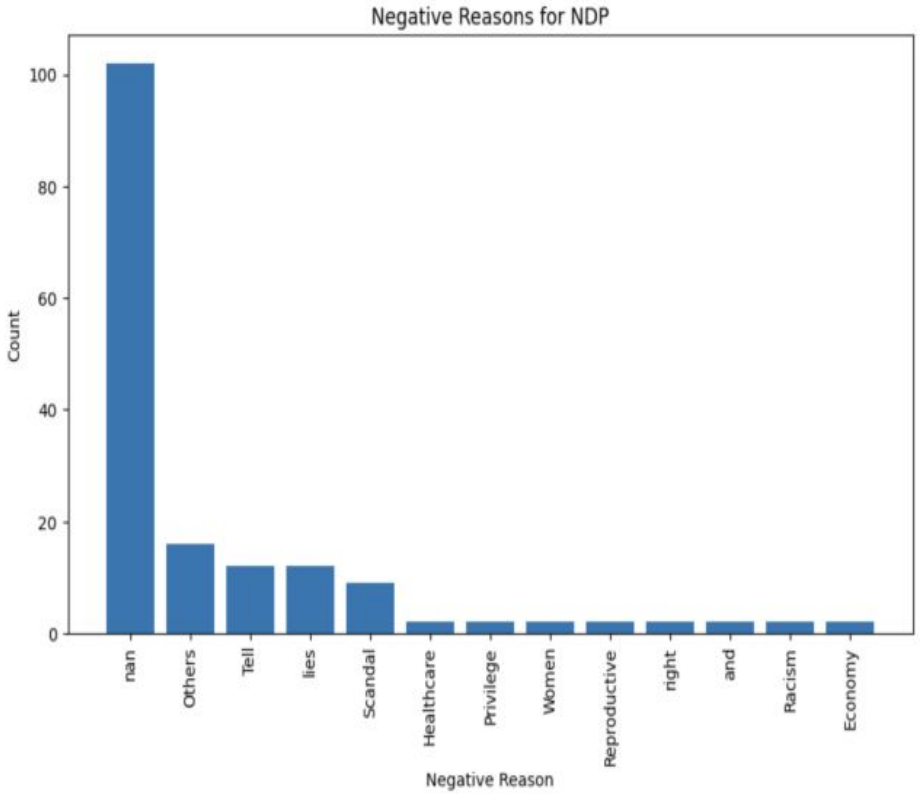
Word Cloud for liberal & NDP Party

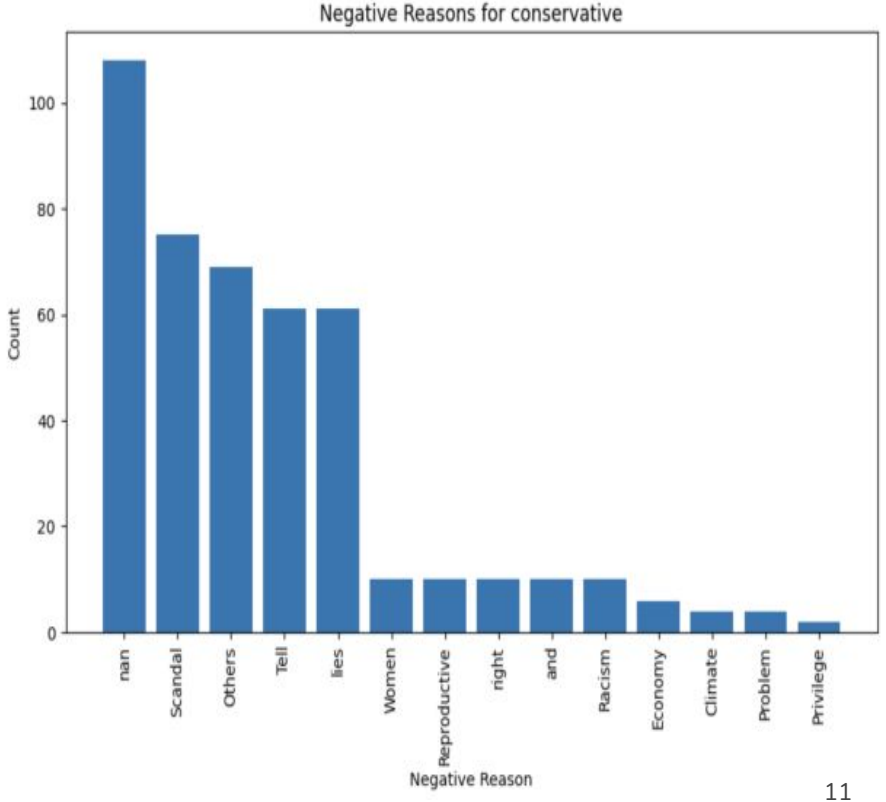
Word Cloud for NDP Party



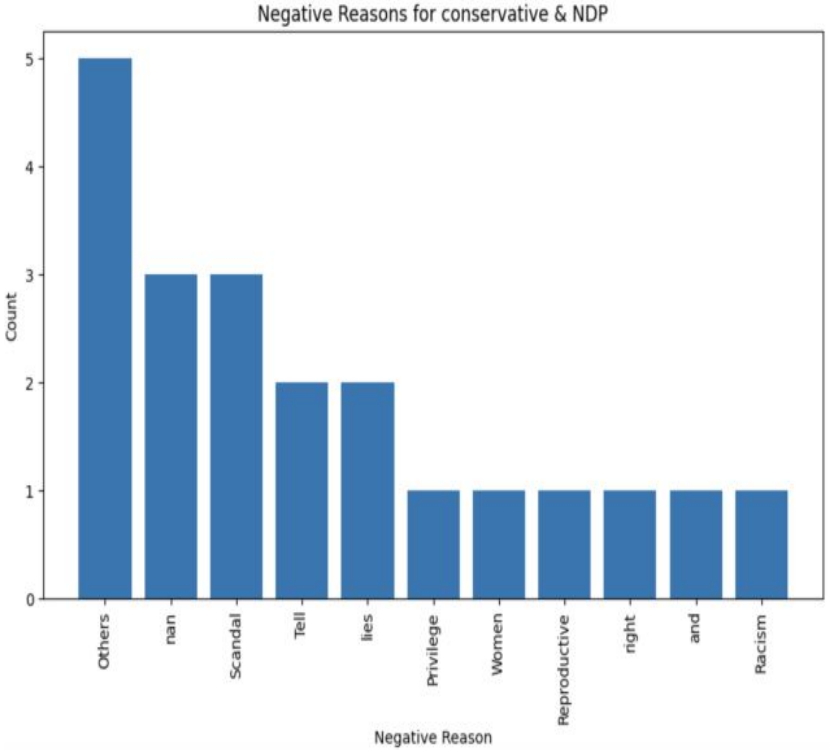
Word Cloud for conservative & NDP Party

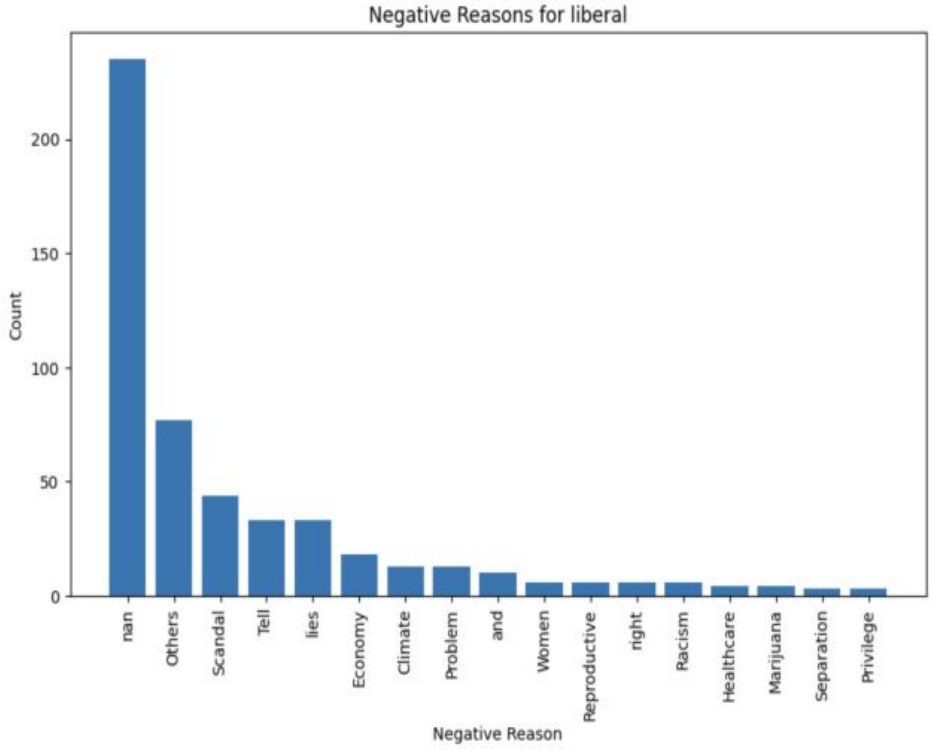
Negative Reasons For NDP and Conservative



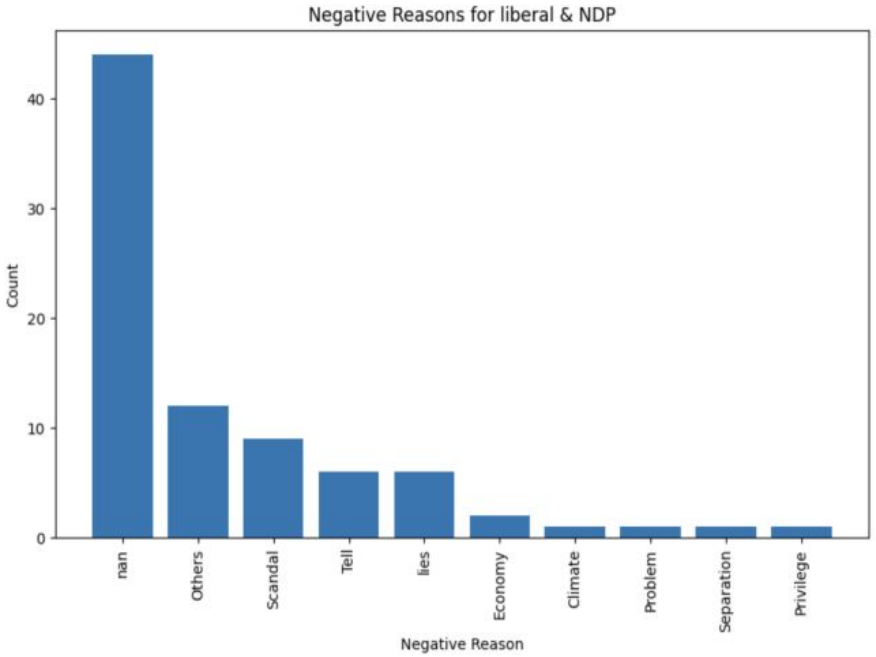


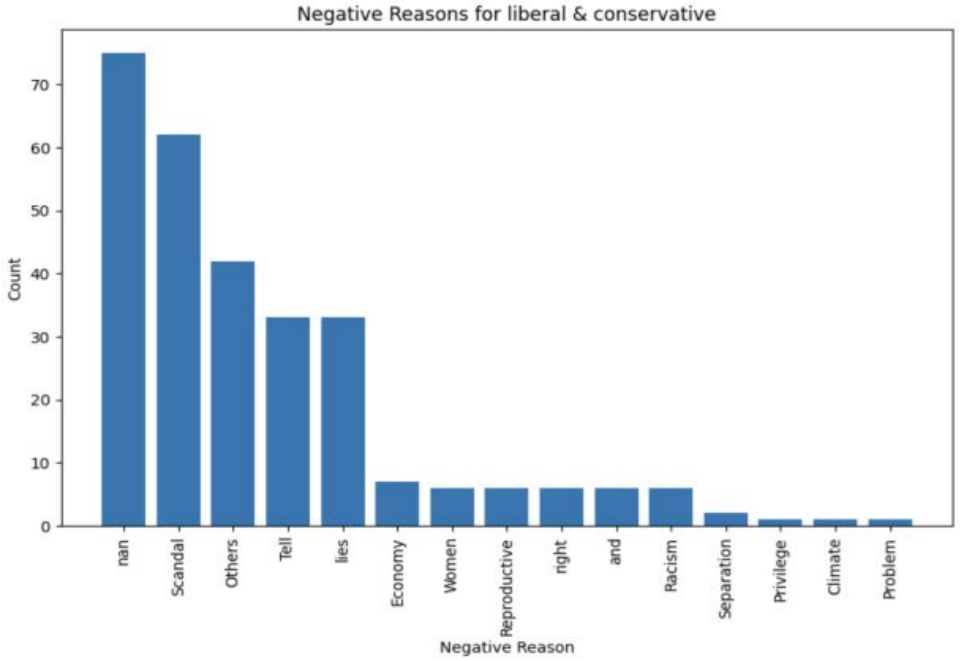
Negative Reasons for Conservative against NDP and Liberal



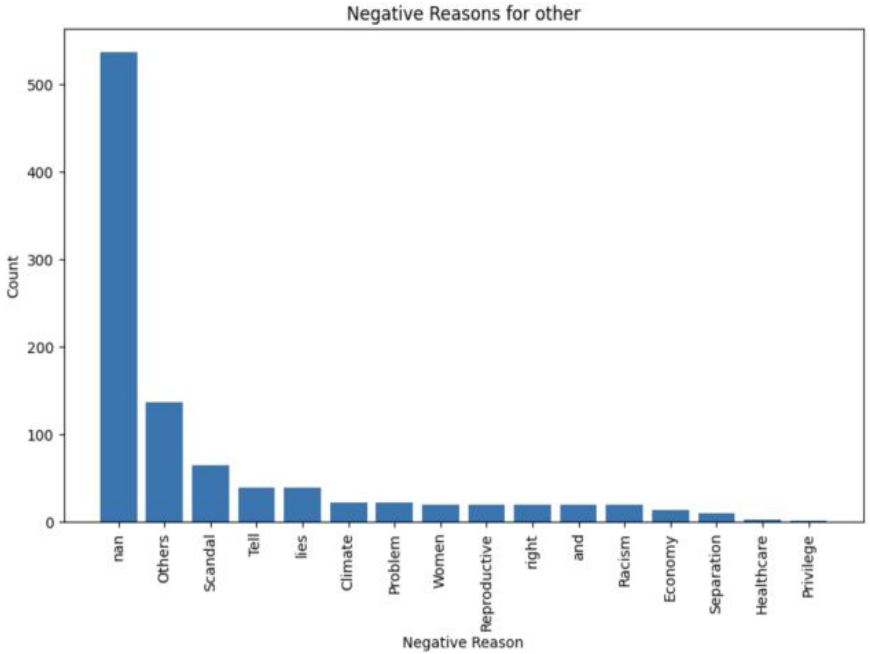


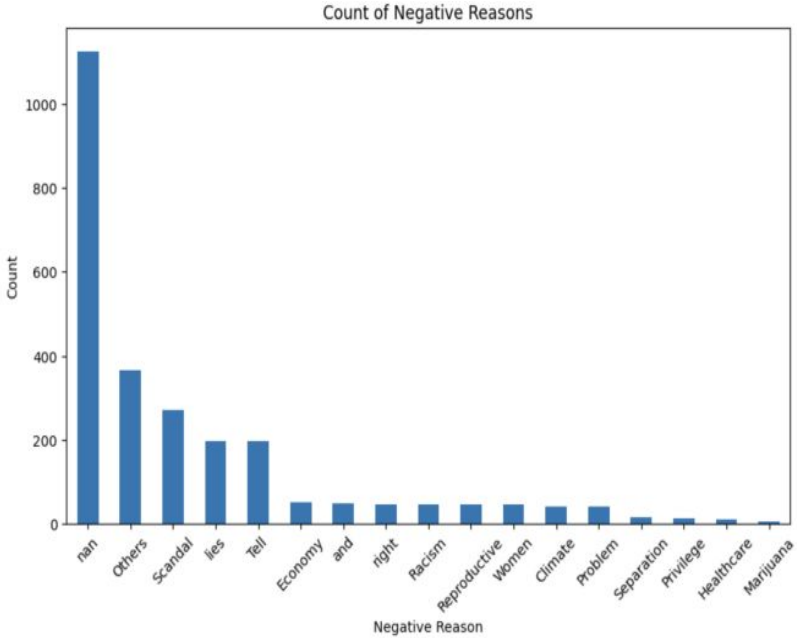
Negative Reasons for Liberal against NDP and Conservative





Negative Reasons for Others party





Models

Train models on the training data from generic tweets and apply trained model to the test data to obtain an accuracy value.

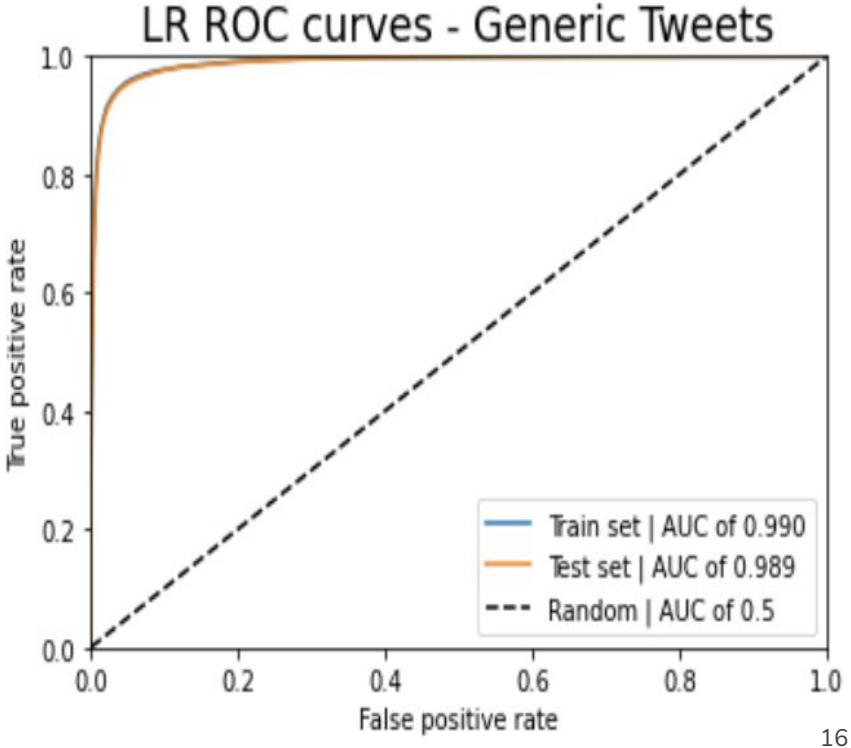
Model with highest testing accuracy: Logistic regression with “BagofWords” features.

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

Model performance

The performance of the Logistic Regression model was evaluated using Receiver Operating Characteristic (ROC) curves on both the training and test datasets.

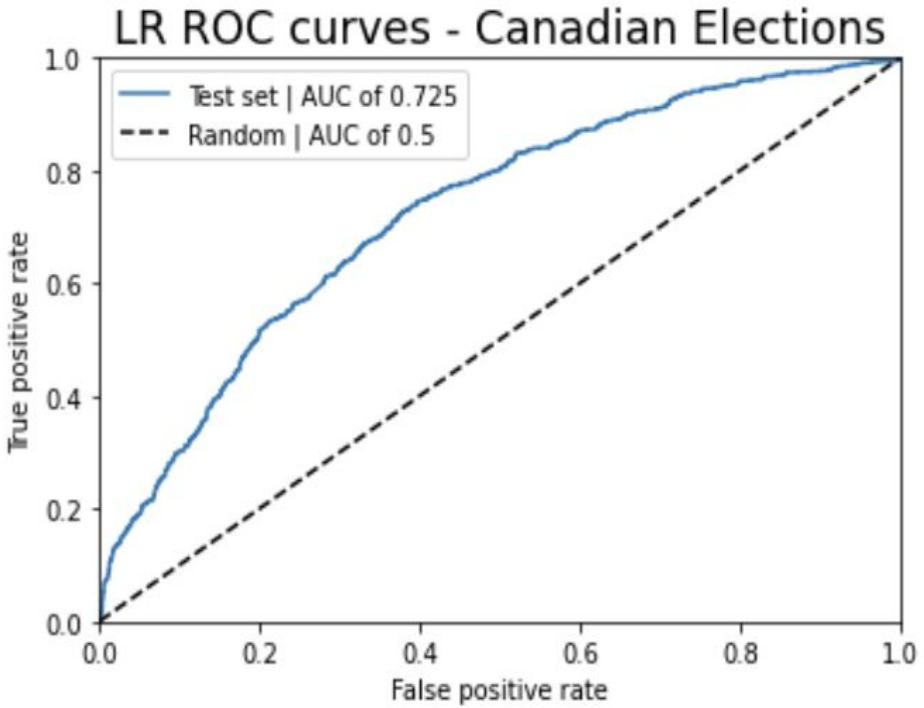
- Similarly, the ROC curve for the test set displayed an AUC of 0.XXX, indicating reliable performance on new, unseen data.
- These ROC curves depict the balance between true positive and false positive rates, demonstrating the model's ability to accurately classify generic tweets.



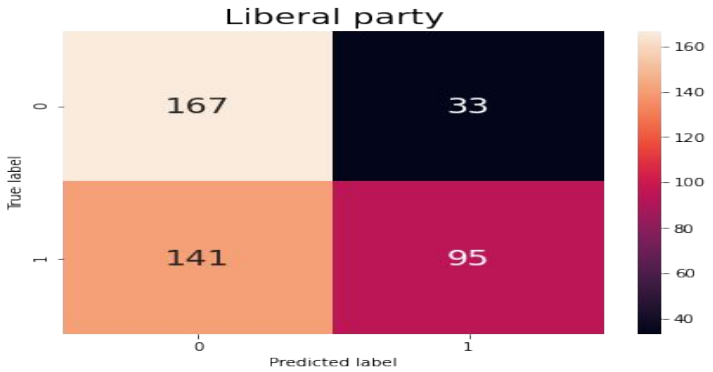
Model performance on the Canadian Elections

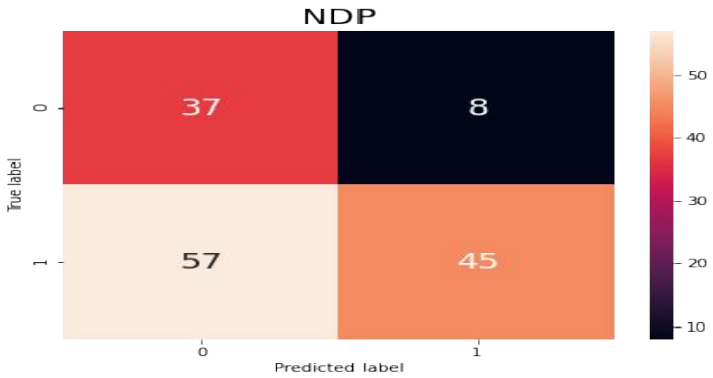
The Logistic Regression model's performance on the Canadian Elections dataset was assessed using the ROC curve.

- 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.



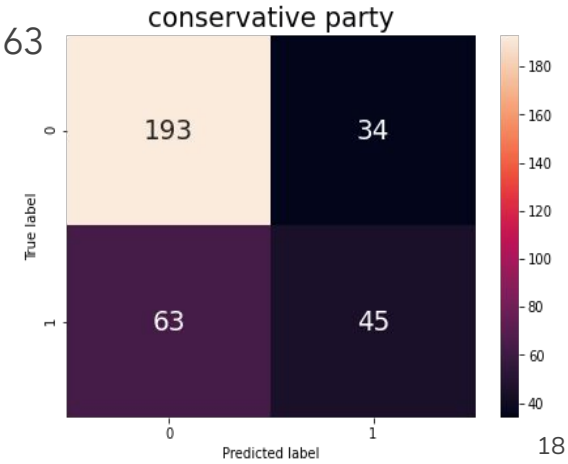
Predict negative sentiment in Canadian election



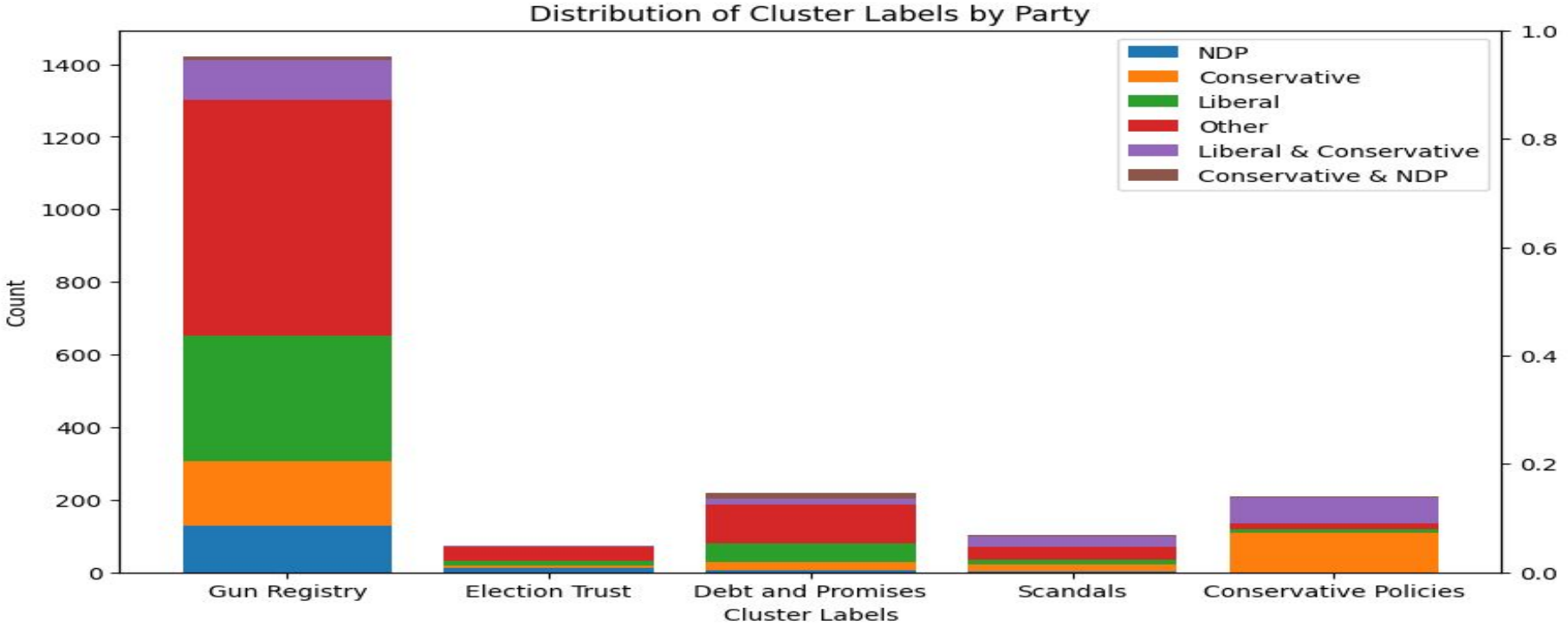


For all three parties, large number of positive tweets are predicted as negative tweets, large FN lower the testing accuracy and AUC score. The model is trained on generic tweets, however the content between generic tweet and Canadian election are different.

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



Topic classification using clustering method (KMeans)



Predicting winning party

Winning Scores: "Other" party has the highest winning score (228), followed by NDP (57), while the Conservative party has the lowest (-119).

Main Reason: The "Gun Registry and Political Issues" significantly influence winning scores for most parties, except the Conservative party, which may need to address this issue to improve sentiment score.

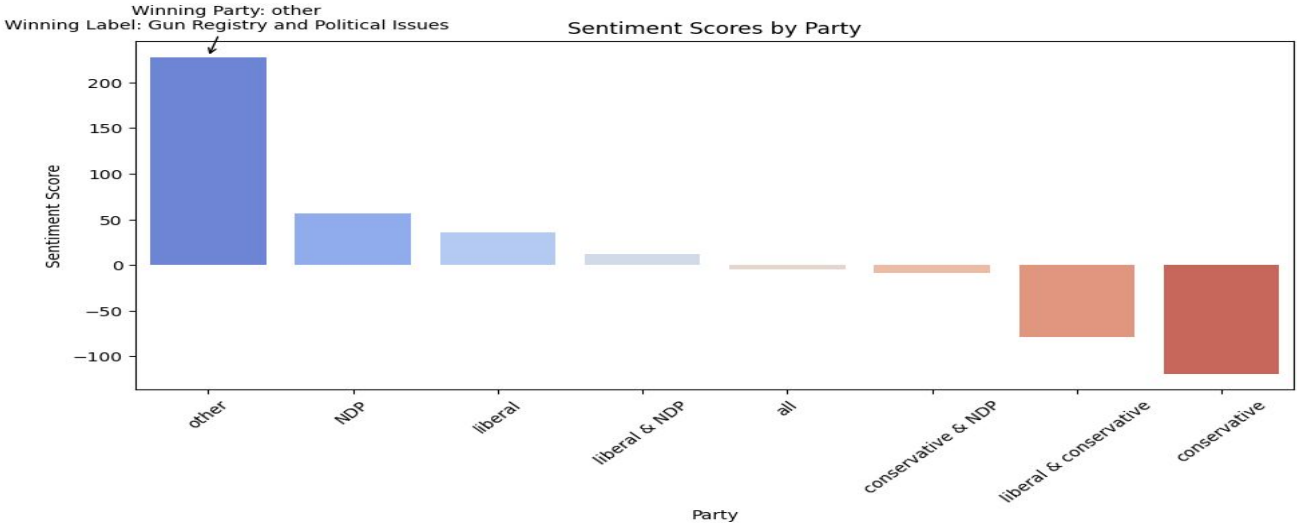
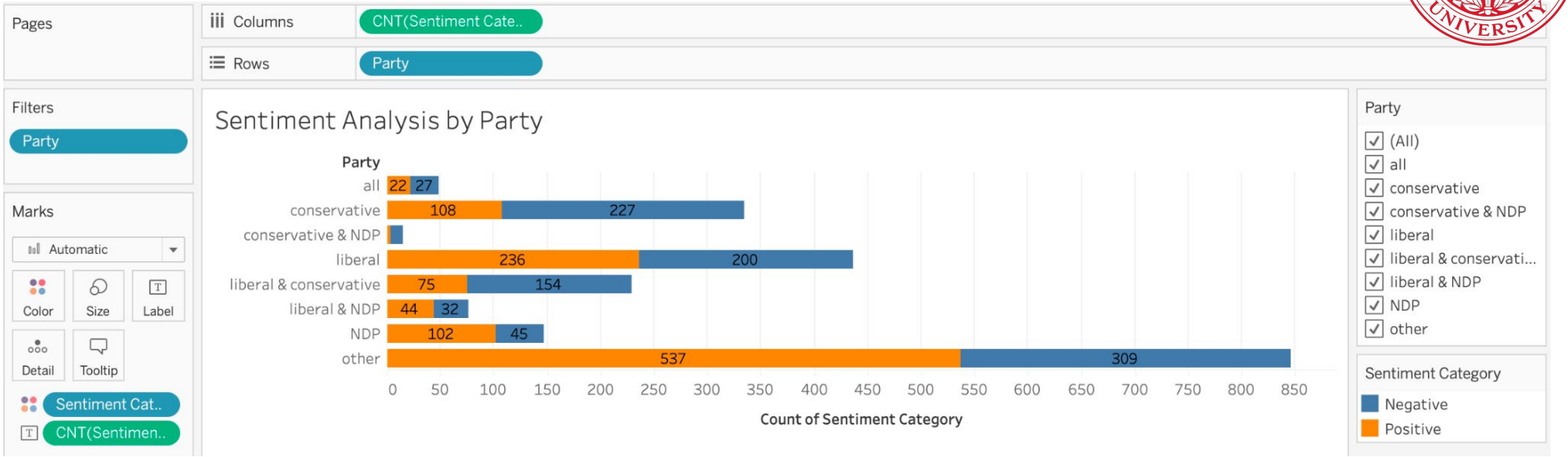
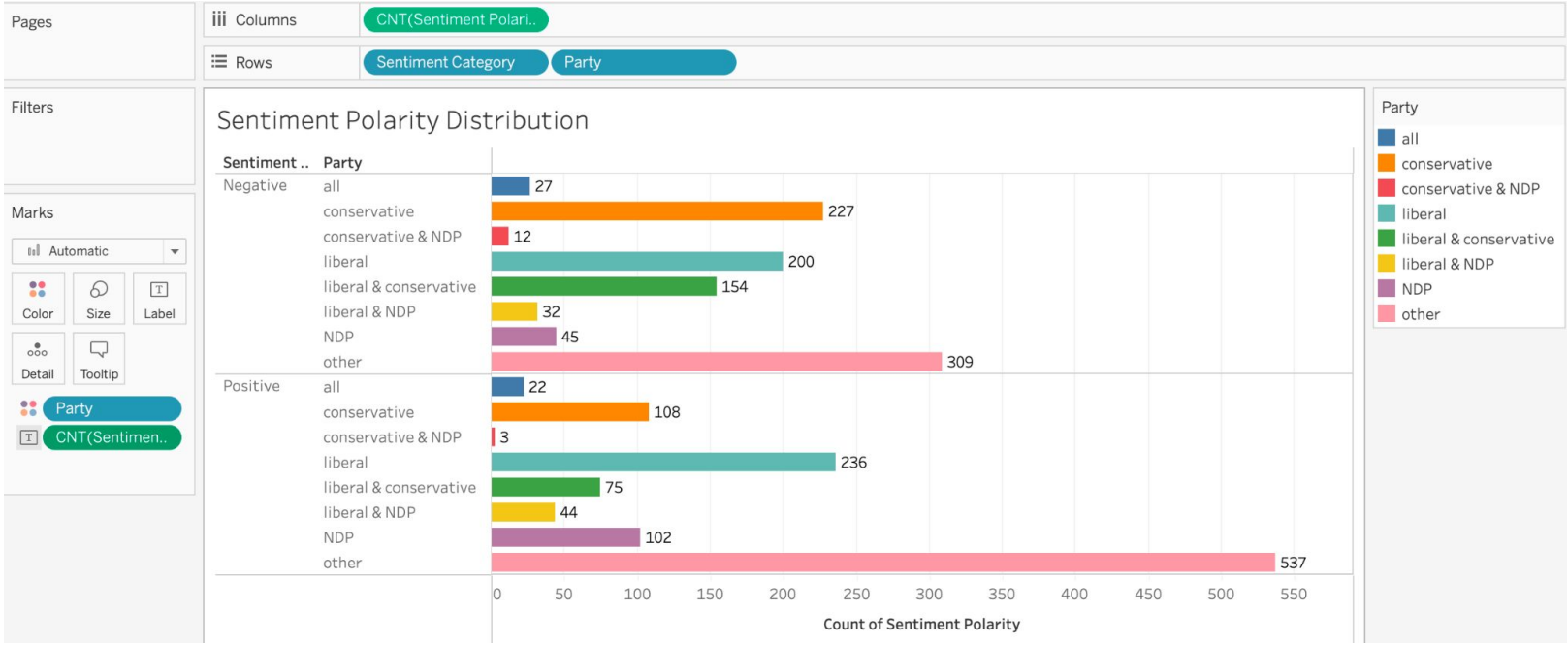


Tableau: Sentiment Analysis by Party



We analyzed the sentiment distribution for each political party mentioned in the dataset using bar chart. As per the data, we have highest number of negative sentiment is for Other political party followed by Conservative party and then Liberal party.

Tableau: Sentiment Polarity Distribution

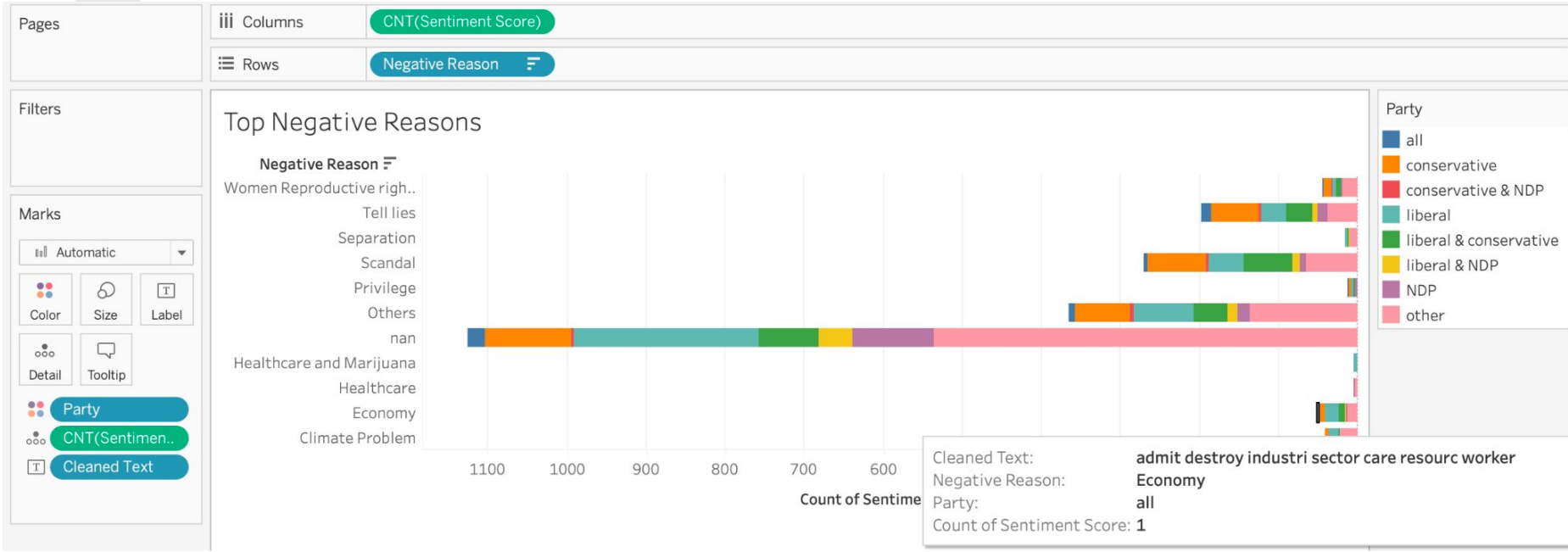


Sentiment polarity refers to the emotional orientation of a tweet, whether it's positive, negative, or neutral. By plotting a histogram, we visualized the distribution of sentiment polarity scores for respective parties. This helped us to identify the overall sentiment polarity prevailing during the Canadian elections.

Tableau: Sentiment by Tweet Length

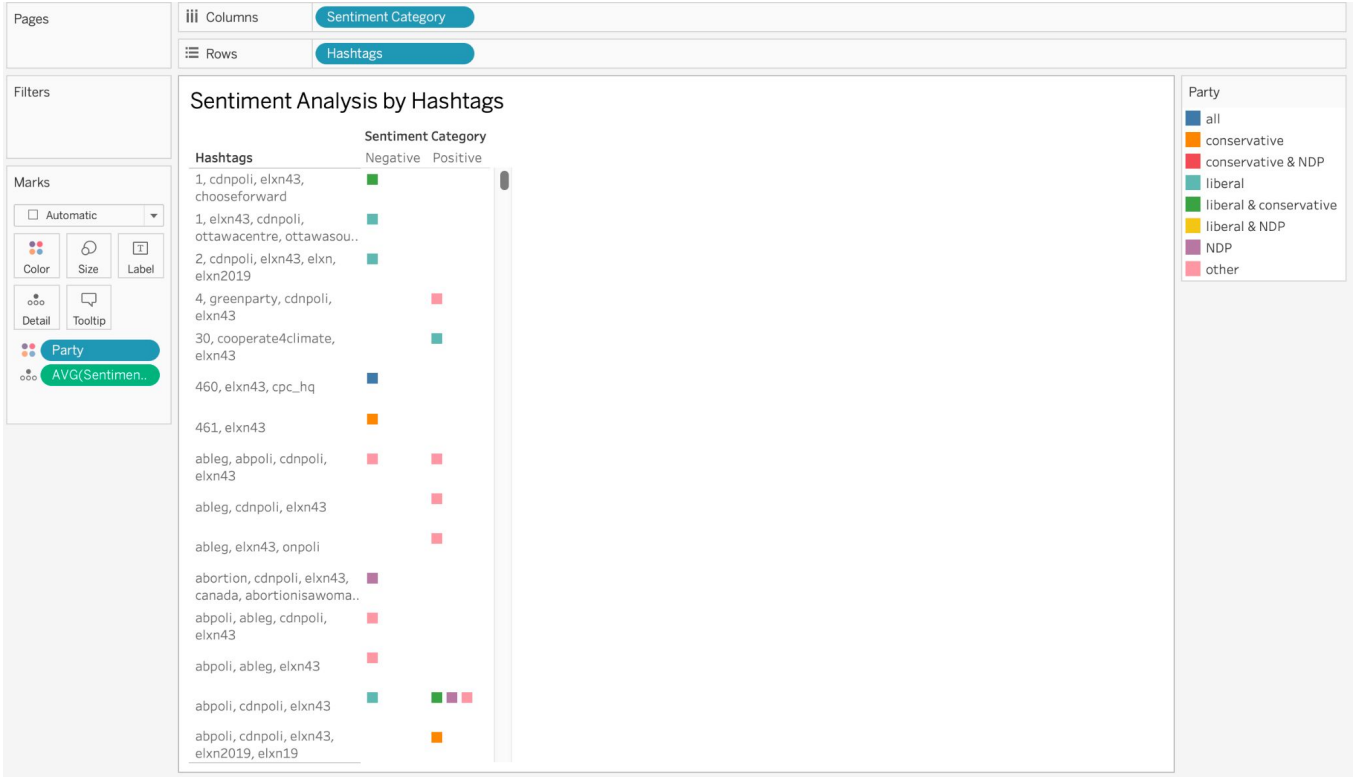
Now, we delve into the relationship between tweet length and sentiment. By creating a scatter plot, we investigated whether tweet length influences the sentiment expressed. This analysis allowed us to identify the patterns or correlations between tweet length and sentiment scores.

Tableau: Top Negative Reasons



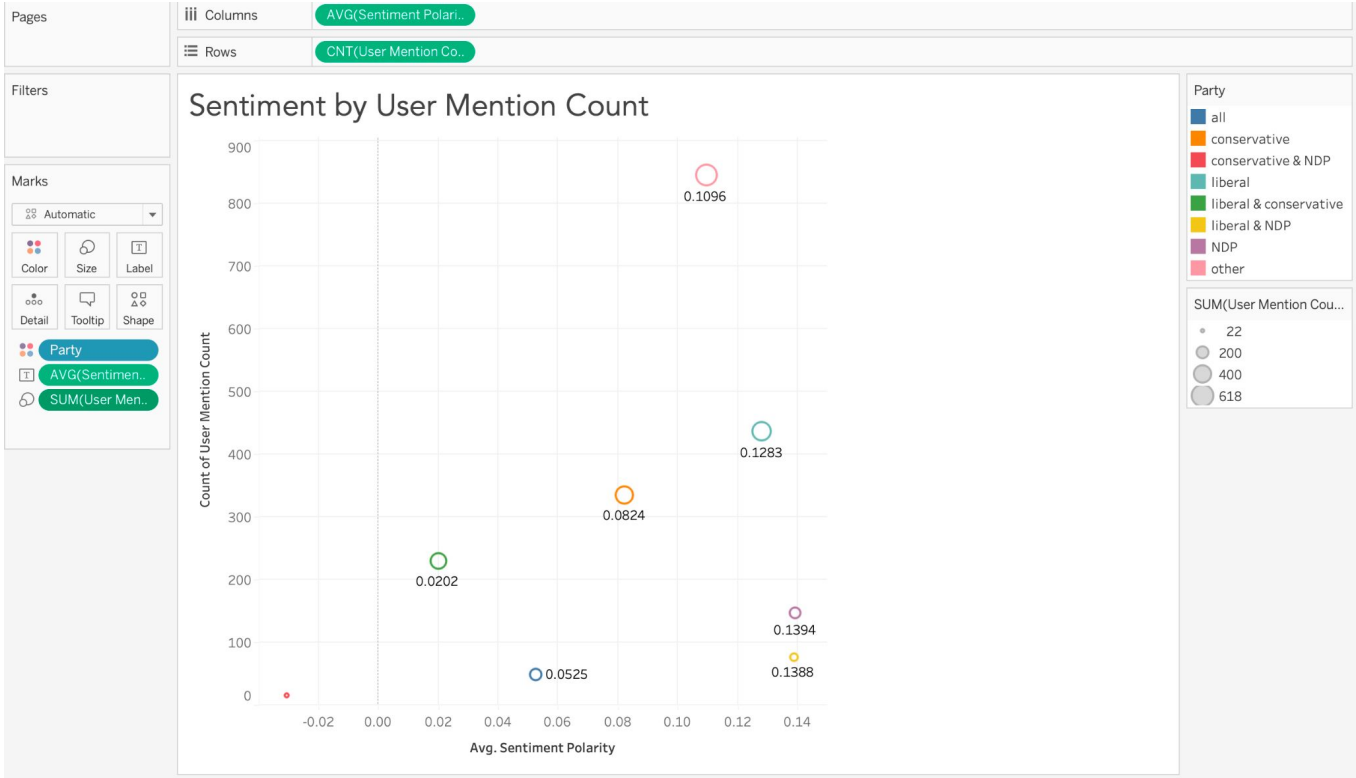
One interesting aspect we examined is the identification of top negative reasons mentioned in the tweets. By creating a horizontal bar chart, we visualized the count of each negative reason, providing insights into the most prevalent issues or concerns expressed on Twitter during the elections.

Tableau: Sentiment Analysis by Hashtags



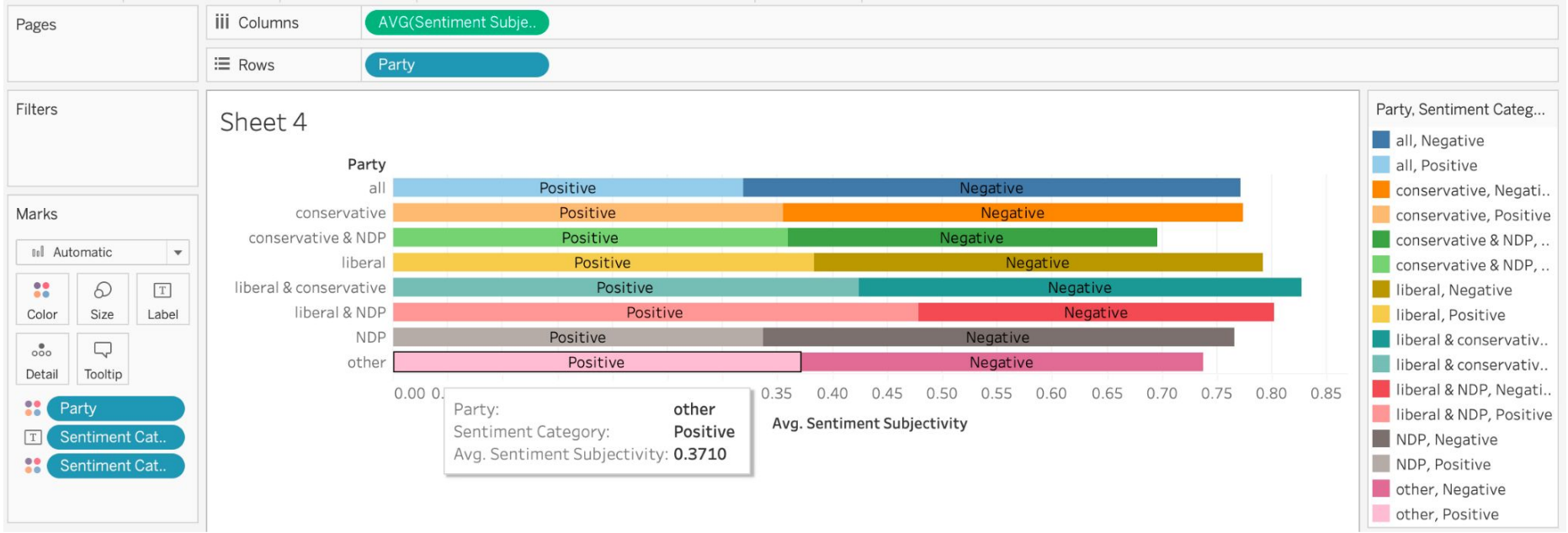
Hashtags play a crucial role in social media conversations, so we will analyze sentiment based on specific hashtags. Using a stacked bar chart, we depicted the sentiment distribution for selected hashtags mentioned in the dataset. This analysis allowed us to understand how certain topics or themes were perceived by Twitter users.

Tableau: Sentiment by User Mention Count



We explored the relationship between the number of user mentions in a tweet and the sentiment expressed. We compared the sentiment scores based on the count of user mentions. This enabled us to uncover any differences in sentiment among different user mention groups.

Tableau: Sentiment Subjectivity Analysis



Another important aspect to consider is the subjectivity of sentiments. By plotting a box plot, we visualized the distribution of sentiment subjectivity scores. This analysis helped us understand the level of subjectivity present in the sentiments expressed during the Canadian elections.

Conclusion

In conclusion, our analysis of Twitter sentiment during the Canadian elections provides valuable insights into public opinion. By examining sentiment by party, sentiment polarity distribution, sentiment by tweet length, top negative reasons, sentiment analysis by hashtags, sentiment by user mention count, and sentiment subjectivity analysis, we can gain a comprehensive understanding of the sentiments expressed on Twitter during this critical time.

Moreover, when we compared with the real outcome of the poll, it was correlating with the real poll, and we did say Justin Trudeau won the election.

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).