Kaggle Project: Classification of Tweets of Politicians from Northern Europe

Arunaggiri Pandian Karunanidhi
Georgen Institute for Data Science,
University of Rochester, NY
akarunan@ur.rochester.edu

Shubham Shailesh Tamhane
Georgen Institute for Data Science,
University of Rochester, NY
stamhane@ur.rochester.edu

***Abstract* - In this study, we explore an extensive dataset from a Kaggle challenge that comprises tweets from general users across seven Northern European countries: Belgium, Denmark, Iceland, Ireland, Netherlands, Norway, and Sweden. These tweets, summing up to 509,031 in number, are segmented into a training set (407,223 tweets) and a test set (101,808 tweets). Our objective is to harness these tweets to predict the political inclinations of the users. By utilizing various preprocessing methods and implementing a logistic regression model, we navigate the complexities of the dataset to derive meaningful insights. This report outlines our approach, techniques, and results in an easily comprehensible manner.**

I. Introduction

With the digital revolution, platforms like Twitter have taken a central role in political discussions, giving individuals a space to express their political ideologies openly. Analyzing the vast sea of tweets can offer a profound understanding of the dominant political ideologies, discourse patterns, and overall sentiments across regions.

The dataset in focus, obtained from a Kaggle challenge, offers a granular view of tweets from users spread across seven Northern European countries. Each of these countries, with their distinct political landscapes, contributes to the heterogeneity of the dataset. For a systematic analysis, the data has been split into training and test sets following an 80-20 distribution for each nation.

In the following sections of this report, we will delve into the specifics of data preprocessing, feature extraction, and model selection. We will detail the logistic regression model's rationale, its performance metrics, and the insights derived from the analysis.

II. Data

A. *Diving into Hashtags*

The backbone of our investigation is the rich dataset from a Kaggle challenge, capturing tweets from users spanning seven Northern European countries. This dataset offers a goldmine of information, serving as a reflection of the prevalent discussions, sentiments, and themes that resonate in the political discourse of these nations.

To delve deeper into the political narrative, we turned our focus to hashtags. In the digital realm, hashtags serve as markers, signposting major themes and topics. Their concise nature and widespread use make them an invaluable tool for extracting insights.

Our methodology was straightforward. After importing the 'training_data.xlsx' using the pandas library, we faced a challenge: multiple hashtags often resided in a single row under the 'hashtags' column. To overcome this and get a clearer view, we split these rows, ensuring each hashtag occupied its own individual row. The ensuing step was to group the hashtags by the user's country, tallying the occurrences of each. To present our findings, we crafted pie charts for each country, showcasing the top 10 hashtags. This visualization technique offered a vivid representation of the dominant subjects in each nation's discourse.

Our perusal of the pie charts bore several insights. A striking observation was the recurrence of certain hashtags across countries. For instance, the ubiquity of #COVID19 in countries like Belgium, Ireland, Netherlands, and Sweden was testament to the pandemic's overarching influence. Similarly, the hashtag #Ukraine featured prominently in Belgium, Denmark, Ireland, and Norway, hinting at potential geopolitical discussions or concerns surrounding Ukraine. Other recurrent themes included defense or alliance-centric discussions encapsulated by #NATO in Denmark and Norway, and European Union-related dialogues indicated by #EU in Belgium and Sweden.

Yet, it wasn't just about universal themes. The dataset was rife with nation-specific narratives as well. Belgium's dialogue seemed focused on its government and the Walloon region, as seen from hashtags like #begov and #Wallonie. Denmark had a slew of 'dk' prefixed hashtags, such as #dkpol and #dkgreen, pointing directly to Denmark-centric topics. Iceland's unique hashtags, like #12stig, likely tied to specific Icelandic events or topics, while Ireland's discourse touched on Brexit's ramifications and rural Ireland themes, as evidenced by #Brexit and #OurRuralFuture.

In sum, the hashtag analysis offered a panoramic view of the political landscape in Northern Europe, revealing both shared concerns and individual country narratives. As our exploration continues, the next layers of analysis promise to add more depth and nuances to these insights.

B. *Unveiling Political Views through Stacked Bars*

Having tapped into the rich narratives provided by hashtags, our next venture was to gauge the political inclinations across the seven Northern European countries in our dataset. With this objective, we turned to the ever-effective stacked bar chart visualization, an apt tool for this investigation given its capability to provide segmented insights within the same visual space.

To get started, our dataset was grouped by two main criteria: the country of the user and their specified political viewpoint. These groupings resulted in a two-dimensional matrix, where each cell indicated the count of users with a particular political affiliation in a specific country.

However, numbers can be misleading in isolation, so we normalized these values. This transformation translated the raw counts into percentages, enabling us to make proportional comparisons between countries, irrespective of the sheer volume of data from each nation.

To bring our insights to life, we embraced Python's seaborn and matplotlib libraries. The palette we employed was pastel-based, ensuring that each political view had a distinct, easily discernible shade. Upon plotting, our stacked bar chart showcased political distributions for each country, stacked atop one another, providing a clear visualization of predominant views.

The ensuing observations were multifaceted:

1. **Dominant Political Views**: Several countries exhibited clear inclinations. The Netherlands had a 'Center' leaning with 62% of its data falling under this category. Both Iceland and Norway prominently leaned 'Left' with 58% and 52.77%, respectively.

2. **Countries with Balanced Views:** On the other hand, nations like Belgium, Denmark, and Sweden did not toe a singular line. In Belgium, for instance, 'Left' accounted for 33.83%, while 'Right' took up 45.49% of the discourse. Similarly, Denmark oscillated between 'Left' (41.12%) and 'Center' (25.93%), and Sweden had a mix of 'Left' (48.36%) and 'Center' (28.53%).

3. **Least Represented Views:** It was evident that the 'Independent' political view was underrepresented. Barring Iceland, which had a notable 14.49% segment under this view, other nations had virtually zero representation. The Netherlands, interestingly, had a mere 3.80% data under the 'Right' category, a noteworthy anomaly.

4. **Specific Observations:** Countries like Iceland and Ireland stood out. While Iceland boasted a significant 'Left' presence, it had an almost negligible 'Center' representation. Ireland, conversely, was balanced with 'Left' and 'Right' views almost neck-to-neck.

In light of these observations, it becomes evident that the political canvas of these nations is variegated. Some lean decidedly in one direction, while others exhibit a more even-handed political discourse. The scant representation of the 'Independent' view is intriguing and warrants a deeper dive, questioning whether it's an authentic snapshot of the ground reality or an outcome of the data collection techniques employed.

C. *Understanding Gender Distribution Across Countries*

Following our exploration into political perspectives and the trending hashtags by country, we turned our eyes to another integral part of the data: gender distribution. By employing a stacked bar chart, we've provided a snapshot of the gender distribution across several countries. Stacked bars allow viewers to quickly comprehend and contrast proportions across categories, making them an apt choice for this analysis.

To give you a more detailed rundown:

- **Belgium**: Here, we noticed a male-heavy distribution, with males accounting for a dominant 81.9% and females forming 18.1% of the total.
- **Denmark**: Denmark showed a more balanced gender distribution, although males (60.5%) still outnumber females (39.5%).
- **Iceland**: Iceland is nearing a balanced representation with females at 45.7% and males slightly higher at 54.3%.
- **Ireland**: This country follows a similar trend to Belgium with a high male representation of 78.9%, whereas females constitute 21.1%.
- **Netherlands**: The distribution here is skewed towards males with 77.1%, with females accounting for 22.9%.
- **Norway**: Norway sees 68.5% males and 31.5% females, again leaning towards male predominance.
- **Sweden**: Interestingly, Sweden breaks from the pattern observed in most other countries. Here, females take the lead with 56.9%, and males form 43.1% of the representation.

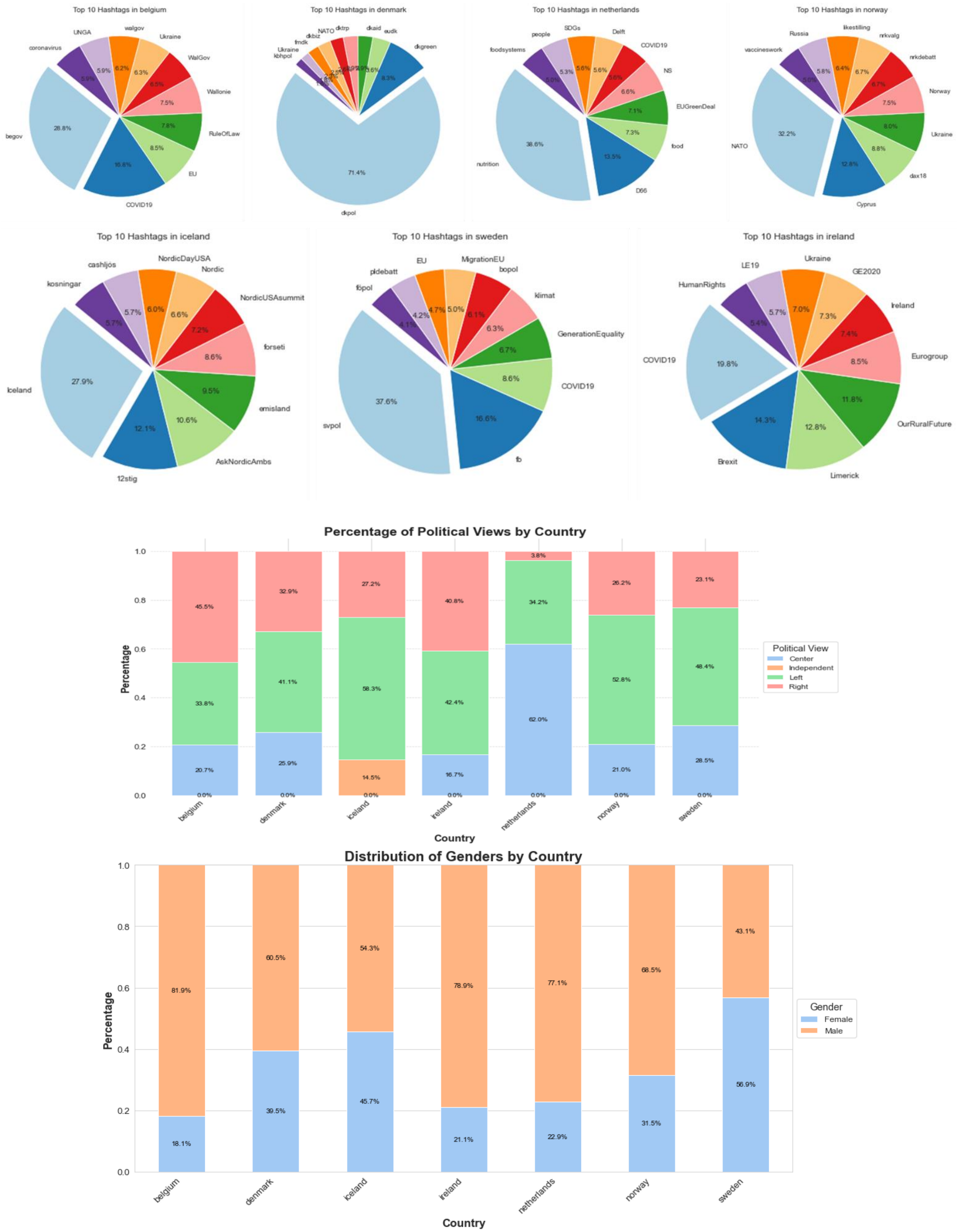

Fig 1. Figures show the pie chart distribution of hashtags, political views and genders by country

This gender distribution insight sheds light on the user demographics of our data, offering a clearer picture of the gender dynamics across these European countries. These proportions might be influenced by various factors such as the platform's user base, regional gender-based online activity, or even the nature of discussions captured in our dataset.

We also performed LDA and Non-negative matrix factorization for topic analysis. After running various iterations of Non-negative Matrix Factorization (NMF), we noticed consistent yet slightly different results. By comparing the graphical outputs, we identified differences in value magnitudes and occasionally, the ranking of significant words. On the other hand, Latent Dirichlet Allocation (LDA) gave quite distinct outcomes. But, when we looked at the bigger picture, both NMF and LDA models pointed to similar topics. So, even with variations in individual outcomes, the primary themes from both methods aligned well.

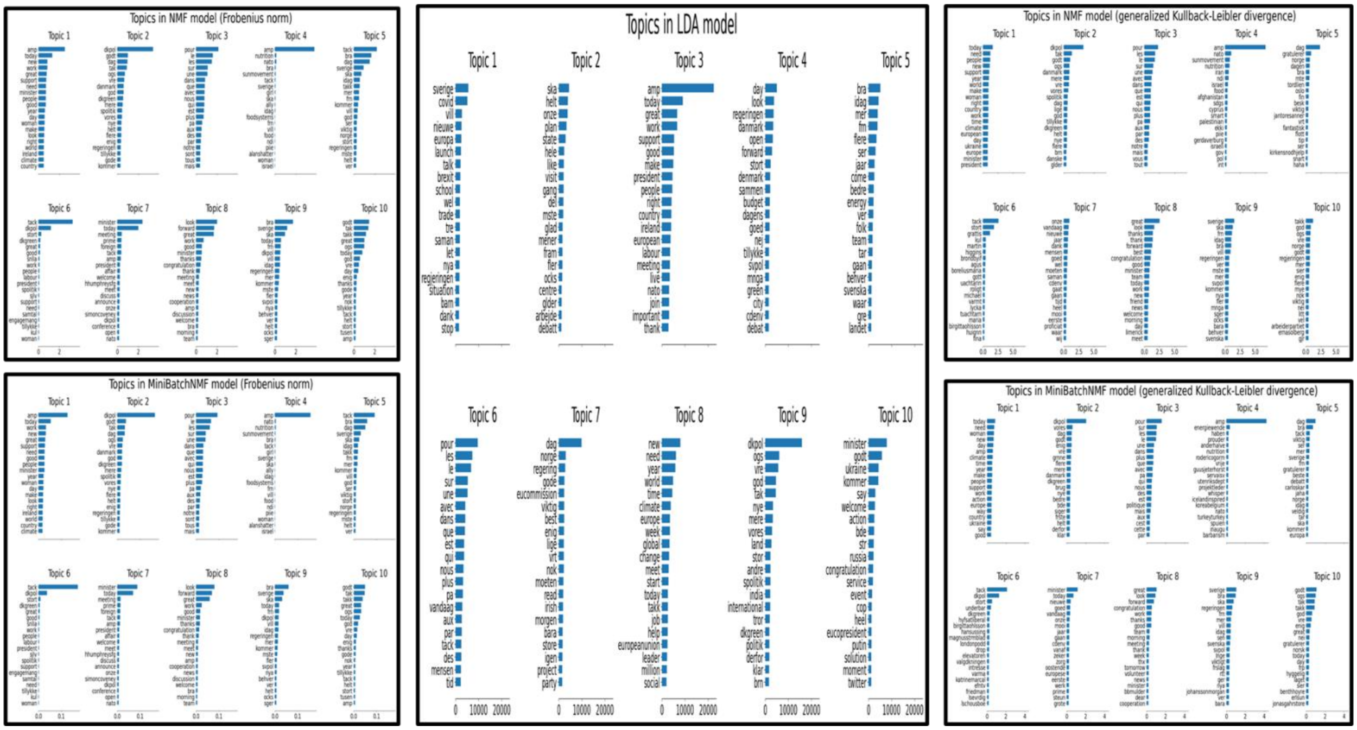

Fig 2. Topics obtained through LDA and non-negative matrix factorization

III. Methodology

A. Data Cleaning

Within the methodology section, an essential component we navigated was data cleaning. Data cleaning is pivotal for preparing our tweets for any advanced processing or analysis. Here's a more comprehensive breakdown of the approach:

1. **Essential Libraries:** We initialized by importing crucial libraries. re aids in regular expressions for pattern searching, emoji helps in managing and converting emojis in our tweets, and nltk is indispensable for natural language processing tasks.

2. **Dealing with Stopwords:** Words like 'and', 'the', or 'is' often don't impart meaningful sentiment or context in analyses. To address these, we assembled a set of "stopwords" from several languages: English, Dutch, Danish, Norwegian, and Swedish. By doing this, we ensured that the tweets, irrespective of their origin, were adequately filtered.

3. **Tweet Refinement:** The heart of our data cleaning process is the “tweet_cleaner” function, where we performed multiple operations on our tweets to make them analysis-ready:

   a) Byte-string Handling: Some tweets might be prefixed as byte strings. These were identified and decoded to regular strings.

   b) Emoji Management: Emojis in tweets were converted to their textual descriptions using the “emoji.demojize” function. Subsequently, these textual descriptions were removed to maintain the textual purity of the tweet.

c) Hyperlinks Removal: Any presence of hyperlinks, which usually don't contribute meaningful information for our purpose, were stripped off.

d) Pattern Removal: We eradicated common patterns such as retweet prefixes. While mentions and hashtags can sometimes offer context, we kept the flexibility to remove them if needed.

e) Special Characters & Numbers: Tweets can often contain a variety of special characters or numbers, which might not be relevant to sentiment or context. Hence, we cleansed the tweets off any non-word characters, non-English letters, and numeric digits.

f) Lowercasing & Filtering: Every word in the tweets was converted to lowercase to ensure uniformity. Further, we pruned words present in our stopwords set and those shorter than three characters, as they usually don't provide substantial sentiment.

g) Whitespace Management: Finally, excessive spaces or unwanted leading or trailing spaces were removed, leaving a clean, continuous string of words.

4. **Lemmatization:** With our tweets now refined, it was essential to ensure that variations of a word were represented as their base form. For instance, 'running' would become 'run'. This process, known as lemmatization, was executed using the WordNetLemmatizer. To accomplish this, words were tagged by their part-of-speech, ensuring accurate lemmatization. Our custom function lemmatize_tweet took care of this intricate process.

5. **Batch Processing:** Once our data cleaning functions were defined, the next step was to apply them across our dataset. For this, we employed tqdm, a tool that provides a visual progression bar, ensuring every tweet was processed. The end result was a new column, 'lemmatized_text', filled with tweets that had undergone rigorous cleaning and processing.

By the close of this intensive data cleaning phase, our tweets were now in a pristine state, primed for potential machine learning applications.

B. Model Building

Once our data was cleaned and preprocessed, the next significant phase in our methodology involved building a machine learning model. The objective was to utilize the cleaned tweets to predict political spectra, and here's how we approached it:

1. **Data Pruning:** We began by dropping the first three columns from our dataframe, which were presumably non-essential for the modeling process. Additionally, we tackled missing values by counting and subsequently eliminating rows with NaN values in the lemmatized_text column. We further ensured that rows with empty strings after lemmatization were also discarded, ensuring our dataset was devoid of any 'empty' content.

2. **Data Splitting:** Before modeling, we segregated our data into training and test sets using train_test_split from scikit-learn. The features were stored in X, and the target variable, pol_spec_user, in y. An 80-20 split was chosen, allocating 80% of the data for training and the rest for validation.

3. **One-Hot Encoding:** Some of our categorical variables, namely country_user and gender_user, were transformed into a format suitable for machine learning algorithms. We used one-hot encoding to convert these categorical variables into binary columns.

4. **Text Vectorization:** Since machine learning models can't inherently understand text data, the tweets in lemmatized_text needed to be transformed into a numerical format. We chose the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique, which essentially assigns importance scores to each term in our tweets.

5. **Data Aggregation**: With our data in multiple parts (textual data in vectorized format and encoded categorical variables), we required a unified format for model training. We combined these diverse datasets using Scipy's hstack method, ensuring our final training and test datasets had both textual and categorical information.

6. **Model Selection and Training:** We opted for the Logistic Regression model due to its efficacy in binary and multi class classification tasks. It was trained on our final training data. The newton-cg solver was chosen for optimization, and the maximum iteration was set to 10,000 to ensure convergence.

7. **Model Evaluation:** Upon training the model, we proceeded with predictions on the test set. Evaluation metrics from the classification_report provided insights into how well our model performed, and the accuracy score gave us the model's overall accuracy rate.

8. **Visualization with a Confusion Matrix:** To have a clear visual on the model's performance, we plotted a confusion matrix. This matrix illustrates the true positives, true negatives, false positives, and false

negatives, giving a comprehensive view of where the model's predictions stood against the actual values.

In essence, the model-building phase was meticulously crafted to ensure our cleaned and processed tweets were efficiently utilized to predict political spectra, yielding actionable insights and a solid foundation for further analyses or refinements. We then transitioned to the crux of our research: building a predictive machine learning model. However, as clarified, we wouldn't be splitting the data into training and test sets; rather, we'd be training our model on the entire dataset and then submitting our predictions on a separate test set to Kaggle.

C. Prediction on Test Set

To evaluate our model's performance on unseen data, we proceeded to predict the political affiliation (pol_spec_user) on Kaggle's test dataset.

1. Preprocessing the Test Data:

We loaded the Kaggle's test set from the test_data.xlsx file. Applied the same preprocessing steps as we did for the training data. This ensured that the test data was consistent with the format the model was trained on.

2. Predictions using the Trained Model:

With the test set now transformed into the required format, we fed it into our trained Logistic Regression model. The model then analyzed the features of each tweet in the test set and predicted whether the author leans more towards a particular political affiliation.

3. Creating the Submission File:

To track each prediction to its corresponding tweet, we paired each prediction with the tweet's unique ID. These paired IDs and predictions were organized and saved into a submissions.csv file, structured per Kaggle's submission guidelines. Our predictions were then finalized and ready for submission to Kaggle, allowing us to see how well our model performs in a real-world scenario.

IV. Results

Our model achieved an accuracy of 74.35% on the test set. The confusion matrix, presented as an accompanying visual, sheds light on the classification across the political categories: center, independent, left, and right.

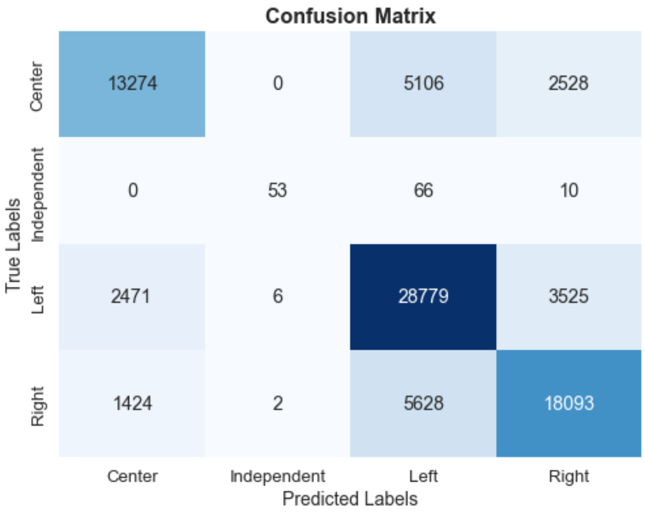

Fig 3. Confusion Matrix

From the Confusion Matrix, we can observe that,

1. The model demonstrates a strong capability in identifying "left" and "right" tweets.
2. The most notable misclassifications occur between "center" and the other categories, especially with "left".
3. The "independent" category, though having limited data, shows minor confusion with "left".

The matrix visually highlights areas where the model excels and potential points for refinement. The accuracy, combined with insights from the confusion matrix, underscores the model's overall reliability and areas for future enhancement.