Classification of Tweets of Politicians from Northern Europe

Aradhya Mathur and Richa Yadav

***Abstract*— Twitter is a social media platform used by millions of people worldwide to share their thoughts and opinions. The sheer amount of tweets generated on the platform can be utilized to gain potential insights and recommendations for various topics. The objective of this project is to classify a Twitter user from seven European countries as Right, Left, Center, or Independent based on tweets, hashtags used, gender, and country.**

I. INTRODUCTION

The aim of this research is to correctly classify politicians' political inclinations by mining their tweets on Twitter. Since the tweet text data is prone to data quality issues, emphasis is given to data cleaning, processing, and feature engineering. Data exploration reveals that country and gender have a significant impact on the political inclinations of a politician. Hence, the classifiers are trained on each country and gender combination separately. Since the distribution of the target variable is not even, random sampling is used to account for the imbalances in the output class. We deployed several robust classifiers like Logistic Regression, Naive Bayes, and SVM and also combined these classifiers to design an ensemble approach using a voting mechanism to get the best predictions.

II. DATA

A. File Description

1. training_data.xlsx - 407,223 unique tweets about coming from seven different Northern European countries). The categorical outcome variable is called 'pol_spec_user'.
2. test_data.xlsx - 101,808 unique tweets for which the 'pol_spec_user' variable will be predicted. 'Id' variable shows the index values associated with each tweet.

B. Dataset description

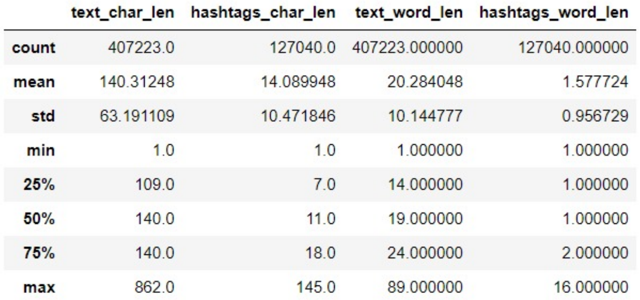
- hashtags: The list of hashtags included in the tweet
- full_text: The text of tweet (including emojis, htmls, hashtags)
- in_reply_to_screen_name: The Twitter screen name of the user the owner of the tweet is replying to (if any).
- country_user: Country of the owner of the tweet
- pol_spec_user: Political view of the owner of the tweet (found only on the training dataset)
- Id: An index number associated with tweets (found only on the test dataset)

Before we delve into the data analysis, let's try to understand these different ideologies -

- Right: It typically means advocating for conservative ideologies. They often oppose government regulation and favor free-market principles.
- Left: It generally means advocating for progressive ideas. Left-leaning individuals and parties often favor government programs to address social and economic inequalities.
- Center: It is characterized by a more moderate and balanced approach. They support a mixed economy with some government intervention, as well as social policies that aim to address societal challenges.
- Independent: They do not align themselves with either the left or right on the political spectrum. They prioritize issue-based decision-making rather than strict adherence to a particular ideology.

C. Data Exploration

This dataset contains tweets from users in 7 northern European nations. These tweets are intended to show the political leanings of the tweeter, whether they are discussing "Left," "Right," "Center," or "Independent" political parties' philosophies. We would try to comprehend the information from several aspects, and would then lastly carry out the data preparation for modeling and forecast of people's political preferences based on tweets text.



	text_char_len	hashtags_char_len	text_word_len	hashtags_word_len
count	407223.0	127040.0	407223.000000	127040.000000
mean	140.31248	14.089948	20.284048	1.577724
std	63.191109	10.471846	10.144777	0.956729
min	1.0	1.0	1.000000	1.000000
25%	109.0	7.0	14.000000	1.000000
50%	140.0	11.0	19.000000	1.000000
75%	140.0	18.0	24.000000	2.000000
max	862.0	145.0	89.000000	16.000000

Fig. 1: Dataset Description

1) Section A: Understanding the text's extent is crucial. Hence, we computed the character and word counts for both tweet texts and hashtags across all users and countries. To accomplish this, we excluded empty hashtags to ensure a precise statistical analysis, noting that there were no empty tweet texts. Further, we analyzed the top 10 hashtags in each country [Fig.2], and found some useful inferences like:

- Belgium and COVID-19 Focus: Belgium's top hashtags prominently feature COVID-19-related tags, with

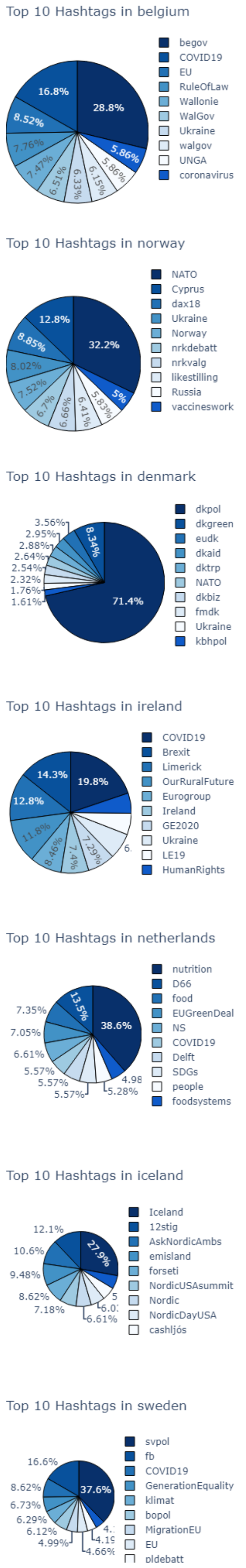


Fig. 2: Distribution of hashtags across countries

"COVID19" being one of the most frequently used hashtags. This suggests that discussions related to the pandemic are prevalent in Belgian social media discourse.

- Denmark's Political Engagement: Denmark's hashtags reflect a strong engagement in political discussions with hashtags like "dkpol" (Danish politics) and "dkgreen" (Danish green initiatives) appearing as the most frequently used. This indicates active political participation and environmental awareness in Denmark.
- Iceland's Cultural and Political Engagement: Iceland's hashtags include a mix of cultural references and political discussions, with hashtags like "Iceland," "12stig" (12 points, possibly related to Eurovision), and "AskNordicAmbs" (asking Nordic ambassadors) indicating diverse social media engagement.
- Ireland and Brexit Concerns: Ireland's hashtags prominently feature "Brexit," indicating the country's significant interest and concerns related to the United Kingdom's departure from the European Union. Additionally, "Limerick" suggests regional discussions.
- Netherlands' Focus on Nutrition: The Netherlands' hashtags are centered around nutrition and sustainability, with "nutrition," "food," and "foodsystems" appearing among the top hashtags. This reflects a focus on healthy eating and sustainable food practices in the country.
- Norway's International Engagement: Norway's hashtags include international topics such as "NATO," "Cyprus," and "Ukraine," indicating the country's engagement in global affairs. Additionally, hashtags related to gender equality ("likestilling") and vaccines ("vaccineswork") show a commitment to important social issues.
- Sweden's Political Discourse: Sweden's top hashtags include "svpol" (Swedish politics) and "pldebatt" (planning debate), highlighting active political discourse and urban planning discussions. "COVID19" also appears, reflecting ongoing pandemic-related discussions.

According to the distribution of political spectrums across all countries [Fig.3], we found the following inferences:

- 'Left' Orientation: The 'Left' political orientation is prominent in multiple countries, including Denmark, Iceland, Norway, and Sweden, with proportions ranging from approximately 41% to 58%. This indicates a prevalent left-leaning political trend in these nations.
- Significant 'Right' Leanings in Belgium and almost negligible in Netherlands : Belgium has highest proportions of individuals identifying as 'Right,' with percentages of approximately 45%. This suggests a substantial right-leaning population in Belgium, Whereas, in Netherlands only 3% individuals identify as 'Right'.
- Strong 'Center' Orientation in the Netherlands: The Netherlands has a dominant 'Center' orientation, with around 62% of individuals aligning with centrist political ideologies.
- Absence of 'Independent' Identifications: Across all the

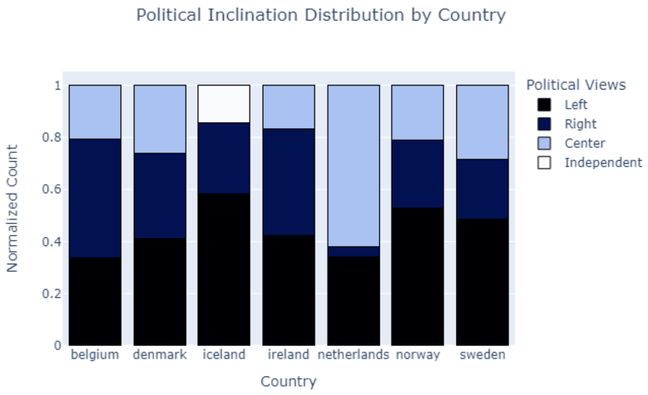


Fig. 3: Political spectrum across countries

countries except Iceland, there are no reported identifications as 'Independent.' This suggests that the surveyed individuals in these countries do not align themselves with an independent or non-affiliated political stance, or it may indicate that the survey did not include an 'Independent' category.

According to the distribution of gender across all countries [Fig.4], we found the following inferences:

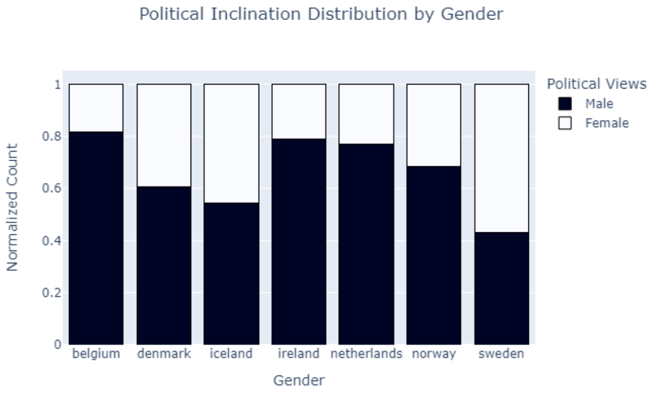


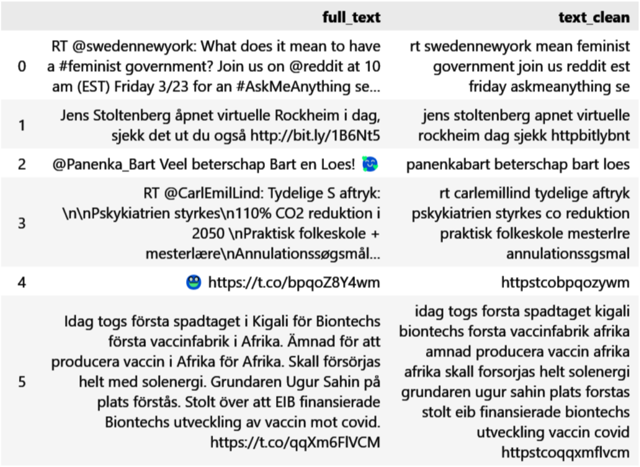
Fig. 4: Political spectrum inclination by gender

- In Belgium, the male proportion appears to dominate, as is the case in the Netherlands, Ireland, and Norway.
- Sweden stands out with a female proportion of 0.568556%.
- Belgium, Ireland, and the Netherlands exhibit roughly 18 to 22% female representation.

2) Section B: Several procedures are undertaken to preprocess and cleanse the tweet text:

- Firstly, the tweet data has been treated by a robust stemming process to ensure that we simplify the data and reduce variations of words to a common form. Since, the tweets are from multiple languages, we used the Snowball Stemmers. Snowball Stemmers are available for multiple languages apart from English like French, German, Spanish, etc., making it a versatile tool for stemming in various linguistic contexts. After getting the stemmed words in place, further cleaning is performed using the lemmatizer.py file as follows:

- Changing all text to lowercase.
- Eliminating stopwords in multiple languages, including Danish, English, Dutch, Norwegian, Swedish, French, and German.
- Standardizing all Unicode characters to UTF-8 encoding.
- Omitting words that are too short (less than 3 characters).
- Eliminating hyperlinks.
- Removing emoticons.
- Discarding punctuation marks.
- Excluding numbers.
- Cleaning up extra spaces, as well as leading and trailing spaces.
- Filtering out empty text entries.



	full_text	text_clean
0	RT @swedennewyork: What does it mean to have a #feminist government? Join us on @reddit at 10 am (EST) Friday 3/23 for an #AskMeAnything se...	rt swedennewyork mean feminist government join us reddit est friday askmeanything se
1	Jens Stoltenberg åpnet virtuelle Rockheim i dag, sjekk det ut du også http://bit.ly/1B6Nt5	jens stoltenberg apnet virtuelle rockheim dag sjekk httpbitlybnt
2	@Panenka_Bart Veel beterschap Bart en Loes! 🌏	panenkabart beterschap bart loes
3	RT @CarlEmilLind: Tydelige S aftryk: \n\nPskykiatrien styrkes\n110% CO2 reduktion i 2050 \nPraktisk folkeskole + mesterlære\nAnnulationssøgsmål...	rt carlemillind tydelige aftryk pskykiatrien styrkes co reduktion praktisk folkeskole mesterlre annulationssgsmal
4	😃 https://t.co/bpqoZ8Y4wm	httpstcobpqozywm
5	Idag togs första spadtaget i Kigali för Biontechs första vaccinfabrik i Afrika. Ämnad för att producera vaccin i Afrika för Afrika. Skall försörjas helt med solenergi. Grundaren Ugur Sahin på plats förstås. Stolt över att EIB finansierade Biontechs utveckling av vaccin mot covid. https://t.co/qqXm6FlVCM	idag togs forsta spadtaget kigali biontechs forsta vaccinfabrik afrika amnad producera vaccin afrika afrika skall forsorjas helt solenergi grundaren ugur sahin plats forstas stolt eib finansierade biontechs utveckling vaccin covid httpstcoqqxmflvcm

Fig. 5: Examples of tweets pre and post cleaning

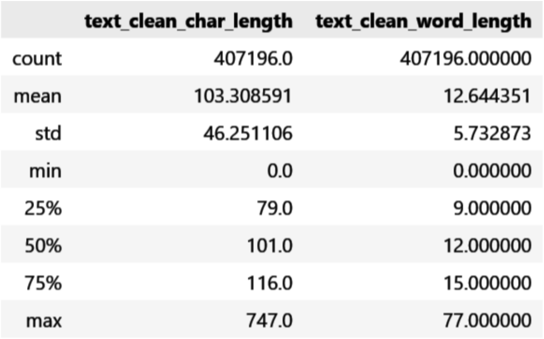
As depicted in Figure 5, the text has undergone substantial cleaning. The summary statistics for the newly created "text_clean" column (Figure 6) indicate the following:

• Both character and word lengths have significantly decreased, dropping from approximately 140 characters and 19 words to nearly 91 characters and 11 respectively, after cleaning the text. This means almost a 40% reduction.

• After cleaning the text, many rows had to be removed due to being empty. There has been around 0.7% reduction in number of rows.

• Also, there has been almost a 13% reduction in the maximum character and word length after cleaning the text.

We also performed topic modeling on the cleaned tweets data to get an idea of grouping various words under multiple topics. For this, we compared two algorithms - Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). We found that:



	text_clean_char_length	text_clean_word_length
count	407196.0	407196.000000
mean	103.308591	12.644351
std	46.251106	5.732873
min	0.0	0.000000
25%	79.0	9.000000
50%	101.0	12.000000
75%	116.0	15.000000
max	747.0	77.000000

Fig. 6: Statistics of clean words and character length

• The NMF model, utilizing the Forbenius norm, exhibits a skewed distribution of words, where a significant proportion is concentrated among the top 3-4 words. However, this skewness is notably improved when we apply the LDA model, indicating that the LDA model effectively normalizes the data. Additionally, the NMF model with Kullback-Leibler divergence (Fig. 7) performs even better in normalizing the word distribution.
• Comparing the scale of the plots, we observe that the NMF model employs a scale of (0, 2) to represent the frequency of the top 10 words, while the LDA model (Fig. 8) utilizes a scale of (0, 20000). This implies that the frequency of words within a single topic is much higher when the LDA model is employed, effectively covering a more extensive portion of the dataset.
• However, it's worth noting that the advantage of using the LDA model comes at the expense of time. The LDA model takes more than 12 minutes to complete the topic modeling process, whereas the NMF model achieves the same results in just about a minute or two.

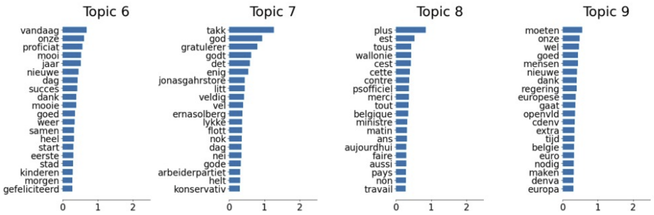


Fig. 7: NMF (Kullback-Leibler divergence) output

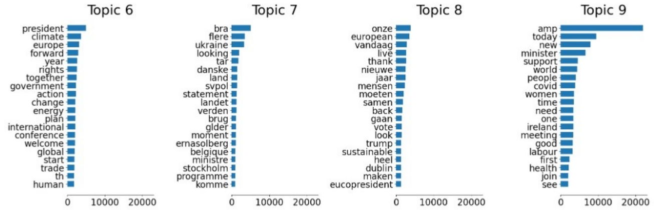


Fig. 8: LDA model output

III. Methods

Our approach comprises several key phases, including data analysis, data preprocessing, model development, and model ensemble techniques. We iterate through this process multiple times to enhance our predictive performance.

A. *Data Analysis*

During the initial data analysis phase, we extract two crucial insights from the training data, which subsequently contribute to improved prediction accuracy. These insights were previously outlined in the Data section. The first insight pertains to the distinct political views distribution among countries. It is imperative to incorporate this insight into our models, as a person's political views are significantly influenced by the country they reside in.

The second insight revolves around the distinct gender distributions observed in various countries. We recognize that one of the factors contributing to the differing political views across countries could be the distribution of genders. As such, we take this factor into account when refining our models.

B. *Data Preprocessing*

As outlined in the Descriptive Analysis section (Section B), we initiate the data preprocessing by cleansing the text data. Subsequently, we transform this cleaned text data into a Count Vectorizer function, also known as a CountVectorizer, which converts the cleaned tweets into a numerical format that can be processed by machine learning algorithms. This vectorized tweet is then converted into a TF-IDF feature matrix before supplying it to our training data for modeling. TF-IDF helps identify the importance of individual terms within a document relative to a collection of documents. It assigns higher weights to terms that are frequent within a document but relatively rare across the entire corpus of documents. We have identified a significant class imbalance issue within the dataset. Specifically, there is a notably low number of individuals with an independent political view. To address this concern, we employed random oversampling techniques. Ultimately, this effort resulted in an equalized distribution of data for each class, including 'Left,' 'Center,' 'Right,' and 'Independent.'

C. *Model Training*

We constructed several models, including the XGBoostClassifier, Naive Bayes Classifier for Multinomial Models, Linear Support Vector Machine (SVM), and Logistic Regression. Initially, we divided our training data into training and validation sets to evaluate the models' generalization performance. These models were trained individually using the training data, and their performance was assessed on the validation data. Notably, the model accuracies were quite similar, with the SVM achieving the highest accuracy.

Subsequently, we aimed to incorporate additional features related to both the country and gender into our models. Instead of simply adding these features as columns, we opted to create distinct models for each combination of country and gender. Following this, we implemented a voting mechanism, effectively establishing an ensemble model. This involved developing Naive Bayes Classifier, SVM, and Logistic Regression models for each unique country-gender pair. We then generated predictions for both the validation and test datasets using the corresponding country-gender model. The final prediction for each text was determined by selecting the prediction that received the most votes. For a detailed breakdown of the validation accuracy of our ultimate ensemble model for each country-gender combination, please refer to Figure 9. The overall validation accuracy of our final ensemble model reached 87%.

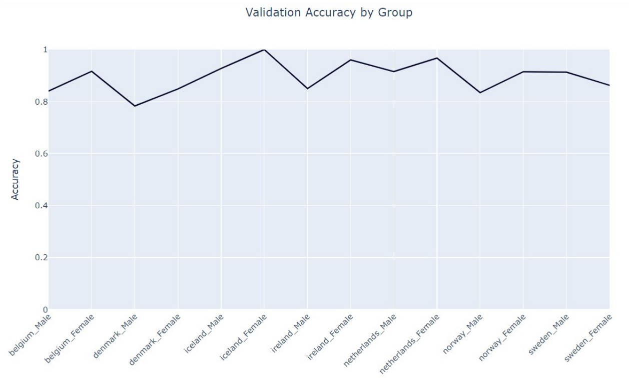
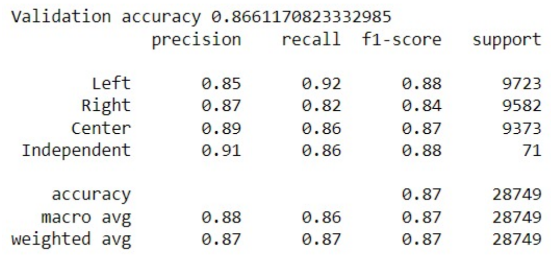


Fig. 9: Validation accuracy across groups

Furthermore, Figure 10 provides a visualization of the confusion matrix for our final ensemble model on the validation data, offering insights into its classification accuracy and any potential misclassifications.



```
Validation accuracy 0.8661170823332985
              precision    recall  f1-score   support

        Left       0.85      0.92      0.88      9723
       Right       0.87      0.82      0.84      9582
      Center       0.89      0.86      0.87      9373
 Independent       0.91      0.86      0.88        71

    accuracy                           0.87     28749
   macro avg       0.88      0.86      0.87     28749
weighted avg       0.87      0.87      0.87     28749
```

Fig. 10: Validation accuracy across groups

IV. Results

Our ultimate test data accuracy stands at 0.80479. We constructed models quickly in contrast to the time typically needed for developing deep learning models, achieving a high level of accuracy on the test dataset. We believe that the key insight is not necessarily creating a complex model, but rather crafting the simplest model that possesses the necessary complexity to process the provided features effectively. In this regard, the emphasis lies more on data analysis and feature engineering than on the complexity of the model itself. Additionally, we have learned that there are instances where we shouldn't expect the model to independently handle all the given features.

For our analysis, we developed distinct models for various subgroups categorized by attributes such as country and gender and also accounting for the multiple languages that the dataset contains. Looking ahead, we can explore alternative word vectorization techniques and deep learning methods like BERT as potential areas for further improvement.

References

[1] https://michael-fuchs-python.netlify.app/2021/05/22/nlp-text-pre-processing-i-text-cleaning/
[2] https://www.nltk.org/_modules/nltk/stem/snowball.html
[3] https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/
[4] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html
[5] https://towardsdatascience.com/use-voting-classifier-to-improve-the-performance-of-your-ml-model-805345f9de0e
[6] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html