Kaggle Challenge Report: Classification of Tweets of Politicians from Northern Europe

Mai Pham and Ciarrah Wang

***Abstract*— In this project, we explored the classification of tweets from Northern European politicians based on their political views using supervised machine learning models. Leveraging a comprehensive dataset, we implemented extensive text preprocessing, including cleaning, lemmatization, and feature extraction using both count-based and term-frequency-inverse-document-frequency (TF-IDF) methods. Our model of choice, the Linear Support Vector Classifier (SVC), demonstrated robust performance with an average cross-validation accuracy of 77.3% and a Kaggle competition accuracy of 78.219%. We also conducted topic modeling using Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) to analyze textual themes. While the model performed well, particularly for the Left and Right political views, misclassifications and underperformance in the Independent category underscored the challenges posed by class imbalance and overlapping language patterns. This project highlights the potential and limitations of machine learning in analyzing nuanced political discourse.**

I. INTRODUCTION

The goal of this project was to classify tweets from Northern European politicians into four political spectrums: Left, Center, Right, and Independent. By applying machine learning techniques, we aimed to predict the political view of users based on the textual content of their tweets. The dataset included features such as hashtags, tweet text, and metadata, providing a rich foundation for analysis. Our process involved data preprocessing to clean and standardize the text data, followed by feature extraction using both count and TF-IDF methods. Among the machine learning models tested—Linear SVC, Logistic Regression, and Random Forest—Linear SVC emerged as the best-performing model. Cross-validation confirmed the robustness of the Linear SVC model, achieving consistent accuracy scores across multiple folds. Despite its strong overall performance, the model struggled with the Independent category due to class imbalance and misclassifications, particularly between adjacent political views such as Left and Center. These findings underscore the challenges of textual data classification in politically nuanced contexts. To improve accuracy, future work could focus on addressing class imbalance through techniques like oversampling or adjusting class weights, and incorporating additional features such as sentiment analysis or user demographics.

II. DATA

In thid project, we used the provided training dataset training_data.xlsx and aim to use the test_data.xlsx to test our model. The percentages of tweets in the training dataset associated with different political views are 42.87% (Left), 25.75% (Center), 31.21% (Right), and 0.2% (Independent). The training dataset includes the following features:

- hashtags: The list of hashtags included in the tweet
- full_text: The text of the tweet (including emojis, HTMLs, and 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 any analysis, we created a table that contains information on minimum, average, median, and maximum for the tweet length (characters and words) and hashtag length (characters and words), as shown in Table I.

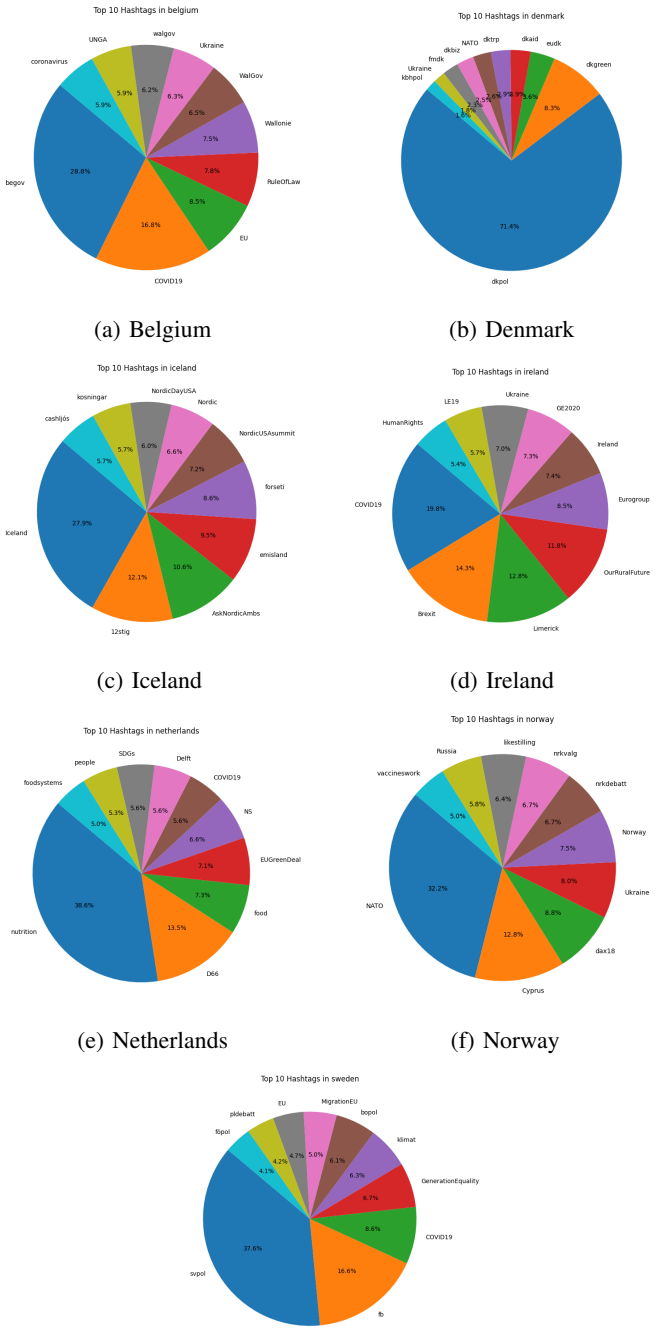
	Minimum	Average	Median	Mean
Tweet Char Length	4	167.304121	156.0	2994
Tweet Word Length	1	20.141102	19.0	89
Hashtag Char Length	1	6.459694	3.0	145
Hashtag Word Length	1	1.180230	1.0	16

TABLE I: Statistics Table for Tweet and Hashtag Lengths

A. Most Commonly Used Hashtags

Next, we found the most commonly used hashtags in each country and obtained the following pie charts in Fig. 1.
Cross-references to subfigures 1b, 1d, and 1g of figure 1.

From these hashtags, we found some Common Themes: (1) Political Themes: Countries like Belgium and Sweden often used hashtags about climate policies, such as #EUClimate or #UNGA. (2) Social Issues: Scandinavian countries had hashtags like #HumanRights or #Equality, showcasing their focus on inclusivity and human rights. (3) Global Topics: Certain hashtags, like #COVID19 or #ClimateCrisis, appeared across multiple countries, emphasizing shared global concerns. We also observed some patterns. From a Global versus Local Focus, we found that countries with active participation in global affairs (e.g., Belgium, Sweden) exhibited hashtags about climate, diplomacy, and international collaborations. Conversely, countries like Iceland and Denmark had a mix of global and localized issues, showing their dual focus on domestic and international relevance. From the perspective of event-driven trends we found hashtag popularity often reflected recent or ongoing events, such as political summits (#UNGA) or global crises (#COVID19), and some countries with upcoming elections or public movements showed hashtags centered on their political landscape.



(a) Belgium (b) Denmark

(c) Iceland (d) Ireland

(e) Netherlands (f) Norway

(g) Sweden

Fig. 1: Top 10 Hashtags in Each Country

B. Political Views by Country

In addition, we investigated the percentage of political views associated with each country and created a stacked bar chart shown in Fig. 2. From this chart, we observed (1) Dominance of Specific Political Views: Belgium, Denmark, and Sweden show a significant representation of left-leaning political views. The Left category occupies more than 50% in these countries, indicating a strong inclination towards progressive or liberal policies. Iceland has a higher proportion of Independent political views compared to other countries, though the Left category is also prominent. This suggests a unique political dynamic or potentially less alignment with traditional political ideologies. Norway and the Netherlands exhibit a balanced distribution between Left and Center, with the Right being less dominant in comparison to other countries. (2) Right-Wing Representation: Sweden and Denmark show higher proportions of right-leaning views compared to countries like Iceland and Norway. This may reflect ongoing debates or cultural factors contributing to a notable conservative presence in these regions. (3) Minimal Representation of Independents: Except for Iceland, the Independent category is almost negligible in other countries, typically comprising less than 5%. This indicates that most political engagement in these countries aligns with traditional Left, Center, or Right categories.

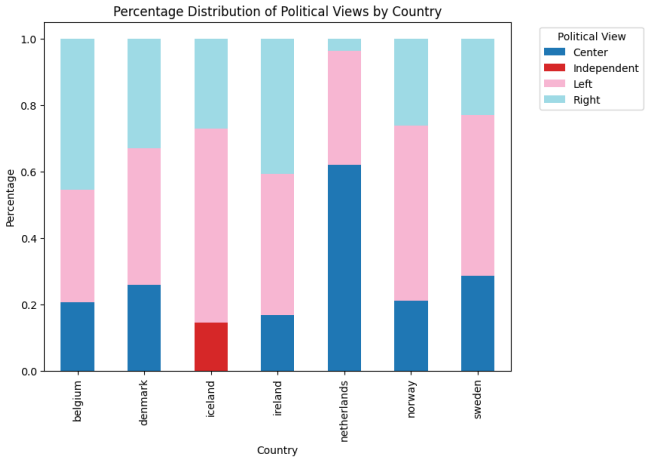


Fig. 2: Normalized Stacked Bar Chart of Political Views by Country

C. Gender Distribution by Country

Finally, we created a stacked bar chart for the distribution of genders by country too, as shown in Fig 3. In this chart, we observed (1) Male-Dominated Representation: All countries show a significantly higher percentage of male representation compared to females. This trend is evident across Belgium, Denmark, Iceland, Ireland, the Netherlands, Norway, and Sweden. The male percentage consistently exceeds 70%, reaching close to or over 80% in some countries like Belgium and Denmark. (2) Lowest Female Representation: Belgium and Denmark have the least female representation among the countries, suggesting a strong male-dominated dataset for these regions. (3) Highest Female Representation: Iceland and Sweden show relatively higher proportions of female representation, with females accounting for around 30-40% of the total. This is notably higher compared to other countries and might reflect better gender diversity in political or social participation for these regions. These findings highlight the need for increased gender balance in political representation, as the current trends still show

a significant underrepresentation of women in most of the countries analyzed.

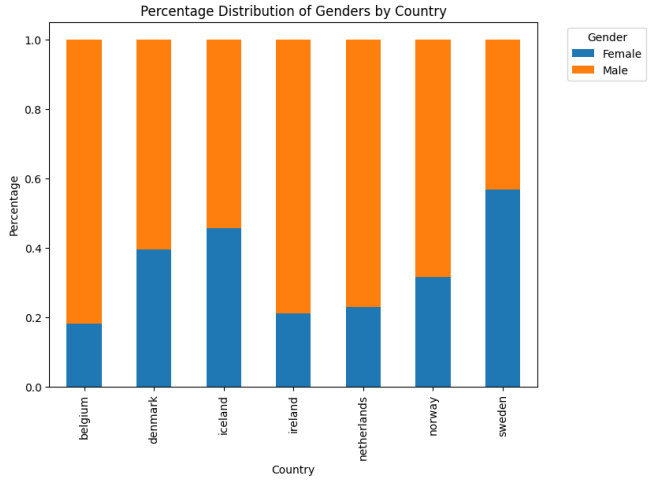


Fig. 3: Stacked Bar Chart of Gender Distribution by Country

D. Text Cleaning and Preprocessing

In this section, we implemented a text_cleaner function in the lemmatizer.py file to preprocess the full_text column in the training dataset. This function performs the following tasks. (1) Stopword Removal: Using the nltk.corpus.stopwords library, we removed common stopwords in English to focus on meaningful words. (2) Word Length Filtering: Words shorter than 3 characters were excluded to reduce noise in the data. (3) Link Removal: All links (starting with http) were removed using regular expressions to eliminate irrelevant content. (4) Emoji and Punctuation Removal: Emojis and punctuation were removed using Unicode regular expressions and the string.punctuation module.

After text cleaning, we applied a lemmatizer function to the preprocessed text to convert words into their base form. We expanded the previous descriptive statistics table by including statistics for the text_clean column, the expansion as shown in Fig 4.

	Minimum	Average	Median	Maximum
Tweet Char Length	4	167.304121	156.0	2994
Tweet Word Length	1	20.141102	19.0	89
Hashtag Char Length	1	6.459694	3.0	145
Hashtag Word Length	1	1.180230	1.0	16
Text Clean Char Length	0	92.311031	90.0	725
Text Clean Word Length	0	12.755716	12.0	77

Fig. 4: Updated Summary Table after lemmatization

E. Topic Analysis: LDA and NMF

We applied Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) on the text_clean column to extract ten top topics in an unsupervised manner. We extracted tf-idf features for NMF, and tf features for LDA. We fit the NMF model (Frobenius norm) with tf-idf features, using n_samples=400000 and n_features=1000, and get the result shown in Fig. 5.

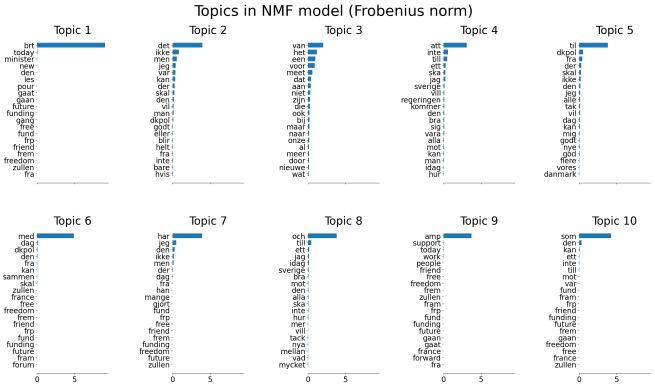


Fig. 5: Topics in NMF model (Frobenius norm)

Then, we fit the NMF model (generalized Kullback-Leibler divergence) with tf-idf features with the same number of samples and features, and get the result shown in Fig. 6.

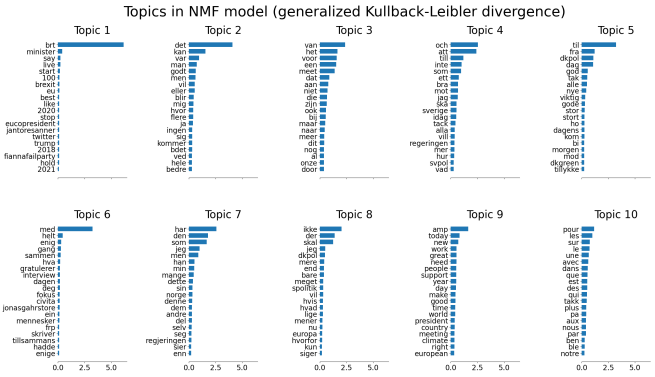


Fig. 6: Topics in NMF model (generalized Kullback-Leibler divergence)

In addition, we fit the MiniBatchNMF model (Frobenius norm) with tf-idf features, using n_samples=400000 and n_features=1000, and batch_size=128. The result is shown in Fig. 7.

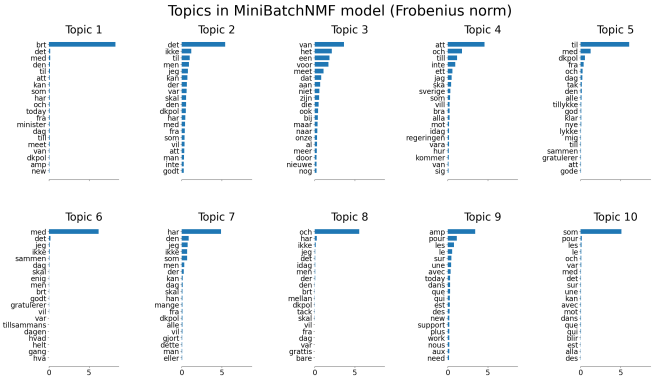


Fig. 7: Topics in MiniBatchNMF model (Frobenius norm)

We also fit the MiniBatchNMF model (generalized Kullback-Leibler divergence) with tf-idf features using the same sample, feature, and batch size, and get the result shown in Fig. 8.

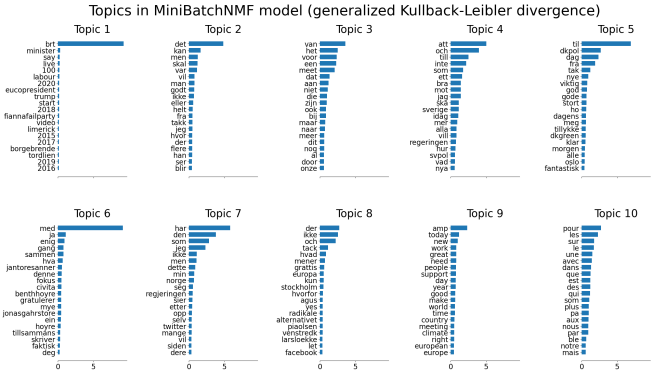


Fig. 8: Topics in MiniBatchNMF model (generalized Kullback-Leibler divergence)

Last but not least, we fit LDA models with tf features, using n_samples=400000 and n_features=1000, and get the result shown in Fig. 9.

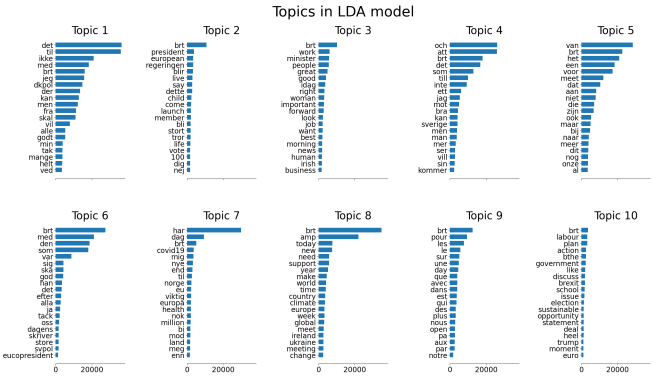


Fig. 9: Topics in LDA model

From the results, the LDA model and NMF models (with both Frobenius norm and Kullback-Leibler divergence) yield distinct patterns in how topics are structured and represented. LDA, a probabilistic model, captures a broader spectrum of frequent terms across topics, indicating a mix of language from different contexts, as seen in repetitive and generic terms like "brt," "amp," and "today." In contrast, NMF models, which are deterministic, tend to produce topics with clearer separations between terms, focusing more narrowly on specific themes or clusters. For example, in MiniBatchNMF, terms within topics are more cohesive, such as "fra," "dkpol," and "til" in political discussions. Additionally, the MiniBatchNMF with Kullback-Leibler divergence highlights contextual nuances in political and societal conversations, revealing underlying latent themes better than LDA in some instances. Overall, while LDA captures general trends, NMF models offer sharper topic distinctions, making them potentially more interpretable for nuanced textual datasets like this one. However, overlapping key terms across all methods, such as "brt" and "minister," suggest shared importance of certain discussions. This underscores the importance of selecting topic modeling techniques based on the desired granularity and interpretability for the project goals.

III. METHODS

For the mean purpose of the project, we focused on predicting the "pol_spec_user" variable using supervised machine learning models. The process began with importing the clean training dataset and handling missing values. Specifically, missing values in relevant columns were replaced with empty strings to ensure a consistent input for text processing. The text data from multiple columns were then concatenated into a single string to serve as the input for feature extraction.

We explored three machine learning models for classification: Linear Support Vector Classifier (Linear SVC), Logistic Regression, and Random Forest. Due to the high computational cost of Random Forest, which took more than an hour to train, we decided to interrupt the process and exclude it from further consideration. Between the two remaining models, Linear SVC outperformed Logistic Regression in terms of accuracy, so we selected Linear SVC as the final model.

To optimize feature representation, we constructed a FeatureUnion to combine two text feature extraction methods: svc_count_vector and svc_tfidf_vector. This approach ensured that the model leveraged both frequency-based and term importance-based textual features. Next, we built a pipeline comprising the FeatureUnion and the Linear SVC classifier to streamline preprocessing and model training.

The pipeline was then fitted to the training data, and predictions were generated for the testing dataset. This method ensured a robust and scalable process for handling textual data, improving both the efficiency and accuracy of the predictive model.

To evaluate the strength of the model, we used cross-validation to calculate its accuracy. The cross-validation process provided a detailed assessment of the model's performance across different data splits, yielding accuracy scores: [0.77301246, 0.77241083, 0.77075327, 0.77367271, 0.77454447]. These scores demonstrated consistent performance, with an average accuracy of approximately 77.3%. The consistency across folds highlighted the model's robustness and generalizability, confirming its suitability for the task.

IV. RESULTS

The results of the Linear SVC model on the training dataset demonstrate a robust performance with an average cross-validation accuracy of approximately 77.43% and an accuracy score of 78.219% through Kaggle. This indicates that the model is effective in predicting the political spectrum labels (pol_spec_user) for the tweets.

The confusion matrix (Fig. 10.) provides deeper insights into the model's strengths and weaknesses. The matrix reveals that the model performs well for the Left and Right

categories, with the highest number of correct predictions for these labels. For instance, 29,140 Left tweets and 19,038 Right tweets were correctly classified. However, misclassifications are evident, particularly between Center and Left, as well as between Right and Left, which suggests some overlap in the features or language patterns used in these categories.

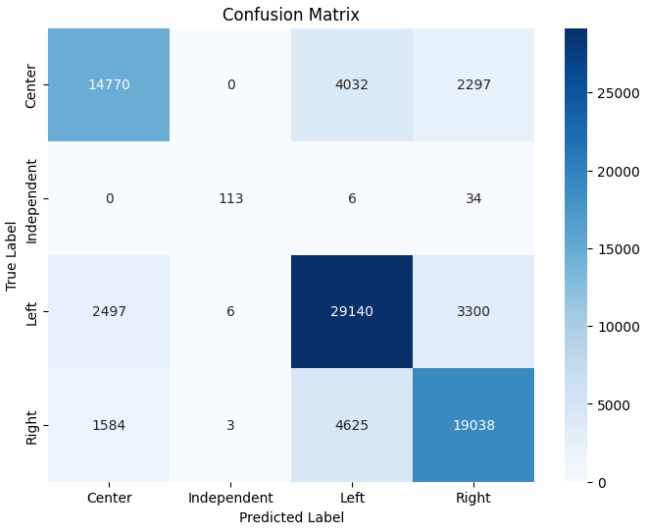


Fig. 10: Confusion Matrix

The Independent label is the smallest class and is poorly predicted, with only 113 correctly classified examples. This could be attributed to class imbalance, as the Independent class constitutes less than 1% of the dataset. Although class weighting was incorporated into the model, it may not have been sufficient to counteract the imbalance, leading to poor generalization for this class.

Despite the model's decent accuracy, there is room for improvement. One limitation of the approach is the reliance on linear methods and the absence of advanced feature engineering specific to linguistic nuances in political discourse. Incorporating domain-specific knowledge or experimenting with non-linear models like ensemble methods or neural networks could potentially enhance predictive performance. Additionally, the preprocessing steps, such as text vectorization and sentiment analysis, were standard and might not have captured all relevant contextual information in the data.

From a critical standpoint, our team's approach focused primarily on optimizing the model with existing features. However, more exploratory analysis and feature engineering tailored to the dataset's specific characteristics might have yielded better results. Future iterations should address class imbalance more effectively and explore more sophisticated NLP techniques, such as contextual embeddings, to improve the model's understanding of textual data.

V. CONCLUSION

This project demonstrated the utility of machine learning in classifying political views based on tweet content, achieving robust results with the Linear SVC model. By employing rigorous text preprocessing and leveraging feature extraction techniques, the model achieved an average cross-validation accuracy of 77.43%. While the model performed well for the Left and Right categories, it struggled with the Independent class, highlighting the challenges of class imbalance and overlapping language features. The confusion matrix revealed areas of misclassification that suggest opportunities for further refinement, such as incorporating domain-specific knowledge or advanced natural language processing (NLP) techniques. Future work could explore ensemble models or neural network approaches to improve classification performance, especially for underrepresented categories. Overall, this project provides valuable insights into the potential of machine learning for political text analysis while acknowledging the complexities inherent in such tasks.

REFERENCES

[1] University of Rochester, "Kaggle Project: Classification of Tweets of Politicians from Northern Europe", Fall 2024: Data Mining,