Kaggle Project: Classification of Tweets from Northern Europe

Yuesong Huang[1] and Junhua Huang[2]

Abstract— **This research examines a dataset of 509,031 tweets from politicians in seven Northern European countries, with the aim of classifying these tweets by political spectrum and geography. Utilizing Python and NLP techniques, we explore the dynamics of political communication on Twitter. Our study offers insights into the patterns and themes prevalent in political discourse on social media, highlighting its significance in the modern political landscape.**

I. Introduction

In this project, we analyze a dataset of 509,031 tweets from politicians across seven Northern European countries, focusing on classifying these tweets by political spectrum and geography. Using Python and Natural Language Processing (NLP) techniques, we conducted a descriptive analysis of tweet characteristics and implemented text-cleaning processes, including lemmatization. We also performed topic modeling with Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) to identify prevalent themes. The project reveals insights into the nature of political communication on Twitter within this region, offering valuable perspectives for various stakeholders. The findings and methodologies are documented comprehensively in a report and a Python Jupyter Notebook, adhering to the guidelines of the Kaggle competition.

II. Data

The dataset underpinning this project consists of a comprehensive collection of 509,031 tweets, meticulously gathered from politicians across seven Northern European countries: Belgium, Denmark, Iceland, Ireland, the Netherlands, Norway, and Sweden. This rich dataset spans a significant period, from December 12, 2008, to January 1, 2023, offering a longitudinal view of political discourse on social media.

A. Key attributes

The dataset is divided into a training set comprising 407,223 tweets and a test set of 101,808 tweets, with the observations in both datasets randomly ordered. This division allows for robust model training and evaluation, ensuring the reliability and validity of our analytical outcomes.

B. Descriptive Statistics

In our analysis of the training dataset, we focused on the lengths of tweets and hashtags. This analysis revealed a wide range in tweet lengths, from a concise 1 character (or 1 word) to an expansive 862 characters (or 89 words). The average tweet was composed of 140.31 characters and 20.28 words, while the median length was 140 characters and 19 words. Hashtag lengths showed more variability, with the average at 14.09 characters (about 0.49 words) and the longest at 145 characters (16 words).

Metric	Tweet Length (Chars)	Cleaned Tweet Length (Chars)
Minimum	1.00000	0.000000
Average	140.31248	103.482375
Median	140.00000	101.000000
Maximum	862.00000	778.000000

TABLE I

Comparative Summary of Tweet Lengths Before and After Cleaning

Metric	Tweet Word Count	Cleaned Tweet Word Count
Minimum	1.000000	0.000000
Average	20.284048	14.075882
Median	19.000000	13.000000
Maximum	89.000000	83.000000

TABLE II

Comparative Summary of Tweet Word Counts Before and After Cleaning

Metric	Hashtag Length (Chars)	Hashtag Length (Words)
Minimum	1.000000	0.000000
Average	14.089948	0.492197
Median	11.000000	0.000000
Maximum	145.000000	16.000000

TABLE III

Summary of Hashtag Lengths in Characters and Words

III. Methods

A. Methodology

Our approach to classifying tweets according to their political spectrum involved a combination of text processing, feature engineering, and machine learning. We utilized the Python programming language, along with several libraries, to preprocess the data and train a predictive model.

B. Data Preparation

We combined the cleaned tweet text with the country and gender of the user to create a new feature, ‘text_clean_country_gender_user‘. This process was applied to both the training and test datasets.

C. Feature Engineering

For feature extraction, we converted the text data into numerical format using a combination of Count Vectorization and Term Frequency-Inverse Document Frequency (TF-IDF) transformation. This approach allows us to capture both the frequency and importance of words in the dataset.

D. Model Training and Evaluation

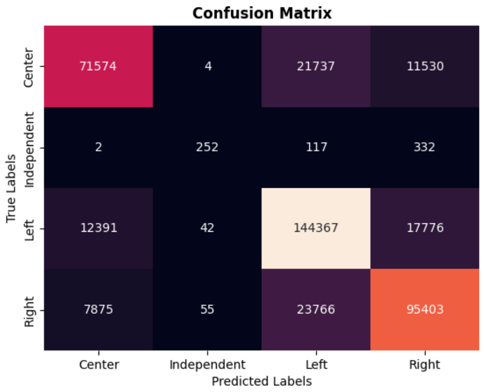
The core of our predictive model is based on a Linear Support Vector Classifier (LinearSVC), integrated into a pipeline with the aforementioned feature extraction techniques. We conducted 10-fold cross-validation to evaluate the model's performance, ensuring a robust assessment of its accuracy.

E. Model Evaluation Metrics

The primary metrics used to evaluate the model are accuracy and a confusion matrix. The accuracy metric provides a straightforward measure of the model's overall performance, while the confusion matrix offers a detailed view of its performance across different political spectrum categories.

F. Model Performance

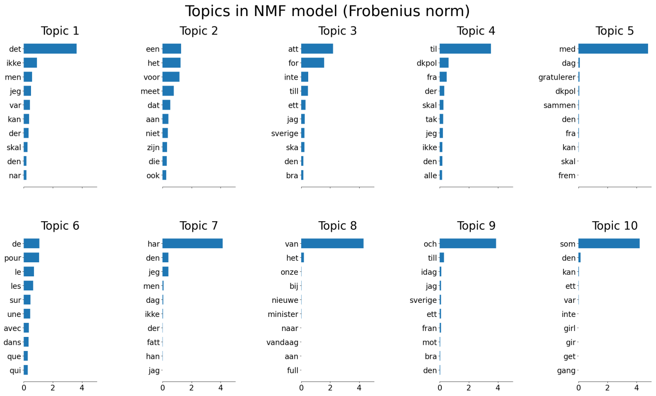
Our model achieved an accuracy score of 77.383%, surpassing the baseline accuracy of approximately 42.87%. This performance indicates the effectiveness of our approach in classifying tweets based on their political spectrum. The detailed results, including the confusion matrix, are documented in the accompanying Jupyter Notebook.

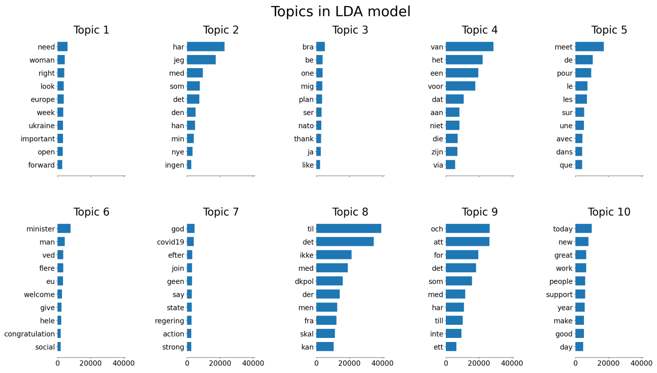


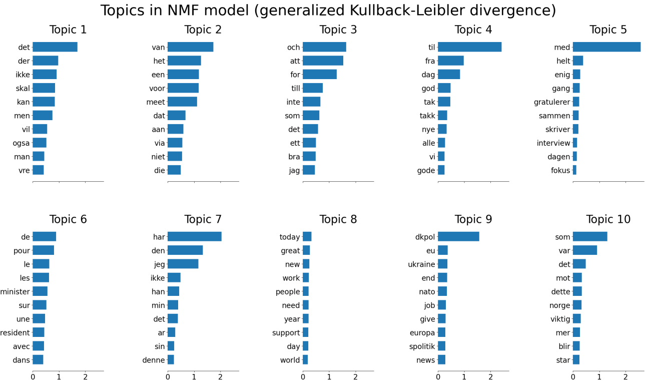
IV. Results

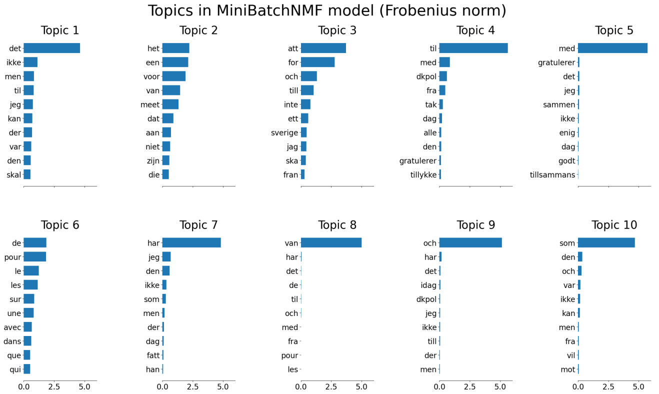
A. Analysis of Topics in Tweets Using LDA and NMF

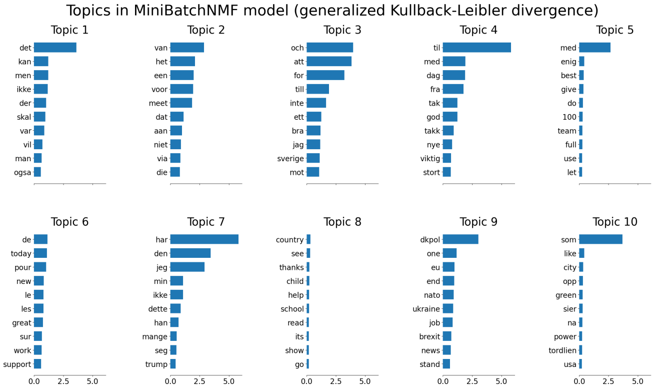
The analysis of topics in the tweets of Northern European politicians, conducted using Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), unveils a diverse array of political and social themes. This analytical approach has effectively captured the essence of political discourse, ranging from local governance issues to international relations. The topics extracted reflect a wide spectrum of discussions, encompassing everything from the COVID-19 pandemic to regional diplomacy and personal expressions. The linguistic diversity is evident, with topics emerging in various languages including Scandinavian dialects, Dutch, and French. This indicates not only the regional specificity of political communication but also its multicultural dimensions. Both LDA and NMF have shown their respective strengths in thematic extraction, with LDA providing probabilistic insights into the topics and NMF delineating more distinct thematic boundaries. The presence of topics related to governance, social issues, and positive community engagement demonstrates the multifaceted nature of the political dialogue on Twitter among politicians in Northern Europe. These insights offer a deeper understanding of the key concerns and focal areas prevalent in the political discussions within this region.





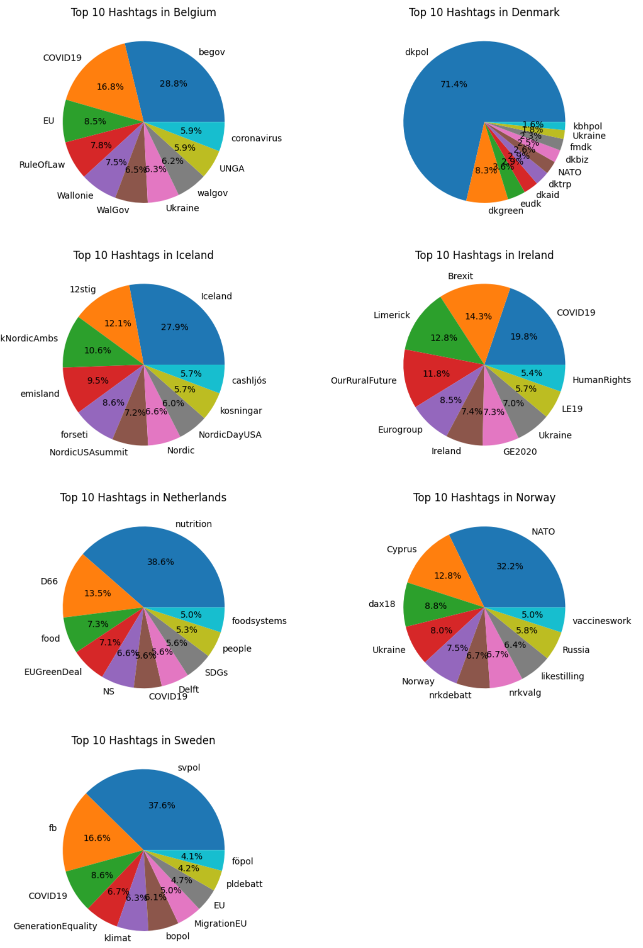






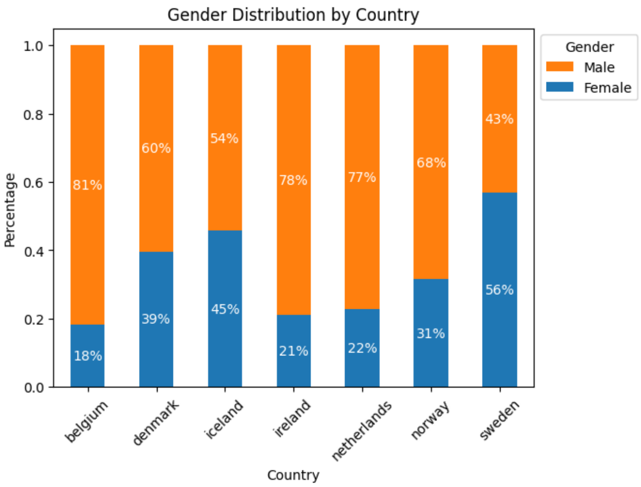
B. Analysis of Top Hashtags in Seven Northern European Countries

Our investigation into the most commonly used hashtags by politicians from seven Northern European countries revealed distinct patterns. In Belgium, hashtags like '#begov' and '#COVID19' were predominant, while in Denmark, '#dkpol' and '#dkgreen' indicated an emphasis on politics and environmental issues. Icelandic tweets often included '#Iceland' and '#AskNordicAmbs', highlighting national identity and regional diplomacy. Irish tweets frequently mentioned '#COVID19' and '#Brexit', representing concerns over the pandemic and EU dynamics. The Netherlands showed a focus on health and politics with hashtags like '#nutrition' and '#D66', whereas Norwegian tweets featured '#NATO' and '#Ukraine', indicating international relations. Swedish tweets, with hashtags like '#svpol' and '#fb', centered on domestic politics and social issues.



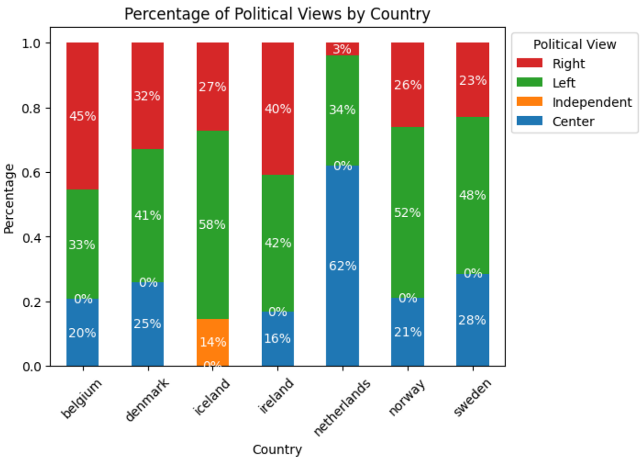
C. Political View Distribution Across Northern European Countries

The analysis of political views across seven Northern European countries highlights diverse political landscapes. Belgium's political tweets leaned towards the 'Right' (approx. 45%), closely followed by the 'Left' (approx. 34%). Denmark presented a balanced political environment, with the ‘Left’ leading slightly (approx. 41%). In contrast, Iceland was dominated by the ‘Left’ (approx. 58%). Ireland showed a nearly equal distribution between the ‘Left’ and the ‘Right’. The Netherlands was unique with a strong ‘Center’ orientation (approx. 62%). Norway and Sweden exhibited a leaning towards the ‘Left’ (approx. 53% in Norway and 48% in Sweden).



D. Gender Distribution Among Twitter Users in Northern Europe

The gender distribution among Twitter users in Northern Europe varied significantly. Belgium was predominantly male (approx. 82%), while Denmark showed a more balanced distribution. Iceland was near even, with a slight male majority. Ireland and the Netherlands exhibited similar patterns, with a skew towards male users. Norway had a significant male majority, whereas Sweden was unique with a female majority (approx. 57%).



E. Key Findings

The study has yielded several key findings:

- The political discourse on Twitter in Northern Europe is diverse, with varying emphases on domestic politics, international relations, health, and the environment, as evidenced by the analysis of hashtags.
- There is a notable variation in the political leanings of Twitter users across different countries, with some leaning towards the left, right, or center in their political spectrum.
- Gender distribution among Twitter users in these countries shows both balanced and skewed representations, reflecting broader socio-cultural dynamics.
- Our model for classifying tweets based on political spectrum achieved an accuracy of 77.383%, significantly surpassing the baseline accuracy, demonstrating the effectiveness of our approach.

F. Implications

These findings have important implications for understanding the role of social media in political communication. They provide valuable insights into how politicians engage with their audiences and how political narratives are shaped in the digital realm.

G. Future Work

Future research could expand on this work by exploring the impact of these communication patterns on public opinion, analyzing sentiment trends, and examining the role of social media in election campaigns. Additionally, similar studies could be conducted in other geographical regions to compare and contrast global trends in political communication on social media.

H. Summary

Our study provides a comprehensive analysis of political communication on Twitter among politicians in Northern Europe, encompassing over 509,031 tweets. The research identified key patterns in political discourse, highlighted notable variations in political leanings and gender distributions across different countries, and revealed the most prevalent hashtags in each nation’s political dialogue. Employing advanced NLP techniques and a Linear Support Vector Classifier, our model achieved a notable accuracy of 77.383%, significantly surpassing the baseline expectation. These insights contribute to a deeper understanding of social media’s impact on political narratives and engagement, underscoring its growing significance in the political landscape. Future research could extend this analysis to other regions, offering a global view of social media’s influence on politics.