Political Ideology: Prediction of DW-Nominate Scores

Eugene Ayonga[1] and Homayra Tabassum[2]

Abstract— **In this study, we examine the ideological positioning of U.S. Congressional members by utilizing DW-Nominate scores and Twitter data. These scores measure lawmakers' ideological stances across two dimensions: the first capturing the liberal-conservative spectrum, and the second reflecting regional and historical differences. Our dataset comprises 469,740 tweets from Congressional members. The models employed are H2O AutoML Stacked Ensembles of Gradient Boosting Machines (GBM) and Generalized Linear Models (GLM), achieving an RMSE of 0.29886. By predicting the ideological dimensions from tweet data, this research provides insights into how public communication on social media correlates with Congressional voting behavior, enhancing the understanding of legislative polarization and partisanship.**

I. INTRODUCTION

This study aims to explore the relationship between Congressional members' public communication and their ideological positions, as measured by DW-Nominate scores, using social media data from Twitter. DW-Nominate scores are a widely used tool in political science to map U.S. legislators on a two-dimensional ideological space. The first dimension reflects the liberal-conservative spectrum, while the second dimension captures historical and regional distinctions. Given the importance of these scores in understanding partisanship and legislative behavior, predicting these dimensions from public data, such as tweets, offers a novel approach to analyzing political polarization.

H2O AutoML Stacked Ensemble models, combining Gradient Boosting Machines (GBM) and Generalized Linear Models (GLM), were used to predict the DW-Nominate scores. Two of the most significant findings were:

Model Performance: The model's performance, as evaluated by average RMSE, achieved a score of 0.29886, significantly improving upon the baseline RMSE of 0.36. By leveraging the intersection of social media behavior and political ideology, this research sheds light on the evolving role of digital platforms in shaping public perception and understanding legislative polarization in the U.S. Congress.

DW-Nominate Scores: The analysis revealed a noteworthy relationship between congressional members and their political ideology, by means of the DW-Nominate scores. The model's ability to correctly predict DW-Nominate1 scores demonstrates its competence in detecting the primary political leanings of legislators, as left-leaning tweets were correctly classified in the range -1 to 0; and right-leaning tweets were correctly classified in the range 0 to 1. This makes sense as the dimension typically explains the majority of roll-call voting behavior. DW-Nominate2 results were less varied, reflecting divisive issues that divide or bring together legislators beyond the liberal-conservative axis, such as regional interests, military aid to foreign allies or even socioeconomic factors.

*This work is in partial fulfillment of the DSCC 483 course requirement

[1]E. Ayonga is an MS. Data Science Candidate at the Goergen Institute for Data Science, University of Rochester, 1209 Wegmans Hall, Rochester, NY - 14627, USA Email: eayonga@ur.rochester.edu

[2]H. Tabassum is an MS. Data Science Candidate at the Goergen Institute for Data Science, University of Rochester, 1209 Wegmans Hall, Rochester, NY - 14627, USA Email: htabassu@ur.rochester.edu

II. DATA

A. Structure

The dataset comprises 469,740 tweets from U.S. Congressional politicians, with a training set of 333,987 tweets and a test set of 135,753 tweets. The tweets span from 2008 to 2020, offering a rich source of data to assess how political communication reflects voting behavior. Key features include tweet text, retweet and favorite counts, hashtags, and year of posting. The outcome variables are the first and second dimensions of the DW-Nominate scores, with a prediction task centered on achieving accurate estimates of these ideological positions.

B. Congressional Tweet Hashtags

The congressional twitter data covers a timeframe that spans the time President Barack Obama was elected to the Covid-19 pandemic. The top ten tweet hashtags were graphed as shown below:

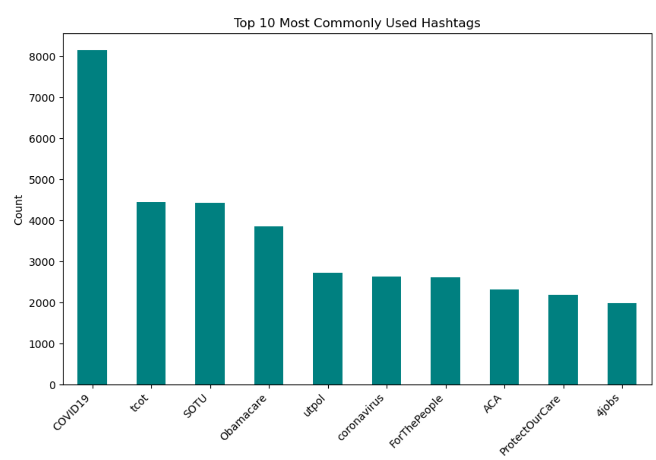

Fig. 1. Top Ten Commonly Used Tweet Hashtags

In reference to the graph above, COVID19 and coronavirus hashtags are related to the global COVID-19 pandemic, indicating a significant focus on discussions around this topic.

tcot stands for 'Top Conservatives on Twitter' a hashtag often used by conservative users. This makes sense since a consertative president won the 2016 election. Discussions around the State of the Union (SOTU) are seen to be of statistical significance as congressional members are usually charged around this event in either supporting the presidential address or countering the issues being addressed. The hashtags Obamacare and ACA stand for Affordable Care Act, also known as Obamacare, reflecting discussions on President Obama's flagship healthcare policy during his eight-year administration. This policy became contentious as President Trump's campaign ran on the platform of ending ACA, but eventually fell by a no-vote from conservative senator John McCain. The hashtag ForThePeople refers to the campaign slogan of the then U.S Senator from California, Kamala Harris when she vied for the 2020 presidential election. Lastly, 4jobs relates to discussions around job creation and economic policies. These hashtags indicate a strong focus on healthcare, economic policies, and the COVID-19 pandemic, reflecting the major issues of the time when these tweets were made.

C. DW-Nominate Score Group Hashtags

Group 1 represents a conservative stance on both dimensions, showing a strong focus on hashtags related to conservative politics. The most frequently used hashtag is tcot (Top Conservatives on Twitter), indicating the group's active participation in conservative online discourse. Other notable hashtags include Obamacare, which conservatives tried to repeal, and COVID19, suggesting that healthcare and the pandemic were prominent issues within this group. The presence of hashtags like TaxReform further reinforces their focus on fiscal policies.

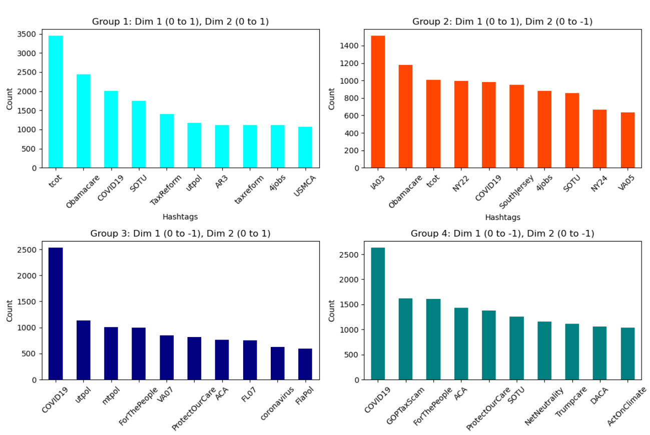

Fig. 2. Hashtags by DW-Nominate Score Groups

Group 2 reflects a conservative stance on the first dimension, but a more liberal stance on the second dimension, which often reflects regional and urban-rural divides. This is evident in their frequent use of the hashtag IA03, focused on Iowa's 3rd congressional district — a specific regional area with both urban and rural characteristics and occasionally flips to both parties interchangeably. The group's focus on such political hashtags suggests an engagement with regional politics, balancing conservative values with an awareness of local and district-specific issues. Additionally, hashtags like Obamacare and COVID19 indicate that healthcare and the pandemic are significant topics for this group.

Group 3 represents a liberal stance on the first dimension and a conservative stance on the second, highlighting a focus on healthcare and governance issues. The most frequently used hashtag is COVID19, indicating that the pandemic is a central topic of discussion. Other prominent hashtags include utpol (Utah Politics) and ForThePeople. Additionally, healthcare-related hashtags like ProtectOurCare and ACA point to a focus on protecting and improving healthcare policies rather than repealing such policies.

Group 4 depicts a liberal stance on both dimensions, focusing heavily on progressive issues. The most frequently used hashtag is COVID19, showing the widespread importance of pandemic-related discussions across the political spectrum. Other top hashtags include GOPTaxScam and ForThePeople, indicating criticism of conservative policies and a push for governance reforms. Additionally, hashtags like NetNeutrality and DACA suggest that this group is engaged in conversations about environmental and immigration issues.

COVID-19-related hashtags are prominent across all groups, reflecting the significant impact of the pandemic on political discourse. Political and policy-related hashtags (tcot, SOTU, Obamacare) are commonly used, indicating that political events and healthcare policies are central to discussions within these groups. Conservative groups focus more on issues like tax reform and district-specific politics, while liberal groups emphasize healthcare, social justice, and governance reforms.

D. Political Ideology

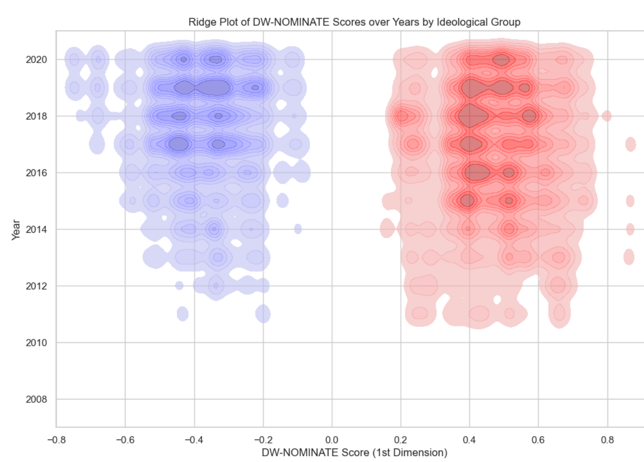

Fig. 3. Ideological Spectrum by DW-Nominate 1 Scores

Conservatives (Red): The distribution of conservative DW-Nominate scores remains concentrated around the 0.4 to 0.6 range throughout the years depicted. There is a slight trend of scores becoming more concentrated around these positive values over time, suggesting that conservative lawmakers have become more consistently conservative in their voting behavior.

Liberals (Blue): The liberal group's DW-Nominate scores are more spread out, generally occupying the -0.2 to -0.6 range. The distribution appears to become more condensed around these negative values in more recent years, which might indicate a trend towards a more unified liberal voting behavior.

Observations Over Time: Both groups show continuity in their ideological positions over the years, with no dramatic shifts observed in the distribution of scores between 2008 and 2020. There is, however, some increase in polarization, as the central peak of conservative values (around 0.4) seems to be further from the peak of liberal values (around -0.4), indicating that the ideological divide between conservatives and liberals has widened slightly.

Polarization: The two distributions remain largely separate, with minimal overlap between the conservative and liberal groups. This separation suggests that there is a clear ideological divide between the two groups, which is a key indicator of polarization. Over the years, the separation remains stable, reflecting ongoing polarization where each group maintains its distinct ideological position without significant overlap.

Stability in Ideology: The ridge plot does not indicate a dramatic shift in either direction, which suggests that while the ideological divide is clear, it has not been exacerbated further in the recent period. This stability may indicate that external events have not significantly altered the core ideologies of lawmakers in these groups during the period shown.

E. Topic Modelling

Latent Dirichlet Allocation (LDA) Model: Topics generated are generally more coherent in the sense that they seem to form more interpretable groups of words. For example, Topic 3 focuses on healthcare-related terms while Topic 6 is about political office and duties. The model is based on the assumption that each document (tweet) is a mixture of topics, and each topic is a mixture of words. This probabilistic approach tends to yield topics that are fairly easy to interpret, though there can still be noise or overlap between topics.

Some topics share common words, suggesting overlap or similarity. For example, the words: work, help, and need appear in multiple topics, indicating these are common themes across several discussions. Despite some commonality, the themes in LDA are distinct enough to differentiate between broad areas like healthcare, government activities, and discussions around leadership or governance.

Non-negative Matrix Factorization (NMF) Model: The topics generated by NMF are also coherent but appear to be more distinct from each other compared to LDA. NMF tends to create sharper distinctions between topics because it is an additive model, meaning it only combines parts (words) in a positive way, which can result in more distinct topics. For instance, Topic 1 in NMF focuses on words like: help, need, and family, which suggests a theme around assistance or support, while Topic 6 centers on healthcare, with words like health, care, access, and affordable.

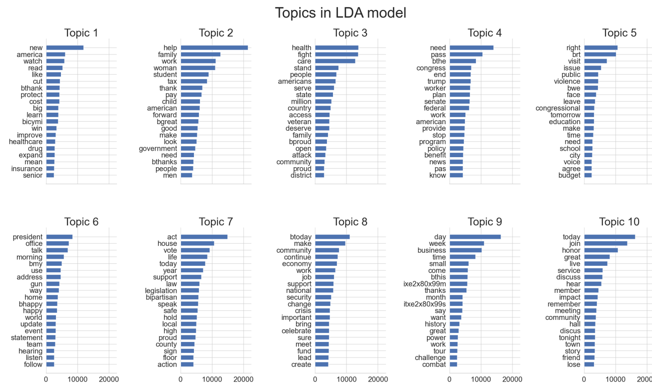

Fig. 4. LDA Topic Modelling

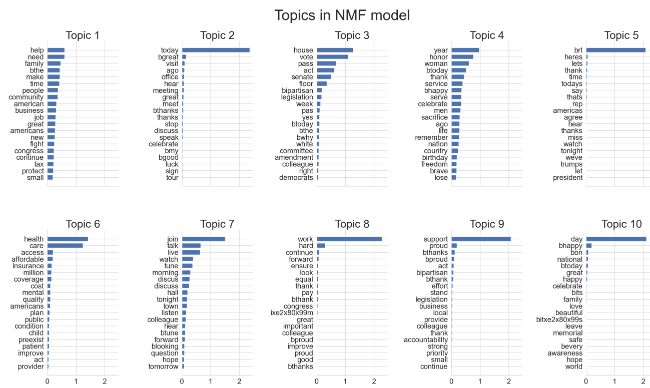

Fig. 5. NMF Topic Modelling

Compared to LDA, the topics in NMF seem to have fewer overlaps, meaning that words are less likely to appear in multiple topics. This makes NMF a good choice for applications where clearly defined topics are more valuable than broadly defined ones.

Comparatively, both LDA and NMF produce coherent topics, but the coherence in LDA is driven by the probabilistic distribution of words across topics. In contrast, NMF's coherence is driven by the additive combination of features, resulting in distinct topic separations. NMF tends to produce more distinct topics with clearer boundaries, while LDA may produce topics that overlap more, which could be interpreted as capturing the complex, multifaceted nature of discussions within the corpus.

LDA may be easier to interpret when you want to understand the mixture of themes within documents. NMF, on the other hand, may be more useful when you want to identify and separate out distinct themes that do not overlap significantly. In LDA, some words appear across multiple topics, reflecting shared themes or concepts. NMF tends to have less of this, which might be better for identifying unique aspects of the data. In summary, LDA is suitable for understanding the broader themes while NMF is useful for identifying distinct categories within the data.

III. METHODS

Data Pre-processing

The initial phase of our methodology entailed a meticulous data pre processing procedure. The first step was to import all necessary Python packages that would be required throughout the modeling process, downloading nltk packages for natural language processing (NLP) tasks, and installing the h20 AutoML package.

The training and testing datasets were then loaded and inspected for missing values. The full text column was passed through an NLP pipeline to remove stopwords, short words that had no additive meaning, url links, emojis and punctuation. This new text was then lemmatized so as to break down words to their root meaning, then a new column with clean text was built. Topic modelling was then done on the cleaned text using both Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), to arrive at optimal topics.

Features columns were then separated from the main dataframe. This bifurcation was essential to independently handle the features and the target variables: dim1 nominate and dim2 nominate.

Model Setup and Training

Post pre-processing, training data was partitioned into training and cross-validation sets, adhering to a conservative 95/5 split to ensure a robust training process.

A h20 AutoML model was chosen and initialized for its powerful ensemble learning capabilities and its proficiency in handling various types of machine learning models. The raining dataframe was then converted into a h20 frame. The hyper parameters such as maximum runtime, and stopping metrics were thoughtfully selected to give the model a substantial iteration base to learn from. Two models were trained to predict both DW nominate scores individually, and then cross-validated to ascertain the robustness of the trained models. The models with the lowest RMSE scores were then selected from the computed leaderboards for the model evaluation stage.

Model Evaluation

The strength of the models were assessed using the average Root Mean Square Error (RMSE). RMSE offers a clear indication of the average distance between the predicted and actual values.

The same pre processing steps were applied to the unseen test dataset to ensure consistency. The trained models was then applied to this data to generate nominate score predictions.

A new DataFrame was constructed consisting of the index and the predicted DW nominate scores for both dimensions, formatted into a CSV file, and then submitted to the Kaggle competition. The model's performance was validated externally with an average RMSE score of 0.29886, illustrating a high level of accuracy and predictive strength by the models.

IV. RESULTS

The results obtained from the predictive modeling of DW-Nominate scores in congressional members' tweet data between 2008 and 2020 yielded an average RMSE score of 0.29886. This low value indicates that the models predictions are closer to the actual data points, and therefore more accurate in predicting the target variables; in this case the DW-Nominate scores for both dimensions.

TABLE I
DW NOMINATE 1 BEST MODEL METRICS

Mean Squared Error (MSE)	0.1211
Root Mean Squared Error (RMSE)	0.3452
Mean Absolute Error (MAE)	0.2780
Root Mean Squared Logarithmic Error (RMSLE)	0.3784
Mean Residual Deviance)	0.1211

TABLE II
DW NOMINATE 2 BEST MODEL METRICS

Mean Squared Error (MSE)	0.0589
Root Mean Squared Error (RMSE)	0.2408
Mean Absolute Error (MAE)	0.1870
Root Mean Squared Logarithmic Error (RMSLE)	0.3560
Mean Residual Deviance)	0.0590

Accuracy Evaluation

The RMSE (Root Mean Squared Error) serves as the primary metric for evaluating model accuracy in this study. With an average RMSE score of 0.29886, the model demonstrates strong predictive performance, indicating that the predicted DW-NOMINATE scores are generally close to the actual values. While the RMSE is an absolute measure of accuracy, it is worth noting that RMSE values should be interpreted in relation to the scale of the target variable. In this case, the low RMSE suggests minimal prediction error.

Further metrics, such as Mean Squared Error (MSE) and Mean Absolute Error (MAE), provide additional context to assess the model's error distribution and robustness. The best-performing model for DW Nominate 1 yielded an RMSE of 0.3452, while for DW Nominate 2, the RMSE was 0.2408, which underscores the model's capacity to handle both dimensions of the ideological spectrum effectively. Given that these scores are considerably below the baseline RMSE of 0.36, the model's accuracy is further validated.

Despite the absence of the R^2 metric in this analysis, the low RMSE scores imply a high degree of alignment between the predicted and actual values, signaling a strong model fit in the context of the Kaggle competition.

Self-Evaluation and Critique

Upon reflection, the approach to model development was well-structured and aligned with sound machine learning principles. The selection of H2O AutoML for its ensemble learning capabilities proved advantageous, as it streamlined the modeling process and enabled the use of a range of algorithms, including Gradient Boosting Machines (GBM) and Generalized Linear Models (GLM). The pre-processing steps, including thorough text cleaning and topic modeling, were crucial in ensuring that meaningful features were derived from the tweet data.

However, certain aspects of the methodology could have been refined for better results. Firstly, while the chosen 95/5 split for the training and validation sets provided ample data for training, it may have left the validation set too small, potentially leading to overfitting. A more balanced split, such as 80/20, might have enhanced the model's generalizability. Additionally, although Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were applied for topic modeling, further hyperparameter tuning or a more comprehensive exploration of other topic modeling techniques could have yielded even more informative features.

One critique of the model-building phase is the reliance on RMSE as the sole performance metric. While RMSE provides valuable insights into prediction error, it does not capture other important aspects of model performance, such as how well the model ranks or classifies data points in comparison to actual rankings. Including metrics like R^2, or analyzing residuals in more detail, could have offered a more rounded evaluation of model performance.

Lastly, while the external validation through Kaggle was useful, further experiments with out-of-sample validation could have ensured that the models generalize more effectively to unseen data. The omission of multicollinearity checks and residual analysis also leaves room for potential improvement in diagnosing the model's underlying assumptions.

In conclusion, while the methodology was solid and the model performed well, future iterations could focus on more robust validation techniques, better balance in training/validation splits, and the inclusion of additional performance metrics to ensure comprehensive accuracy and generalizability.

References

[1] GeeksforGeeks, "Removing Stop Words with NLTK in Python," [Online]. Available: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/. [Accessed: Sep. 02, 2024].

[2] OpenAI, "ChatGPT," OpenAI, 2024. [Online]. Available: https://chat.openai.com. [Accessed: Sep. 20, 2024].

[3] Scikit-learn, "Topic extraction with Non-Negative Matrix Factorization and Latent Dirichlet Allocation," [Online]. Available: https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html. [Accessed: Sep. 03, 2024].