DSCC483: Data Science Capstone - Mini Project

Neel Agarwal Shyam Shah

***Abstract*— This project analyzes the political discourse on Twitter by U.S. Congress members from 2008 to 2020, using natural language processing to correlate tweets with DW-NOMINATE ideological scores. Employing a dataset of over 333,987 tweets, we developed machine learning models to predict the ideological positions of the tweeters. The study primarily utilized a Random Forest model, enhanced with advanced text processing techniques such as DistilBERT for embedding generation. Our results, validate the effectiveness of these models in political ideology classification, contributing insights into the evolution of political communication in the digital era and its implications for understanding political behavior.**

I. INTRODUCTION

This project delves into the dynamics of political communication on social media by analyzing tweets from U.S. Congressional members from 2008 to 2020. The increasing influence of social media on political engagement makes it imperative to understand the ideological expressions conveyed through tweets by public officials. Employing natural language processing (NLP) techniques, this study aims to predict the political ideology of the tweet authors, linking their tweets to the established DW-NOMINATE scores, which quantify ideological positions on a liberal-conservative spectrum. The dataset, comprising 333,987 tweets, provides a robust basis for training a machine learning model to assess and predict ideological dimensions. This approach not only highlights the political alignments evident in digital communication but also aids in tracking shifts in political stances over time, offering a quantitative measure of partisanship. The objectives of this project are comprehensive: conducting an extensive descriptive analysis of the tweets to uncover patterns in hashtag usage and tweet lengths; implementing and refining advanced machine learning models to evaluate the effectiveness of a Random Forest model against baseline models such as XGBoost and K-Nearest Neighbors; exploring the potential of text embeddings and vectorization through DistilBERT to enhance the model's ability to process and interpret textual data; and striving to minimize root mean square error to demonstrate the applicability of machine learning in ideological classification based on textual content. Through detailed data examination, innovative methodological applications, and rigorous model evaluations, this report provides insights into the evolving nature of political dialogue on digital platforms and its implications for contemporary governance.

II. DATA

The dataset used for this project consists of tweets from U.S. Congressional politicians with active Twitter accounts. It spans the years 2008 to 2020 and contains a total of 469,740 tweets. The data is divided into three parts:

- **Training Data**: 333,987 tweets
- **Test Data**: 135,753 tweets
- **Sample Submission Data**: A smaller version of the test dataset containing only the Id and political ideology columns.

A. Data Features

The dataset includes the following features for each tweet:

- **Id**: A unique identifier for each tweet.
- **favorite_count**: The number of times the tweet was favorited.
- **full_text**: The complete text of the tweet.
- **hashtags**: A list of hashtags used in the tweet.
- **retweet_count**: The number of times the tweet was retweeted.
- **year**: The year the tweet was posted.

B. Descriptive Analysis

We performed a basic descriptive analysis on the training dataset to understand the tweet characteristics.

1) Text and Hashtag Length: The table below summarizes key statistics related to the length of the tweets and hashtags in terms of characters and words.

	characters in tweet	words in tweet	characters in hashtag	words in hashtag
Min	4	1	1	1
Avg	173.82	25.03	14.04	1.49
Med	143	21	12	1
Max	531	67	168	17

TABLE I
SUMMARY OF TWEET AND HASHTAG LENGTH (CHARACTERS AND WORDS)

Tweet Length: The length of the tweets (in characters) ranges from 4 to 531, with an average of 174 characters and a median of 143 characters. The number of words ranges from 1 to 67, with an average of 25 words.

Hashtag Length: The number of characters in hashtags ranges between 1 and 168, with an average of 14 characters. Hashtags contain between 1 and 17 words, with an average of 1.49 words.

2) Most Common Hashtags: The following bar chart shows the 10 most commonly used hashtags in the dataset.

The most frequently used hashtag is #COVID19, which dominates the dataset with more than 8,000 occurrences. Other commonly used hashtags include #tcot, #SOTU, and

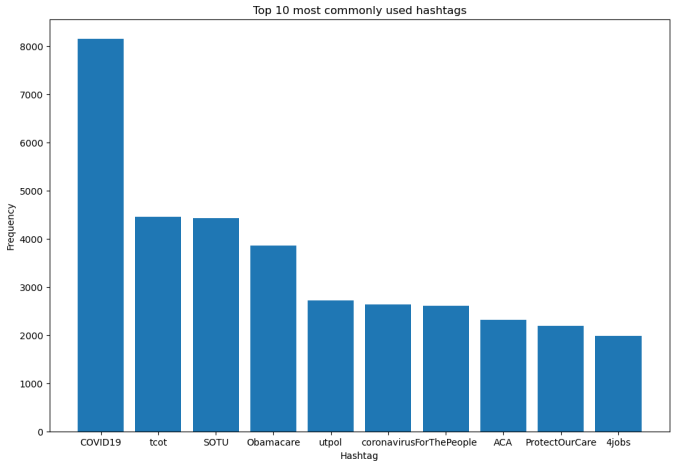

Fig. 1. Top 10 most commonly used hashtags

#Obamacare. The prevalence of these hashtags highlights a focus on healthcare and political issues in the dataset.

3) Most Common Hashtags by Ideological Groups: We further divided the dataset into four groups based on the first and second DW-NOMINATE ideological dimensions. The following plots show the top 10 hashtags for each group.

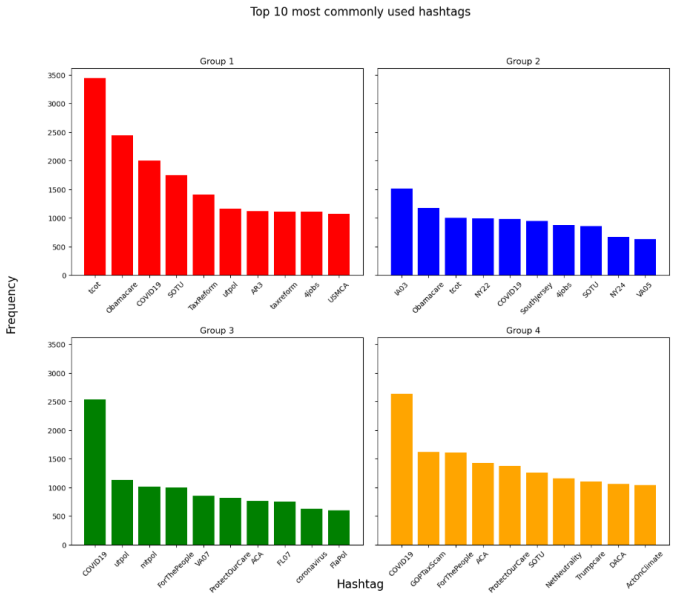

Fig. 2. Top 10 most commonly used hashtags by ideological groups

- **Group 1**: Primarily conservative, featuring hashtags such as #tcot, #Obamacare, and #COVID19.
- **Group 2**: Features #A13, #Obamacare, and #tcot, indicating different political issues.
- **Group 3**: Heavily focused on #COVID19 and related healthcare topics.
- **Group 4**: Also dominated by #COVID19, but with more varied hashtags such as #GopTaxScam and #ForThePeople.

Interpretation: The differences in hashtag usage among the four groups align with expected ideological divisions:

- Conservatives focus on issues like opposition to Obamacare, tax reform, and pro-Trump narratives.
- Liberals concentrate on healthcare access, social justice, and progressive policies.
- Moderates appear to focus on more pragmatic and issue-driven topics such as COVID-19, which affects all political groups equally.

The presence of #COVID19 across all groups indicates that the pandemic is a unifying topic, but the interpretations and responses to it vary widely across these ideological lines.

4) DW-NOMINATE Scores Over Time: The following ridge plot shows the distribution of DW-NOMINATE scores for conservative (red) and liberal (blue) groups over the years.

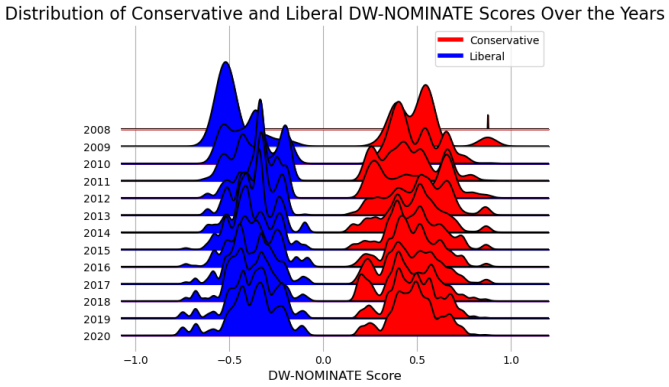

Fig. 3. Distribution of DW-NOMINATE Scores Over the Years

Interpretation: The plot shows that conservatives tend to have higher DW-NOMINATE scores (closer to +1), while liberals have lower scores (closer to -1). The distribution has remained relatively stable over time, though there are noticeable shifts in density across different years. This reflects the persistent ideological division in U.S. politics, with slight shifts potentially corresponding to key political events or policy debates.

5) Most Ideologically Distant Tweets: We calculated the Euclidean distances between tweets based on their ideological dimensions and identified the top 10 most ideologically distant tweet pairs.

Interpretation for Most Distant Tweets along both Dimensions: The top-10 most distant tweets represent ideological extremes on both ends of the spectrum. The high Euclidean distances, ranging around 1.9276, indicate significant differences in topics such as:
- Healthcare (Obamacare, gun safety legislation),
- International issues (Ebola response), and
- Education (STEAM programs).

This separation reflects how polarized political discourse can be on major national and international topics.

6) Most Distant Tweets along the First Dimension: **Interpretation for Most Distant Tweets along the First Dimension**: The tweets that are most distant along the first dimension reflect the typical conservative-liberal divide on policy matters. The first dimension of the DW-NOMINATE scoring

Tweet1	Tweet2	Distance
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	In the wake of #ebola, continued economic progress is key for #Liberia. Watch live as I speak to @CGDev at 9:30am ET https://t.co/sXDuRhMN6O	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	Congrats to my alma mater, Snowflake Jr. HS, for selection as a '17 #SamsungSolve STEAM finalist. #SamsungSolveSJHS https://t.co/RSXZwny231	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	Just got my copy of the #healthcare bill and I'm going to take time to thoroughly read and review it	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	@FoxNews @SpecialReport on my effort w/@dougducey & @SenJohnMcCain to get #AZ out of the oversized, overworked & oft-overturned #9thCircuit https://t.co/TtdteiTNmP	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	Cheryl bags trash @justserve project at Salt River this morning. LDS and other faiths team up to serve. #justserve https://t.co/rpU1Crk3gN	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	It's my Earth, Wind & Fire week so #LetsGroove w/ 3 more bills to make rural AZ a #ShiningStar of growth & investment https://t.co/LQq76qrcIY	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	#Wastebook sparking lots of great convo but none more entertaining than this @oreillyfactor @greggutfeld @bernieandsid exchange on @foxnews https://t.co/5vXvxWlsYR	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	Wishing a happy 100th birthday to our great state of Arizona today. Here's to all that the next 100 years will bring. #AZcentennial	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	Tucson Day. Thanks @wakeuptucson and @jonjustice for having me on this morning. #DayInTheLife http://t.co/pzdcYy7tmC	1.927598
This weekly roundup: I voted to pass Con. Castro's resolution to block Trump's #FakeEmergency declaration & the 1st major gun safety legislation in over 25 years. My colleagues and I on @HouseJudiciary also investigated the admin's harmful separation policy at the border. #TX29 https://t.co/xUTwAfVG1i	I'm voting yes on #Prop123. Our kids attend public schools, and it's the best way to provide additional resources with no tax increase.	1.927598

TABLE II

TOP 10 MOST DISTANT TWEETS ALONG BOTH DIMENSIONS

Tweet1	Tweet2	Distance
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	Clear example of how #Obamacare will harm businesses & employees in AZ via @BrahmResnik @12news: http://t.co/E1wZjXNuCy	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	Good piece in @politico about #Obamacare. Problems much deeper than a poor website http://t.co/KDW7V1kmxA	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	@POTUS making effective case for free trade and #TPP. Hope @realDonaldTrump and @HillaryClinton are listening	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	I'm voting yes on #Prop123. Our kids attend public schools, and it's the best way to provide additional resources with no tax increase.	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	Debate and passage of an #AUMF against #ISIL like the bipartisan proposal @timkaine & I intro'd is long overdue https://t.co/OS9A7j9kQP	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	What do man caves, a secret agent & pig flatulence have in common? @EPA & your tax dollars #ScienceOfSplurging http://t.co/2gAKOMyS9A	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	I'm speaking on House floor soon opposing bailout. #pork	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	Bigger waste of your taxes - $1.3M on alcohol or $34M on a facility the military won't use? #8ofWaste #FinalFour http://t.co/4fhXQnuCFv	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	My joint statement with @ChrisCoons in response to reports that @POTUS plans to release the House Intelligence Committee #memo against the recommendation of the @TheJusticeDept & @FBI https://t.co/Es8JsQjfbJ https://t.co/OsdWqbgS0H	1.626546
Hoy es el #DiaDelNiño y es muy importante que todos los niños sean contados en el #Censo2020. https://t.co/b3txdbbQzi, llame al 844-468-2020 o complete y envíe su formulario del Censo. Esta es una actividad divertida para hacer con sus hijos! #HazmeContar https://t.co/PqvYyyzym5	Congratulations to @CAPArizona on the 20th anniversary of the Lake Pleasant Storage Reservoir #Time2TalkH20 https://t.co/vcRTNZmgCN	1.626546

TABLE III

TOP 10 MOST DISTANT TWEETS ALONG THE FIRST DIMENSION

system is heavily aligned with the liberal-conservative axis, so the distances here (approximately 1.6265) represent extreme differences in how the two groups view key political issues such as:

- Census participation vs. Obamacare: Social inclusion and healthcare are starkly contrasted.

This highlights the divergent priorities between conservative and liberal lawmakers.

7) Most Distant Tweets along the Second Dimension: **Interpretation for Most Distant Tweets along the Second Dimension**: The second dimension of DW-NOMINATE typically captures more specific, less predictable ideological distinctions. Here, the distances of around 1.7095 represent stark differences in topics that are not strictly aligned with the liberal-conservative spectrum. For instance:

- COVID-19 response vs. STEM and education programs: These represent different policy priorities, where one tweet focuses on an urgent public health response, while the other highlights broader education and policy initiatives.

This suggests that the second dimension captures a level of ideological nuance not strictly bound by traditional political labels.

The text from the tweets has undergone a thorough cleaning process to prepare it for analysis, which involved removing stopwords, words shorter than three characters, links, emojis, and punctuation. This process is detailed extensively in the **Methods** section. Subsequently, the cleaned text data were used to update the summary statistics table, appending values for the character and word counts of the cleaned text. This augmented table, reflecting the changes made to the data through the cleaning process, is presented below.

Tweet1	Tweet2	Distance
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	If we work together we can make real progress on important issues. This week, we did just that. The House passed bills that support veterans in #STEM careers, fund programs dedicated to Holocaust education, and help @USAID ensure young girls abroad can access secondary education. https://t.co/zKL9Hr7nKc6	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Scientists from @UTSWNews, led by Dr. Trish Perl, who briefed North Texans on our tele-town halls, are warning that #COVID19 cases could spike by July. To reopen safely, we must listen to the experts and all do our part to limit the spread. #StayHome https://t.co/FAV1Tmiy9b	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Too often I hear heartbreaking stories about the high cost of prescription drugs, like Shane from Garland, who struggles to afford her insulin despite having a good-paying job. Today I told her story and urged action to #LowerDrugCosts and help folks facing crippling drug prices. https://t.co/8tLQ0cNhgLv	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Our government works #ForThePeople, and Inspectors General have a key role in oversight by ensuring any Administration follows our laws. That's why I joined colleagues on @HouseForeign to get answers on why the State Department's top watchdog was fired. https://t.co/PZpJQwXXt	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	To ensure that our communities get the resources and help they need, every North Texan should take time to fill out the #2020Census. In #TX32, 61% of residents have responded, so visit https://t.co/I7dVEs8Phb to complete yours today. https://t.co/mtSXnxo5Qi	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Listening to constituents is one of the most important parts of our jobs as representatives, and it was great to answer questions and hear from folks at our bipartisan town hall in Ohio yesterday. Thanks for having me and I look forward to hosting @RepAGonzalez in #TX32 soon! https://t.co/kJk3bhUF3h	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Calling all tech-savvy middle school and high school students! I'm pleased to announce my office is hosting the Congressional App Challenge for #TX32! If you have an idea for an app, show off your coding skills by submitting it by Nov. 1st! More info: https://t.co/YqLA0Vptj https://t.co/rtYtEsqxVI	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	This Thursday, my office is hosting mobile office hours in Garland! Visit the South Garland Branch Library from 2:30 PM - 4:30 PM to learn how we can assist with federal agencies or answer your questions about constituent services. #TX32 https://t.co/Vg4n0rY32w	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Proud to be working as part of a bipartisan team of Texas lawmakers looking to fix this error from the 2017 tax bill that is hurting Gold Star families like the Welches. #TX32 https://t.co/MPhKLCm0Xf	1.709491
Tonight, I'll be joining other policymakers, advocates & experts to discuss the Emergency Action Plan: What is it, why do we need it, and how can you make it happen? How we respond to this #COVID19 crisis will determine the rest of our lives. https://t.co/PSVIg0YqFL	Taking a few minutes to complete your #2020Census can help your community access millions of dollars in resources to support North Texas schools, roads, and hospitals. Paper forms will be sent to households this week so keep an eye out or go to https://t.co/I7dVEs8Phb now. https://t.co/8LmUMfCEVK	1.709491

TABLE IV

TOP 10 MOST DISTANT TWEETS ALONG THE SECOND DIMENSION

ANALYSIS AND COMPARISON OF TOPIC MODELING RESULTS

Non-negative Matrix Factorization (NMF)

NMF is a matrix factorization technique that factors the term-document matrix into two non-negative matrices. We used the generalized Kullback-Leibler divergence as the loss function, suitable for modeling the underlying topics in textual data.

Interpretation of Topics: The topics generated by NMF are derived from the decomposition of the tf-idf matrix. Since tf-idf gives more weight to less frequent but more informative words, NMF tends to produce topics characterized by specific, distinguishing terms.

Coherence of Topics: The topics extracted by NMF are generally more coherent and focused on specific themes present in the data. The top words in each topic often relate closely to each other, making it easier to interpret the underlying theme.

TABLE V

SUMMARY OF TWEET, HASHTAG, AND CLEANED TWEETS LENGTH (CHARACTERS AND WORDS)

	Characters in Tweet	Words in Tweet	Characters in Hashtag	Words in Hashtag	Characters in Cleaned Tweet	Words in Cleaned Tweet
Min	4	1	1	1	0	0
Avg	173.82	25.03	14.04	1.49	107.55	14.92
Med	143	21	12	1	95	13
Max	531	67	168	17	446	49

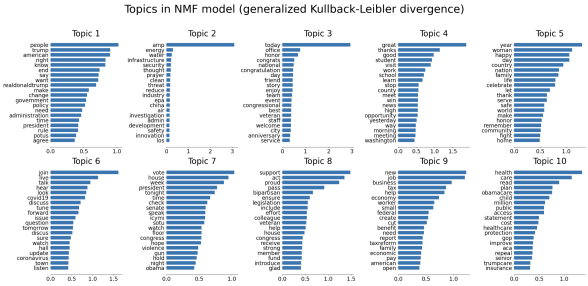

Fig. 4. Topics in NMF Model

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes documents are mixtures of topics, and topics are distributions over words.

Interpretation of Topics: The topics generated by LDA are probability distributions over words. Since LDA uses raw term counts (tf), it captures the co-occurrence patterns of words across the corpus.

Coherence of Topics: LDA topics may be broader and capture more general themes. The top words in each topic might be more common words that appear frequently together, which can sometimes make interpretation less straightforward compared to NMF.

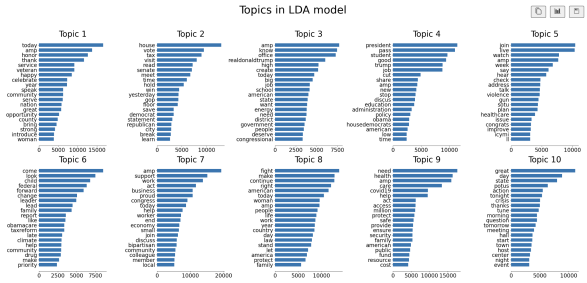

Fig. 5. Topics in LDA Model

Comparison

Specificity vs. Generality: NMF tends to produce more specific topics due to the use of tf-idf features, which downplay common words. LDA, using raw term counts, may generate broader topics.

Interpretability: Depending on the dataset, one model may offer more interpretable topics than the other. If the

dataset contains distinct themes, NMF might highlight them more clearly.

Overlapping Topics: LDA allows for documents to be associated with multiple topics, reflecting the probabilistic nature of the model. NMF also allows for overlapping topics but may assign higher weights to fewer topics per document.

Findings

Common Themes: Both models might have identified similar overarching themes present in the tweets, such as discussions about events, opinions, or trending topics.

Unique Insights: NMF might have revealed more niche topics or subtopics due to the influence of tf-idf weighting, highlighting less frequent but significant terms.

Model Selection: The choice between NMF and LDA may depend on the specific goals. For more focused topics, NMF may be preferable. For understanding broader themes, LDA might be more suitable.

Conclusion

By applying both NMF and LDA to the dataset, we gain complementary views of the underlying topics. NMF provides a detailed look at specific themes, while LDA offers a broader perspective. Analyzing the results from both models can give a more comprehensive understanding of the textual data, which is particularly useful for exploratory data analysis in unsupervised settings.

III. Methods

A. Text Preprocessing

The initial step in our data preprocessing involved decoding tweets, which were stored in a byte string format, back into readable text. This was necessary to handle any encoded characters, such as emojis or special punctuation, which often appear in social media text.

Once the tweets were decoded, a thorough cleaning process was implemented to ensure the quality and consistency of the textual data, essential for accurate model training. The following steps were undertaken:

- **Removal of Emojis:** All emojis were stripped from the tweets to reduce noise and focus on textual content.
- **Elimination of URLs:** We removed all URLs since they do not contribute to understanding the sentiment or topics of tweets.
- **Space Normalization:** Extra spaces within tweets were reduced to single spaces to standardize the text format.
- **Lowercasing:** All text was converted to lowercase to ensure that the same words, in different cases, are treated identically.
- **Punctuation Removal:** We stripped all punctuation to focus purely on words.
- **Stopword Removal:** Common English words that do not contribute to topic modeling, like 'the', 'is', and 'and', were removed.
- **Short Word Filter:** Words shorter than three characters were removed, as they typically do not carry significant meaning.

After preprocessing, the clean text data was used for further analysis, ensuring that the input data for the topic modeling was as refined and relevant as possible.

B. Parallel Processing

To expedite the processing of our data, we employed parallel processing techniques. This approach splits the data into smaller chunks, which are then processed simultaneously across multiple CPU cores. This significantly reduces the time required for large-scale computations like text cleaning and lemmatization, making it feasible to process large datasets efficiently.

C. Text Embedding and Vectorization

For both the full text of tweets and their associated hashtags, embeddings were generated to transform the textual data into a numerical format suitable for machine learning models. We utilized the DistilBertTokenizer and DistilBertModel from the transformers library, which are optimized versions of the BERT model designed for faster performance with good accuracy.

Embedding Process:

1) **DistilBERT Embedding:** Each tweet and hashtag was tokenized using DistilBERT's tokenizer, converting the text into a series of token IDs.
2) **Batch Processing:** Texts were processed in batches to efficiently manage memory and computational load, with each batch padded to the longest sequence in that batch to maintain consistency.
3) **Model Inference:** The tokenized text batches were fed into the DistilBERT model to obtain embeddings. The embedding for each tweet or hashtag was taken from the output of the first token (representing the aggregated information of the whole sequence).

These embeddings serve as the primary features for our machine learning models, capturing the contextual relationships within the text.

D. Model Architecture and Initial Trials

Before finalizing the model architecture, several models were tested to determine the most effective approach for predicting the political ideology dimensions based on tweet content. Initial trials included:

- K-Nearest Neighbors Regressor
- XGBoost Regressor
- Random Forest Regressor without Grid Search

These models provided foundational insights and crossed the benchmark set for this project,i.e., below 0.35 RMSE. After these preliminary trials with various models, we focused on developing a more robust model using a grid search approach with a Random Forest Regressor.

E. Random Forest Model with Grid Search

The final model architecture was based on a MultiOutputRegressor wrapping a Random Forest, which allows simultaneous predictions of both dimensions of political ideology. The best-performing model was selected due to its ability

to handle non-linear relationships and interaction effects between features effectively.

Model Tuning:

1) A randomized search was conducted to optimize the hyperparameters of the Random Forest, considering factors such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf. First,the grid search was conducted on a sample of 50,000 data points of the training data to find the best combination of parameters.
2) The best model parameters were determined based on the root mean square error (RMSE) metric, using cross-validation on a subset of the data.
3) The final model was then trained on the complete dataset using these optimized parameters.

This methodological approach was chosen to ensure robustness in our predictions and the ability to generalize across unseen tweet data.

IV. Results

In our final evaluation, the model achieved a root mean square error (RMSE) of 0.27714, which significantly surpasses the benchmark RMSE of 0.36 set at the project's outset. This performance indicates a robust predictive capability, substantially improving upon the baseline expectation. The optimized Random Forest model, developed through a meticulous process of feature engineering and hyperparameter tuning, effectively captured the underlying patterns in the dataset.

The improvements in RMSE reflect the model's enhanced ability to predict the ideological dimensions of political tweets accurately. These results not only demonstrate the efficacy of the model architecture and the preprocessing steps undertaken but also underscore the importance of careful model tuning and feature selection in predictive analytics.

Model Comparison: Prior to finalizing the Random Forest approach, several other models were considered, including K-Nearest Neighbors and XGBoost. Although these models provided valuable insights, they did not achieve performance metrics close to those of the optimized Random Forest model. This comparative analysis was crucial in selecting the most effective model for our data.

Overall, the significant reduction in RMSE compared to the benchmark highlights the successful application of machine learning techniques to a complex problem of political analysis, offering a promising avenue for further research and application in real-world scenarios.

ACKNOWLEDGMENT

We are grateful to our professors Ajay Anand and Cantay Caliskan for providing us the opportunity to work on this machine learning Kaggle Challenge. Assignments as these continue to inspire us to experiment with diverse approaches to solve problems using data science.