DSCC483: Mini Project

Jayant Patil Jaimin Shah

***Abstract*—This study investigates the relationship between political discourse on Twitter and the ideological leanings of U.S. Congressional members. Utilizing a dataset of 333,987 tweets from 2008 to 2020, we engineered features by leveraging the capabilities of a fine-tuned DistilBERT model. Specifically, we used fine-tuned the distilbert-base-uncased model on the m-newhauser/senator-tweets dataset, which comprises tweets made by U.S. senators during the first year of the Biden administration. The primary objective was to classify our tweets based on Democratic or Republican sentiment. Additionally, we employed BERT to generate semantic embeddings of the tweets, which were combined with numerical features and input into an XGBoost classifier to predict the ideological stance of each tweet. The model demonstrated robust predictive performance, underscoring the efficacy of our approach. These findings contribute to the understanding of the evolving dynamics of political communication on digital platforms and its implications for partisan discourse analysis.**

I. INTRODUCTION

Social media platforms have become pivotal arenas for political discourse, with public officials increasingly utilizing these channels to communicate their views and engage with constituents. Twitter, in particular, has emerged as a prominent platform for political engagement, offering real-time insights into the thoughts and actions of policymakers. This study aims to analyze the ideological expressions conveyed through tweets by U.S. Congressional members and to predict their political leanings using advanced natural language processing (NLP) and machine learning techniques.

We focus on two ideological dimensions derived from the DW-NOMINATE scores, which quantify political positions on a liberal-conservative axis. By correlating tweet content with these scores, we seek to develop predictive models that can accurately classify the ideological stance of tweet authors. To achieve this, we employ a multifaceted feature engineering approach that integrates traditional TF-IDF vectors, BERT embeddings for deep semantic analysis, and a novel Prediction Bias feature.

This hybrid feature set, combined with the robust predictive capabilities of XGBoost, enables us to achieve a nuanced understanding of political communication on Twitter. Our approach not only identifies the ideological leanings of individual tweets but also provides a framework for tracking shifts in political discourse over time. By leveraging advanced NLP techniques, this study contributes to the growing body of research on digital political communication and offers a quantitative tool for analyzing the complex landscape of online political engagement.

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 liked.
- **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	184	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 184, 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 indicates that a significant number of tweets were related to the COVID-19 pandemic. Hashtags like tcot (Top Conservatives on Twitter) and SOTU (State of the Union) reflect the use of Twitter for political discussions. Health-related hashtags such as Obamacare, coronavirus, and ACA show a focus on healthcare topics. ForThePeople and ProtectOurCare suggest

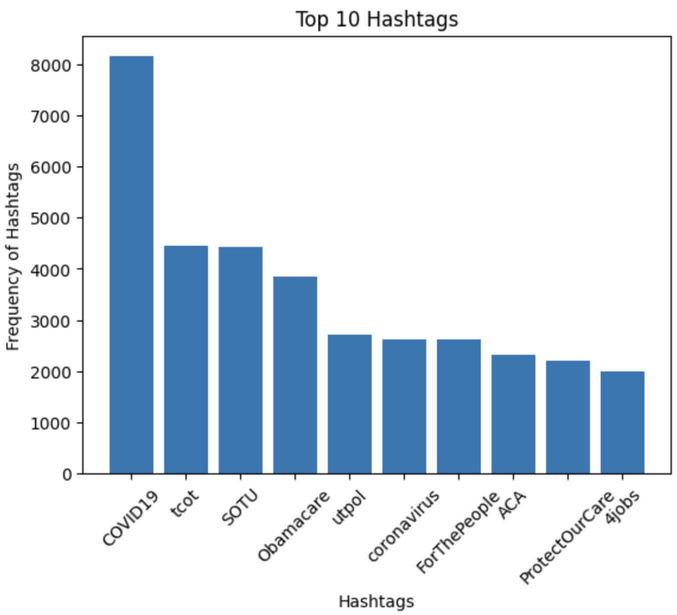
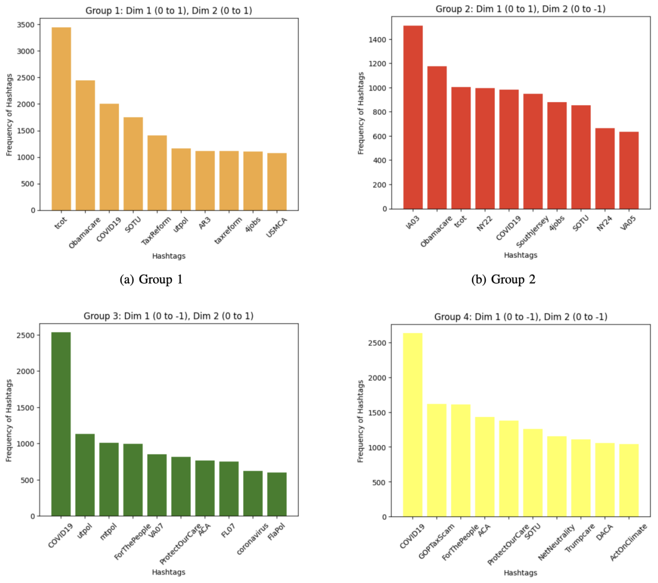


Fig. 1: Top 10 most commonly used hashtags

discussions around people's rights and healthcare protection. The hashtags illustrate a mix of healthcare issues, political events, and discussions relevant to the political landscape during the dataset's period.

3) Most Common Hashtags by Ideological Groups: To explore the hashtag preferences of different ideological segments, we segmented the dataset into four groups according to the first and second DW-NOMINATE dimensions. The visualizations below highlight the top 10 hashtags employed by each group.



(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

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

- **Group 1**: The most frequent hashtags include tcot, Obamacare, and COVID19. This group predominantly features conservative and healthcare-related discussions.
- **Group 2**: Common hashtags are IA03, Obamacare, and tcot, indicating a mix of state-specific political discussions and broader conservative topics.
- **Group 3**: The most frequent hashtags are COVID19, utpol, and mtpoli. This group shows a focus on pandemic-related and regional topics, with a generally liberal leaning
- **Group 4**: Dominant hashtags include COVID19, GOPTaxScam, and ForThePeople. There is a noticeable emphasis on opposition to conservative policies and pandemic-related issues.

Patterns and Comparisons:

COVID19 is a common hashtag across all groups, highlighting the universal concern regarding the pandemic. Groups 1 and 2 share a significant overlap with hashtags like tcot and Obamacare, suggesting common conservative discourse. Groups 3 and 4 show liberal-leaning hashtags, such as ForThePeople and ProtectOurCare, which emphasize healthcare and social justice issues.

The frequency and type of hashtags indicate that political and healthcare topics dominate the discourse across all groups, but the tone and stance differ according to the ideological leanings represented by the groups.

4) DW-NOMINATE Scores Over Time: The ridge plot displays the distribution of DW-NOMINATE scores over the years from 2008 to 2020, distinguishing between two ideological groups: conservative (red) and liberal (blue). The scores are plotted along the x-axis, where negative values indicate a more liberal stance and positive values indicate a more conservative stance. The y-axis represents the years.

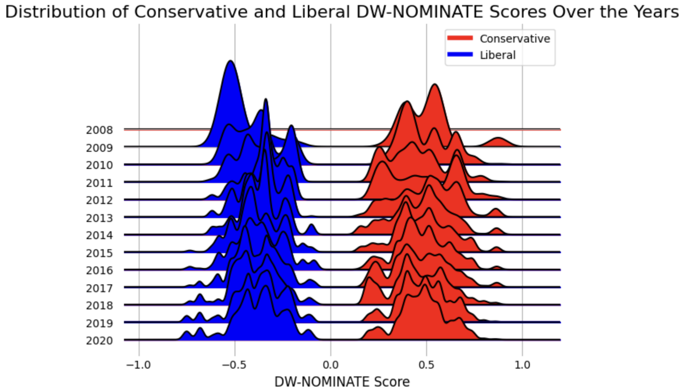


Fig. 3: Distribution of DW-NOMINATE Scores Over the Years

Observations: Overall Trend: There is a noticeable polarization between the conservative and liberal groups, with each group maintaining distinct DW-NOMINATE score distributions over the years. The plot illustrates a clear ideological divide, with conservatives tending towards higher DW-NOMINATE scores (+0.5) and liberals leaning towards lower scores (-0.5). The polarization has been relatively consistent, with the two groups maintaining their separate ideological positions without significant overlap. This suggests that the ideological divide has remained stable, with no significant convergence or divergence in the score distributions of the two groups. The plot illustrates a clear and consistent ideological divide between conservatives and liberals over the period from 2008 to 2020. There is no significant indication of a reduction in polarization, as the distributions for each group remain distinct.

5) Most Ideologically Distant Tweets: The table below presents the top-10 tweet pairs that are ideologically most different from each other, calculated using the Euclidean distance along both dimensions (Dim1 and Dim2).

Tweet1	Tweet2	Distance
This weekly roundup: I voted to pass Con. Castro's resolution 2 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
In Discovery Green, took a knee with @RepAlGreen, @houmayor Sylvester Turner, and thousands of Houstonians to protest police brutality, racism, and white supremacy in America.'s heart is with GeorgeFloyd and his family this afternoon. https://t.co/0bO3FWjifG	Congrats to my alma mater, Snowflake Jr. HS, for selection as a '17 SamsungSolve STEAM finalist. SamsungSolveSJHS https://t.co/RSXZwny23l	1.927598
Literally changing lives ForThePeople!!! https://t.co/RR6YzdKkIc	Just got my copy of the healthcare bill and I'm going to take time to thoroughly read and review it	1.927598
Roland Gramajo was arrested by @ICEgov on Thursday.'s a leader in the Houston Latino community and a father of five American children.call me radical for calling to AbolishICE.://t.co/WPOC90OJP0	@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 week, my constituent Anna Alvarez joined me for SOTU2020 to highlight why we must RaiseTheWage.impeachment trial came to an end. While acquitted, Trump was NOT exonerated., we also passed bills such as the PROAct that protect workers' rights to join a union. https://t.co/rI0C4eDxad	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
@HoustonTX launched its Census2020 YES! to the Census campaign because everyone must be counted.Census will determine how billions in federal funds will be allocated across the country.must make sure our community gets its fair share. https://t.co/5buMKb3WLw	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
DontLookAway from the inhumane conditions that Trump is forcing asylum seekers to live in.anti-immigrant agenda continues to risk the lives of people who just want an opportunity at the American Dream.policies are inconsistent with our values as Americans.	Wastebook sparking lots of great convo but none more entertaining than this @oreillyfactor @greggutfeld @bernieandsid exchange on @foxnews https://t.co/5vXvxWlsYR	1.927598
Here's your weekly update! This week, we celebrated VeteransDay, fought for our Dreamers, passed a resolution on the ERA, marked up bills in the Financial Services Committee that will protect consumers, and I also introduced a new bill, the Stop EITC and CTC Seizures Act.	Wishing a happy 100th birthday to our great state of Arizona today. Here's to the all that the next 100 years will bring. AZcentennial	1.927598
We have a responsibility to our teachers and students, which is why I led a letter urging TEA Commissioner @MikeMorath to reverse his decision not to directly disburse CARESAct funds to local education agencies.	Tucson Day. Thanks @wakeuptucson and @jonjustice for having me on this morning. DayInTheLife	1.927598
America needs all people to wear a face mask. MaskUp	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

Interpretation for Most Distant Tweets along both Dimensions: All tweet pairs in the dataset exhibited an identical distance value of 1.927598, the maximum recorded distance. These tweet pairs underscore how differing focuses—ranging from national policy criticism to community support—can reflect broader ideological divides. The distance metric effectively captures the substantial variation in content, tone, and political engagement between tweets, providing insights into the nature of ideological discourse on social media.

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 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:

- The tweet pairs exhibiting the greatest distance along the first dimension (approximately 1.6265) highlight the stark contrast between conservative and liberal perspectives on key political issues. For instance, the contrast between calls for census participation and opposition to Obamacare exemplifies the divergent priorities of these two groups on social inclusion and healthcare.

Tweet1	Tweet2	Distance
.@VP here's an idea: You can honor the victims of today's shooting with action. Urge McConnell to allow a vote on HR 8 and HR 1112. Your thoughts and prayers won't undo the violence in #Texas. #DoSomething #texasshooting #odessashooting https://t.co/h2LvnZGo0o	What do man caves, a secret agent & pig flatulence have in common? @EPA & your taxdollars #ScienceOfSplurging http://t.co/2gAKOMyS9A	1.626546
The arc of the moral universe is long, but today it was further bent towards justice for #LGBTQ Americans. In America, no one should be fired for their sexual orientation or gender identity. https://t.co/EbMeUkFMHx	RT @Jason_Samuels: Sen. @JeffFlake isn't messing around at the #NPCBee. Taps @ScrippsBee champ Gokul Venkatachalam for advice #npcb https://…	1.626546
Here's the latest on #Imelda, more heavy rain is expected. Make sure to stay informed and don't let your guard down. #houwx #txwx #HouNews https://t.co/FmQLiQQHU7	Watching Dallin's football scrimmage at 2:30pm. 109 *F. #sweatfest http://t.co/e3KDXRvuaX	1.626546
More awesome #GoTexanDay spirit and a fired up group of seniors at Denver Harbor MSC. I ran into an intense game of bingo and had to play along! #TX29 https://t.co/7rAHPCPhEN	@franklinblog, #438days was an INCREDIBLE read. Great escape from politics. Thanks.	1.626546
I am proud to join @RepDMP and the rest of my colleagues on the @HispanicCaucus Women's Task Force to express my support for the #LatinaProsperity Principles. https://t.co/jt3UWqXCig	Having trouble with a federal agency? Stop by my staff's #Scottsdale office hours tomorrow to see if we can help. http://t.co/kLrwyZktKt	1.626546
His testimony has only enabled Trump's #CultureOfCorruption. I took an oath to support and defend the #Constitution. On today's #ConstitutionDay, we are holding #CoreyLewandowski, @realDonaldTrump, and his administration accountable because no one is above the law.	Enjoyed meeting with @DodieLondenEIPS National Excellence in Public Service participants from #AZ this morning https://t.co/UHxy6Fe3KG	1.626546
.@hcphtx has launched a #COVID19 Dashboard that will provided regularly updated information about the number of active #Coronavirus cases in Harris County. View it now: https://t.co/NTNJJkaPa7 https://t.co/j42OkWovDJ	With my friend President Loeak of the Marshall Islands. Take me back to the islands! #RivalSurvival @MartinHeinrich http://t.co/WuUk6RGKJ4	1.626546
Gracias @USHCC for having the opportunity to share with our strong Hispanic #Entrepreneurs our stories during the In Her Footsteps Panel. The work you do for Hispanic businesses is more important than ever! https://t.co/RukvJN4CLa	Great run by US Women's Ice Hockey Team in #Sochi! Proud to see Arizonan Lyndsey Fry @fry_X_cycle awarded a silver medal.	1.626546
Melrose Park Civic Club was a good crowd last night. #TeamSylvia, @RepWalle and others gave updates on community happenings. #TX29 https://t.co/knqSH6ZclD	One of the best Thanksgiving traditions #turkeybowl https://t.co/mgkkhMFFjr	1.626546
Estos son los distritos escolares del área de #Houston que cancelan actividades por causa de #Imelda. #txwx #houwx https://t.co/ngLD9iihFA	So, if "low level" staffers at the IRS can target #teaparty and #patriot, what's next? Party registration? Donation history?	1.626546

TABLE III: Top 10 most distant tweets along both dimensions

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.

Tweet 1	Tweet 2	Distance
My mommy, Sandy, was beaten for being too pretty, too ugly, too smart, too dumb, too black. Let us reject the myth that strong women, bold women, independent women, do not find themselves in the throes of violence at the hands of someone who claimed to love them. #VAWA #VAWA19 https://t.co/9dXeUsCKTd	As arguments begin in the Trump Administration's lawsuit to invalidate the ACA, I joined colleagues to tell the story of Rob & Debbie Rose of Rowlett, who were able to purchase insurance thanks to the ACA. We must work together to #ProtectOurCare and help families like the Roses. https://t.co/JRDOfb8iuT	1.709491
Black youth in America are facing a #MentalHealth crisis. They face disproportionate barriers to accessing quality, affordable mental health care, and startling new data reveals that they are dying by suicide at higher rates than their white peers. https://t.co/fRqzgIWidn	As we all reflect on #IndependenceDay, and safely spend time with loved ones, let us remember the health care workers and many others who are sacrificing so much to keep us healthy and save lives. #HappyFourthofJuly https://t.co/kXFZ5LjtJp	1.709491
From #PublicCharge to attempts to end #MedicalDeferredAction, the cruelty is the point. As the occupant of the White House continues attacking our immigrant neighbors, I'm working with my #TriCaucus colleagues to oppose this cruelty. #NoToPublicCharge https://t.co/GNjqGIfRR9	As your Representative, one of my top priorities is to stay connected to North Texas. Make sure to follow me on social media and subscribe to my newsletter to stay updated on all my latest work in Congress for #TX32. https://t.co/4AvsLhKbMZ	1.709491
As a child to witness the abuse & degradation of the person who is your world, your everything - it is an image, a feeling, which never leaves. Today's passage of #VAWA shows the nation that we say NO MORE. #VAWA19	Had a great discussion today with folks at the @NDCC on my work in Congress for #TX32 and how we can grow our economy so that it benefits everyone in our region. https://t.co/LqLwkWJswE	1.709491
.@TheBlackCaucus's 21st Century Infrastructure Principles are a bold step toward ensuring our infrastructure investments center sustainability, accessibility & community-connectedness - affirming transportation & housing #justice for all. https://t.co/s3urGA8dyL	Our small businesses are hurting during this pandemic, which is why I voted for the #CARESAct to help bring resources and aid to our community. Dallas County recently expanded its efforts to help and some may qualify for more assistance. Learn more here: https://t.co/fbaZ48HoUt	1.709491
RT @NotoriousVOG: C'Mon! I so AGREE with THIS, all of it, every single word! @RepPressley #UrbanAgenda #BOSpoli #MApoli #Blacktwitte…	Welcome news for North Texans hurting from the economic fallout of #COVID19. @PUCTX will suspend shutoffs for Texans impacted. If you are unable to pay your water or power bills, please make sure to call your provider and ask for assistance. https://t.co/EYAq1fx6aR	1.709491
#BlackLivesMatter. Period. I commend @JessJTang & @BTU66 for recognizing our young people as leaders in the civil rights movement. We must remain committed to lifting the goals of @BlmBoston during this week & all year round as we work to actualize justice for all. https://t.co/qMYMpfjfFU	Thanks to everyone who joined my #TX32 telephone town hall tonight. Listening to you and answering your questions is such an important part of my job. If you want to join future town halls or stay updated on my work in Congress, sign up for my newsletter https://t.co/jhSKTinCEm https://t.co/U0V97a50I8	1.709491
Starving a child is violence. Punishing a mother and her family is violence Contempt for poverty is violence. This administration's attacks on #SNAP would threaten benefits for more than 100000 ppl in Massachusetts which includes 72000 children. This is child abuse. Period. https://t.co/gBfVQZcYJf	Today is the 85th anniversary of Social Security, a bedrock program that has helped millions of seniors live with financial stability and dignity after retirement. That's why I am dedicated to strengthening and protecting this earned benefit for the years to come. #SocSec85 https://t.co/PZTHDMrVP5	1.709491
Joined my partner in good & fellow Bay Stater @RepMcGovern to lift the voices of #grandparents who have graciously taken on the role of caring for their grandchildren. Caregivers come in many forms & it's past time that the federal gov't honors them https://t.co/F8Mb0ArwMP https://t.co/wABXqvYkvn	#MemorialDay is a chance for a grateful nation to thank the heroes we have lost and their families. Today I met Phyllis, a Gold Star mom whose son Johnny was killed in Iraq in 2006. He is not forgotten. Thank you, Phyllis, for sharing your son's story & for your family's service. https://t.co/sy0Hz57HRe	1.709491
My #WCW goes out to my sisters in service who faithfully fight for #justice, #fairness & #equality every single day. We are the backbones of our families, our communities & our democracy. @EleanorNorton @RepKatiePorter @RepRashida @RepUnderwood @RepSpeier #ERANow https://t.co/jqJWYZ9ZsB	It was an honor to spend time with Madison of Highland Park HS & Cullen of Lake Highlands HS before they leave for the @NavalAcademy! Congrats to all the #TX32 students on your admission to our nation's service academies. And thank you for your willingness to serve. https://t.co/s8FAihVjP4	1.709491

TABLE IV: Top 10 most distant tweets along both dimensions

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.

TABLE V: Summary of cleaned tweet length (characters and words)

	Character Count	**Word Count**
Min	3	1
Max	430	43
Mean	106.93	14.18
Median	94	12

Evaluation and Comparison of Topic Modeling Results

We used a couple of topic modeling methods to analyze the tweets. They are described in detail below:

Latent Dirichlet Allocation (LDA)

LDA is a popular technique for topic modeling. It assumes that each document is a mixture of several topics, and each topic is a distribution over words. LDA uses a generative model to infer the underlying topics from a given corpus of documents. This is done by iteratively assigning words to topics and updating the topic distributions until a satisfactory model is obtained.

Interpretation of Topics: LDA produces topics as probability distributions of words, capturing the co-occurrence patterns of terms within the corpus (using raw term frequencies). This allows LDA to identify complex and interrelated themes within the dataset, making it a powerful tool for topic modeling.

LDA topics tend to be broader and encompass more general themes. The top words associated with each topic are often more common words that appear frequently together, which can sometimes make interpretation more challenging compared to NMF.

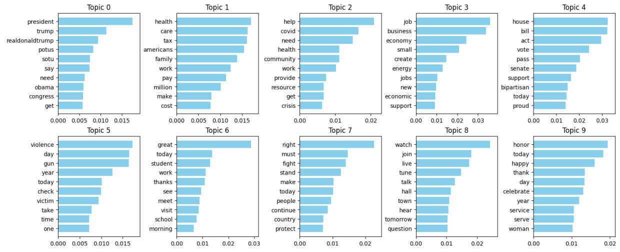


Fig. 4: Topics in LDA Model

Non-negative Matrix Factorization (NMF)

It's a popular technique used in topic modeling. It decomposes a matrix (typically representing documents and their word counts) into two non-negative matrices: a topic matrix and a document-topic matrix. The topic matrix represents the distribution of words in each topic, while the document-topic matrix indicates the contribution of each topic to a given document.

NMF generates topics by decomposing the tf-idf matrix, which gives greater weight to less common but more informative words. This results in topics that are characterized by distinctive, specific terms. NMF's non-negative nature ensures that the topics are interpretable.

Compared to other topic modeling techniques, NMF typically produces topics that are more cohesive and centered around specific themes within the data. The top words associated with each topic often have a strong thematic connection, making it easier to understand the underlying meaning.

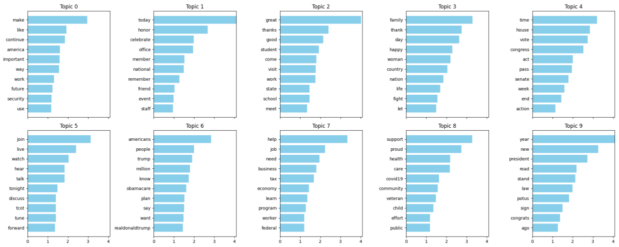


Fig. 5: Topics in NMF Model

Analysis of Topic Modeling output

Common Themes: Both models captured themes related to politics, healthcare, economic concerns, and community support. Discussions on legislative actions, COVID-19, and presidential topics are prominent.

Unique Insights:

NMF: Offers more focused topics on forward-looking ideas, education, and specific political opinions. It tends to present a more coherent narrative around each theme. *LDA*: Provides a broader view of topics with overlapping themes, such as violence and gun control, media participation, and activism.

Model Selection:

NMF: Best for scenarios requiring clear and distinct topics, where emphasis on specific narratives is needed. *LDA*: Suitable for exploratory analysis where a comprehensive understanding of interconnected themes is desired.

The choice between NMF and LDA will depend on the analytical goals—whether the focus is on detailed, distinct topics or a broad overview of thematic interrelations.

Conclusion

Using both NMF and LDA on the dataset allows us to obtain a more complete understanding of the underlying topics. NMF offers a detailed view of specific themes, while LDA provides a broader perspective. By analyzing the results from both models, we can gain a more comprehensive understanding of the textual data, which is particularly valuable for exploratory data analysis in unsupervised settings.

III. Methods

A. Text Preprocessing

The initial stage of our data preprocessing involved decoding the tweets, which were stored in a byte string format, into human-readable text. This step was crucial to ensure that any encoded characters, such as emojis or special punctuation, were properly interpreted, allowing for accurate analysis of the textual content.

After decoding the tweets, a rigorous data cleaning process was implemented to ensure the quality and consistency of the tweet data, which is crucial for effective model training. The following steps were undertaken:

- **HTML Decoding:** HTML entities in the tweet text are decoded to their corresponding characters using the html module.
- **URL Removal:** Links and URLs are removed from the tweet using a regular expression that matches patterns starting with http, https, or ftp followed by any sequence of characters.
- **Emoji Removal:** Emojis are filtered out using a regex pattern that matches a range of Unicode characters representing various emoticons, symbols, and pictographs.
- **Tokenization and Stopword Removal:** The tweet is tokenized into individual words. Words that are common with little semantic value(stopwords) are removed from the tokenized list.
- **Punctuation Removal:** All punctuation marks are removed from the tweet replacing each punctuation character with an empty string.
- **Short Word Filter:** Words shorter than three characters were removed due to their relavative insignificance.

Following the preprocessing steps, the cleaned textual data was utilized for subsequent analysis, ensuring that the input data for the topic modeling process was refined and relevant.

B. Feature Engineering

A novel feature, Prediction Bias, is introduced to classify tweets based on their political alignment. This feature is generated using the pre-trained language model *distilbert-political-tweets*, which analyzes the content of tweets to predict their likelihood of being written by a Democrat or a Republican. By incorporating this feature, the model enhances its ability to discern the political bias present in the textual data, providing a deeper understanding of the tweet's context and sentiment.

C. Text Embedding and Vectorization

The *TfidfVectorizer* was employed to transform the tweet texts into a sparse matrix of Term Frequency-Inverse Document Frequency (TF-IDF) features. This method quantifies the importance of each word in a tweet relative to its occurrence in the entire dataset, effectively representing the textual data. The resulting TF-IDF features highlight significant terms while minimizing the influence of less relevant words, reducing noise. This provides a complementary representation of the tweets' content, which, when combined with BERT embeddings, enriches the feature set for more nuanced analysis and prediction.

The pre-trained BERT model (bert-base-uncased) is utilized to generate sentence-level embeddings for each tweet, capturing rich semantic information essential for understanding the context and meaning of the text. The process involves:

1) **Tokenization:** Each tweet and hashtag was tokenized using BERT's tokenizer, transforming the text into a sequence of token identifiers.
2) **Sentence Embeddings:** The tokenized tweets are passed through the BERT model to generate embeddings that summarize the context and semantic meaning of each tweet.

These embeddings are the primary features used by our machine learning models, representing the contextual relationships in the text.

To leverage both traditional and deep learning-based text representations, the TF-IDF vectors, BERT embeddings, char$_{c}ountandPredictionBiaswereconcatenatedintoasinglefeaturesetusinghstack.$

D. Modeling Approach

For predicting the two NOMINATE dimensions, an XGBoost classifier was employed. XGBoost, a powerful gradient boosting algorithm, was chosen for its scalability, flexibility, and efficiency in handling large datasets and complex feature sets. Given the high dimensionality of the combined feature set (TF-IDF + BERT embeddings), XGBoost's ability to process such data efficiently, combined with its support for GPU acceleration, made it an ideal choice.

To capture the nuances of each NOMINATE dimension, separate XGBoost models were trained, each utilizing the same combined feature set. This approach allowed for tailored predictions of both dimensions.

IV. Results

The proposed model achieved a root mean square error (RMSE) of 0.26580, significantly outperforming the initial benchmark of 0.36. This substantial improvement demonstrates the model's robust predictive capabilities and its ability to accurately capture the underlying patterns in the political tweet data.

The optimized XGBoost model, developed through careful feature engineering, effectively extracted relevant information from the text, leading to the improved RMSE. This result emphasizes the importance of effective data preprocessing and feature selection in machine learning applications.

The reduction in RMSE indicates that the model has successfully learned to predict the ideological dimensions of political tweets with greater accuracy. This achievement underscores the potential of machine learning techniques for analyzing complex political discourse and offers promising avenues for further research and real-world applications.

Acknowledgment

We would like to express our gratitude to our professors, Ajay Anand and Cantay Caliskan, for providing us with the opportunity to participate in this machine learning Kaggle Challenge. Such assignments continue to motivate us to explore various approaches to solving problems using data science.