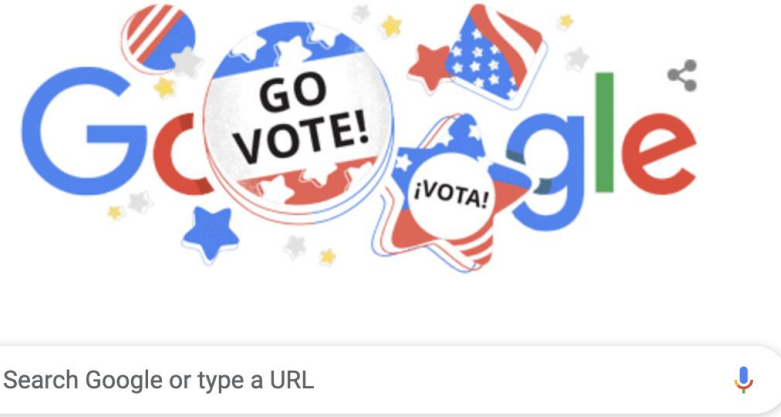
DECODING TWEETS TO FORECAST MID-TERM ELECTIONS

15.072 Advanced Analytics Edge

Team members: Nikos Galanos, Arushi Jain, Anant Vashistha



MIT MANAGEMENT SLOAN SCHOOL

MOTIVATION

Are elections important country and global topic?

Is social media more than social media?

"Twitter" is a powerhouse of information, can we use that information?

Can we predict election outcome using social media?

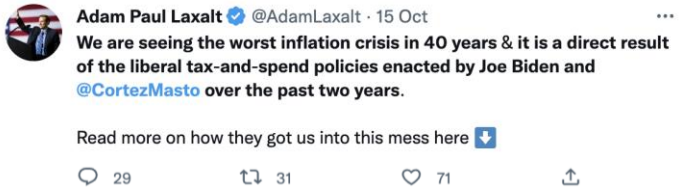
- Twitter is a widely recognized social media that dissipate information.
- Understanding the incidents taking place in country
- Attempt to build a model which can understand public sentiments on candidate's tweets to predict US elections

MIT MANAGEMENT SLOAN SCHOOL

OUR APPROACH

1. Scraping Tweets (Data Extraction)
2. Exploratory Data Analysis
3. Topic Modeling : LDA
4. Sentiment Analysis : Textblob, Flair, DistilBert
5. Feature Engineering
6. Prediction Models: Results
7. Conclusion
8. Challenges

SCRAPING TWEETS



Features	Extracted Value
Tweet ID	1581363660588400000
Text	We are seeing the worst inflation crisis in....
Username	AdamLaxalt
Quotes	0
Favorites	71
Replies	29
Retweets	31
Created at	Sat Oct 15 19:17:31 + 0000 2022
Hashtags	[]

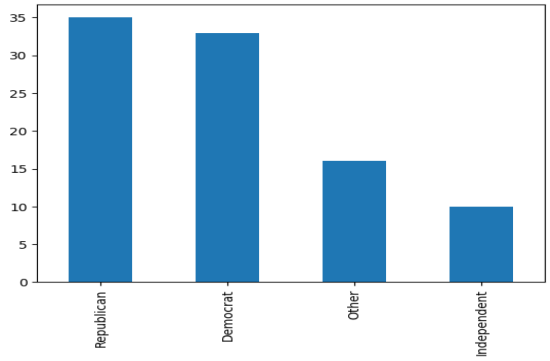
Extracting candidate's tweets information including replies to understand public sentiments

- Extracted candidate's tweets information and replies data using Tweepy for detailed analysis.
- The time-period for the candidate's tweets is **Sep 15 to Oct 15, 2022**, whereas replies time period is **Oct 08 – Oct 15, 2022.**
- The dataset contains up to **200 tweets** per candidate and up to **1000 tweets** per Candidate.
- Candidate's tweets dataset contains around 9.6K observations and the replies dataset contains around 41.6K observations.

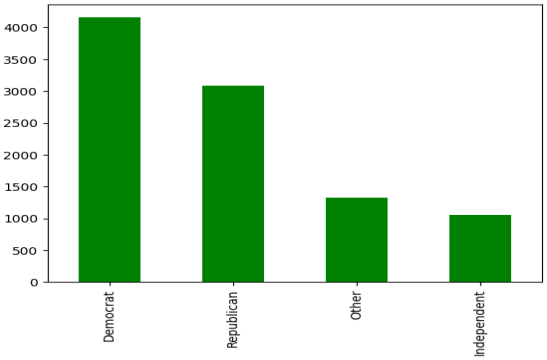
MIT MANAGEMENT SLOAN SCHOOL

EDA: CANDIDATE TWEETS VIEW - PARTY & STATE WISE

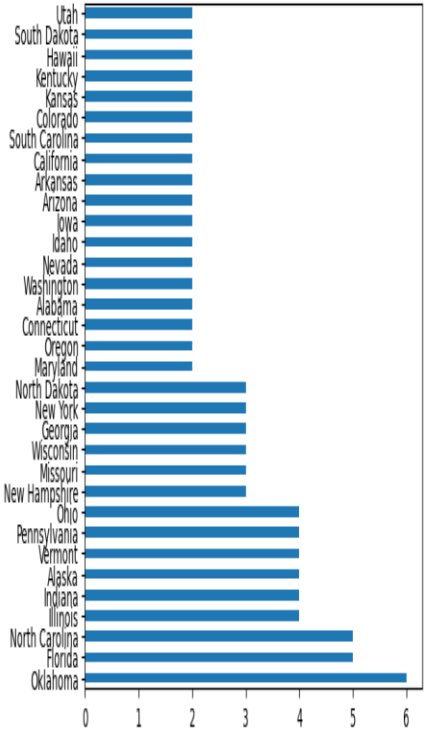
of Senators per Party



of Tweets per Party



of Candidates per State

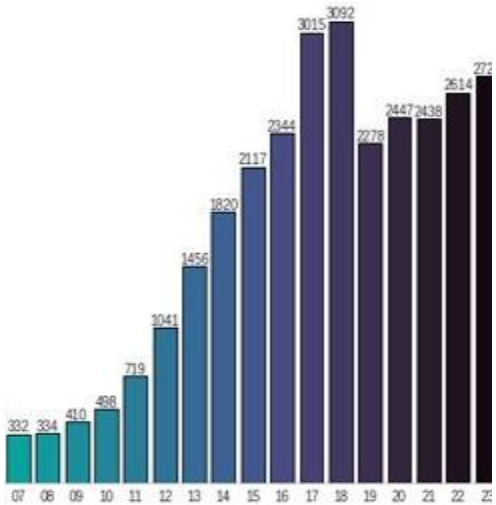


of Tweets per State

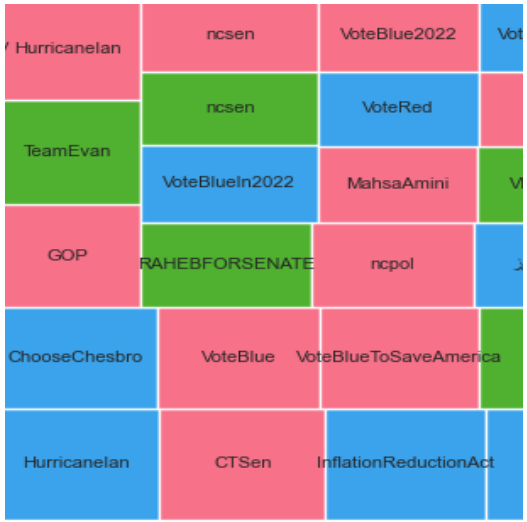
MIT MANAGEMENT SLOAN SCHOOL

EDA: REPLIES AND HASHTAGS ANALYSIS

of Replies per Hour of the Day



Popular Hashtags per Party



Democrats Wordcloud

Republicans Wordcloud

TOPIC MODELING

LATENT DIRICHLET ALLOCATION (LDA) MODEL WITH TF-IDF

Term Frequency (TF) Inverse Document Frequency (IDF)

- TF (w,d) = Occurences of w in document d/ Total words in document d
- IDF (w, D) = ln(Total documents in corpus D/ Number of documents containing w)
- TF-IDF (w, d, D) = TF (w, d) * IDF (w, D)

Relevant Keywords are given higher importance in each document

Latent Dirichlet Allocation (LDA)

- Classify tweets (documents) into multiple topics and topics into multiple words
- **Document** = Weighted Collection of Topics
- **Topic** = Weighted Collection of Keyword Combinations

Finding **hidden ("latent") topics** from the document term matrix AND getting the **topic distribution** across documents

Document Term Matrix

	W1	W2	W3	W4	W4	W6	W7
D1	0.2	0.3	0	0	0.1	0.2	0.1
D2	0.3	0.1	0.3	0.4	1.3	0	0
D3	0	0	0.4	2.5	0.1	0	0.2
D4	0.2	1.4	0	0	0.1	0.2	0.4

Document Topic Matrix

	T1	T2	T3
D1	0.2	0	0.8
D2	0.3	0.1	0.6
D3	0	0	1
D4	0.2	0.4	0.4

Topic Word Matrix

	W1	W2	W3	W4	W4	W6	W7
T1	0.3	0.4	0	0	0	0.2	0.1
T2	0	0.1	0.3	0.4	0	0	0.2
T3	0	0	0.4	0.5	0.1	0	0

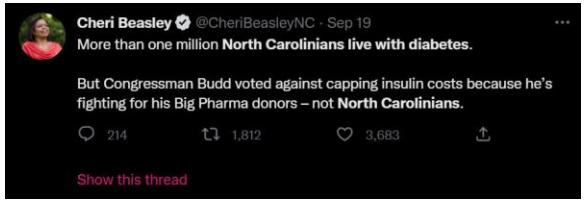
MIT MANAGEMENT SLOAN SCHOOL

TOPIC MODELING: LDA IN ACTION

Topic 1 = 0.012*social + 0.011*cost + 0.011*medicare + 0.010*security + 0.010*senior + 0.009*insurance + 0.009*discus + 0.009*prescription + 0.008*insulin

Topic 1 Contribution = 79.4%

Topic 1 Contribution = 78.7%



Topic 1 Contribution = 73.2%

Topic 2 = 0.008*corporation + 0.007*profit + 0.007*created + 0.007*resolution + 0.007*location + 0.007*inflation + 0.006*energy

Topic 2 Contribution = 69.3%

Topic 2 Contribution = 69.8%

Topic 2 Contribution = 68.3%

MIT MANAGEMENT SLOAN SCHOOL

TOPIC MODELING: LDA IN MORE ACTION

Topic 3 = 0.015*abortion + 0.014*woman + 0.014*choice + 0.013*libertarian + 0.013*freedom + 0.012*health + 0.011*care + 0.010*right + 0.009*access

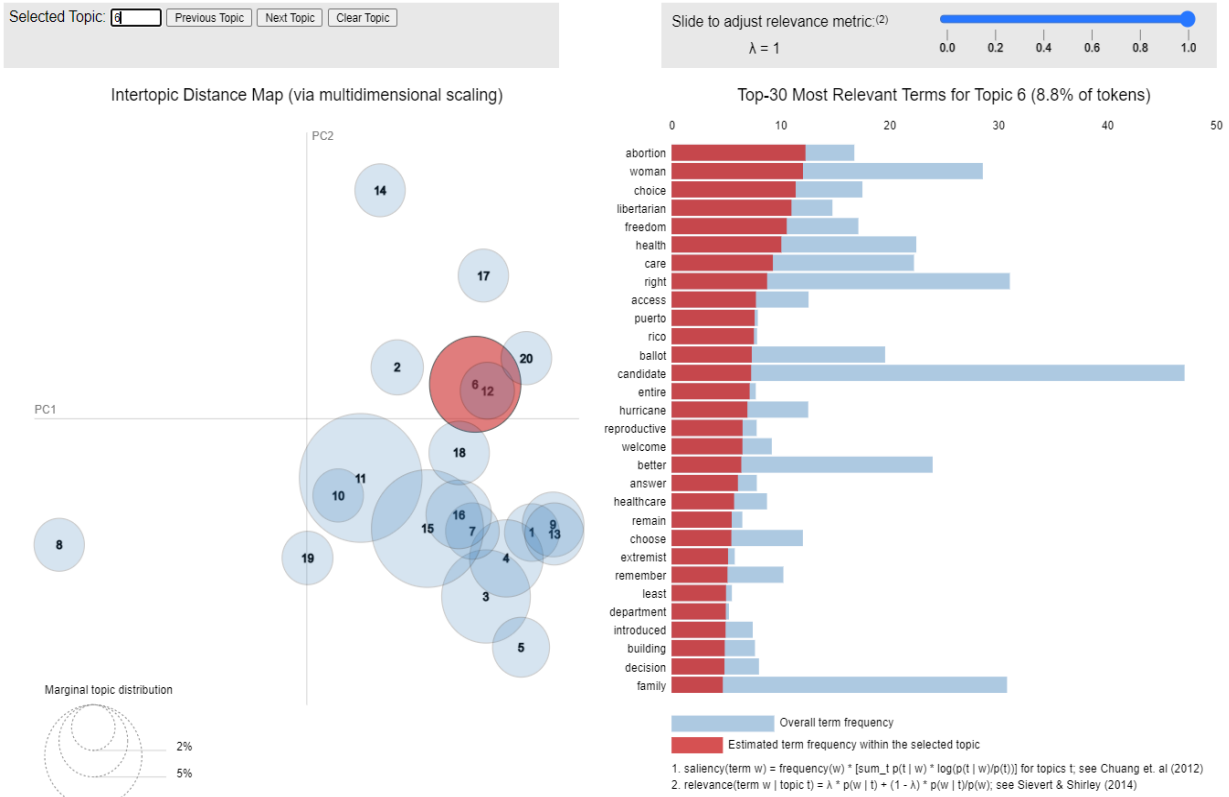
Topic 3 Contribution = 78.9%

Topic 3 Contribution = 71.8%

Topic 3 Contribution = 77.9%

MIT MANAGEMENT SLOAN SCHOOL

TOPIC MODELING: MERGING SIMILAR TOPICS



LDA with 20 topics

- More topics give better keywords
- Cannot have 20 features for 93 candidates

Manually merge topics based on:

- PCA Distance Map
- Top Topic Keywords
- Relevant terms
- Example tweets for topics

Reduced 20 topics to 7 topics

- General Election Campaign
- Healthcare
- Crime/ Police
- Inflation/ Economy
- Abortion/ Reproductive Rights
- School/ Education
- Current Events (Hurricanelan, Iran, Ukraine)

New topic weightage

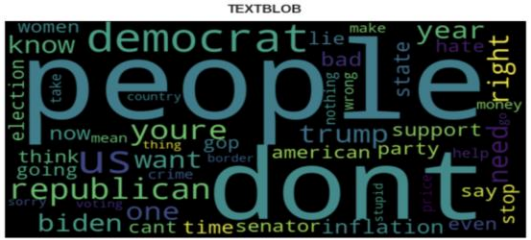
- Sum of merged topics weightage for each tweet

MIT MANAGEMENT SLOAN SCHOOL

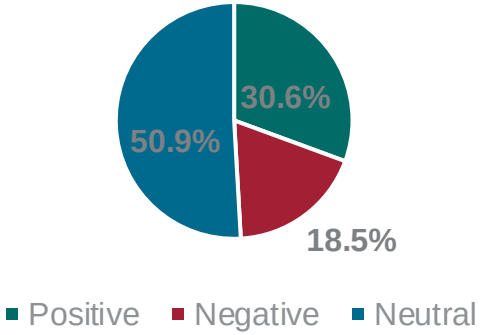
TEXTBLOB BASED SENTIMENT ANALYSIS

Methodology

- TextBlob is a library that allows textual data processing and can perform tasks such as sentiment analysis, classifying texts to Positive, Neutral and Negative.



Sentiments Distribution



Reply Tweet	Score
@TheOtherMandela The Republican Party is the party of Parental rights, family values and Law and order. Ron Johnson will continue to support policies that help working class families rise in life. He is a Truth Teller who exposes corruption in Washington. He is the perfect candidate for WI. us	1
@SenatorLankford Nothing on leader Pelosi's husband? Latest reporting was he was going to break her kneecaps. You and your goon party are responsible for this. And you can't even make a rudimentary comment because your goons will turn on you. Pathetic.	-0.1

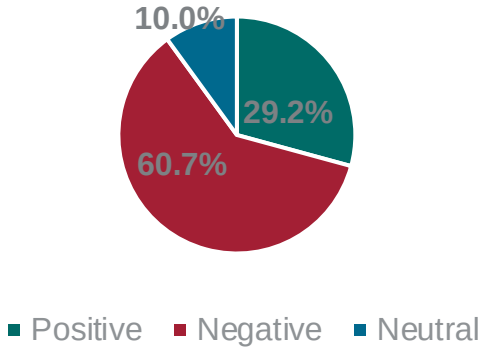
MIT MANAGEMENT SLOAN SCHOOL

FLAIR BASED SENTIMENT ANALYSIS

Methodology

- Flair is a powerful library for Natural Language Processing with a framework which builds directly on Pytorch.
- Flair for sentiment analysis classifies text as Positive or Negative hence we assigned a Threshold
- Tweets having score more than 0.6 have been classified as positive, between 0.6 to –0.6 as neutral, and less than –0.6 as negative

Sentiments Distribution



Reply Tweet	Score
@TheOtherMandela The Republican Party is the party of Parental rights, family values and Law and order. Ron Johnson will continue to support policies that help working class families rise in life. He is a Truth Teller who exposes corruption in Washington. He is the perfect candidate for WI. us	0.9916
@SenatorLankford Nothing on leader Pelosi's husband? Latest reporting was he was going to break her kneecaps. You and your goon party are responsible for this. And you can't even make a rudimentary comment because your goons will turn on you. Pathetic.	-0.999

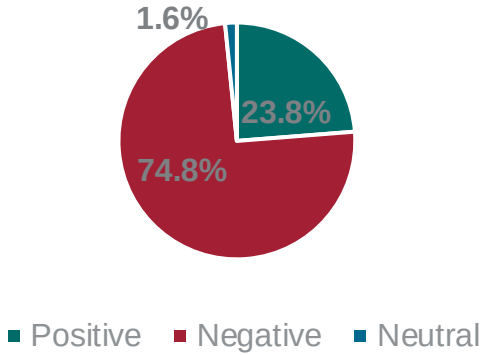
MIT MANAGEMENT SLOAN SCHOOL

DISTILBERT BASED SENTIMENT ANALYSIS

Methodology

- DistilBERT is based on BERT architecture (a pre-trained model)
- Analyzed the sentiments of replies dataset using the model
- Tweets having score more than 0.6 have been classified as positive, between 0.6 to –0.6 as neutral, and less than –0.6 as negative

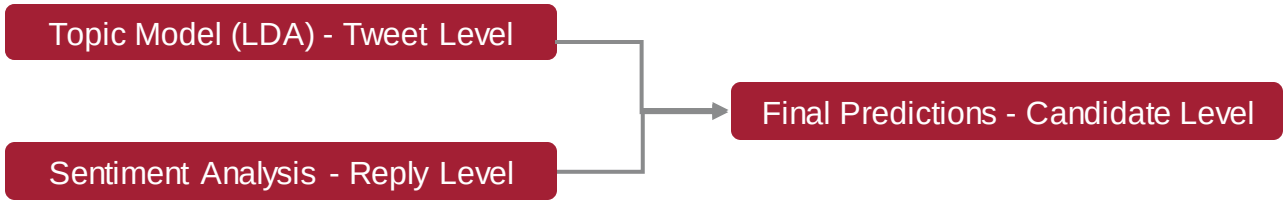
Sentiments Distribution



Reply Tweet	Score
@TheOtherMandela The Republican Party is the party of Parental rights, family values and Law and order. Ron Johnson will continue to support policies that help working class families rise in life. He is a Truth Teller who exposes corruption in Washington. He is the perfect candidate for WI. us	0.997
@SenatorLankford Nothing on leader Pelosi's husband? Latest reporting was he was going to break her kneecaps. You and your goon party are responsible for this. And you can't even make a rudimentary comment because your goons will turn on you. Pathetic.	-0.99

FEATURE MODELLING

COMBINING TOPICS AND SENTIMENTS



Missing Reply Data for Most Tweets due to Scrape Limits! :(

Interpolating Topic and Sentiment Features to a Candidate Level

Assumption: Candidates receive hate proportional to the number of comments for each tweet

Additional Twitter Metadata per Candidate: Average Likes, Average Retweets, Followers Count

Additional Electoral Metadata: Party (Democrat/ Republican/ Independent/ Other), Currently Serving

Total Features = 7 topic sentiments + 3 twitter metadata + 2 party metadata = 12 features for 93 candidates

3 Datasets – Blob, Flair, Bert

CandidatemeanSentiment (Candidate) = mean(replySentiment) for Candidate

tweetSentiment (tweet) = CandidatemeanSentiment * tweetreplyCount

tweetTopicSentiment (topic, tweet) = tweetTopicContribution * tweetSentiment

CandidateTopicSentiment (Candidate, topic) = mean(tweetTopicSentiment) for Candidate and topic across all tweets


ELECTION
RESULTS

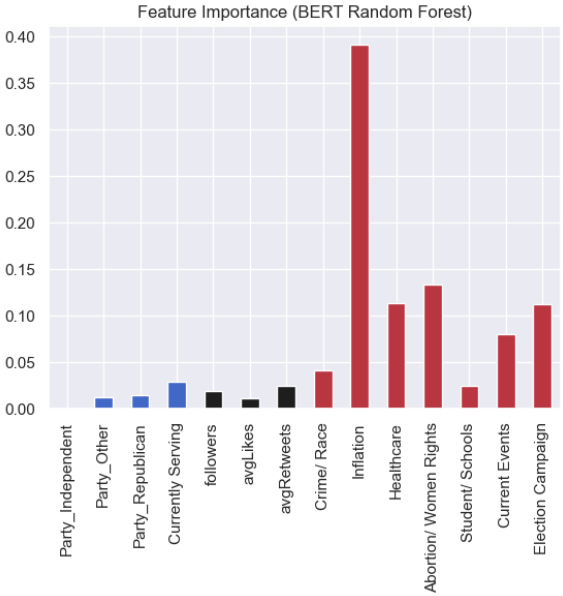
PREDICTING THE ELECTION RESULTS!

- **Train/ Test Split:** 75-25 stratified by Party and Elected
- **Hyperparameter Tuning:** 5-fold Cross Validation
- **Dependent Variable:** Vote Share
- **Winning Candidates:** Candidate with highest vote share for each state

DistilBert gives the best out-of-sample performance with **CART and Random Forest**

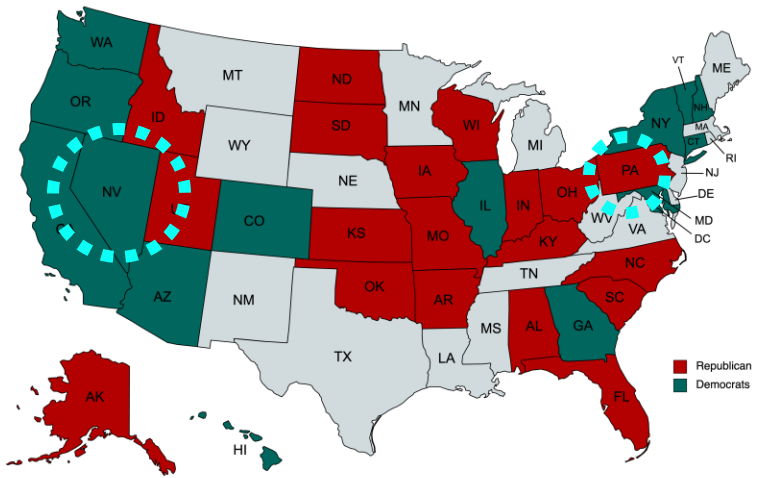
Baseline (Predicting everyone as not-elected): Train = 42/65, Test = 18/28						
Model	TextBlob		Flair		DistilBert	
	Train	Test	Train	Test	Train	Test
Linear	59/65	22/28	59/65	22/28	60/65	25/28
Lasso	60/65	23/28	59/65	22/28	60/65	25/28
Ridge	61/65	24/28	59/65	22/28	60/65	25/28
CART	62/65	23/28	56/65	21/28	61/65	26/28
Random Forest	62/65	21/28	64/65	23/28	61/65	26/28
Bagging	57/65	20/28	53/65	22/28	56/65	23/28
AdaBoost	59/65	22/28	63/65	24/28	55/65	22/28

MOST IMPORTANT VARIABLES:
Inflation, Abortion, Healthcare

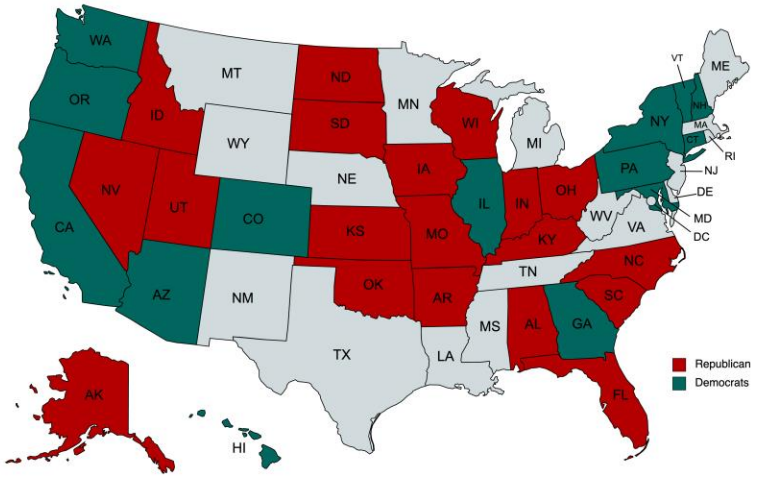


RF BERT MODEL CORRECTLY CLASSIFIED THE WINNING PARTY IN 31 OUT OF 33 STATES

Predicted Outcome

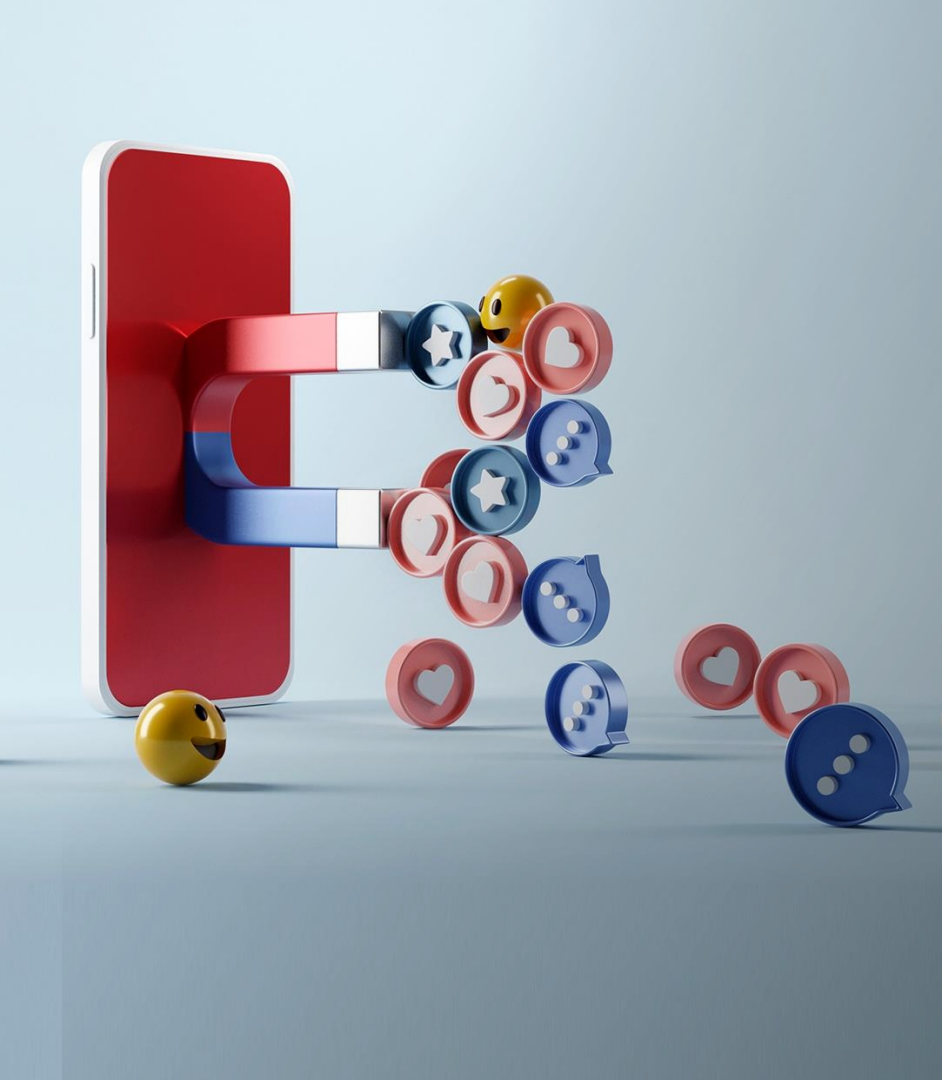


Actual Outcome



- RF BERT model correctly classified the winning party in 31 states out of 33 states
- Nevada (NV) and Pennsylvania (PA) are classified incorrectly, there was a strong tie in NV between Republican and Democrats

MIT MANAGEMENT SLOAN SCHOOL

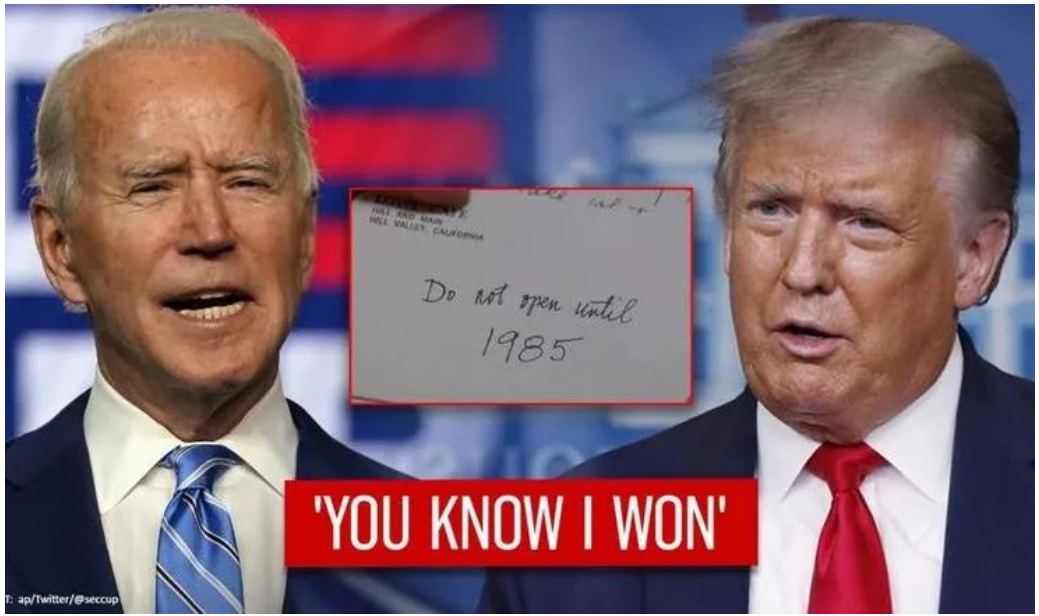


CHALLENGES

1. Replies and tweets data is limited to two months
2. No twitter accounts for 36 candidates
3. Absence of geographical data of replies
4. Merging topics manually
5. Extrapolation of reply sentiments for tweets that don't have replies
6. Not using hyperlinks, mentions and photos in tweets as features

MIT MANAGEMENT SLOAN SCHOOL

CONCLUSION



- Predicted **87 candidates correctly out of 93 candidates** (61/65 in-sample and 26/28 out-of-sample)
- Best Performing Model**: DistilBert-based RandomForest**
- Most important election topics: **inflation, abortion, and healthcare,** aligned with the recession, Roe vs. Wade, high healthcare cost
- Correctly classified winner party in **31 states out of 33 states** (Nevada and Pennsylvania classified incorrectly)

MIT MANAGEMENT SLOAN SCHOOL