

TABLE OF CONTENTS

01

OUR DATASET

Overview of our dataset, audience and purpose

04

OPTIMIZATION

Steps we took to optimize for increased accuracy

02

ETL AND LANGUAGES

Description of our ETL practices and chosen languages

O5
CONSIDERATIONS &
LIMITATIONS

Limitations we encountered

03

TRAINING THE MODEL

Overview of our model training

06

CONCLUSION

Final thoughts and questions

kaggle

OUR DATASET

- We used two datasets
 - Collections of tweets pertaining to Trump and Biden
 - Tweets were collected from October 15th, 2020 to November 8th, 2020

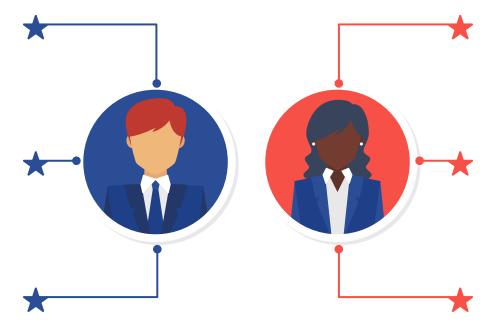


AUDIENCE AND PURPOSE

POLITICIANS

RESEARCHERS AND JOURNALISTS

GEN. POP. AND SOCIAL MEDIA USERS



CAN SOCIAL MEDIA BE USED TO PREDICT ELECTIONS

DOES SOCIAL MEDIA ATTENTION HAVE AN EFFECT ON CAMPAIGN AND PERCEPTION

VIEWS OF ELECTION



ETL & LANGUAGES

- Python
 - **NLTK**
 - Vader lexicon
- Pandas
- MathPlotLib
- Seaborn
- Amazon Web Services
 - o S3FS
- **Word Cloud**
- Plotly















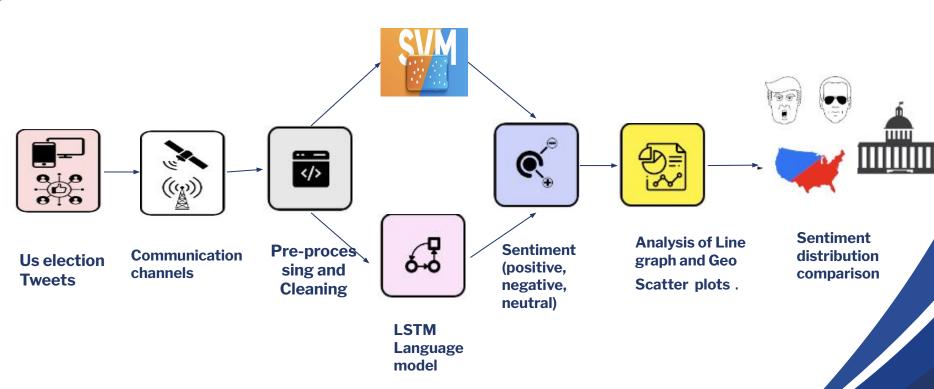






OUR PROCESS

SVM



ETL CHALLENGES AND SOLUTIONS



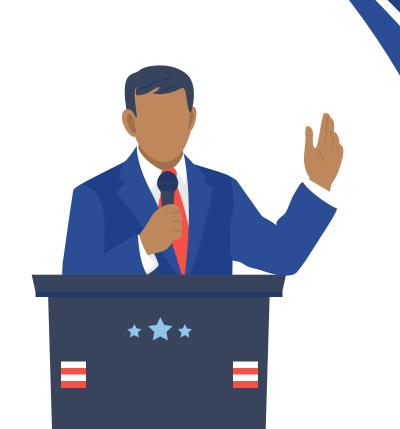




- Original Trump csv file was large with 971,158 rows
 - Encountered errors trying to load csv into a dataframe
- Read it in 5,000-row chunks to manage memory use
 - Switched to 1,000-row chunks to address malformed data upon parsing errors
 - Chunks are then combined into a single DataFrame for analysis
- Cleaned files were still large so Git Large File
 Storage (Git LFS) was utilized

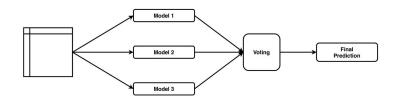
TRAINING THE

MODEL



COMPARISON OF THREE ML MODELS FOR SENTIMENT ANALYSIS

- Support Vector Machine (SVM)
- LSTM
- Ensemble model combining predictions from SVM and LSTM using a hard voting mechanism / Ensemble model combining predictions from SVM and LSTM using soft voting with 92% accuracy



* TRAINING THE MODEL *

OPTIMIZATION =

Using SVM

Accuracy: 0.8905238801113562

CIASSILICACIO	m keport.			
	precision	recall	f1-score	support
negative	0.86	0.82	0.84	7686
neutral	0.91	0.95	0.93	8135
positive	0.90	0.90	0.90	11838
accuracy			0.89	27659
macro avg	0.89	0.89	0.89	27659
weighted avg	0.89	0.89	0.89	27659

Using LSTM

Using Ensemble Training

Ensemble		sification	Report:	,230	
		precision	recall	f1-score	support
	0	0.90	0.89	0.89	7686
	1	0.93	0.95	0.94	8135
	2	0.94	0.93	0.93	11838
accui	racy			0.92	27659
macro	avg	0.92	0.92	0.92	27659
weighted	avg	0.92	0.92	0.92	27659

Classification	7.5		20	9
	precision	recall	f1-score	support
negative	0.83	0.58	0.68	7686
neutral	0.35	0.97	0.52	8135
positive	0.00	0.00	0.00	11838
accuracy			0.45	27659
macro avg	0.39	0.52	0.40	27659
weighted avg	0.33	0.45	0.34	27659

SENTIMENT ANALYSIS

How VADER Works:

Scores: VADER outputs
four types of scores:
positive, neutral,
negative, and compound.
The compound score, a
combined measure of
the first three, ranges
from -1 (very negative) to
1 (very positive).





IF COMPOUND SCORE IS BETWEEN -0.05 AND 0.05



► IF COMPOUND SCORE >= 0.05

TWEETS BEFORE AND AFTER CLEANING

#Elecciones2020 | En #Florida: #JoeBiden dice ...
@IslandGirlPRV @BradBeauregardJ @MeidasTouch T..
#censorship #HunterBiden #Biden #BidenEmails #...
In 2020, #NYPost is being #censorship #CENSORE...
FBI Allegedly Obtained Hunter Biden Computer, ...
Comments on this? "Do Democrats Understand how...
In an effort to find the truth about allegatio...
Twitter is doing everything they can to help D...

VOTE FOR #JoeBiden https://t.co/IIROoL5U00

26

this is how made his \n
this is how made his \n
in is being by twitter to manipulate a us ...

fbi allegedly obtained hunter biden computer
d...
comments on this do democrats understand how
r...
in an effort to find the truth about allegatio...

twitter is doing everything they can to help d...

vote for
is tearing up at the over the

TEMPORAL SENTIMENT TRENDS





Date

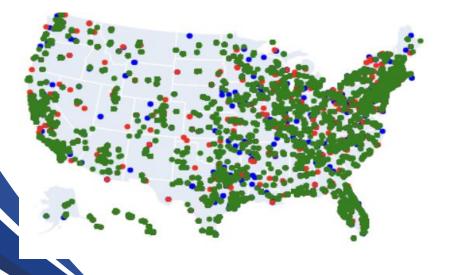


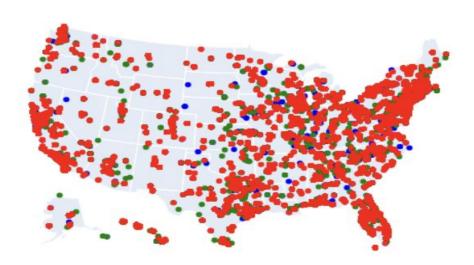
GEOGRAPHICAL SENTIMENT DISTRIBUTION



JOE BIDEN





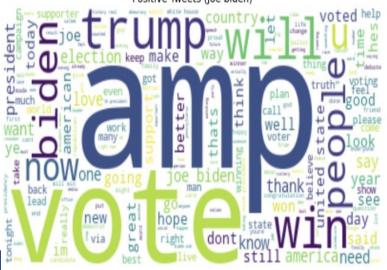


THEMES AND KEYWORDS



JOE BIDEN

Positive Tweets (Joe biden)





DONALD TRUMP

Positive Tweets-Trump



SENTIMENT DISTRIBUTION COMPARISON

DONALD TRUMP









38.05%

25.85%

36.10%

JOE BIDEN









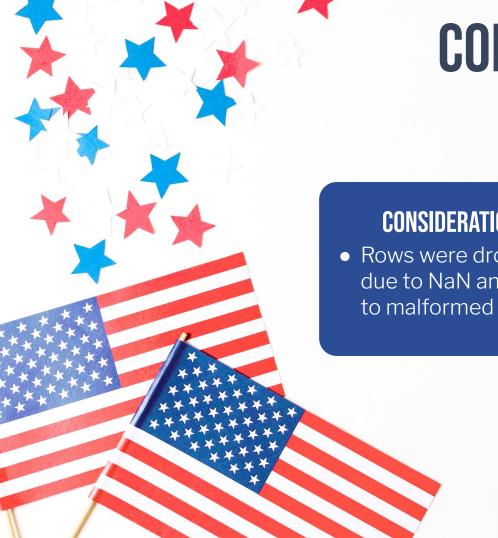
27.52%

29.6%

42.81%

KEY INSIGHTS

- Joe Biden's sentiment was predominantly positive,
 - Indicating stable support and resilience against negative campaigns
- Donald Trump's sentiment fluctuated, with
 - Negative sentiments slightly outweighing positive ones
 - Polarized public opinion
- State-specific analysis showed distinct regional differences in sentiment
 - Indiana showing the highest positive sentiment ratio for Biden and South Dakota for Trump
 - Guam exhibited the highest negative sentiment ratio for both candidates, indicating areas of low support



CONSIDERATIONS AND **LIMITATIONS**

CONSIDERATIONS

 Rows were dropped due to NaN and due to malformed data

LIMITATIONS

- Access to tweets was limited due to Twitter's new API policy
- Lost data because it was malformed

CONCLUSION AND QUESTIONS

- Given our final sentiment scores we can conclude that twitter sentiment analysis is an effective leading indicator of election results.
- Moving forward, candidates should pay close attention to social media sentiment scores over time as they can be a predictor of overall outcome.
- Will this hold true in 2024?



RESOURCES

• Trump and Biden 2020 Tweets