Sentiment Analysis
A Modern Approach to Political Polling

Suleyman Qayum

OUTLINE

- BUSINESS CONTEXT
- DATA
- MODEL DEVELOPMENT
- RESULTS
- EVALUATION
- LIMITATIONS & NEXT STEPS

BUSINESS CONTEXT

- *Gallup*, a global analytics and advice firm, wants to expand their polling capabilities
- Company currently relies primarily on telephone interviews to conduct polls
- *Gallup* has decided to begin exploring the viability of using **sentiment data from Twitter to conduct polls**
    - *provides a far more efficient means of gathering data*
    - *larger sample of the population can be polled*
    - *makes polling data more robust overall*
- The board at *Gallup* requests a demonstration that shows polling via sentiment data is a viable option
    - *should feature a model that can accurately predict the sentiment of a tweet*
    - *the model needs to deliver estimates of Joe Biden's approval ratings with respect to the nation and each one of the 50 stated*

DATA

- Training Data
  - *consists of $\sim 1{,}600{,}000$ tweets*
  - *relevant variables include:*
    - ***text***
    - ***sentiment label***
- Test Data
  - *consists of $\sim 100{,}000$ tweets relating to president Joe Biden*
  - *scraped from Twitter (07/12/2022 - 07/19/2022)*
  - *relevant variables include:*
    - ***text***
    - ***like count***
    - ***user location***
    - ***sentiment label***

DATA

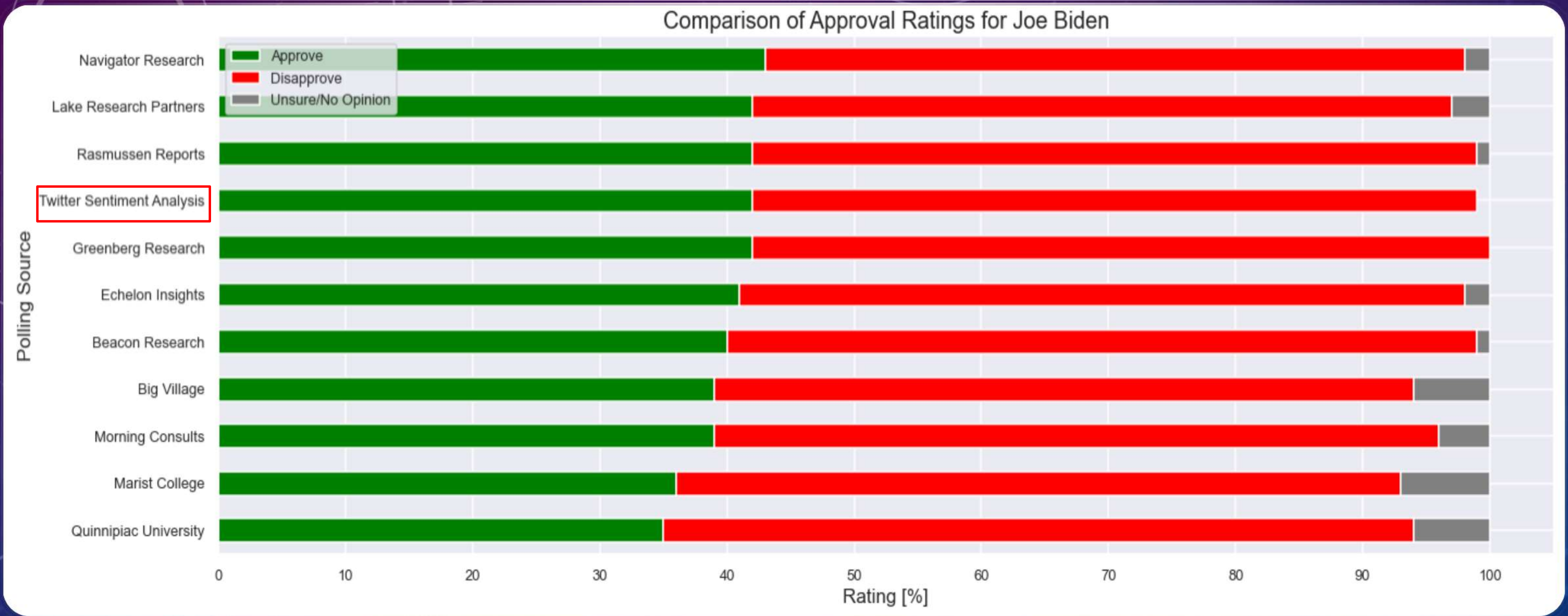
- All sentiment labels were created using *VADER:*
  - *utilizes a set of rules + dictionary that maps lexical features (i.e. word/phrase/emoji) to sentiment scores*
  - *provides a convenient, yet accurate, means of labeling data*
  - *tweets categorized as one of the following:*
    - *POSITIVE*
    - *NEUTRAL*
    - *NEGATIVE*
- **Why don't we just use VADER to classify sentiment?**
  - *VADER cannot adapt as language changes*
  - *Cannot be applied to other languages*
  - *Does not perform well on text containing specialized terms*
  - *Misspelling & grammatical mistakes cause performance to suffer*
  - *Cannot detect more subtle nuances in language*

MODELING

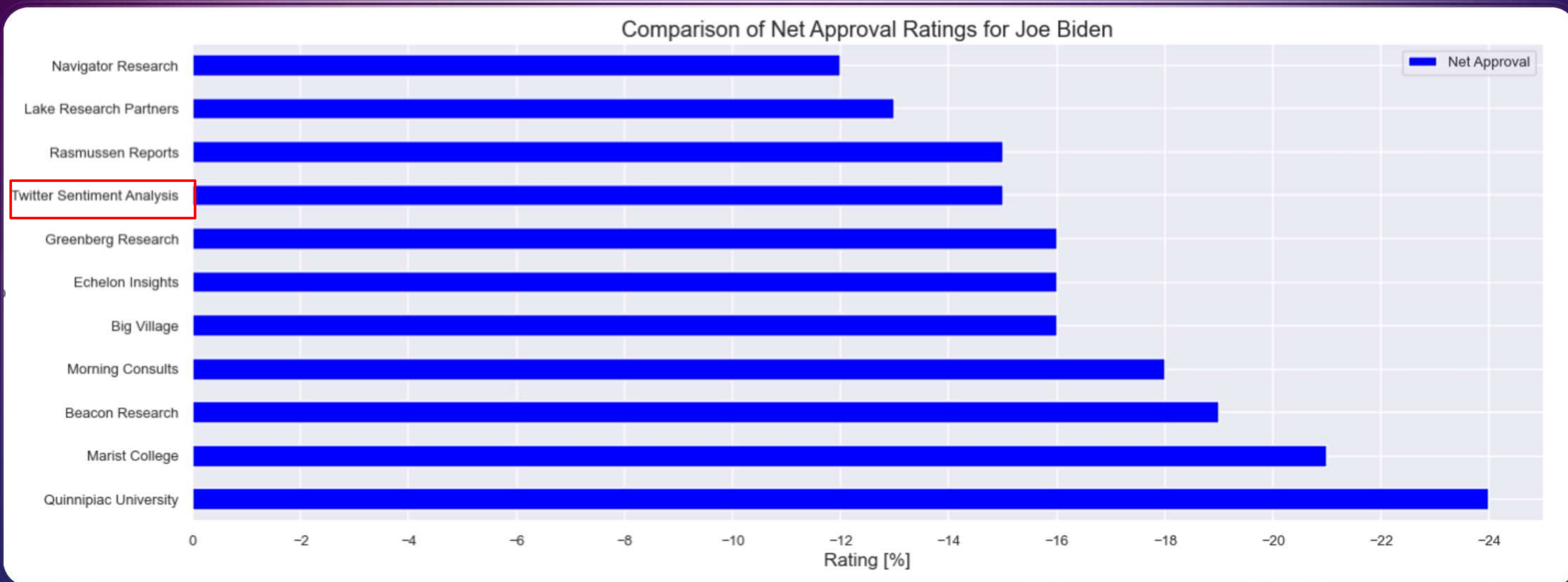
- The following models were trained:
  - *Naive Bayes*
  - *Logistic Regression*
  - *Random Forest*
  - *Recurrent Neural Network*
- After training, models were fed a portion of the Training Set they had not seen before
- Model with best performance was chosen -- **Recurrent Neural Network**
- **RNN** was then fed the Test Data, and these predictions were compared to:
  - *test labels (from VADER)*
  - *real polling data*
- The sentiment labels predicted by the RNN were converted into approval ratings:
  - *NEUTRAL tweets were discarded*
    - ***Approval Rating** → percentage of tweets that were POSITIVE*
    - ***Disapproval Rating** → percentage of tweets that were NEGATIVE*
    - ***Net Approval Rating** → Approval Rating – Disapproval Rating*

RESULTS

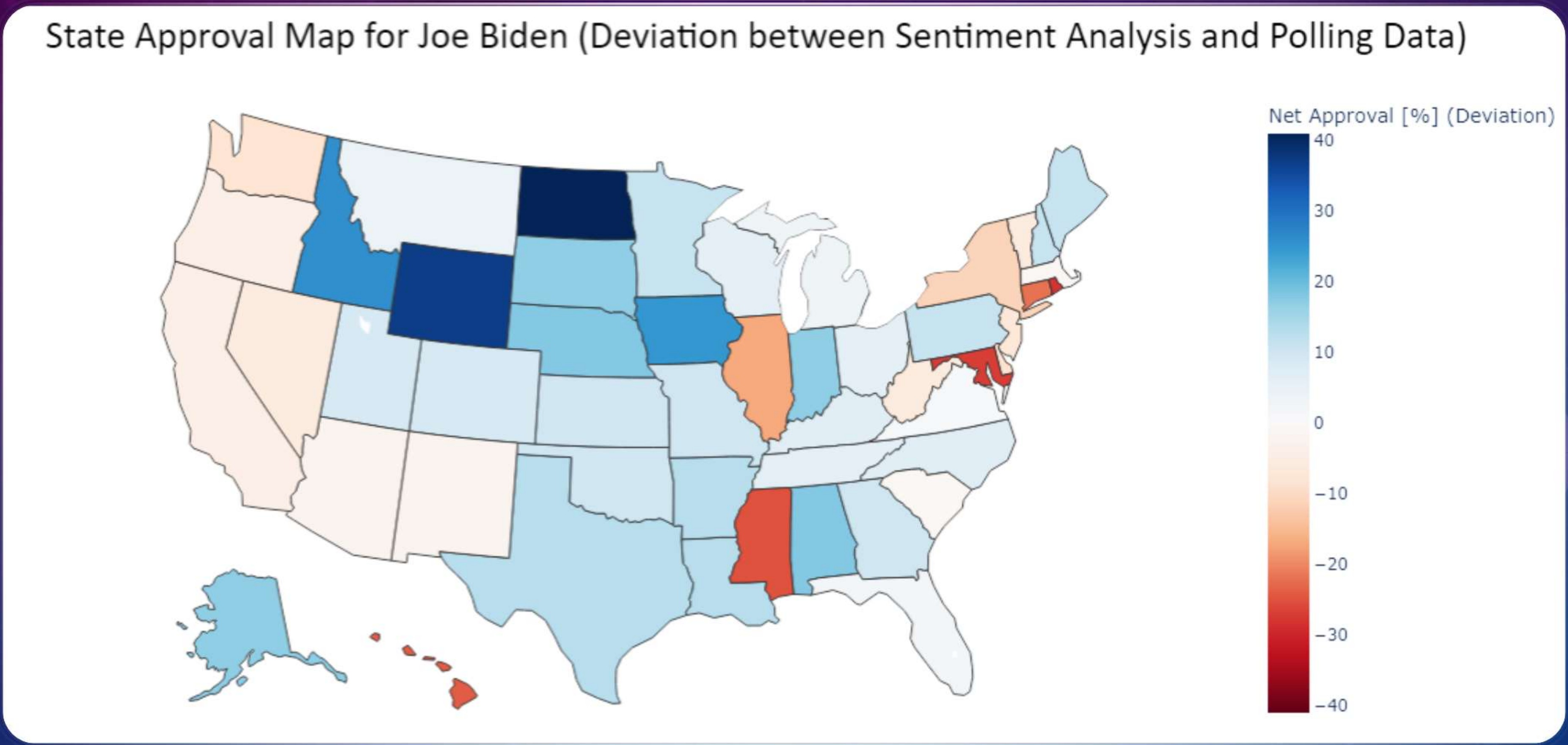
- The **overall approval and disapproval rating** predicted by the RNN model was compared to sets of real polling data
    - *all 10 polls were conducted on, or near, the week all tweets in the Test Set were posted*



❖ The **net approval rating** predicted by the RNN model was also compared to the same 10 sets of polling data



- The **difference in net approval between** the RNN model and traditional polling data was calculated for each state
  - predicted net approval for each state was compared to a single poll from that state



EVALUATION

- ❖ The model shows a great deal of promise
- ❖ Performed very well when compared to overall polling data
- ❖ Although not as accurate, the RNN predicted statewide net approval ratings surprisingly well
  - *there was, on average, only around 435 labeled tweets per state*
  - *predictions from the RNN deviated from polling data by*
    - *≤10% in 27 states*
    - *≤ 15% in 37 states.*
- ❖ The RNN model chosen had a relatively simple architecture
  - ❖ *implies there is a great deal of potential if we choose to develop more complex networks*

LIMITATIONS & NEXT STEPS

- ❖ Rate limits imposed by Twitter API
- ❖ Issues pertaining to location data
- ❖ Fine-grained sentiment analysis
- ❖ Identifying topic of tweets
- ❖ Utilizing emoticon data
- ❖ Utilizing other social media platforms to collect sentiment data

THANK YOU

Email: sqayum33@gmail.com

Github: @sqayum