PPOL 5206 May 8, 2025

Predicting Presidential Approval Rating with Reddit Comment Sentiment

By Daniel Cardenas Sanchez, Holt Cochran, Sam Cohen, Gabriel Soto, & Vaishnavi Singh

Background and Motivation

Many politicos and pundits complain about the accuracy of polls during presidential, state, and federal election cycles (Raffio, 2024). After the 2016 presidential election, for instance, confidence in polling diminished significantly, with the Trump victory many viewed as an upset (Ibid). To account for events like this, polling methods are constantly being changed and updated to account for demographic shifts, statistical insights, and technological advances in the field of AI and machine learning (Keeter and Kennedy, 2024). While issues with public perception of polling accuracy are nothing new, some of the unconventional and non-traditional ways to predict outcomes nowadays are.

For instance, Polymarket–an online betting platform–allows users to not only bet on the outcomes of sports games, but political events as well. In 2024, it outperformed most traditional pollsters in predicting Donlad Trump as the winner of the presidential election (Shaner, 2024). These unorthodox prediction methods apply to other domains as well– machine learning algorithms trained on reddit comments in the "R/WallStreetBets" subreddit were able to predict the price of stock shares relatively accurately (Wang and Luo, 2021).

For our project, we wanted to employ a similar non-traditional prediction method, and see if a model trained on the sentiment of reddit comments about sitting presidents could yield similar approval rating results to actual polls. There is some past research that utilizes sentiment to predict election outcomes: Alvi et. al. observe twitter comment sentiment to see how they compare with election results over time, and observe progress in this subfield (Alvi et. al., 2023). However, because of the enormous amount, constant influx, and dynamic nature of data relating to reddit comments and approval ratings, we add to this field by utilizing big data architecture to make our models scalable. Findings could potentially produce a way to gauge approval rating without the need for extensive polling more broadly, or at the very least, augment pollsters' findings.

Methodology

a. Data

Our data for this project comes from two main sources: the American Presidency Project website, and the Reddit API. The American Presidency Project is a non-profit organization affiliated with UC Santa Barbara that keeps records and data of US presidents (American Presidency Project, 2025). They retain a repository of weekly approval rating data for US presidents, which we were able to scrape using a Python script with the Pandas library. For this project, we decided to focus our efforts on the last three sitting presidents: Obama, Trump, and Biden (i.e. 2009 - present). However, even though we only considered approval rating data from this timeframe in our model, we retrieved all approval rating data from 2000 to present (i.e. Bush onward). Approval rating data for the American Presidency Project is collected by Gallup (Ibid).

Our comment data came from the Reddit API. The Reddit API allows users to pull data and metadata from multiple subreddits, and allows users to hone results based on keywords, subreddits, dates, and interacts. We chose multiple political subreddits to pull comment data from at the daily level, and made sure to stratify our selection by political ideology– that is, we made sure that our selection included a relatively even amount of both left-leaning and right-leaning subreddits in order to minimize any potential bias. Our comment data included comments in these subreddit from 2009 to present day. Once these comments were ingested, we calculated sentiment scores for each using the Textblob sentiment dictionary (with scores ranging from -1 to 1). These scores were then normalized, and aggregated by mean at the daily and weekly levels to provide a rough estimation of daily and weekly sentiment in each respective subreddit. The structure of the data can be found below:

Comment sentiment scores aggregated by mean at the daily level:

	subreddit1	subreddit2	…
Day1	Mean sentiment (float)	Mean sentiment	Mean sentiment
Day2	Mean sentiment	Mean sentiment	Mean sentiment

Comment sentiment scores aggregated by mean at the weekly level:

	subreddit1	subreddit2	…
Week1	Mean sentiment (float)	Mean sentiment	Mean sentiment
Week2	Mean sentiment	Mean sentiment	Mean sentiment

Presidential approval rating by week:

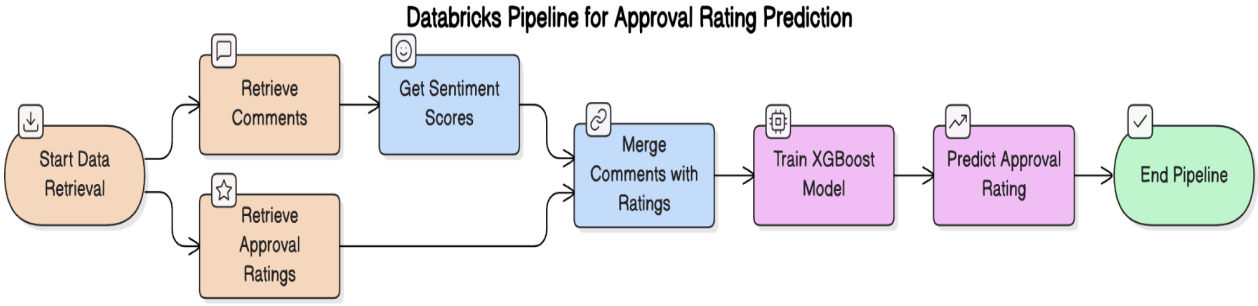
	Approval rating
Week1	Approval (%)
Week2	Approval (%)

b. Cleaning, ingestion, and pipeline in Databricks

We chose Databricks as our main cloud platform. Databricks is both user-friendly and collaborative, and allows users to utilize Spark without having to manually connect to external clusters for distributed computing. This, along with its multitude of options regarding dashboard creation, SQL capabilities, and free trial credits made it an ideal choice.

Databricks also makes pipeline generation relatively streamlined. We created a pipeline which could: (1) Call the Reddit API for new comments at specified intervals; (2) calculated the sentiment scores for each comments; (3) clean, wrangle, and merge these comments with the

presidential approval rating by date or week; and (4) train a model to predict approval rating based on sentiment. A flow chart illustrating this process can be found below:



The structure of the final merged dataset(s) can be found below:

	subreddit1	subreddit2	…	Approval rating
Week1	sentiment (float)	sentiment	…	Approval (%)
Week2	sentiment	sentiment	…	Approval (%)

… and

	subreddit1	subreddit2	…	Approval rating
Day1	sentiment (float)	sentiment	…	Approval (%)
Day2	sentiment	sentiment	…	Approval (%)

c. Modelling

We trained XGBoost models on our comment sentiment data to predict approval rating (our target variable). XGBoost not only outperforms many machine learning algorithms for both classification and regression tasks– it also deals with null values rather well. Since there were multiple subreddits that did not have sentiment values for specific days, some observations were treated as null after ingestion. We specified the existence of null values in the XGboost parameters, and utilized a 80-20 train-test split. After training, the model was used to predict the approval rating taking in newly ingested reddit comments as input.

Results

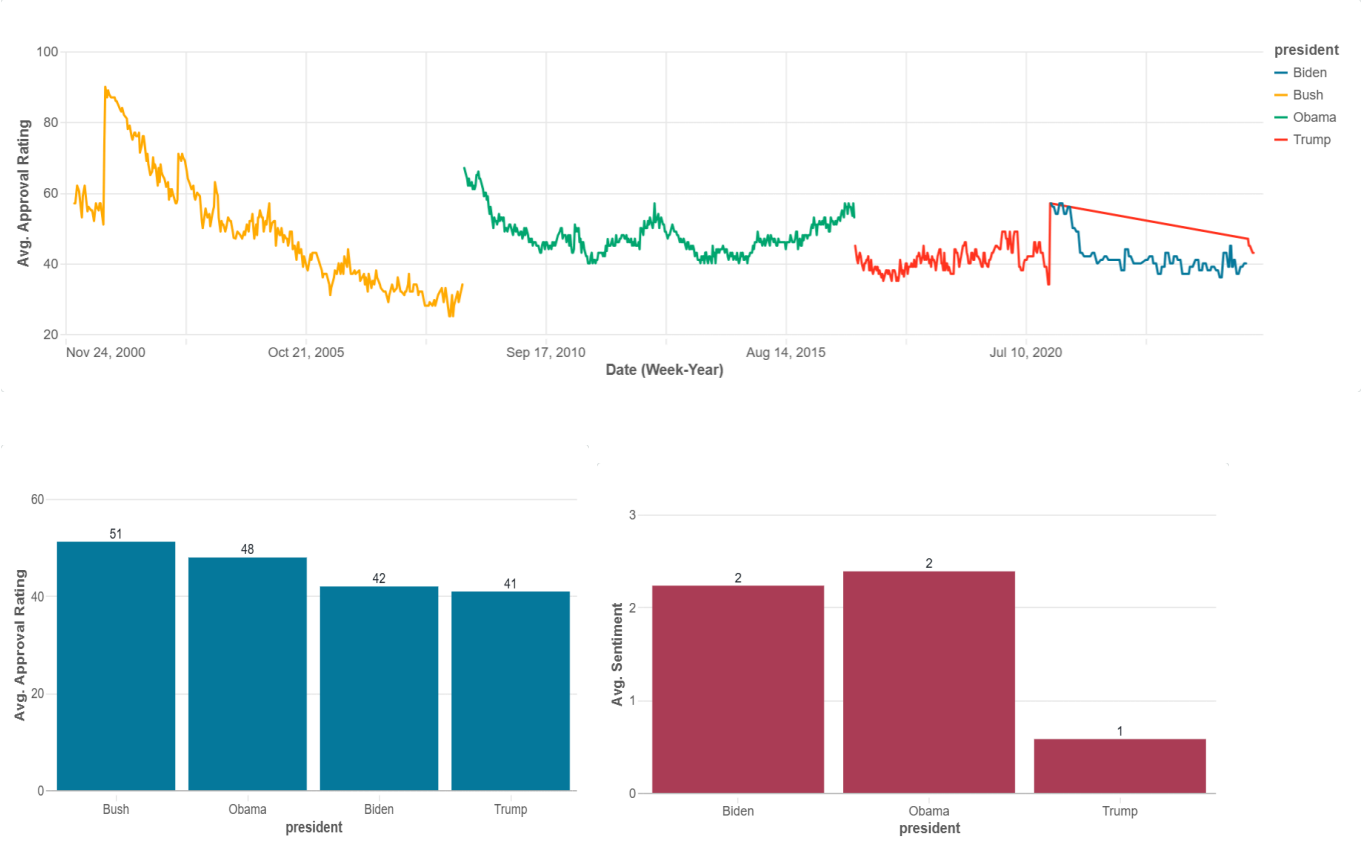
a. Descriptive statistics

Summary and descriptive statistics were calculated using a SQL workbook. Since approval ratings are aggregated at the weekly level, there were 796 weekly observations in this data set. Broken down by day, there were 5,885 observations. The average weekly approval rating from 2009 to 2025 was 45.13 percent. Broken down by party, Democratic presidents had a higher overall average approval rating (46.5 percent) when compared to Republican presidents (40.9 percent). The absolute maximum approval rating during this time frame was 67

percent (Obama's first week in office), and the minimum approval rating (34 percent) occurred only one week before (Bush's last week in office). Not counting Bush, the lowest approval rating was 34 percent for Donals Trump in the wake of January 6, 2021. Regarding reddit comments, there were hundreds of thousands of observations. The average sentiment score was neutral at 0.022, with the absolute maximum and minimum sentiment scores at 1 and -1, respectively.

b. Visuals

Using Databricks, we were able to create informative data visualizations and dashboards regarding the relationship between sentiment and approval.



Above: Weekly approval rating by president (top), Average approval rating by president (bottom left), Average sentiment score by president (bottom right)

c. Model results

We tested two different XGboost models: one predicting weekly approval ratings using comment sentiment aggregated by mean at the weekly level (n=796), and one at the daily level (n=5,885). For the weekly model, we made sure to lag the approval rating by one week to account for any delay in reporting. For the daily approval rating model, we assigned each day of the week the approval rating associated with that week as a whole, as described by Gallup.

Despite limited total observations and a sparse data matrix, our models delivered relatively promising results. The lagged weekly model had an Mean Absolute Error of 3.4, with a Root Mean Squared Error of 4.36. The daily model had a Mean Absolute Error of 3.1, with a Root Mean Squared Error of 4.08. Both had relatively high R-squared values (0.54 and 0.42, respectively), indicating a relatively high proportion of variance explained by the model. Results can be found below:

Lagged Weekly Model (n=796)	*Daily model (n=5,885)*
Mean Squared Error (MSE): 19.0638	Mean Squared Error (MSE): 16.6738
Mean Absolute Error (MAE): 3.4061	Mean Absolute Error (MAE): 3.1338
RMSE: 4.3662	RMSE: 4.0834
R-squared (R^2): 0.5411	R-squared (R^2): 0.4219

Justification and Conclusion

This model and pipeline offer a novel way to gauge public opinion about sitting US presidents. Comments can be pulled in real-time in Databricks to generate estimated approval scores, offering insight without the need for extensive (and often costly) public polling. This project is not without its flaws, however: there are only several hundred to several thousand observations depending on the model, meaning these models are likely not capturing the full picture, and may be overfitting to the training data. In addition, much of the matrix is sparse. Finally, unlike traditional polling methods, it is difficult to get a representative sample due to the anonymity of reddit users.

However, despite these flaws, the scope of this project is promising: it is relatively easy to scale up this pipeline, and to continually enhance the model over time. For instance, as new Reddit comments come in, the model can be re-trained offering more updated insight. The scheduled jobs and streamlined ML pipeline make this process relatively seamless as well. In addition, existing and future models can be hyperparameter tuned for optimal performance.

Overall, despite some drawbacks, this model and future iterations of it can help augment existing polls by offering a new, online discourse-centered metric. In addition, due to the cloud and use of big data architecture through Databricks, it can be scaled upward if necessary. Findings and predictions can lead to more responsive governance, early warnings for political shifts, and improved public engagement.

References

American Presidency Project. (2025). Presidential Job Approval. www.presidency.ucsb.edu. https://www.presidency.ucsb.edu/statistics/data/presidential-job-approval-all-data

Alvi Q, et. al. (2023). On the frontiers of Twitter data and sentiment analysis in election prediction: a review. PeerJ Computer Science. doi: 10.7717/peerj-cs.1517. PMID: 37705657; PMCID: PMC10495957. https://pmc.ncbi.nlm.nih.gov/articles/PMC10495957/

Al Zaabi, Sultan Ali. (2021). Correlating Sentiment in Reddit' elating Sentiment in Reddit's Wallstreetbets with the Stock Market Using Machine Learning Techniques. Rochester Institute of Technology. https://repository.rit.edu/cgi/viewcontent.cgi?article=12195&context=theses

Raffio, Nina. (2024). Can We Still Trust the Polls? USC Today. University of Southern California. https://today.usc.edu/can-we-still-trust-the-polls/

Reddit.com. (2025). reddit.com api documentation. https://www.reddit.com/dev/api/

Ketter, Scott & Kennedy, Courtney. (2024, August 28). *Key things to know about U.S. election polling in 2024*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/28/key-things-to-know-about-us-election-polling-in-2024/

Shaner, Kyle & Shaner, Emily. (2024, December [or specific day if available]). *Election results show potential of prediction markets*. University of Cincinnati. UC News. https://www.uc.edu/news/articles/2024/12/election-results-show-potential-of-prediction-markets.html

Wang, Charlie & Luo, Ben. (2021). Predicting $GME Stock Price Movement Using Sentiment from Reddit r/wallstreetbets. Department of Computer Science, University of Illinois at Chicago. https://aclanthology.org/2021.finnlp-1.4.pdf