Analyzing Political Trends of Indian American

Harshal Talele
University of Rochester
Rochester, US
htalele@ur.rochester.edu

Harshalkumar Loya
University of Rochester
Rochester, US
hloya@ur.rochester.edu

***Abstract*—Indian American community in US has gained a growing political influence in recent times and has come to recognised as a important electorate to cater to for the political parties on both sides of political divide. Given such prominence, no or very little has been done to analyse their political bias specially when the electorate remains divided. We propose to analyze the trends and biases of the Indian-American community from their real-time opinions on social media to get the sense of ground truth in an much robust way.**

***Index Terms*—Political bias, Indian-American, US election, automated approval rating, linguistic analysis, LDA classification, sentiment analysis, Twitter, social identity**

I. Introduction

The art of social decision-making and the division of power in a society is called politics. Political processes affect every part of life and affect every individual. With the people of Georgia deciding to elect a democrat and officially ending a long hard-fought midterm election, there is no dull moment in politics. To the astonishment of many analysts and pollsters, the Democrats have so far fared well in the November 8th midterm elections. In the past, voters have never routinely punished the party in power at the midpoint between presidential elections. However, the Republicans are doubtful of flipping the Senate and are set to take control of the House with one of the smallest swings in years despite high inflation and low approval ratings for Joe Biden. To drill down and analyze the results future, we will consider a specific ethnic group of the Indian American community as Indian Americans are the second-largest immigrant group in the United States. According to data from the 2018 American Community Survey (ACS)—which is conducted by the U.S. Census Bureau—there are 4.2 million people of Indian origin residing in the United States. Although a large proportion of Indian Americans in the United States are not U.S. citizens (38 percent), roughly 2.6 million are U.S. citizens (1.4 million are naturalized citizens and 1.2 million were born in the United States). Indian Americans are now the second-largest immigrant group in the United States. Their growing political influence and the role the diaspora plays significantly in recent elections and thus raises important questions— Did the community change its political affiliation from strongly voting blue as a block to being as polarized as it can be and more and more people increasingly favouring Donald trump and by extension Republican Party as their choice.

Twitter data opens up new horizons for scientists, both as a rich data source in its own right and also as a way of gathering information from the public. Roughly one-quarter of American adults use Twitter. And when they share their views on the site, quite often they are doing so about politics and political issues. A new Pew Research Center analysis of English-language tweets posted between May 1, 2020, and May 31, 2021, [1] by a representative sample of U.S. adult Twitter users, finds that fully one-third (33%) of those tweets are political in nature. and certain demographic groups are especially active contributors to the overall volume of political content on Twitter. Most notably, Americans ages 50 and older make up 24% of the U.S. adult Twitter population but produce nearly 80% of all political tweets. And 36% of the tweets produced by the typical (median) U.S. adult Twitter user aged 50 or older contain political content, roughly five times the share (7%) for the tweets from the typical 18- to 49-year-old.

Given twitter's influence on society, it becomes an obvious choice to gauge public sentiment. We work with various attributes of Twitter data in this study namely – Twitter text, User Description, tweet location, User location, User ID. We majorly focus on the tweet text put out by the user to judge their opinions and eventually political beliefs. Despite microblogging's rising popularity, little is known about how it affects personality. Do users utilise their microblogs to show off their personalities? Can one accurately assess someone's personality based on their microblogs? We will gain a better grasp of the connection between personality and social media as a result of the answers to these questions. Furthermore, microblogs offer a valuable opportunity to investigate personality expression and per-caption in a naturalistic setting. Self-report surveys conducted in controlled, decontextualized settings are frequently used in contemporary personality studies. However, in order to properly comprehend what humans are like, as Barker and Wright (1951) argued, we must observe natural behaviour in ordinary contexts.

II. Related Work

Previous research has shown that people inadvertently leave personality-related "behavioral residue" in their physical and virtual environments. Examples of personality expression have been found in daily conversations, bedrooms and offices, Facebook profiles, and virtual world activities. Since people frequently use microblogs to record their thoughts and activities, it is reasonable to expect that an individual's microblogs will also contain their personality-related residue.

Most research done on gauging the inclination of people is done via a process of manual collection of data using survey or poll-based research. Survey research is defined as ”the collection of information from a sample of individuals through their responses to questions” This type of research allows for a variety of methods to recruit participants, collect data, and utilize various methods of instrumentation. Though it is important to collect data, survey-based research is always criticized for the fact that it has a very less i.e., a fraction of the sample size on which it makes predictions. With the involvement of manual activity, it can also be biased and can become a tool of politically biased organizations. For every survey supporting democrats, it is possible to produce a different one with contrasting results.

The reliability of survey data may depend on the following factors:

- Respondents might not feel motivated to offer truthful, accurate responses.
- Respondents might not feel comfortable answering questions that make them appear unfavourable.
- Respondents may not be entirely aware of the reasoning behind any particular response due to forgetfulness or even boredom.
- Compared to other question categories, surveys containing closed-ended questions could have a lower validity rate.
- Data mistakes resulting from non-answers to questions may exist. Bias results from the possibility that a different proportion of responders than those who choose not to reply to a survey question.
- Because different respondents may perceive different answer choices, this could result in data that is confusing. For instance, the response option ”somewhat agree” may signify different things to various subjects and have a distinct meaning for every respondent. Answer choices of ”yes” or ”no” might also be troublesome. If ”just once” is not an option, respondents may select ”no.”

Thus, working with Twitter data ensures a result closer to ‘ground truth’ and is real-time analysis.

III. Data and Methodology

In this section, we describe the data extraction, data filtering and the methodologies used to analyze the data

A. Data Extraction

To extract the tweets, we used the Twitter developers accounts and generated the API keys to get authentication for data extraction. We then used tweepy to select and collect the tweets.

The official Twitter API enables us to obtain tweets according to different queries using keywords, location, type of tweet as parameters. We used the following query keywords for the tweets -

election OR usaelection OR uselection OR election2024 OR presidentialelection OR Trump OR Biden OR senate OR congressman OR congresswoman OR hindu trump OR modi trump OR modi uselection OR indian us president OR india us president OR modi biden OR indian biden OR demoractic president OR republican president

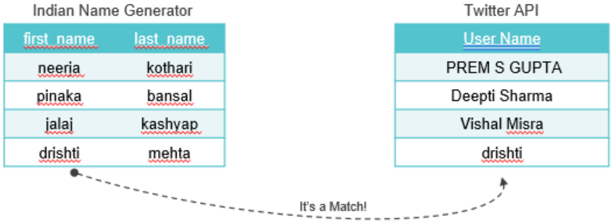
The location for the tweets in the query was given as *place_country:US*

We extracted the data for tweets matching the above query for a duration of two years from November 2020 to November 2022.

B. Data Filtering

We extracted around 3.5 Million tweets using the aforementioned process. That gave us around 267K unique users. However, our focus is on Indian Americans and not the entire population. We used *User Name* to filter the tweets by Indian Americans. We tried using python libraries like *namesor* and *ethnicolr* but some were paid and others did not give specific category of Indian Americans. Hence, we used name matching with a dataset of Indian names from a generator.

We used the python library *indian_names* to get a dataset of randomly generated Indian names. We then used keyword matching of twitter user names with these generated names to filter the tweets by Indian Americans. Ultimately had the data of around 2.8K unique users who were Indian Americans.



first_name	last_name
neeria	kothari
pinaka	bansal
jalaj	kashyap
drishti	mehta

User Name
PREM S GUPTA
Deepti Sharma
Vishal Misra
drishti

Fig. 1. Filtering Indian names by Keyword Matching

C. Data Pre-Processing

The tweet text which is the major attribute of the data has a lot of unwanted strings that we can get rid of. We use Stemming, Lemmatization and removing stopwords to get cleaner and relevant stream of tweet text. We leverage the nltk library in python for this process. *Stemming* is the process of producing morphological variants of a root/base word. *Lemmatization* is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item Both stemming and lemmatization produce the inflected words’ root forms. The distinction is that although lemma is an actual language term, stem may not be. Previous studies have shown use of different words from user descriptions based on their identities. [2] One of the identities was the political parties they support. We used that to assign the *Support Group* to each user. Fig. 2 shows the keywords and their respective polarities from that research.

D. Topic Modelling using LDA

Previous research shows that Latent Dirichlet Allocation topic models to be one of the best option in text categorization

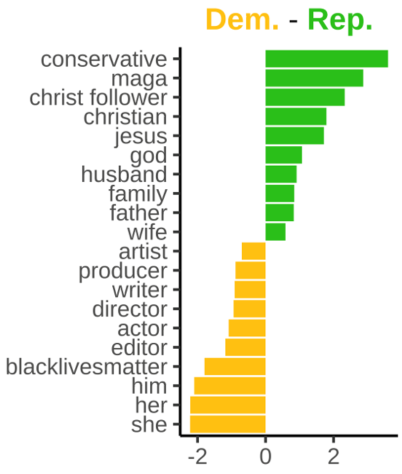


Fig. 2. Words in User Description of respective Support Group

[3]. We use the gensim package to run LDA models over our data and get the top 10 topics that are the most frequent. Fig. 3 show the word clouds for each topic.

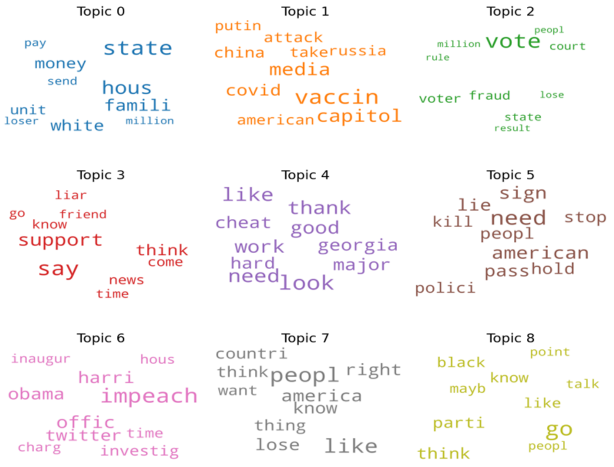


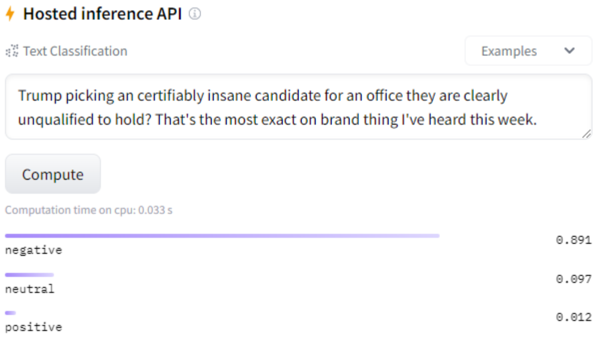
Fig. 3. Frequent Topics - Output of the LDA Topic Modelling

E. Sentiment Analysis

To understand the sentiment of the tweets, we tried using TextBlob library in python. However, it did not give accurate results. Upon research, we found pre-trained models that could be used to get more accurate results for sentiment. We specifically used the Twitter-roBERTa-base for Sentiment Analysis [4] from huggingface.co that had been trained on over 124M tweets to get the sentiment. It assigns a sentiment score in the range of 0 to 1 for each of the sentiment, namely, positive, neutral and negative. The sentiment score for all three add up to 1 and the sentiment with the maximum score is dominant in that particular text. Fig 4 shows an example.

IV. Results and Insights

In this section, we will go through some of the insights and results derived from the analysis.



Trump picking an certifiably insane candidate for an office they are clearly unqualified to hold? That's the most exact on brand thing I've heard this week.

Computation time on cpu: 0.033 s

```
negative    0.891
neutral     0.097
positive    0.012
```

Fig. 4. Sentiment Analysis output for the pre-trained model

- The Fig. 5 shows word clouds of tweets from supporters of both Democratic and the Republican parties. Trump continues to be the talk of the town among both.

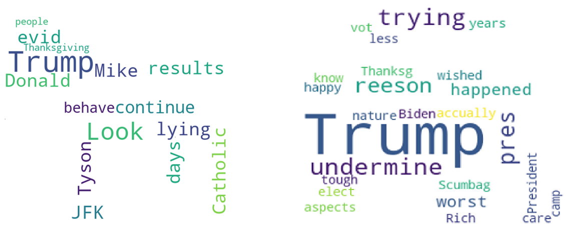


Fig. 5. Word Clouds for tweets from both support groups

- The negative sentiment has been consistently strong for the duration of 2 years, both for Trump and Biden. But, in case of tweets mentioning Trump, the positive sentiment has seen a decline over the 2 years which shows a decrease in good-sentiment towards Trump.

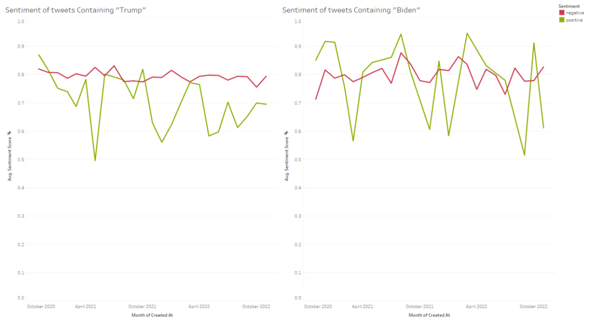
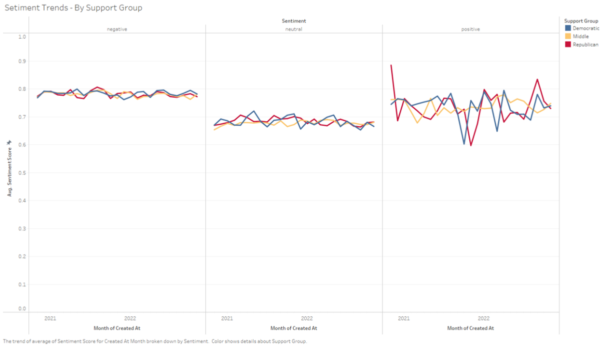


Fig. 6. Average sentiment score trend - Trump vs Biden

- The negative and neutral sentiment have been fairly consistent over the duration of two years. Negative sentiment stays stronger than the neutral. As far as the positive sentiment is concerned, we observe a higher variation in the sentiment score signaling that people do not tend overly positive even when they are not being negative or neutral.

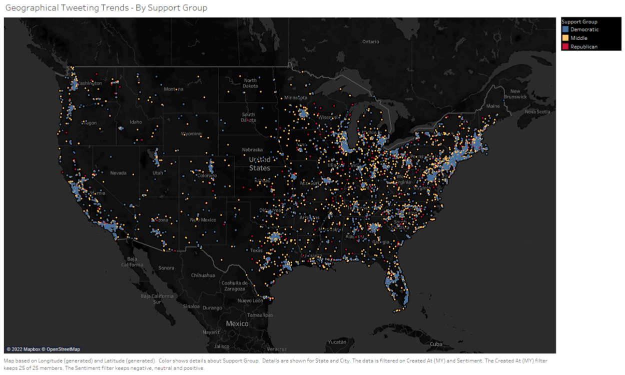


Setiment Trends - By Support Group

The trend of average of Sentiment Score for Created At Month broken down by Sentiment. Color shows details about Support Group.

Fig. 7. Average sentiment score trend - By Support Group

- The eastern part of US seems to have higher density of tweets with some more concentration around the west coast.

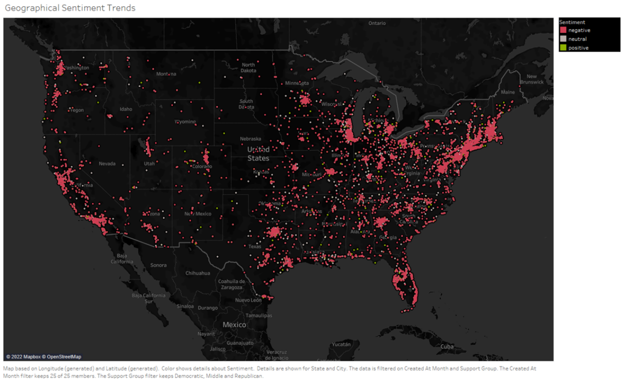


Geographical Tweeting Trends - By Support Group

Map based on Longitude (generated) and Latitude (generated). Color shows details about Support Group. Details are shown for State and City. The data is filtered on Created At (MY) and Sentiment. The Created At (MY) filter keeps 25 of 25 members. The Sentiment filter keeps negative, neutral and positive.

Fig. 8. Geographical Distribution of Tweets

- The negative sentiment overpowers the other two across the entire nation which tells us people potentially use twitter to lash out their political opinions against the opposing party of their support.

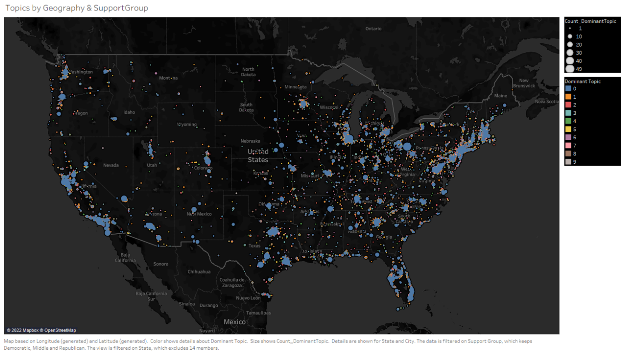


Geographical Sentiment Trends

Map based on Longitude (generated) and Latitude (generated). Color shows details about Sentiment. Details are shown for State and City. The data is filtered on Created At Month and Support Group. The Created At Month filter keeps 25 of 25 members. The Support Group filter keeps Democratic, Middle and Republican.

Fig. 9. Geographical Distribution of Sentiment

- "Topic-0" is the most dominant, it is almost equally spread across the nation. This topic talks about Economic factors.

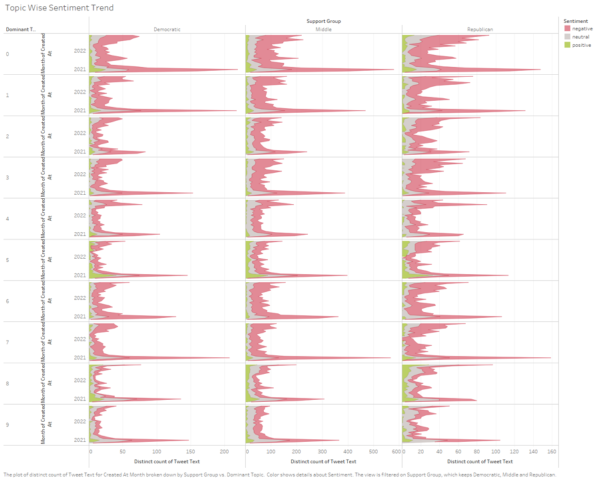


Topics by Geography & SupportGroup

Map based on Longitude (generated) and Latitude (generated). Color shows details about Dominant Topic. Size shows Count_DominantTopic. Details are shown for State and City. The data is filtered on Support Group, which keeps Democratic, Middle and Republican. The view is filtered on State, which excludes 14 members.

Fig. 10. Geographical Distribution of Topics

- The trend of number of tweets across different support groups and sentiments also validates the previous insight of negative sentiment being dominant.



Topic Wise Sentiment Trend

The plot of distinct count of Tweet Text for Created At Month broken down by Support Group vs. Dominant Topic. Color shows details about Sentiment. The view is filtered on Support Group, which keeps Democratic, Middle and Republican.

Fig. 11. Topic wise Sentiment Trends

V. Conclusion

The present work proposes a real-time sentiment analysis of the dynamic public onion on elections and related discussion. We also aim to develop an automated public sentiment tracker which has the potential to replace the poll-based approval rating trackers. We have convincingly shown that the immigrant community, in this case, a south Asian group, believes in keeping their political ideologies to themselves or is apolitical when it comes to having a stand on national issues. This makes the study and predictions of which way the individual swings make it more interesting. As discussed above, the present work is not without limitations. However, we hope that our method, and its public availability, will encourage future mixed methods analyses that dig further into the use of public opinion and the structure of online identity as presented in bios. Researchers and psephologists interested in US elections, sentiment analysis and topic modelling should find our work useful.

VI. Future Scope

We have the data of the unique individual and their opinion on US politics and political leaders, we can, for further fine-grained analysis, divide the user dataset into naturalized and non-naturalized citizens. Naturalized citizens are people who have been born in the US and have US voting rights based on their citizenship via birth. Non- non-naturalized citizen is a person who has not been born on US soil but eventually moved and acquired US citizenship via other legal avenues. The survey-based results show an explicit divide between political affiliation between these two categories.

The timing of the shift raises some interesting questions which can lead us to interesting insights. The Indian American community got more polarized and shifted its political beliefs in the last two elections which coincided with the rise of right-wing nationalist leader Narendra Modi as the prime minister of India. Prime Minister Modi has always shown great affection for president Trump and almost endorsed him as a candidate by stating that Trump is and will be a friend of India at the White House by saying 'Phir Ek Baar Trump Sarkar'(One more chance for the Trump government) at the football stadium rally packed with Indian Americans at Houston, Texas. Do political beliefs of an individual change based on how politics play out 'back home'?

As discussed in the above sections to filter out the tweets and by extension the unique users, we have used the Keyword matching approach. This method has limited utility as Username is not a very robust parameter to get all the individual users available on the public platform. The keyword matching can produce a slight bias in the results as we will have users from different ethnicities that are our target group and hence cannot create a just pitcher. It is much more efficient to use pre-trained machine learning models which have the capability to produce results at this granularity.

When dealing with one's political beliefs, it is imperative to find out what shapes that particular belief. This is where political events and their special coverage in mainstream print and digital media come in. We can complement our study by adding News Data for the time we are interested in the tweets and can find some novel patterns.

References

[1] Pew Research Center - Politics on Twitter: One-Third of Tweets From U.S. Adults Are Political

[2] Pathak, A., Madani, N., & Joseph, K. (2021). A Method to Analyze Multiple Social Identities in Twitter Bios. arXiv. https://doi.org/10.1145/3479502

[3] B. Kane and J. Luo, "Do the Communities We Choose Shape our Political Beliefs? A Study of the Politicization of Topics in Online Social Groups," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3665-3671, doi: 10.1109/BigData.2018.8622535.

[4] Loureiro, D., Barbieri, F., Neves, L., & Anke, L. E. (2022). TimeLMs: Diachronic Language Models from Twitter. arXiv. https://doi.org/10.48550/arXiv.2202.03829