CS 6850

Project Report

Information Spread in Elections

Sidhee Hande (sph75), Atharv Jain (aj479), Hrishik Rajendra (hr332)

Introduction

The Indian General Elections are one of the largest and most complex exercises of democracy in the world. Held once every five years, the results of these elections decide the members of the Lok Sabha, the lower house of the Parliament of India. Elections in India are dynamic events influenced by numerous factors, including socio-political developments, public sentiment, and regional dynamics. In 2014, then election winner, Narendra Modi gained immense popularity and widespread support throughout the country, and this came to be known as the Modi Wave. This was further fueled by the extensive use of news outlets and social media platforms to propagate his message and amplify his reach, and ultimately led to a landslide victory by him and the Bharatiya Janata Party (BJP) in 2014.

Initially motivated by the need to understand how Narendra Modi was able to spread his influence throughout the nation, in this paper, we seek to observe how important public sentiment is to predict the results of the Indian general elections. We use a selection of datasets for this goal, studying public sentiment and opinion through news headlines and Twitter chatter. Our paper seeks to uncover the relationship between the news media and social media discourse, analyzing sentiment and its role in influencing public opinion, ultimately driving election outcomes.

Data

Our project uses three main datasets, the India News Headline Dataset [1], the Indian News Dataset (IND) [2], and the Indian Political Tweets 2019 Dataset [3].

Our first dataset, the India News Headline Dataset, serves as a persistent archive of notable events published by the Times of India, one of India's most popular newspapers, from 2001 to 2022. With approximately 3.6 million headlines , this dataset provided us with a strong foundation to observe and analyze how public sentiment was influenced through news headlines, and the key figures which were associated with a general sentiment. We considered news headlines from a few select politics-related categories, between 1st January 2013 and 16th May, 2014 (day of the Election Results).

The IND Dataset contains news data or articles from the ten most highly rated Indian news websites (i.e, India Times, Firstpost, NDTV, The Indian Express, Times Now, One India, Hindustan Times, India TV, News18 and Zee News), with the main motive that they have news articles with a large number of views or shares which is a good indicator of news popularity

among readers.The dataset is annotated with popularity labels. Articles that have more than 500 reshares are labeled 1 (popular), while articles with less than 500 reshares are labeled 0 (unpopular). The dataset also categorizes the articles according to the genre they follow. We considered articles from all 10 news portals that were marked to be under the “Politics” subject. Lastly, we also decided to observe the public chatter during the 2019 Indian general election by using the Indian Political Tweets dataset. Twitter is a leading social media platform where users share thoughts, opinions and engage in public discourse in real-time. With more than 20 million users in India, political chatter on Twitter during the general elections becomes an important factor which might potentially influence public opinion, voter behavior and electoral outcomes. This dataset contained a sample of around 46000 tweets collected between the hours of 11:00 pm and 12:00 am, from February 14 to May 16, 2019, around when the Indian general elections were taking place.

Methods

1. Sentiment Analysis

The first step of our project involved finding the general sentiment of the news headlines and the tweets. For this, we used an extension of the Multilingual Language Model: XLM-T [4 - https://arxiv.org/pdf/2104.12250.pdf] which was trained to focus on the sentiment pertaining to politics. The model was fine tuned on eight languages: Arabic, English, French, German, Hindi, Italian, Spanish and Portuguese. This aspect of the model was critical for our analysis because there were instances of headlines, articles and tweets which used a mix of multiple languages. Running sentiment analysis on the raw news data and tweets provided us with a metric which measured how positive, negative or neutral the given text was.

2. Named Entity Recognition

In order to build graphs from the datasets we had selected, we decided to extract named entities from the raw news data (both news headlines and news abstracts). Named entities are described as real-life objects which can be denoted with a name and belong to an entity category such as place, person, organization, time, amongst others. These entities would later be used as the nodes for our graph. We used spaCy, the popular Python natural language processing library and only selected the relevant entities belonging to the categories person, norp (nationalities or religious or political groups), organization, event, law or money.

Recognizing entities from tweets proved to be a lot simpler. In this case we recognized an entity as the users/accounts mentioned in the tweet with an '@'. We did not have to explicitly filter for only entities related to politics since the tweet dataset only contained tweets which were political, so we felt that all the mentions in the tweets were appropriate for our project.

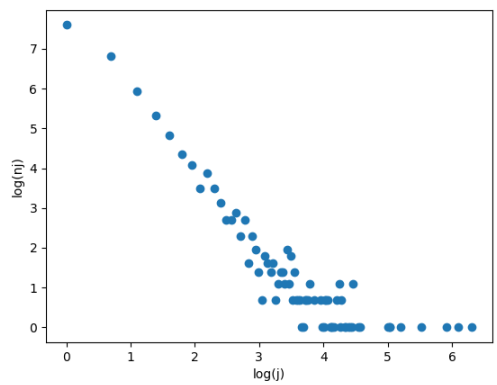
Network Graphs

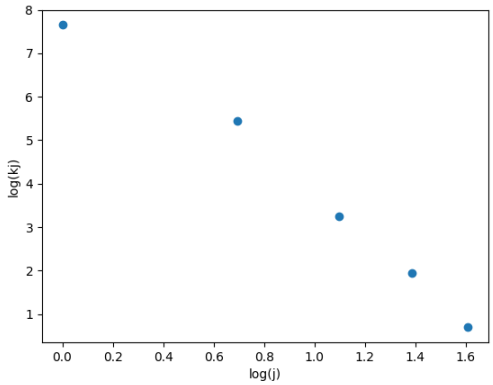
Times of India Headlines Graph

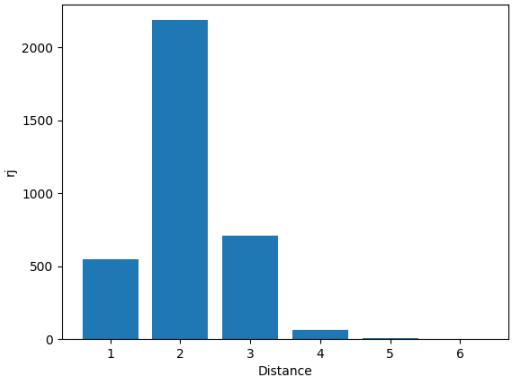
We constructed a graph of the named entities that we extracted from the headline text. The nodes of the graph were tuples of (entity name, average positive sentiment score, average neutral sentiment score, average negative sentiment score). If two entities were mentioned together in the same article, then we would draw an edge between their respective nodes.

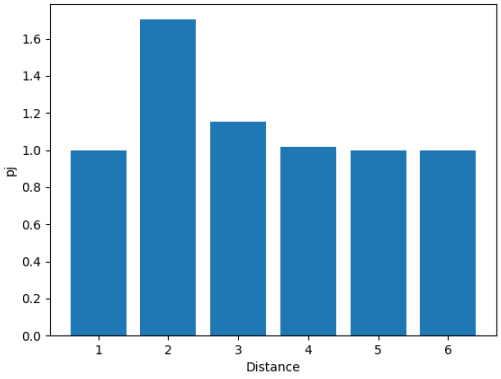
Mathematical Analysis

From the graphs as well as the metrics calculated, we see that the largest connected component accounts for a little more than half of the total network. We also see that the majority of the nodes have a distance of 1-3 to another node. This low distance indicates that when entities are mentioned in headlines, the topics associated with them are targeted. They don't necessarily represent political alliances, rather they show the interaction between these entities during election time. The topics these entities tend to gravitate towards are a concentrated set of categories.









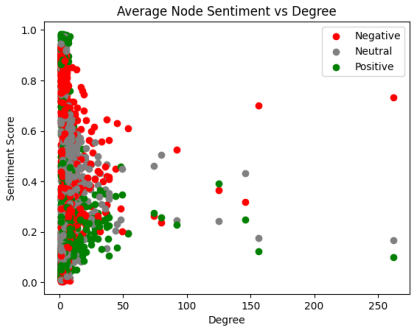
Largest Connected Component	Total Number of Nodes	Amount of Total Network Composition
3519	6205	.5671232876712329

Tweets Graph

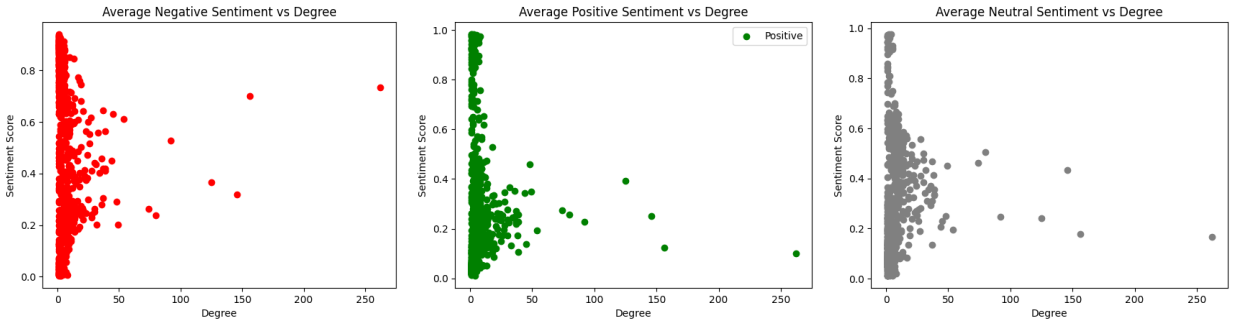
We created a graph for the twitter data in a similar way. Tweets with two or more accounts mentioned would result in edges being formed between the accounts (nodes). We also kept track of the sentiment score of the tweet to give us an idea of the general sentiment associated with each node. After the graph had been fully constructed, we found the average sentiment associated with each node. This is calculated by taking the average of the sentiments (in 3 categories: positive, negative and neutral) of all the tweets it has been mentioned in (i.e. average

of the sentiments for each edge), resulting in the average sentiment of that node. The resulting graph from the Twitter dataset had 8794 nodes with 53798 edges.

We first wanted to observe how a node's degree varied with its general sentiment score for each node for each category (positive, negative and neutral). We also plot it separately to view each sentiment's relation with the node's degree individually.



Relationship between Sentiment and Degree for the Twitter Graph



Relationship between each Sentiment and Degree

The degree of the node reflects the co-occurrence of the entity pair in the same tweet.

As we can see from the above images, there is no one clear trend that we can observe about the sentiment of tweets. However, we can observe that for all the nodes, the average positive sentiment score tends to be the lowest.

Next we tried to dig deeper into finding the most popular entities in the Twitter dataset. We measure popularity in terms of the number of tweets that mentioned the entity, as well as the number of retweets, replies and favorites they received. We came up with this formula to calculate the reachability of entities:

Reachability = 1*Retweets + 0.8*Replies

The weights 1 and 0.8 were decided arbitrarily. We essentially wanted to capture the idea that engaging with a tweet by retweeting it is more effective in increasing visibility than replying to

it. After applying this formula, we calculated the top 10 most influential entities which are listed in the table below.

Election Outcome Prediction Experiment

We ran a small experiment to see if we could predict the outcome of an election based on the popularity of political entities in tweets. We already described how we calculated the popularity of entities mentioned in tweets. Next, we observed all these popular entities and labeled them as either "1" representing a pro-BJP node or "-1" representing a pro-Congress node. Entities that were not part of the Indian political discourse were labeled as "0" implying that they would not affect the voting dynamics.

We made a few assumptions to run this experiment:

1. The reachability reflects the minimum number of people who must have read the tweet.
2. Viewers that read a tweet which has a positive/neutral sentiment about an entity, will lean towards voting for that entity (or the political leaning of the entity).
3. On the other hand, viewers that read a tweet which has a negative sentiment about an entity, will lean against voting for that entity (or the political leaning of the entity).

Our aim is to explore whether we are able to predict the outcome of the election simply based on visibility among tweets. We aggregate the viewership statistics of pro-BJP nodes and those of pro-Congress nodes and see which are higher.

Based on assumptions 3 and 4, we follow this formula for counting the number of voters:

Predicted Number of Voters = (1- Negative Sentiment Weight)*Reachability

This table shows the entities, along with their viewership and predicted voting statistics:

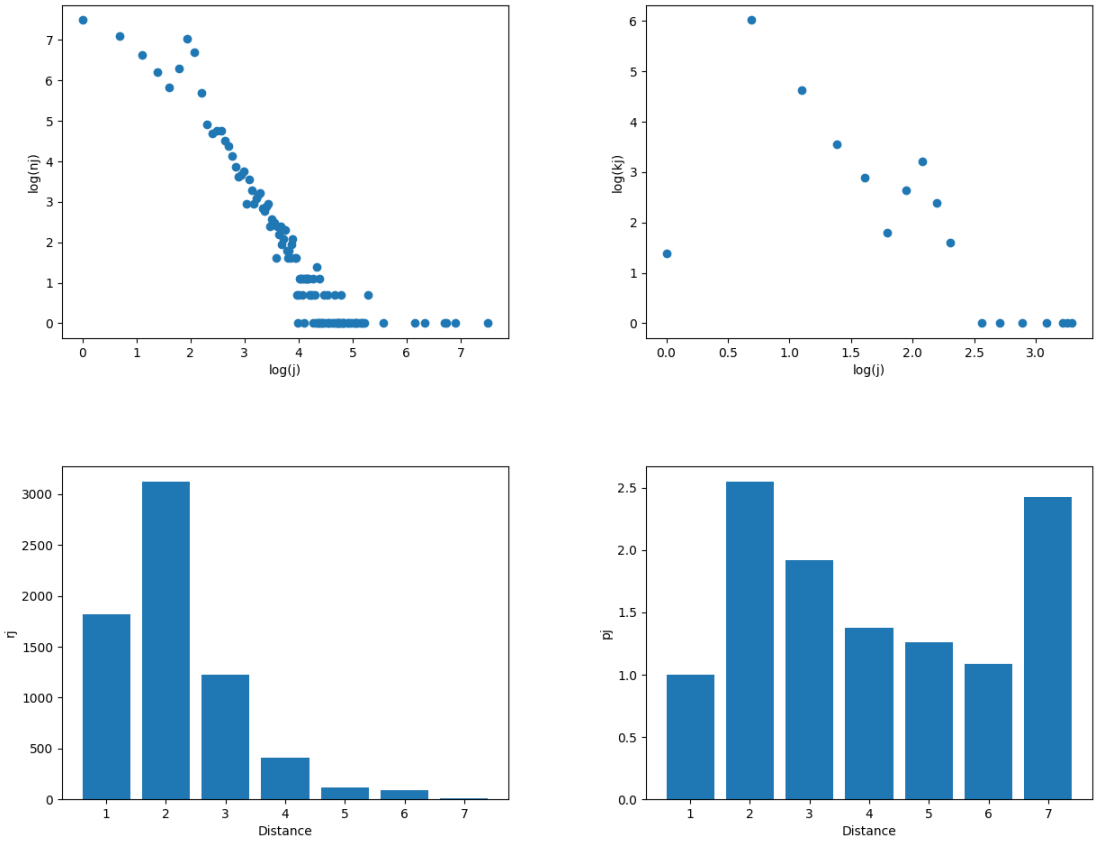
Entity	Annotation	Reachability	Negative Sentiment	Predicted Voters
narendramodi	+1	1148181	0.42	665945
RahulGandhi	-1	613334	0.51	300534
INCIndia	-1	347869	0.52	166977

BJP4India	+1	186914	0.45	102803
AOC	0	186081	0.64	0
realDonaldTrump	0	101317	0.74	0
SenFeinstein	0	91617	0.75	0
AmitShah	+1	67638	0.36	43288
IlhanMN	0	63531	0.7	0
PMOIndia	+1	52841	0.43	30119

Results:

From the above table we can see that the total number of predicted pro-BJP votes is 8,42,155 and the total number of pro-Congress votes is 4,67,511. These numbers are in line with the actual 2019 General Election results where the BJP, led by Prime Minister Narendra Modi won with a landslide majority.

Mathematical Analysis



We can see that most of the nodes have a distance of 1-3 with the largest connected component being 6800 nodes. The social graph is composed of 8794 nodes indicating that the largest connected component accounts for approximately 77.33% of the network.

Largest Connected Component	Total Number of Nodes	Amount of Total Network Composition

6800	8794	.7732544916988856

Indian News Dataset Graph

We created a popularity graph of the entities derived from the Indian News Dataset since we wanted to study the popularity of a news article, and subsequently the popularity of the political entities mentioned in the news article. We also added meta-data to the graph in terms of the positive, neutral and negative sentiment scores that we calculated for each of the articles and their entities. *[**Note**: The sentiment score should not be confused with the public sentiment about the entity. An article may criticize one political party while praising the other and thus have a high overall negative sentiment score but this does not imply that both parties garner negative public sentiment. The sentiment score of an article reflects more on the news reporting style, than it reflects on the public perception of a political entity.]* The creators of the IND dataset manually annotated articles based on the number of reshares they received. If the number of reshares received were greater than 500, then the article was marked as “Popular”, else not. We wanted to observe the relationship between popularity of an article and the political entities it spoke about. Entity mention frequency is also an important aspect since if the same entity is mentioned multiple times in a “popular” news article, then we can say that the effect of the article is amplified when shaping public opinion about that entity. Conversely, it may also be the case that the article itself may get more reshares and thus be more popular if it mentions certain entities. Hence, this relationship was of particular interest to us.

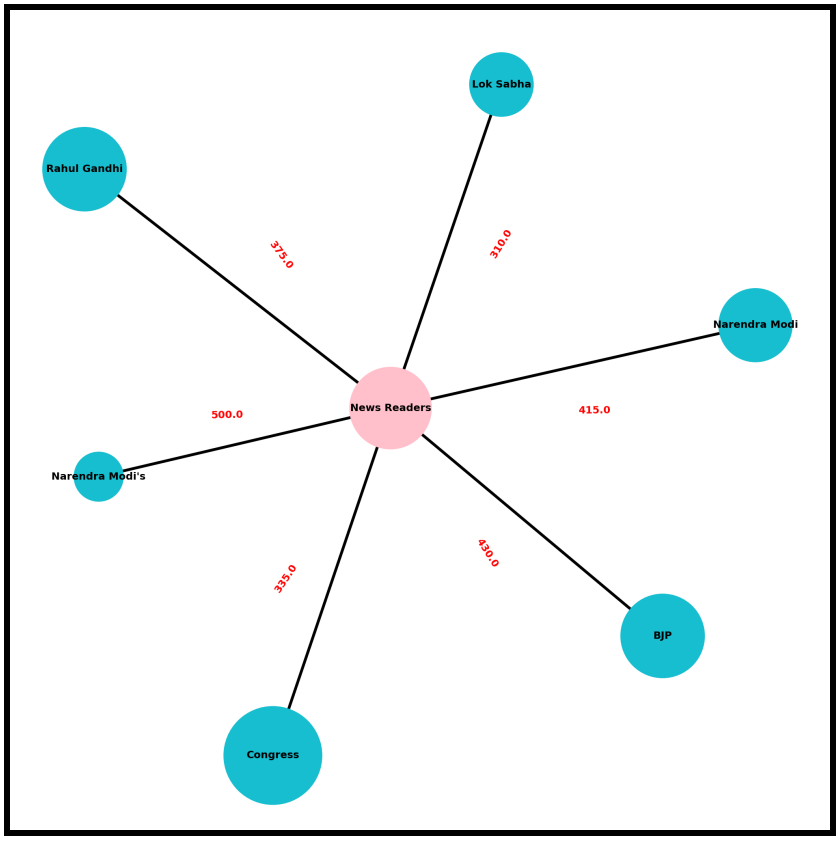
We created 10 graphs for the 10 news portals. The nodes of the graph were the entities mentioned in the articles published by that news portal. We observed these graphs to understand the top 3 most frequently mentioned entities for each news portal. They are summarized in the table below.

News Portal	Top 3 entities
India Times	'Lok Sabha', 'EC', 'Narendra Modi', 'the Election Commission'
First Post	'Congress', 'BJP', 'Sena'

NDTV	'Congress', 'BJP', 'Lok Sabha'
Indian Express	'Congress', 'Lok Sabha', 'BJP'
Times Now	'Lok Sabha', 'Congress', 'BJP'
One India	'Congress', 'Lok Sabha', 'Rahul Gandhi', 'Narendra Modi', 'BJP', "Narendra Modi's"
Hindustan Times	'BJP', 'Congress', 'Thackeray', 'Sena'
India TV	'Congress', 'BJP', 'Lok Sabha'
News Eighteen	'Congress', 'li', 'BJP'
Zee News	'Congress', 'Lok Sabha', 'BJP'

Observations:

- Of all the nodes (entities) mentioned above, the most popular entity in terms of the number of reshares they received was “Narendra Modi”. Every article that mentioned Narendra Modi received more than 500 reshares. While the sentiment of the articles mentioning him ranged from positive to negative depending on the context, he was clearly a popular candidate. This is good evidence in the direction of the actual election results we saw in 2014 with Narendra Modi going on to become the Prime Minister of India.
- The 2 entities which were mentioned by the most number of news portals (9/10) were BJP and Congress, the 2 leading political parties in India. These results are not surprising since these 2 parties do tend to dominate the national political discourse in India.



Most influential nodes for "One India" news portal

Here is a sample negative sentiment graph for articles from the news portal "One India". The size of the nodes reflects the magnitude of the average negative sentiment observed for the entity. The weight of the edge reflects an estimate of the number of reshares the articles mentioning the entity received. These are not exact numbers and only exist to help us make a relative comparison. We assume that reshares reflect viewership. As we can see in this graph, the negative sentiment encompassed by "Congress" and "Rahul Gandhi" are larger than those encompassed by "BJP" and "Narendra Modi". [*Clarification: "Narendra Modi" and "Narendra Modi's" reflect the same person but our NER model detected them separately so we have maintained them as separate nodes*] What is interesting is that despite these nodes having a higher negative sentiment, the total number of re-shares for the opposing party BJP and their leader Narendra Modi were higher. This is also reflective of the popularity of Mr Modi as a political candidate and evidence of the "Modi Wave".

Independent and unbiased media outlets are the pillars of a democracy. While it is difficult to find empirical evidence to ascertain a news outlet's political leanings, we can try to gain evidence of their reporting styles. Do they tend to sensationalize an article with heavy use of “negative” words, or do they try to maintain a neutral reporting style? Since we had a sample of 200 articles from 10 Indian news portals, we tried exploring the average sentiment of the articles that they published. Below is a summary.

News Portal	Average Positive Sentiment Score	Average Neutral Sentiment Score	Average Negative Sentiment Score
First Post	0.105247	0.538871	0.355882
Hindustan Times	0.114270	0.574836	0.310895
Indian Express	0.091363	0.601471	0.307165
India Times	0.131756	0.363126	0.505118
India TV	0.138759	0.627177	0.234065
NDTV	0.139323	0.485084	0.375593
News Eighteen	0.147272	0.626423	0.226305
One India	0.113189	0.634766	0.252045
Times Now	0.095085	0.657732	0.247183
Zee News	0.108851	0.658323	0.232826

All the articles are from around the same time period and hence likely cover similar political topics. However, as we can see the sentiment distribution across news portals is not uniform. We can see that India Times tends to report articles with a negative writing style, since it has the highest average negative sentiment score of 0.50. We can also see that none of the news portals

have a high average positive sentiment score. This reflects not just the lack of a positive writing style, but also the lack of positive sentiment around election topics.We hypothesize that election topics rarely tend to involve any positive publicity, or positive events. We've all heard about how campaign time tends to be a match of mudslinging with politicians taking potshots at each other. The media tends to sensationalize the bad and under-report "good" news. The above data, although taken from a small sample of news articles, reflects this anecdotal evidence.

Election Outcome Prediction Experiment

We ran a small experiment to see if we could predict the outcome of an election based on the popularity of political entities in news article abstracts. The IND Dataset contains information about the number of reshares an article published by the news portal "India Times" received. We observed all the entities detected by our named entity recognition model and labeled them as either "1" representing a pro-BJP node or "-1" representing a pro-Congress node.

We made a few assumptions to run this experiment:

4. The number of reshares an article received reflects the minimum number of people who must have read it.
5. We assume that the sets of people viewing the article for different nodes do not intersect, i.e. the articles reach distinct sets of people.
6. Viewers that read an article which has a positive/neutral sentiment about an entity, will lean towards voting for that entity.
7. On the other hand, viewers that read an article which has a negative sentiment about an entity, will lean against voting for that entity.

Our aim is to explore whether we are able to predict the outcome of the election simply based on visibility in news media. We aggregate the viewership statistics of pro-BJP nodes and those of pro-Congress nodes and see which are higher.

Based on assumptions 3 and 4, we follow this formula for counting the number of voters:

Voters = (1- Negative Sentiment Weight)*Viewer Count

This table shows the entities, along with their viewership and predicted voting statistics:

Entity Name	Annotation	Total Viewers	Negative Sentiment	Predicted Voters
BJP	+1	22099	0.216	17326
Smriti Irani	+1	560	0.65	196
Congress	-1	48218	0.398	29027
Priyanka Gandhi Vadra	-1	16960	0.466	9056
National Democratic Alliance	+1	8700	0.01	8613
Shiv Sena	+1	9550	0.07	8881
Uddhav Thackeray	+1	9550	0.07	8881
Bharatiya Janta Party	+1	8700	0.01	8613
Priyanka Gandhi	-1	756	0.17	627
Rahul Gandhi	-1	17200	0.4	10,320
Kanhaiya Kumar	-1	694	0.04	666
Narendra Modi	+1	479	0.28	345
Mamta Banerjee	-1	278	0.01	275
Punjab Harinder Singh Khalsa	+1	474	0.08	436
DMK	-1	468	0.4	280

Pro-BJP : 60,112 viewers, 53,291 votes

Pro-Congress : 84,574 viewers, 50,251 votes

Although the pro-Congress entities had a higher number of viewers/views, they also had an overall higher negative sentiment associated with them which is why the pro-BJP entities had a higher number of votes.

Results: The BJP would be the winning political party according to this data.

The results of the above experiment are in line with reality, since the BJP did win the 2019 Indian General Elections. However, the margin of their victory is not accurately reflected in our experiment since the real elections did not have such a close margin.

Our assumptions in this experiment would not hold in reality, because votes are decided on many factors other than news article coverage. Hence, we cannot draw overarching conclusions about the direct relationship between media viewership and a political win. However, this experiment does provide evidence to the fact that political parties can not only win elections by positive publicity about their party and leaders, but also with negative publicity about the opposing party and its leaders. In class we discussed finding the influential sets of nodes to observe cascading behavior in information networks. We concluded that this is a computationally hard problem, but we can try to identify broad sub-classes of the models for which good approximation results can be obtained. Our analysis of entity extraction, aggregation of their sentiments and total reshare counts is an attempt to identify and approximate the most influential names/organizations associated with the Indian Elections.

Future Work

We would like to explore more datasets that give viewership statistics for news articles and news videos. It would be interesting to observe how opinions can be shaped through different media platforms. We would also like to improve our implementation of the sentiment score model to separate the sentiment of different entities found within the same article. For example, in the sentence, “BJP defeated Congress in the 2014 elections”, we want to be able to identify the sentiment for “BJP” to be a positive one, while that of the “Congress” to be negative. This will help us study the sentiment of political parties and leaders more accurately. If we are able to

accurately gauge this sentiment, assuming that the news media sentiment also reflects the public sentiment, we want to try and predict who the winner of the next Indian election will be.
We also want to do similarity detection of news articles across media platforms, to see the order in which different news portals cover a particular news. It would be interesting to analyze the viewership of such an information spread, and possibly detect a cascade for news that goes "viral".

Conclusion

Political opinions are heavily shaped by news media portals which publish news in print as well as electronic media. Positive publicity of a favorable party coupled with negative publicity of an opposing party tends to work in the favor of the party with positive publicity. We provided empirical evidence to indicate the same. It is possible to leverage the power of natural language processing and network analysis to detect the most influential entities in an election. News portals tend to adopt a negative reporting style to sensationalize news and garner more views.

Bibliography

[1] Barclay, F. P., Venkat, A., & Pichandy, C. (2015, May 15). *India elections 2014: Time-lagged correlation between media bias and Facebook trend: Global Journal of Human-Social Science*. India Elections 2014: Time-Lagged Correlation between Media Bias and Facebook Trend | Global Journal of Human-Social Science. Retrieved March 27, 2023, from https://socialscienceresearch.org/index.php/GJHSS/article/view/1336/5-India-Elections-2014-Time-Lagged_JATS_NLM_xml#figures

[2] Easley, D., & Kleinberg, J. (n.d.). Networks, crowds, and markets: A book by David Easley and Jon Kleinberg. Retrieved March 29, 2023, from http://www.cs.cornell.edu/home/kleinber/networks-book/

[3] Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 101–110. https://doi.org/10.1609/icwsm.v8i1.14559

[4] Hart, William, et al. "Feeling Validated versus Being Correct: A Meta-Analysis of Selective Exposure to Information." *Psychological Bulletin*, vol. 135, no. 4, 2009, pp. 555–588., https://doi.org/10.1037/a0015701.

[5] Lee, Y. J. (2015, September). *Do I follow my friends or the crowd? information cascades in online ...* Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. Retrieved March 27, 2023, from https://repository.upenn.edu/cgi/viewcontent.cgi?article=1324&context=marketing_papers

[6] Nemanja Spasojevic Lithium Technologies | Klout, Spasojevic, N., Klout, L. T. |, Zhisheng Li Lithium Technologies | Klout, Li, Z., Adithya Rao Lithium Technologies | Klout, Rao, A., Prantik Bhattacharyya Lithium Technologies | Klout, Bhattacharyya, P., Technology, U. of, University, C., University, M., Research, B., Office, A. T., & Metrics, O. M. V. A. (2015, August 1). *When-to-post on social networks: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and data mining*. ACM Conferences. Retrieved March 29, 2023, from https://dl.acm.org/doi/abs/10.1145/2783258.2788584

[7] Tasgin, Mursel, and Haluk O. Bingol. "Gossip on Weighted Networks." *Advances in Complex Systems*, vol. 15, 5 May 2012, https://doi.org/10.1142/s0219525912500610.

[8] Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research, 34*(4), 441–458. https://doi.org/10.1086/518527