Impact of Social Media on Democratic Decision Making

Eliott Chapuis, Arjun Chintapalli, Joy Kimmel, Gowri Nayar, M.S. Suraj, Jia Yi Yan
Team 16

Abstract/Introduction

We analyzed Twitter and Reddit data through word frequency and natural language-processing (NLP) techniques in order to determine voter sentiment and predict results for both Brexit and the India Elections 2014.

The failure of polling data to predict results has been a rising, relevant problem in society. Specifically in regards to the Indian election of 2014 and the British EU referendum, "Brexit", in 2016, the election results did not correspond to the portrayal of the situation by the media and the polling data. We have sought to address this problem through the use of natural language processing on social media to predict election results and track public opinion.

Approaches/Innovations

- A direct comparison between social media data from Twitter and Reddit, and validate against traditional polling data.
- An examination of the predictive power of social media across multiple cultures to prove validity and universality of models.
- NLP analysis through sentiment indexing as well as political categorization

Data

Data Collection

Datasets: Brexit Twitter, Brexit Reddit, India Election Reddit, India Election Twitter

Reddit: Used the PRAW API to retrieve Reddit title, comments, karma, and upvotes

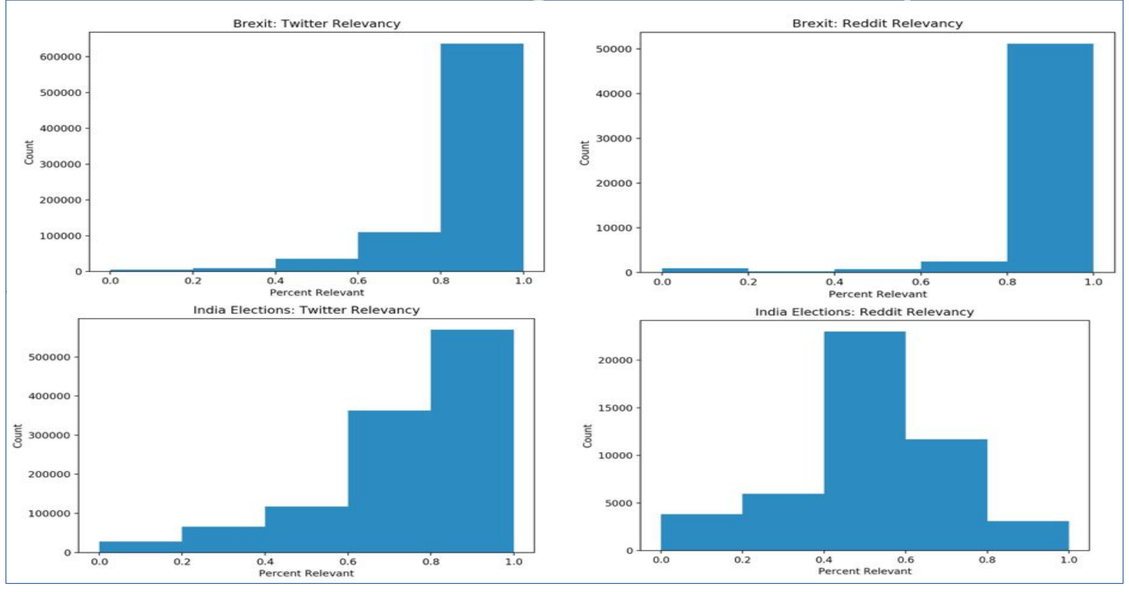
Twitter: Used Get Old Tweets API to get tweets, #tags, and # of retweets for ~800k tweets. We used this API in order to bypass twitter restriction of only returning week old tweets as well as rate limit.

ABP News, India Times, BBC: Poll data on Indian Elections and Brexit for validation

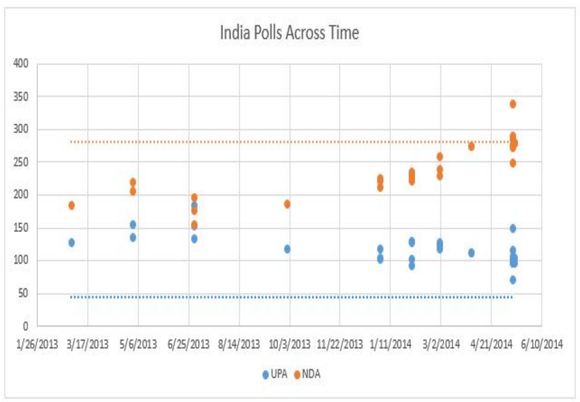
Processing, Clustering and Filtering

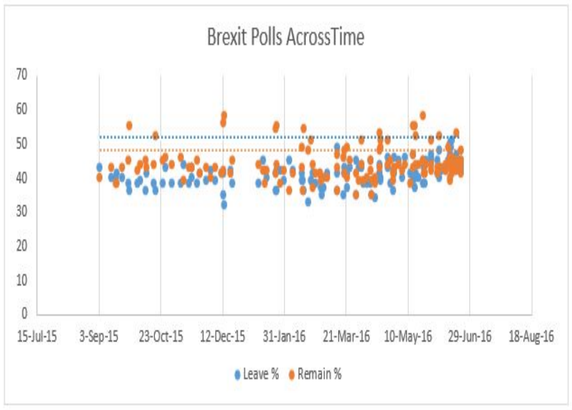
- **Processing:** We used the Word2Vector API to cluster the words through k-means within the data around common themes. Then we manually went through the clusters to determine the relevancy of each cluster. Using these clusters, we calculate the percentage each text originates from a given cluster and removes the tweets/comments with less than 50% relevancy. Then, only the top 10% of relevant tweets/comments per day are used to feed into the NLP model.
- **Relevancy**: analyzed the percentage of words belonging to relevant clusters per tweet/Reddit comment, and observed that Indian posts in general were less relevant to the topic than Brexit posts

Pre-Processing Post Relevancy



Historical Polling Data

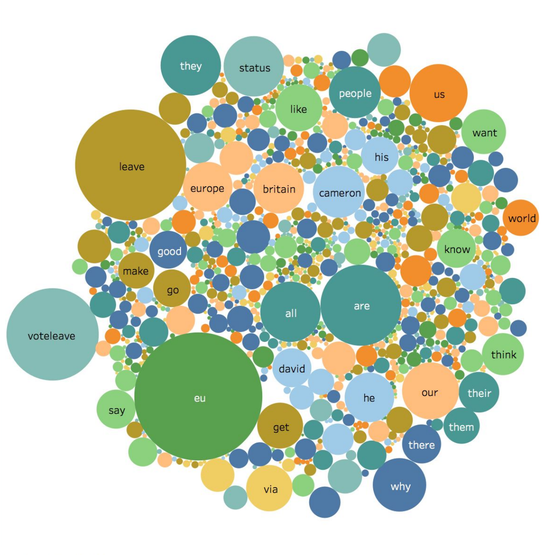




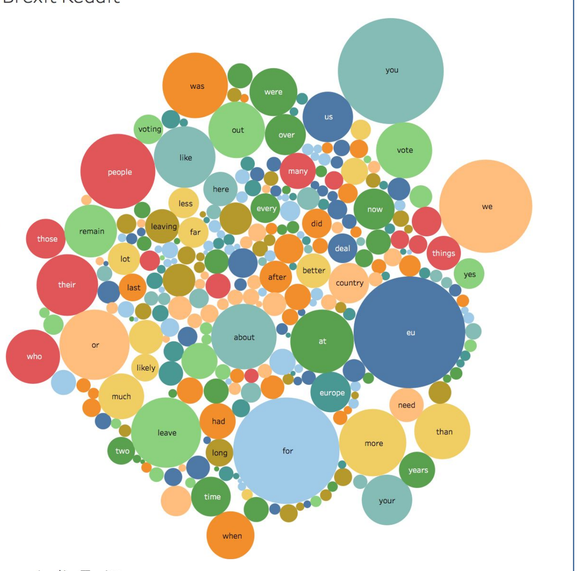
Experiments/Results

- **Word clouds**: visualized most frequent words of the top ten clusters in each dataset

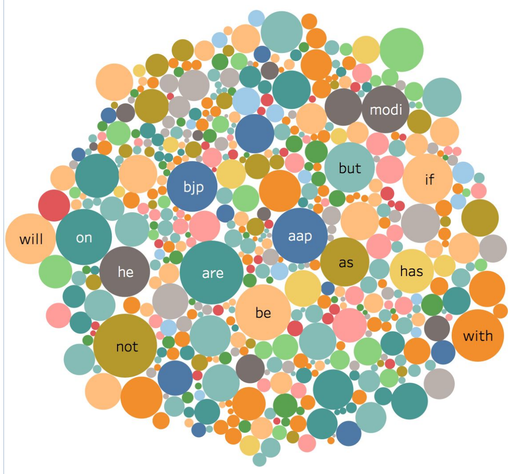
Brexit Twitter



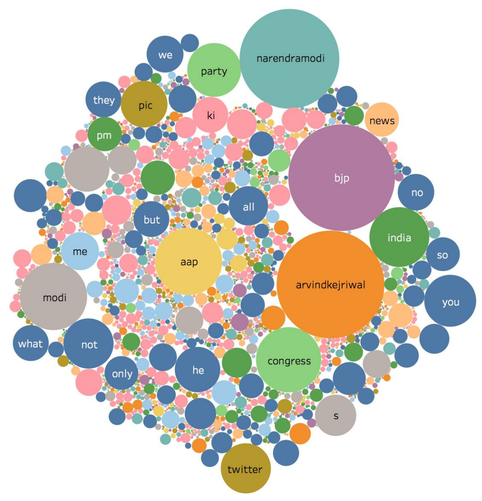
Brexit Reddit



India Reddit

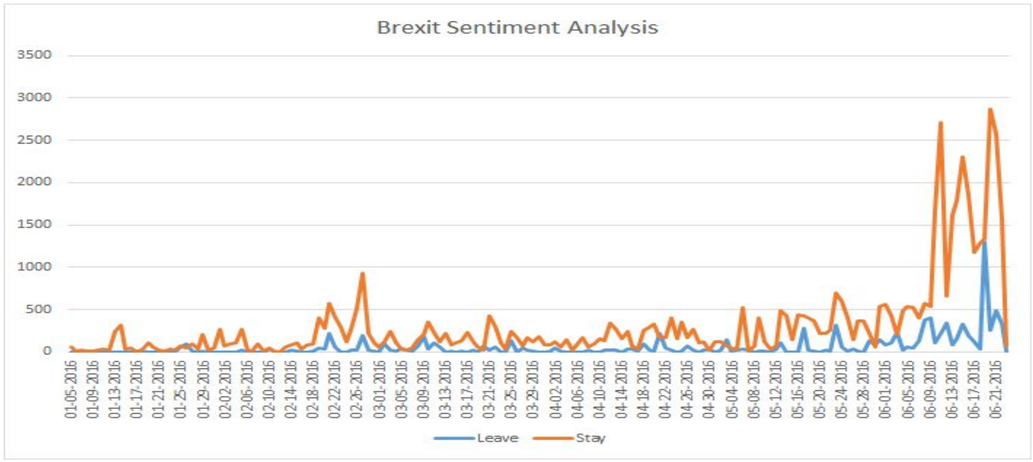


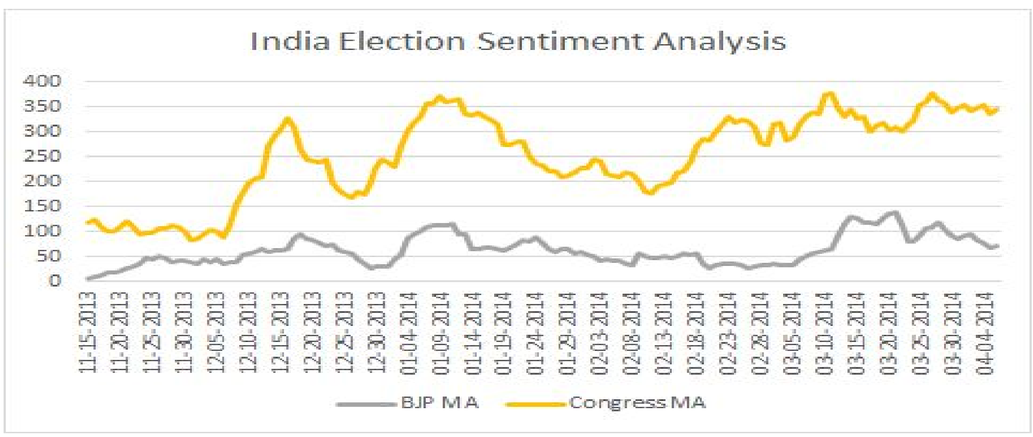
India Twitter



The graphs above show the words proportionate in size to their count and colored by cluster

- **NLP:** Sentiment Analysis on Twitter Data and Political Analysis on Reddit Data
  - Categorized Reddit data into conservative and liberal over time for India and Brexit:
    - Where conservative relates to BJP party and Brexit, liberal to Congress and Stay





Conclusion

As shown above, our sentiment and political analysis proves to be an alternate pathway to better analyze and poll people's political opinions. Current results show an overwhelming younger and liberal bias in social media, thereby demonstrating the same pitfalls as polling data.

Further work is needed to get more representative datasets that include older and more conservative datasets. Examples of such datasets could include Facebook and Linkedin. We could also alleviate this bias by using a weighting factor to shrink the weight of liberal groups and increase the weightage of conservative groups to be in sync with the wider population.

References

1. Ahmed, Saifuddin, Kokil Jaidka, and Jaeho Cho. "The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties."
2. Anstead, Nick, and Ben O'Loughlin. "Social media analysis and public opinion: The 2010 UK general election."
3. Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media."
4. Chadha, Kalyani, and Pallavi Guha. "The Bharatiya Janata Party's Online Campaign and Citizen Involvement in India's 2014 Election."
5. Conover, M. D.; Ratkiewicz, J.; et al. "Political Polarization on Twitter," Center for Complex Networks and Systems Research, 2011.
6. Del Vicario, Michela et al. "The Anatomy of Brexit Debate on Facebook." arXiv:1610.06809 [cs] (2016): n. pag. arXiv.org. Web. 6 Mar. 2017.
7. Fabio Celli, Evgeny A. Stepanov, Massimo Poesio, Giuseppe Riccardi. "Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters". Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 110–118, Osaka, Japan, December 12 2016.
8. Grčar M, Cherepnalkoski D, Mozetič I. "The Hirsch index for Twitter: Influential proponents and opponents of Brexit". In: Proc. 5th Intl. Workshop on Complex Networks and their Applications. Studies in Computational Intelligence. Springer; 2016.
9. Himelboim, Itai; McCreery, Stephen; Smith, Marc. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." Journal of Computer-Mediated Communication, January 2013, Vol. 18, No. 2, 40-60.
10. Howard, Philip N., and Bence Kollanyi. "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum." arXiv:1606.06356 [physics] (2016): n. pag. arXiv.org. Web. 6 Mar. 2017.