Exploration of Social Media Content for Informed Political Understandings

Haley Roberts, Monroe Farris, and Sam Pastoriza

Meet the Team

Monroe Farris

Current Role: Data Science Engineer @ MITRE

Undergrad: University of Virginia

Sam Pastoriza

Current Role: Incoming Data Scientist @ Deloitte

Undergrad: Rose-Hulman Institute of Technology

Haley Roberts

Current Role: Incoming Data Scientist @ Geico

Undergrad: Georgia Tech

Case Study: Georgia's 2022 Gubernatorial Election

The Candidates

D - Stacey Abrams

R - Brian Kemp (Incumbent)

High-Level Goals

- Understand **how social media data manifests into election results**
- Determine if it's possible **to make predictions of election outcomes** based on social media data

Reviewing the Technical Space

The power of prediction with social media

Harald Schoen
University of Bamberg, Bamberg, Germany
Daniel Gayo-Avello
University of Oviedo, Oviedo, Spain
Panagiotis Takis Metaxas
Wellesley College, Wellesley, Massachusetts, USA and Harvard University, Cambridge, Massachusetts, USA
Eni Mustafaraj
Wellesley College, Wellesley, Massachusetts, USA
Markus Strohmaier
Graz University of Technology, Graz, Austria, and
Peter Gloor
MIT – Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Schoen et al. [3]

Key Takeaways

- **Doubt** if social media data can be used for predictions
- If predictions are possible - **statistical models would be used** over prediction market or survey based models

Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France

Andrea Ceron, Luigi Curini, Stefano M Iacus
Università degli Studi di Milano, Italy

Giuseppe Porro
Università degli Studi dell'Insubria, Italy

Ceron et al. [1]

Key Takeaway

- Sentiment Analysis **with human coders** provides promising results for use of social media data for predictive modeling

A survey on the use of data and opinion mining in social media to political electoral outcomes prediction

Jéssica S. Santos[1] · Flavia Bernardini[1] · Aline Paes[1]

Santos et al. [2]

Key Takeaways

Issues with using social media data for predictive insights:

- Labeling data reliably during short period of electoral campaigns
- **Absence of a robust methodology** to collect and analyze data
- Lack of labeled datasets
- **Little understanding of evaluation** of results

Methodology

Goal: Investigate the ability to utilize Natural Language Processing (NLP) to…

a) gain insights into political elections

b) make accurate and informed political predictions

Approach: 3-Pronged Investigative and Exploratory Analysis

1

2

3

Methodology

Goal: Investigate the ability to utilize Natural Language Processing (NLP) to…

a) gain insights into political elections

b) make accurate and informed political predictions

Approach: 3-Pronged Investigative and Exploratory Analysis

1 **Understand the two candidates**

- Exploratory Data Analysis
- Volumetric Analysis

2

3

Methodology

Goal: Investigate the ability to utilize Natural Language Processing (NLP) to…

a) gain insights into political elections

b) make accurate and informed political predictions

Approach: 3-Pronged Investigative and Exploratory Analysis

1 Understand the two candidates

- Exploratory Data Analysis
- Volumetric Analysis

2 Understand the tone of content for each candidate

- Sentiment
- Emotion
- Hate Speech

3

Methodology

Goal: Investigate the ability to utilize Natural Language Processing (NLP) to…

a) gain insights into political elections

b) make accurate and informed political predictions

Approach: 3-Pronged Investigative and Exploratory Analysis

1 Understand the two candidates

- Exploratory Data Analysis
- Volumetric Analysis

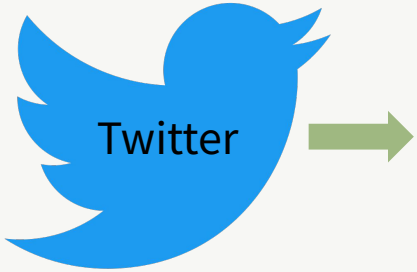
2 Understand the tone of content for each candidate

- Sentiment
- Emotion
- Hate Speech

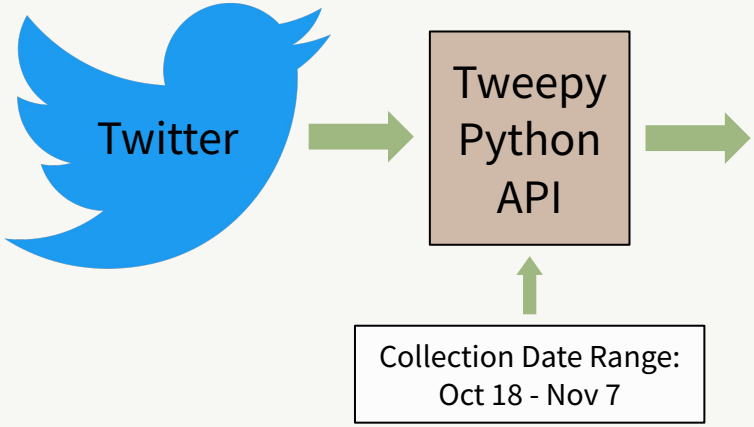
3 Understand the key social topics for each candidate

- Topic Modeling (NMF, LDA)
- Word2Vec

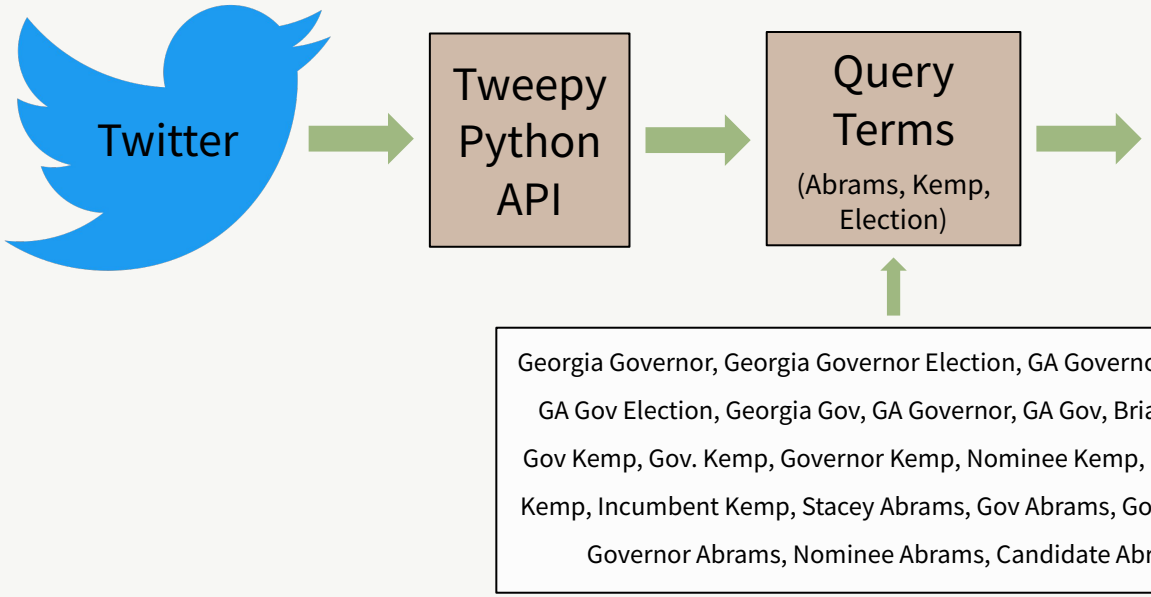
Data Collection



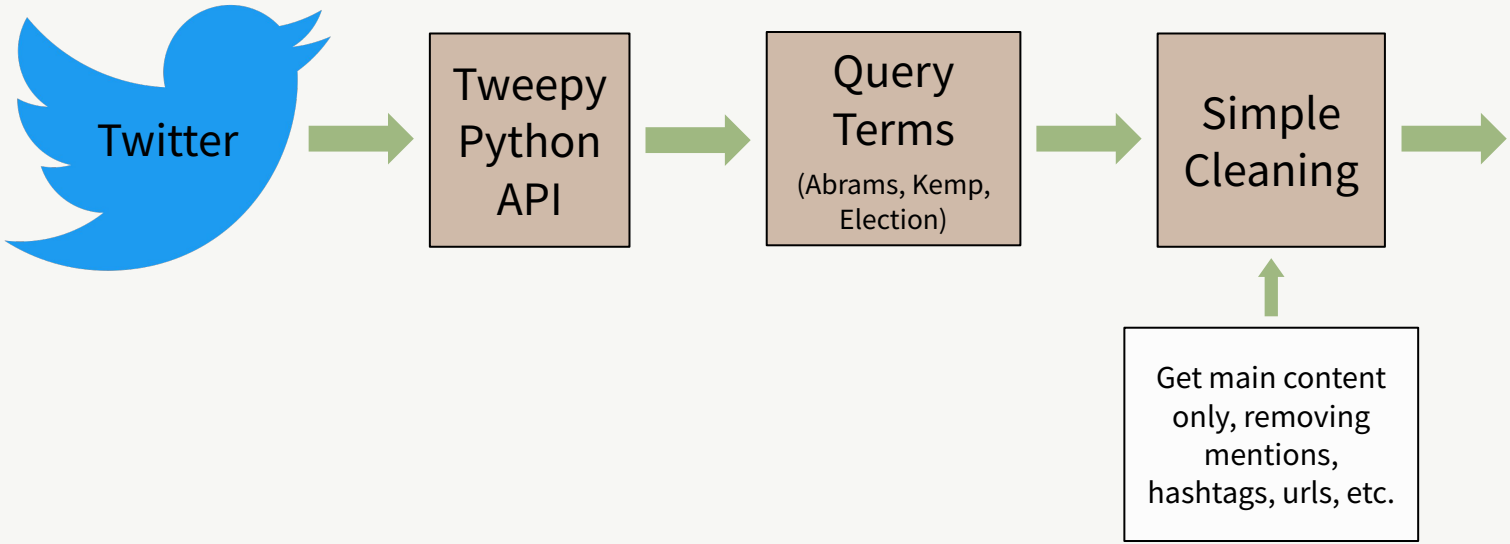
Data Collection



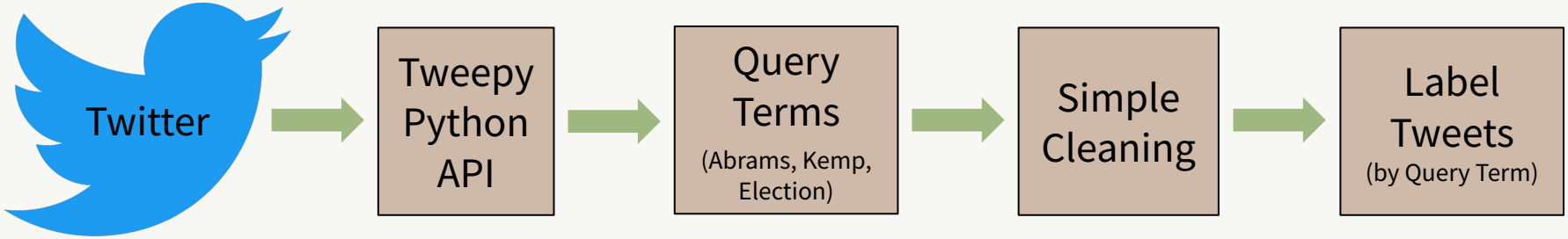
Data Collection



Data Collection



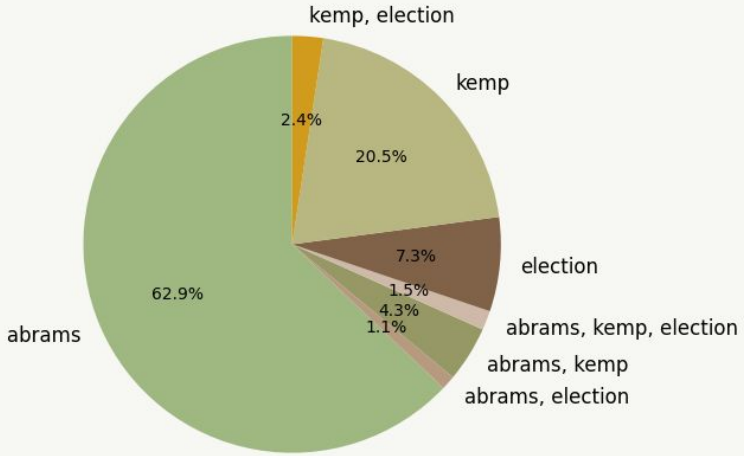
Data Collection



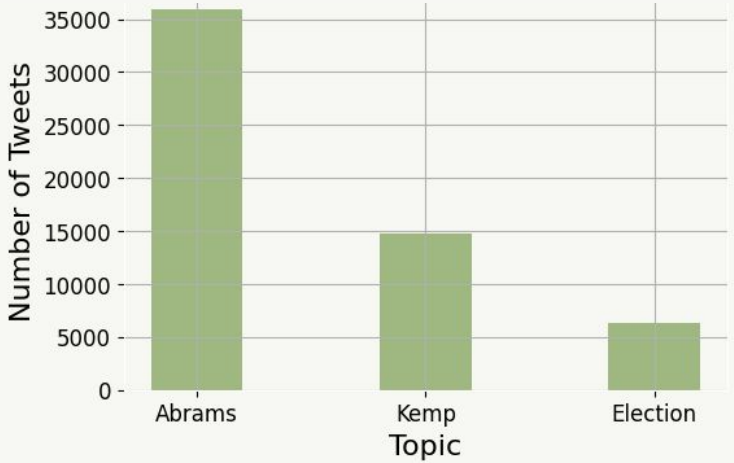
Volumetric Understanding

Goal: Understand the distribution of Tweets about each candidate and the election as a whole

Percent of Tweets of Each Combination of Topics



Number of Tweets of Each Topic

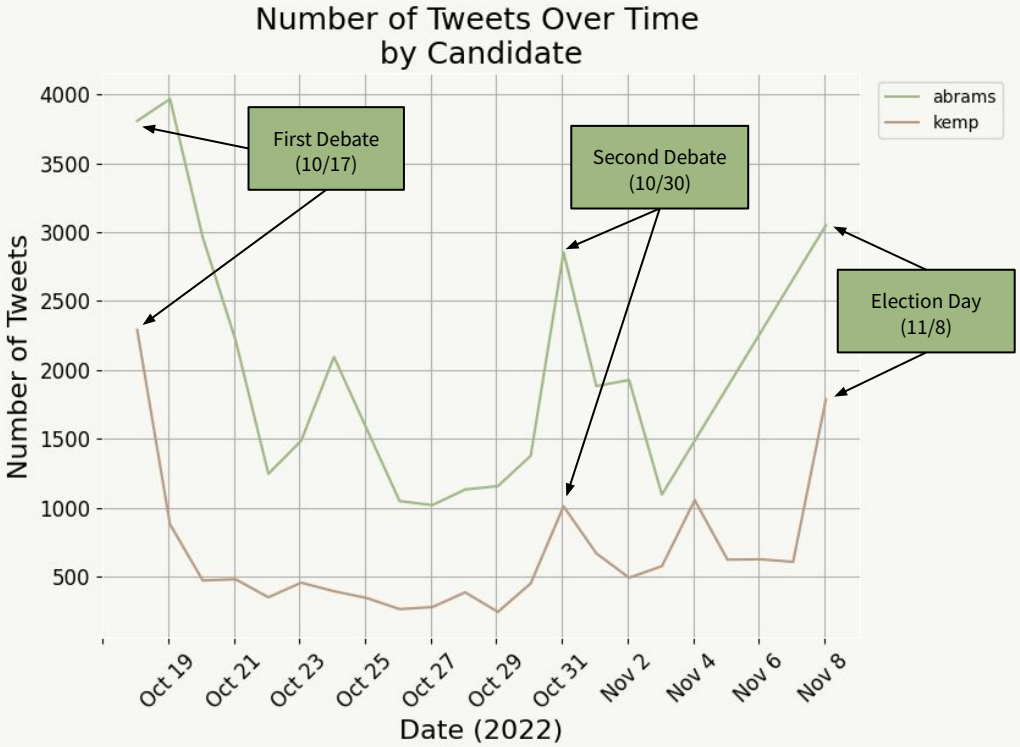


Observations:

- **Far more Tweets about Stacey Abrams** as compared to Brian Kemp
- **Tweets generally focus on a candidate** as opposed to the election in general

Volumetric Understandings (cont.)

Goal: Understand the distribution of Tweets about each candidate and the election as a whole



Observations:

- **More Tweets about Abrams** compared to Kemp
- Peaks in tweet counts for **both candidates** occur at **key points in the election:**
  - Debates
  - Election Day

Model-Based Understanding of Tones/Sentiment of Political Social Media Content

Goal: Leverage pre-trained models to understand tones and sentiments of Twitter posts about the candidates and the election

Preprocessing: Clean text of tweets

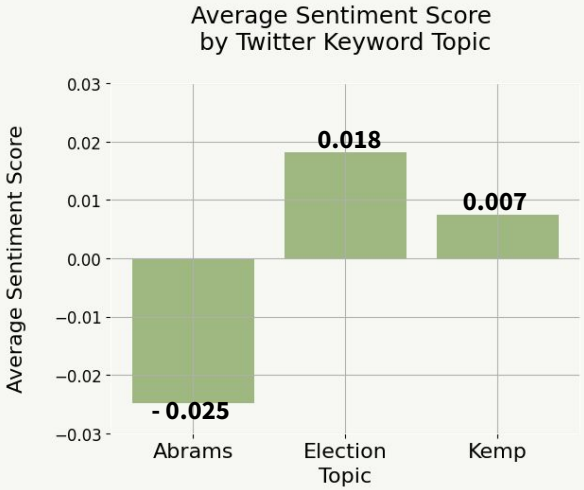
- Removal of punctuation, digits, mentions, etc.
- Removal of any tweets with no text after cleaning

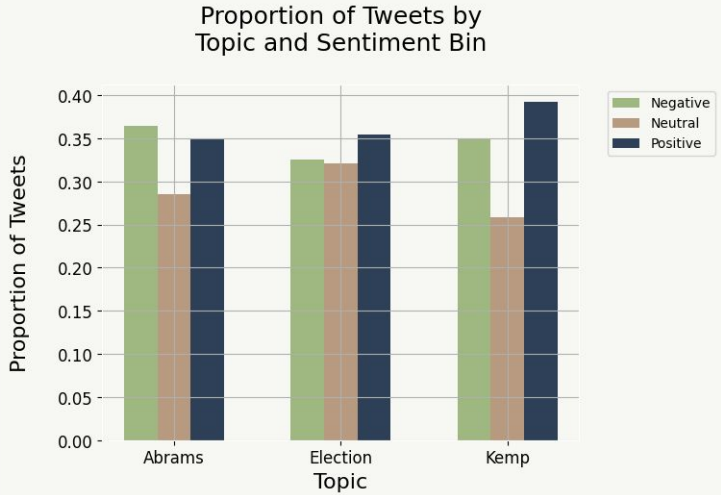
Approach: Use 3 pre-trained models to determine different tones/sentiments of each tweet

1. Sentiment Model (VADER)
2. Emotion Model (pysentimiento)
3. Hate Speech Model (pysentimiento)

Model-Based Understanding of Tones (cont.)

Approach #1: Sentiment Model (VADER)



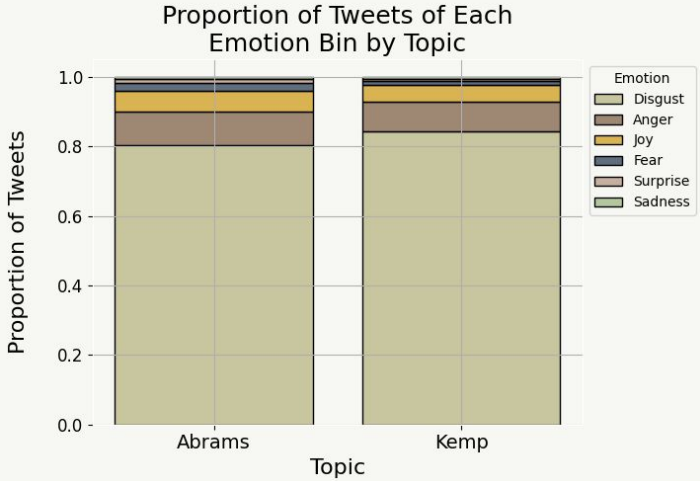


Observations:

- All average sentiment scores fall into a neutral category (-0.05 to 0.05) but **Tweets mentioning Stacey Abrams are slightly more negative** than those mentioning Brian Kemp
- **Proportions of Tweets across sentiment bins are relatively similar** between the candidates

Model-Based Understanding of Tones (cont.)

Approach #2: Emotion Model (pysentimiento)

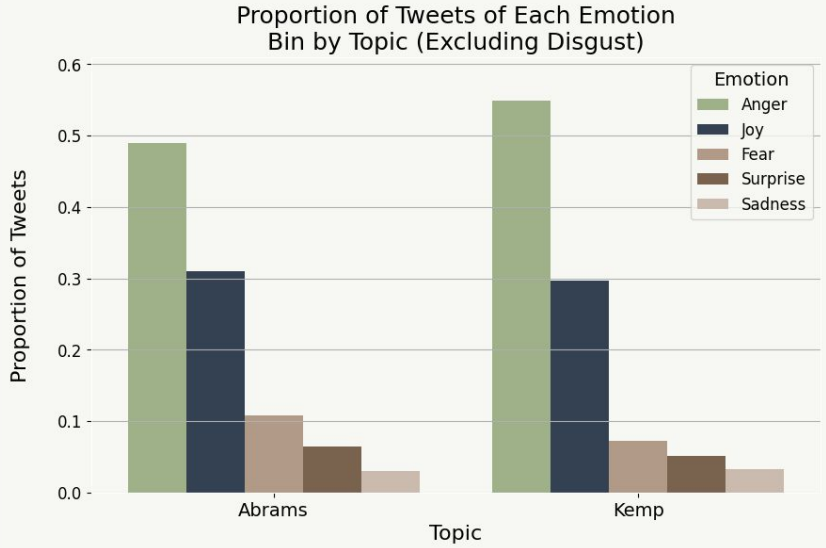


Observations:

- **Generally similar distribution of emotions** between the two candidates
- Slightly larger proportion of tweets about Kemp that have “anger” emotions

Model-Based Understanding of Tones (cont.)

Approach #2: Emotion Model (pysentimiento)

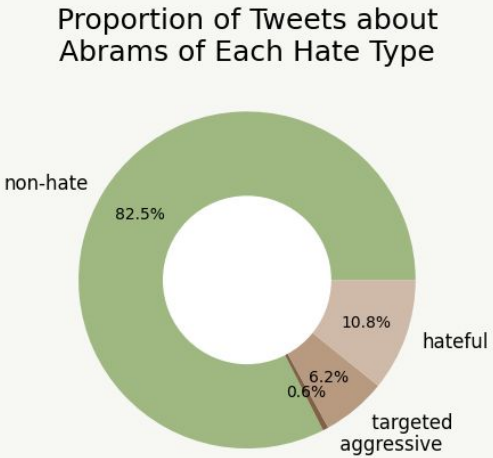


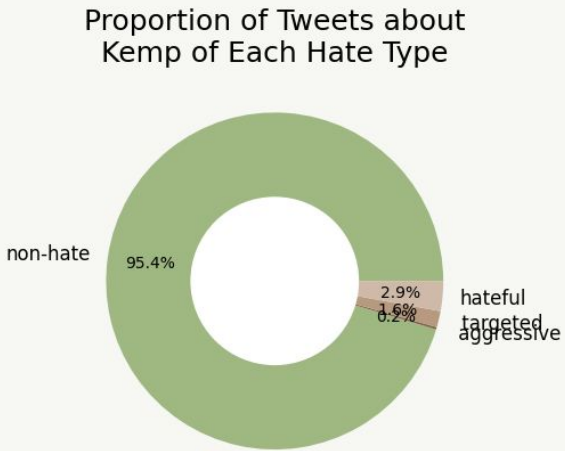
Observations:

- **Generally similar distribution of emotions** between the two candidates
- Slightly larger proportion of tweets about Kemp that have “anger” emotions

Model-Based Understanding of Tones (cont.)

Approach #3: Hate Speech Model (pysentimiento)



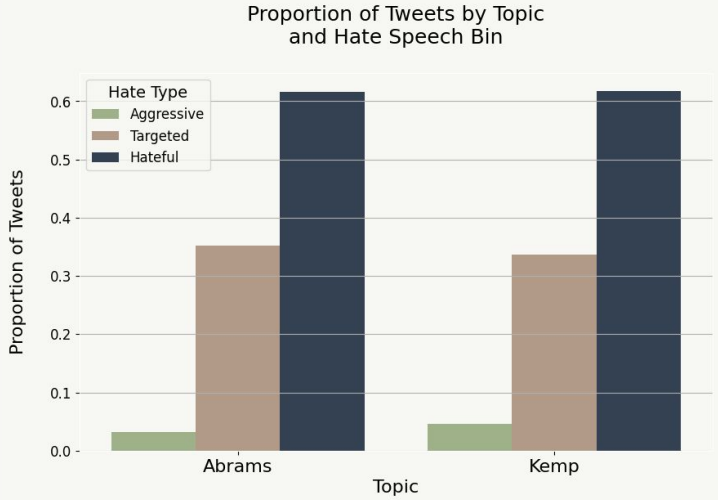


Observations:

- Overall tweets contain **mostly non-hate speech**
- **Larger proportion of tweets about Abrams** that have hate speech
- **Generally similar distribution of types of hate speech** between the two candidates

Model-Based Understanding of Tones (cont.)

Approach #3: Hate Speech Model (pysentimiento)



Observations:

- Overall tweets contain **mostly non-hate speech**
- **Larger proportion of tweets about Abrams** that have hate speech
- **Generally similar distribution of types of hate speech** between the two candidates

Topic Modeling of Political Content

Goal: Leverage NLP techniques to understand key political topics in social media posts about the candidates and the election

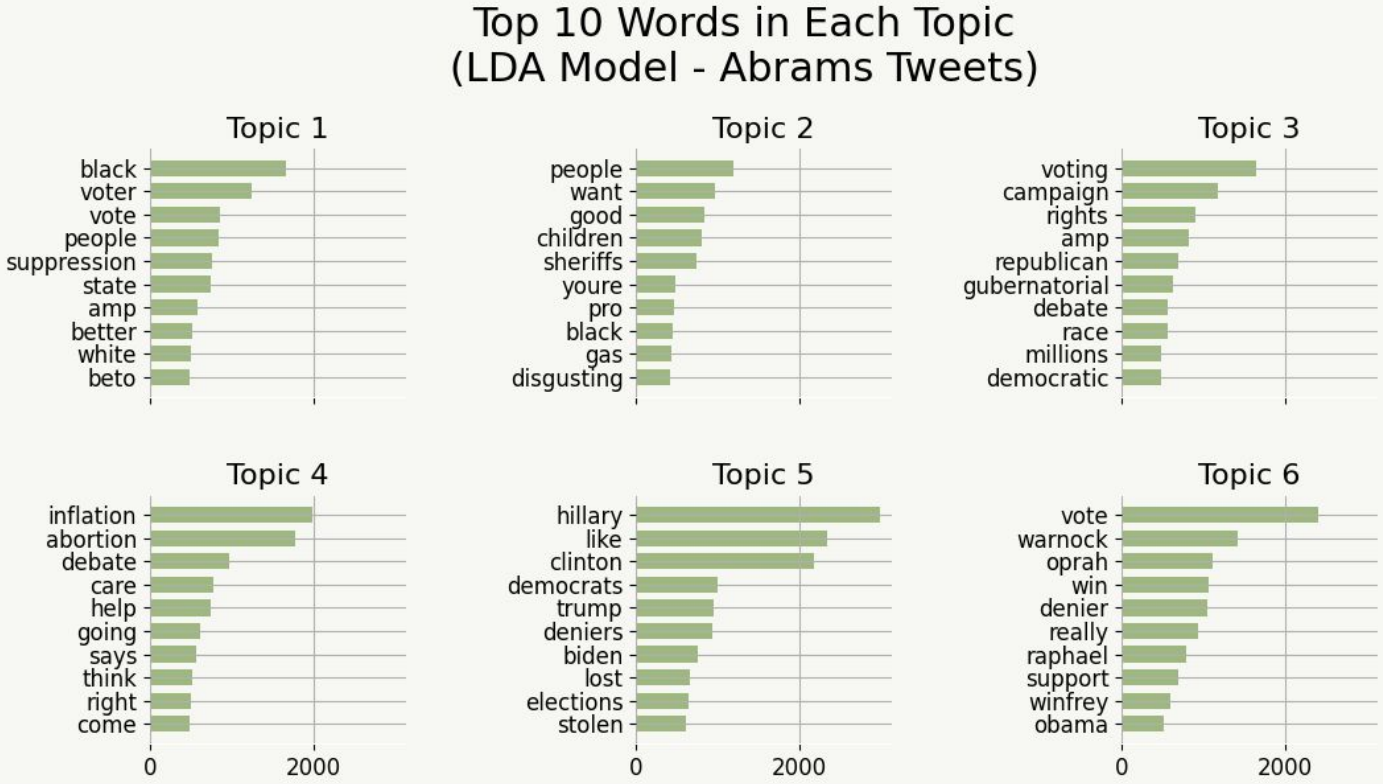
Preprocessing: Clean text of tweets

- Removal of punctuation, digits, mentions, etc. + stopwords and search words
- Removal of any tweets with no text after cleaning

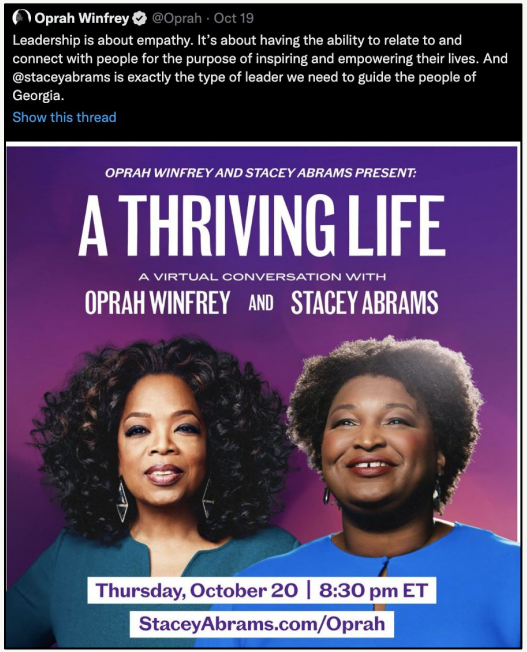
Approach: Use 3 methods to determine important topics of each candidate

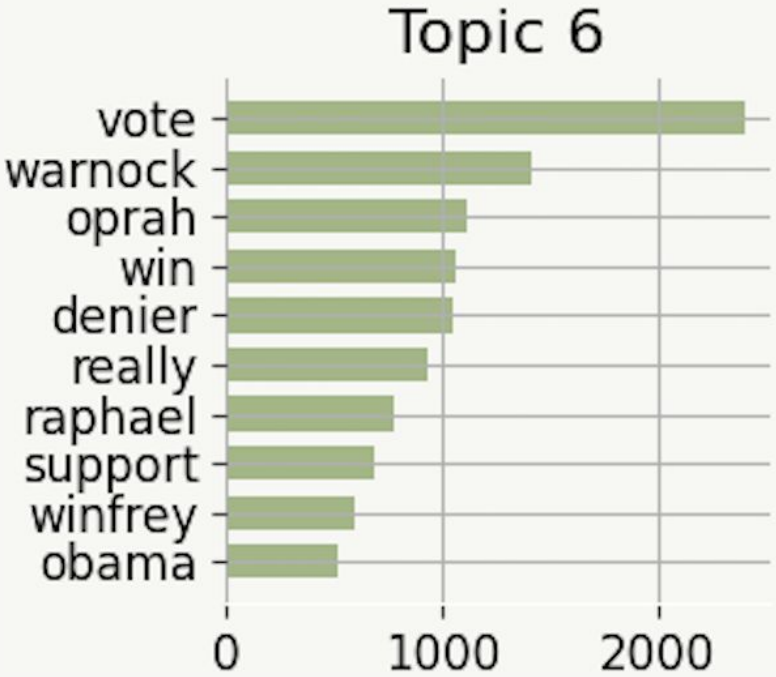
1. Linear Discriminant Analysis (LDA)
2. Non-Negative Matrix Factorization (NMF)
3. word2vec

Topic Modeling (cont.): LDA - Abrams



Topic Modeling (cont.): LDA - Abrams

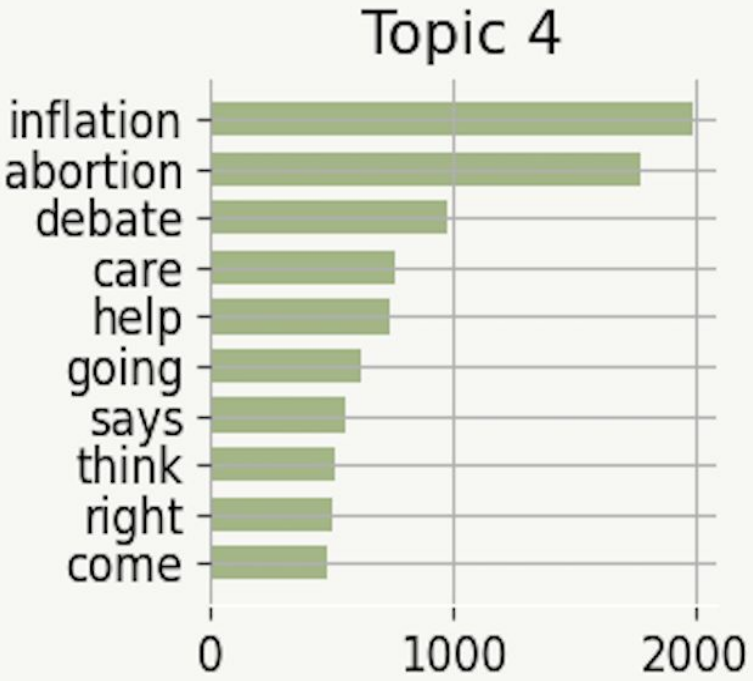




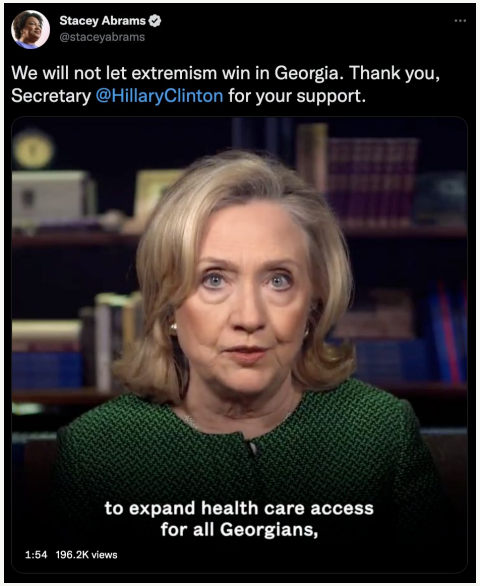
Stacey Abrams
@staceyabrams

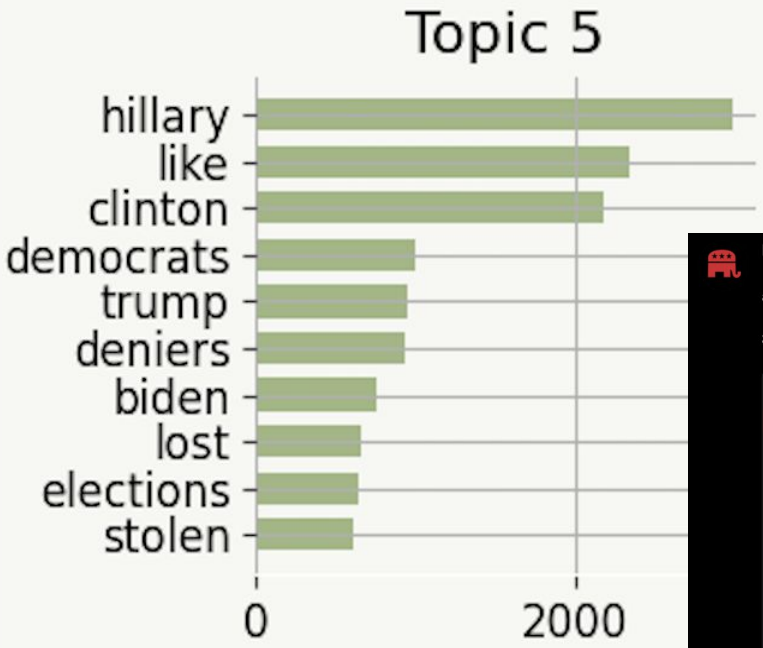
My dear friend, Sen. @ReverendWarnock is an ally in our fight for justice and has been working tirelessly on behalf of Georgians in the U.S. Senate. Join me in supporting Rev. Warnock tonight as he takes the debate stage to show Georgia what real leadership looks like.

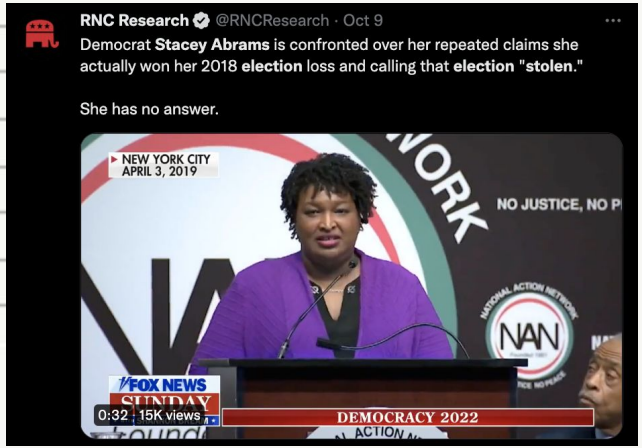
Topic Modeling (cont.): LDA - Abrams



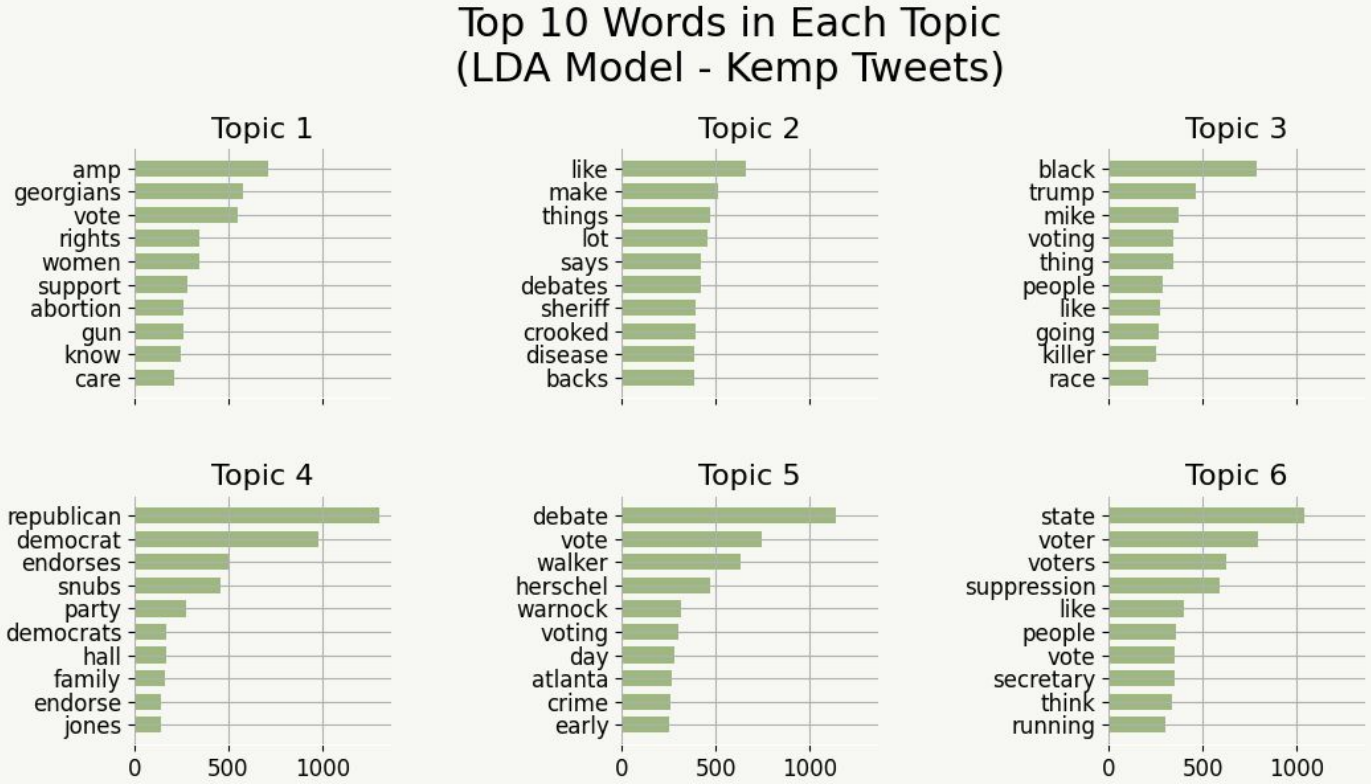
Topic Modeling (cont.): LDA - Abrams



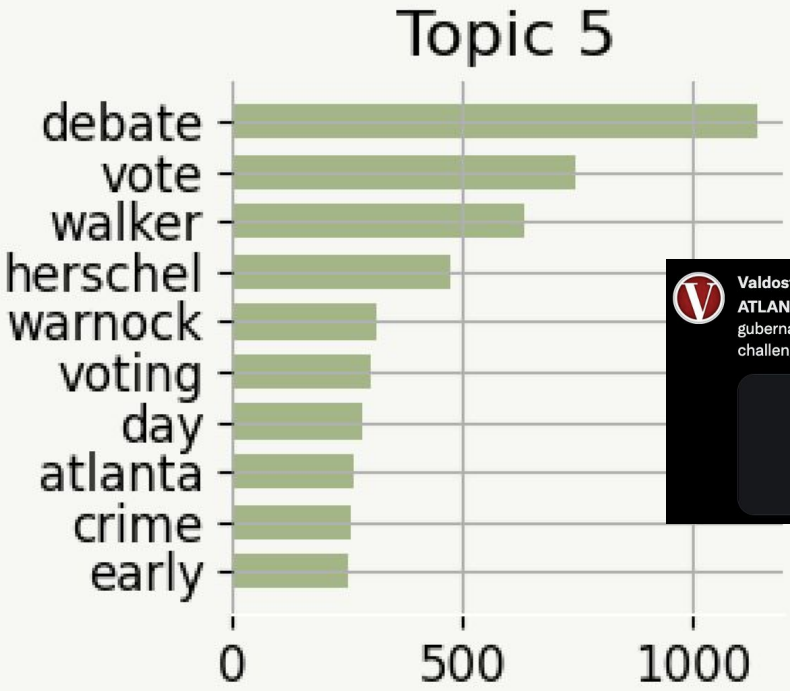




Topic Modeling (cont.): LDA - Kemp



Topic Modeling (cont.): LDA - Kemp



Valdosta Daily Times @TheVDT · Oct 20

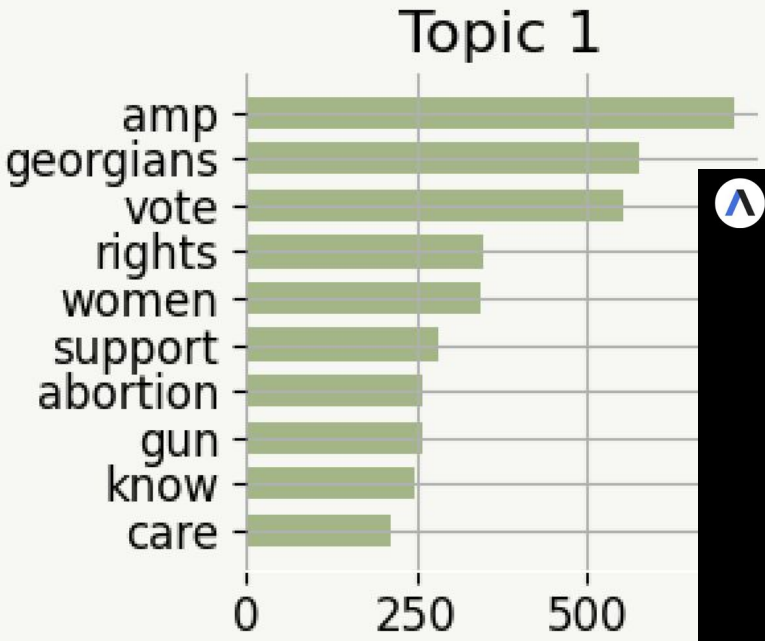
ATLANTA — Crime and education were fiery topics during Monday's gubernatorial debate, which saw a rematch between Gov. **Brian Kemp** and challenger Stacey Abram.

valdostadailytimes.com

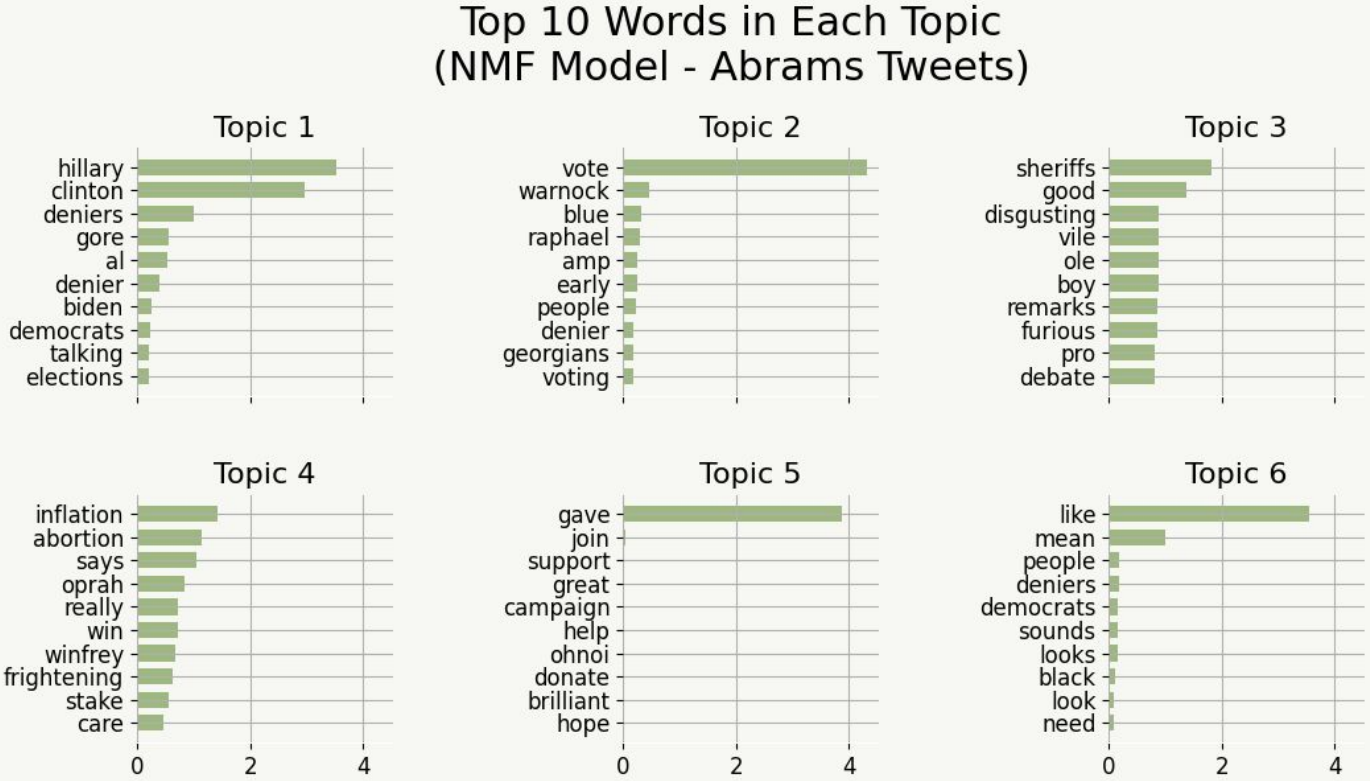
Abrams, Kemp debate crime, education

ATLANTA — Crime and education were fiery topics during Monday's gubernatorial debate, which saw ...

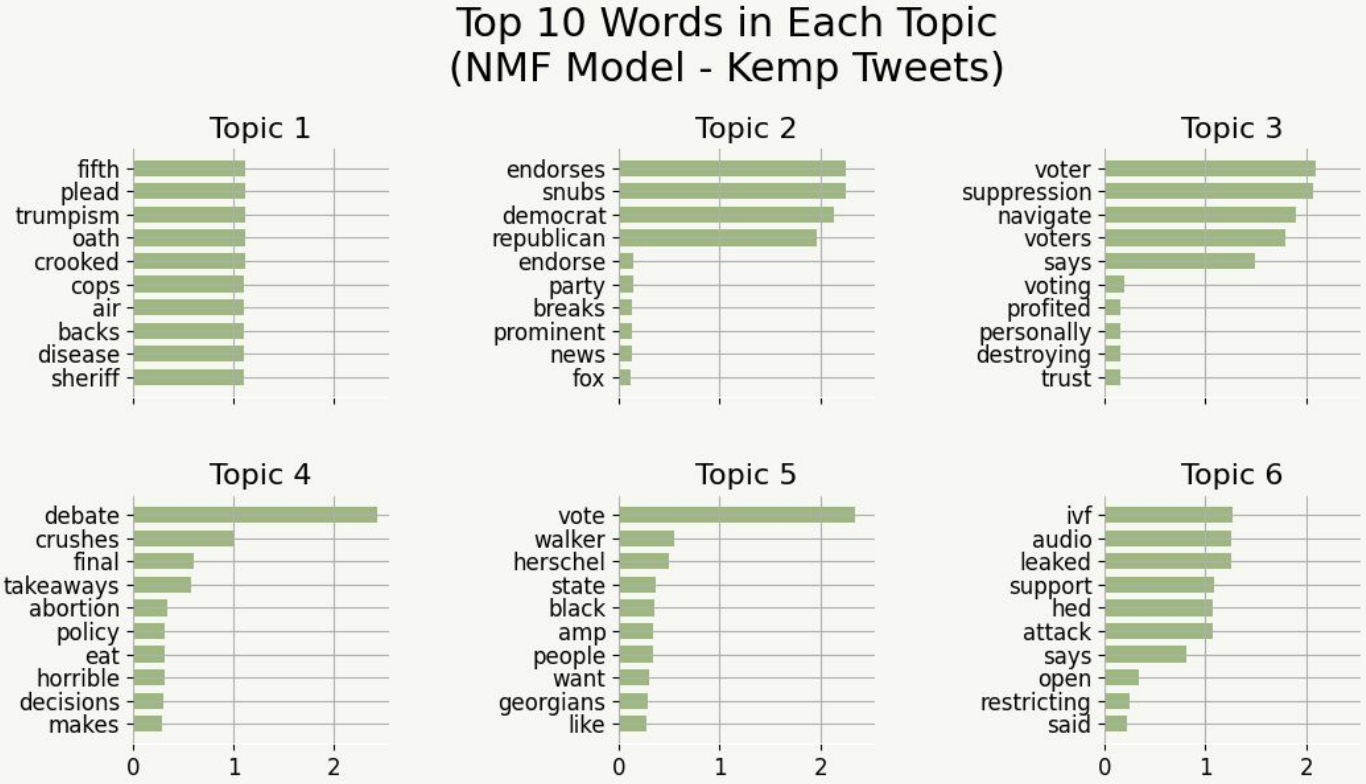
Topic Modeling (cont.): LDA - Kemp



Topic Modeling (cont.): NMF - Abrams



Topic Modeling (cont.): NMF - Kemp



Topic Modeling (cont.): Word2Vec

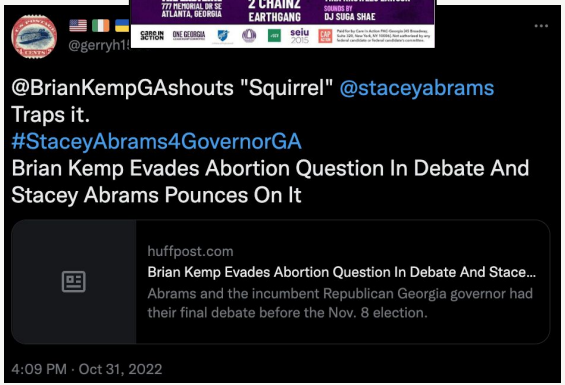
Candidate	Associated Words
Abrams	“Is Disgusting”
Abrams	“Beyonce”
Abrams	“Abortionist”
Kemp	“Most Georgia”
Kemp	“R Georgia Senate”
Kemp	“Squirrel”

Topic Modeling (cont.): Word2Vec

Candidate	Associated Words
Abrams	“Is Disgusting”
Abrams	“Beyonce”
Abrams	“Abortionist”
Kemp	“Most Georgia”
Kemp	“R Georgia Senate”
Kemp	“Squirrel”

Topic Modeling (cont.): Word2Vec

Candidate	Associated Words
Abrams	“Is Disgusting”
Abrams	“Beyonce”
Abrams	“Abortionist”
Kemp	“Most Georgia”
Kemp	“R Georgia Senate”
Kemp	“Squirrel”



Investigative Conclusions

High-Level Goals:

1. Understand **how social media data manifests into election results**
2. Determine if it's possible **to make predictions of election outcomes** based on social media data

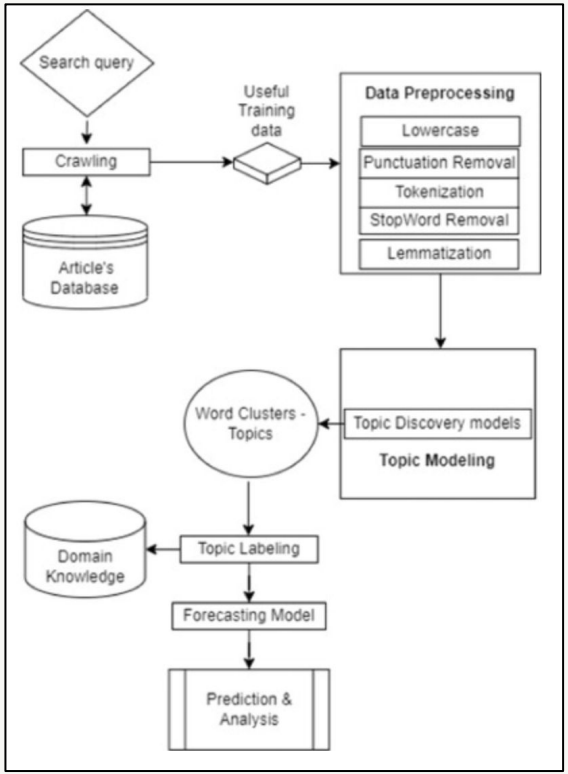
Main Takeaways: Wide range of NLP techniques with social media data related to political elections **does not provide enough signal to make predictions of political outcomes**

- Volumetric analyses showed trends of increases in number of Tweets corresponding to main events in the election such as debates and election day
- Various sentiment models focused on different types of speech but provided negligible insights
- Topic modeling was the most insightful and identified key political and social topics related to each candidate, but predictive power of these insights is lacking

Limitations

1. Amount of data and access to historical Twitter data
    - API provides access to only a 7 day time span of data
    - Would more historical data provide a more robust analysis?
        - Potentially starting data collection on the day a candidate announces their campaign (which can be almost a year in advance)
2. Negligible sentiment insights cast doubt into impact of sentiment or tone on election results
    - Reaffirms literature review - doubt in validity of this type of approach
3. Limited predictive power of NLP approaches

Future Directions



Sample LDA prediction pipeline obtained from Gupta et al.[4]

- Topic modeling was the most insightful of the three analyses
  - Future work could leverage topic modeling as a starting point for a prediction mechanism

- Create a forecasting model to better understand if one could predict the outcome of an election given a clear understanding of how people vote on key issues (gun rights, abortion, etc.)
  - Step 1: Topic Modeling
  - Step 2: Linkage of topics to voters to understand how a stance on a topic may impact how one votes
  - Step 3: Election predictions given stances on key political issues

Questions?

References

[1] Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media & society*, *16*(2), 340-358.

[2] Santos, J.S., Bernardini, F. & Paes, A. A survey on the use of data and opinion mining in social media to political electoral outcomes prediction. *Soc. Netw. Anal. Min.* 11, 103 (2021). https://doi.org/10.1007/s13278-021-00813-4

[3] Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M. and Gloor, P. (2013), "The power of prediction with social media", *Internet Research*, Vol. 23 No. 5, pp. 528-543. https://doi.org/10.1108/IntR-06-2013-0115

[4] Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of Research Trends using LDA based Topic Modeling. *Global Transitions Proceedings*.