Investigating the COVID Conversation through NLP

Patrick Norman

Questions

- What do people talk about when they talk about COVID-19?
- Do people from different regions of the US have different things to say?
- Do these regional conversations map to the severity in their area?

Potential Applications

- Targeted vaccine education
- Predicting adoption of social distancing
- Understanding cultural differences between states

Data Sources

- Twitter
- Covid Act Now API

Methods

- Twint using search queries
- NMF for topic modeling
- TextBlob for sentiment analysis
- Simple K-Means clustering

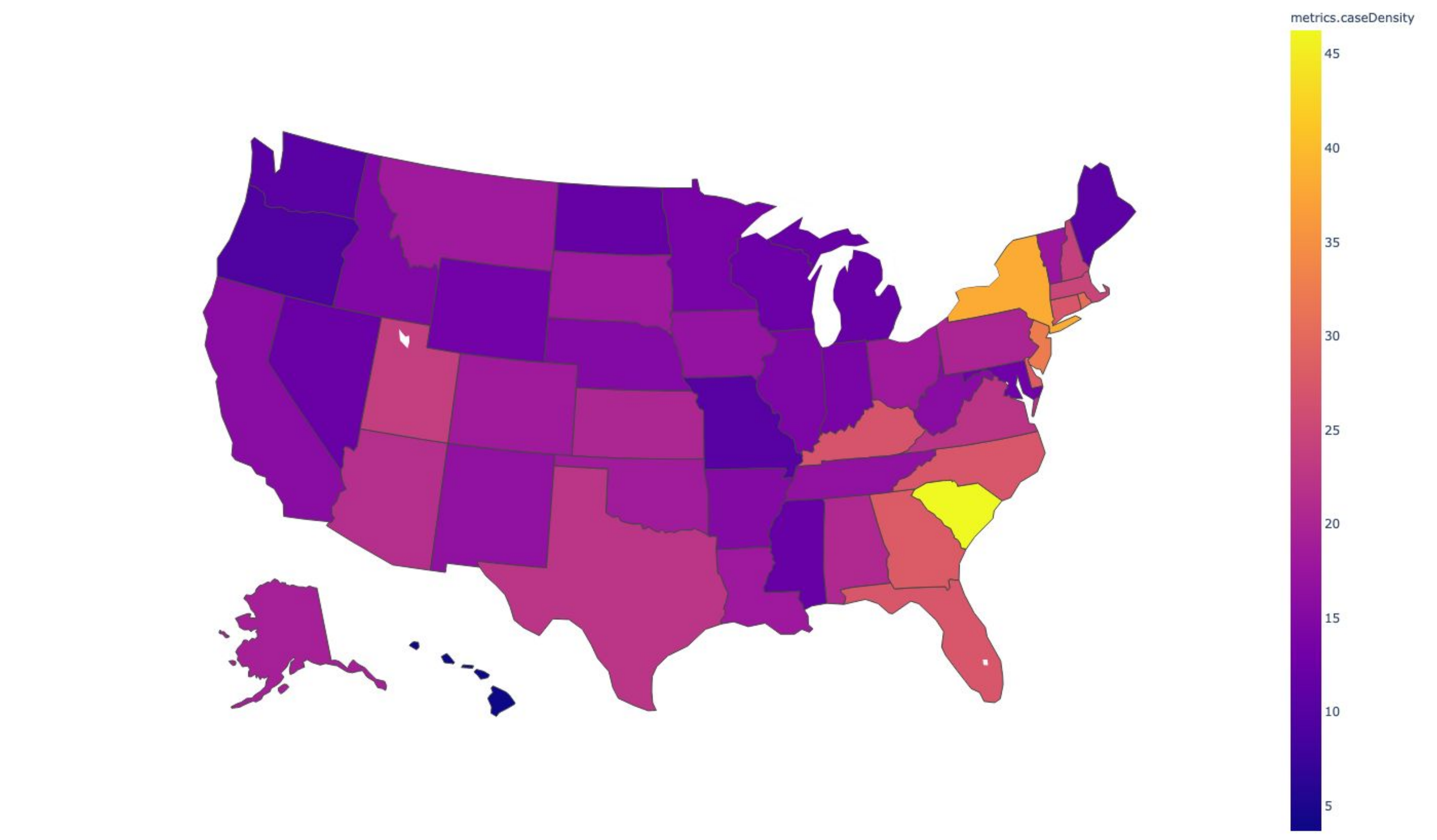
Results: Topics and Clusters

- Uniform between states
- **Politics** (fake, lawmaker, GOP, government, stimulus)
- **Case numbers** (deaths, cases, ICU, beds)
- **Testing / vaccination** (shot, appointment, center)

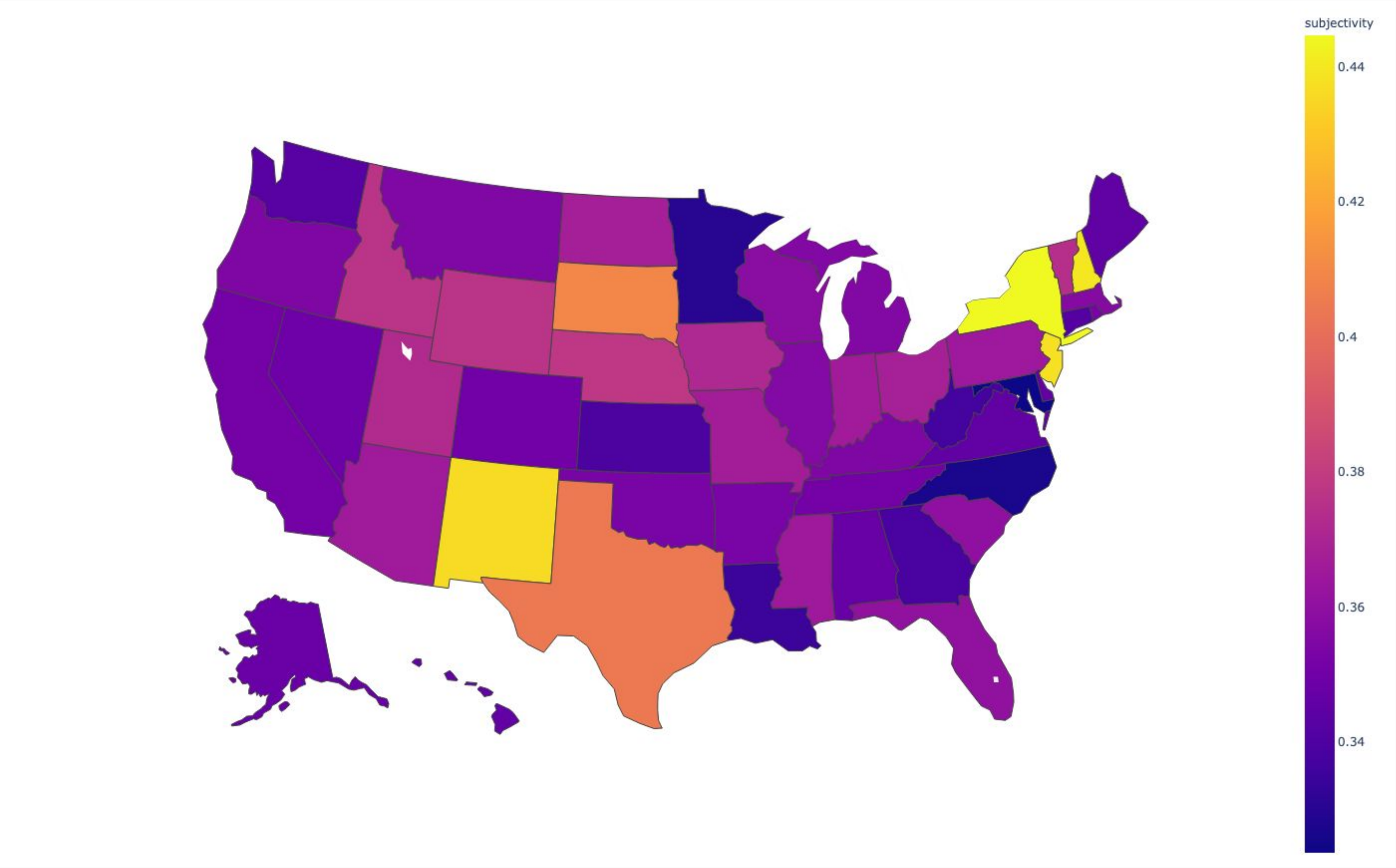
Results: Sentiment

- Polarity vs subjectivity
- Sentiment varies somewhat between states
- No correlation between sentiment and objective situation

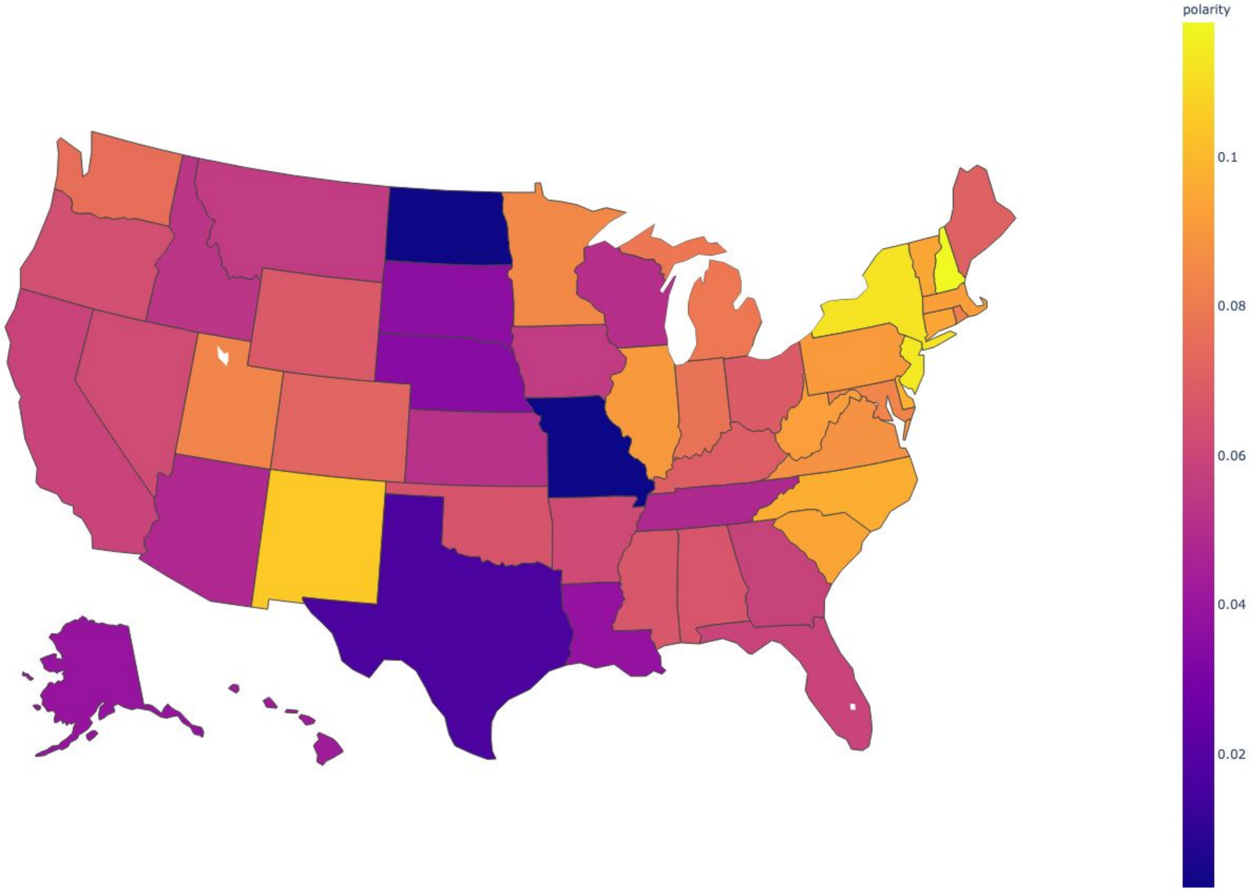
COVID case density is relatively uniform across the US



Subjectivity isn't clearly correlated with case density



Polarity isn’t clearly correlated with case density either



19% case severity variation explained

- This is ok, but not *great*.
- We need more explanatory power to apply this as a tool

Possible Improvements

- Different data sources
- Named entity recognition

Conclusions

- The COVID conversation...
  - Politics
  - Tracking the spread
  - Vaccination and treatment
- We’re more similar than we are different
- Twitter isn’t real life!

Appendix

- Correlation coeff. between tweet sentiment and case density = 9.2e-02
- Linear Regression test score with CV = 0.28

Sources