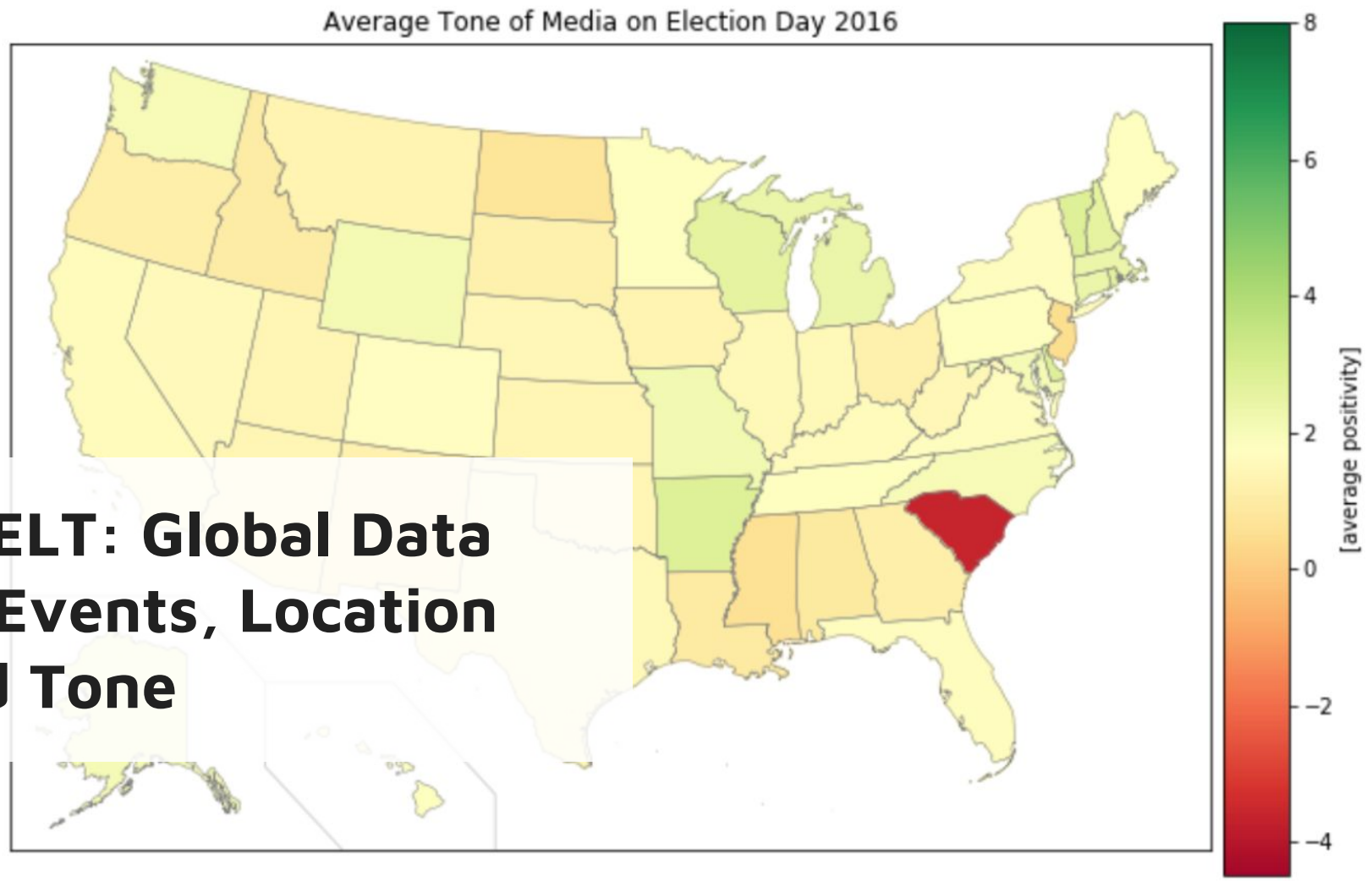
GDELT: Predicting The Tone of Media Reports

Presentation Outline

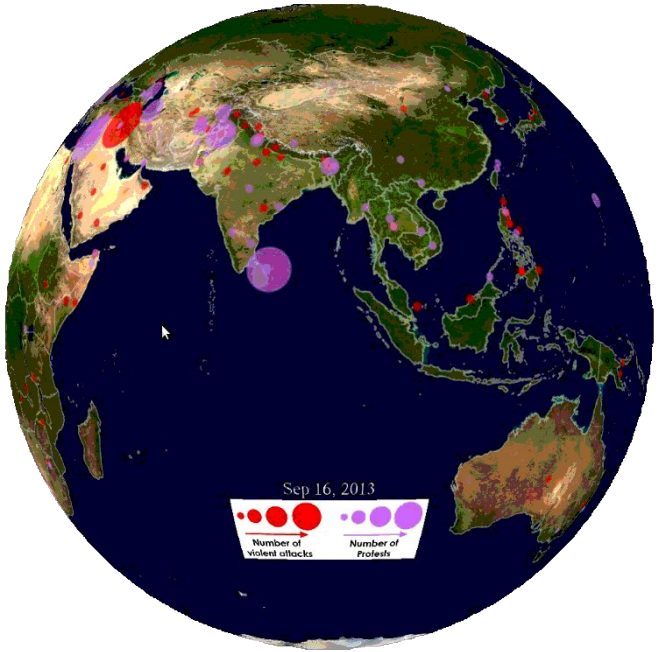
- GDELT Dataset
- Average Tone Prediction
- Modelling Approaches and Results
- Applications
- Future Work



GDELT: Global Data on Events, Location and Tone

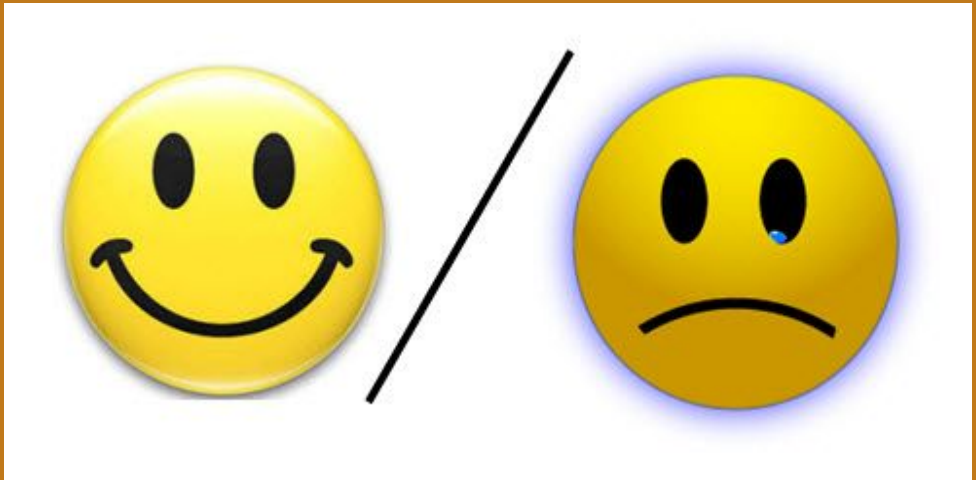
More about GDELT

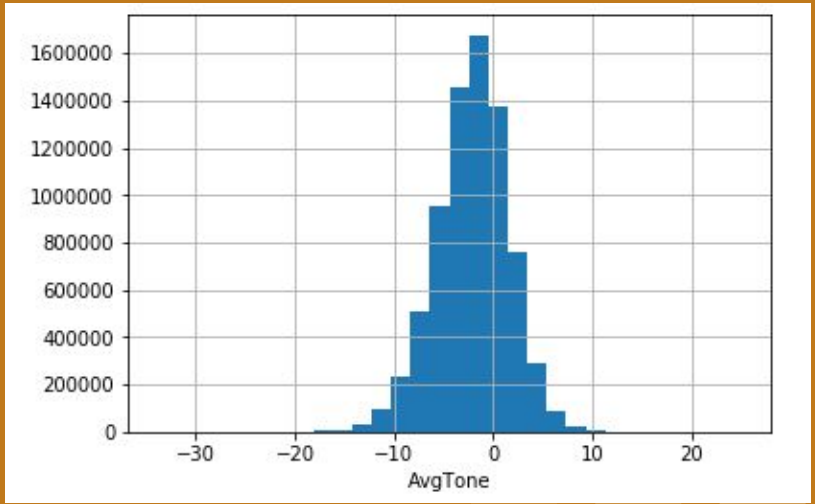
- It's **massive**! GDELT 1.0 is the "small" one and is over 30GB.
- Collects information on events from media reports around the world.
- Uses the CAMEO code system:
    - Stores events as one actor performing a particular verb on another one.



https://www.gdeltproject.org/

Predicting Average Tone





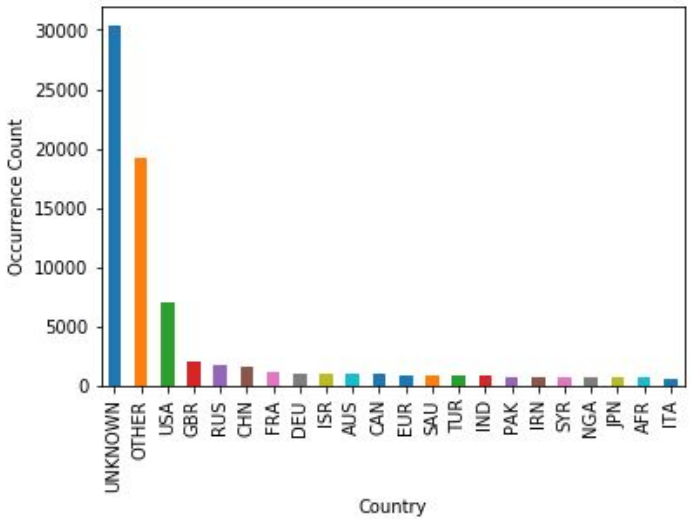
https://www.investingforme.com/commentary/2014/06/look-out-investment-yield-vs.-investment-distributions

Our Objective

- Our objective is to predict the average tone of a newly recorded event.

Features

- Country of origin for actor 1 and actor 2.



- Root cameo code.

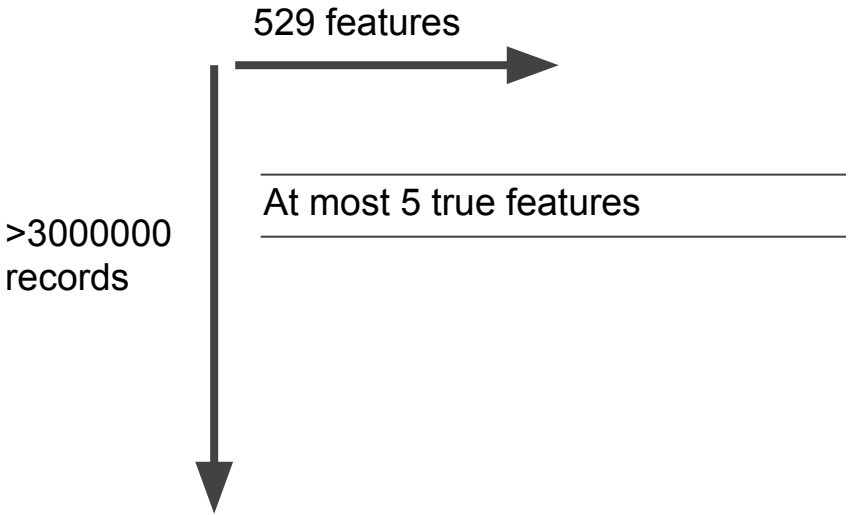
- “Type” or “role” of both actors. “Police Forces” or “Media”, etc.

Appeal
Demand
Fight

http://cliparting.com/free-police-clipart-28848/

Data Characteristics

- Sparse.
- A lot of it.
- High dimensional.
- Majority categorical data.



Preparing for Models

- Few options to deal with having this many features.
    - Use actor 2 country to determine if event is internal.
    - Build models per Actor 1 country.

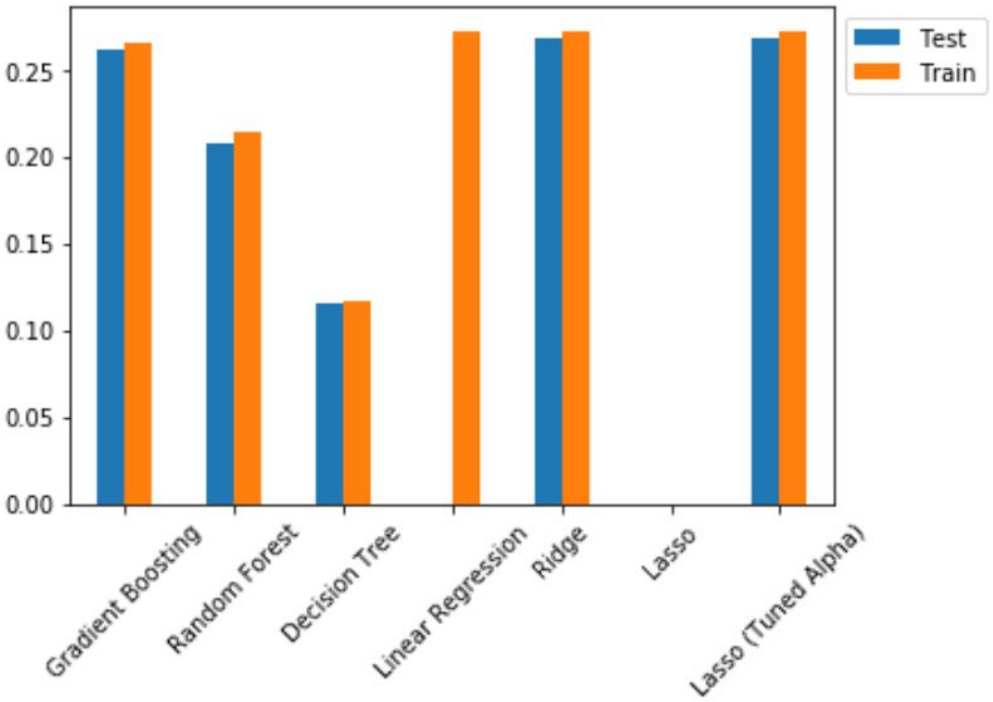
Modelling Approaches

Models Attempted

- Linear Regression
  - Lasso
  - Ridge
- Decision Trees
  - Random Forest
  - Gradient Boosting

Model Comparison On Sample Data

Model	Accurate In-sample?	Accurate Out-sample?	Robust against overfitting?	Quick?	Interpretable?	Works w/o tuning?
Linear						
Lasso						
Ridge						
Decision Tree						
Random Forest						
Gradient Boosting						



Model Comparison On Sample Data

And the winner is...

- Linear Regression with Ridge (L2) regularization.
  - R^2 on test set is .272 when applied to all data.
  - (Unweighted) Mean R^2 on all country by country models is .184.
  - (Weighted) Mean R^2 on all country by country models is .207.

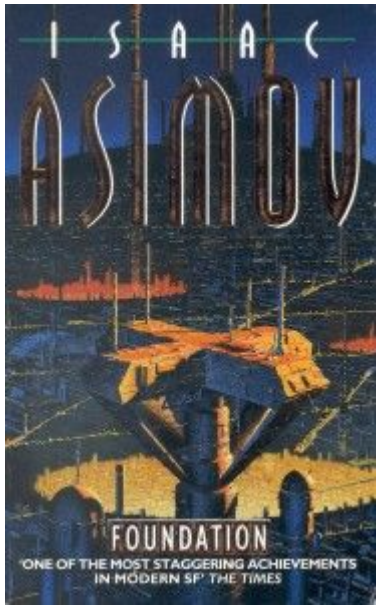
Insights

- Predicting average tone is pretty difficult.
- The event root codes are powerful predictors of tone.
- Type codes are generally stronger than country codes when predicting decreases.

Applications

What can we do with this?

- Help predict when an event might escalate.
- Interpreting the model results could give insights into media perceptions on relationships between countries.



http://www.vavatch.co.uk/books/asimov/found1as.jpg

Future Work

Future Work and Limitations

- Expand the amount of data from GDELT that is used to build the model.
- Building a more robust model that benefits from increased computing power, probably Gradient Boosting.

References

- GDELT Website - https://www.gdeltproject.org/
- GDELT Data Format Documentation - http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf
- CAMEO Code Manual - https://www.gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf
- Full Report on Github - https://github.com/eistre91/thinkful-projects/blob/master/GDELT%20Supervised%20Learning%20Capstone.ipynb

Erik Istre

e.istre91@gmail.com

https://github.com/eistre91

https://www.linkedin.com/in/erikistre/

Image: https://www.pexels.com/photo/green-water-fountain-225769/