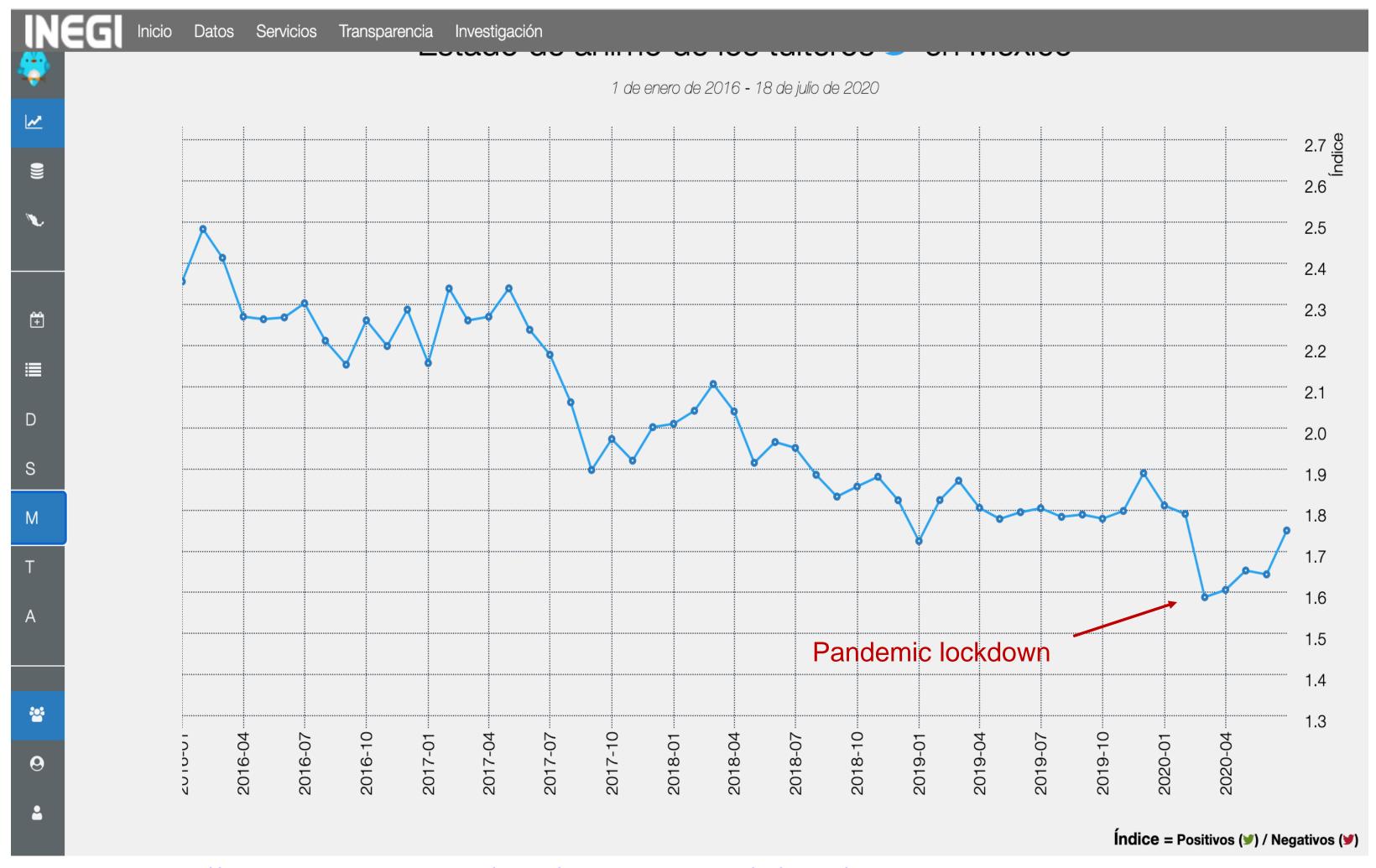


ML at INEGI

Alejandro Ruiz Researcher, Social Scientist

Sentiment Analysis from twitter



- » 10,000 people labeled over 54,000 tweets based on the emotion transmited
 - > Positive
 - Negative
- » 93,000 georreferentiated tweets on a daily basis
- > + 300 millions tweets
- » TFiDF + SMV
- » positivity quotient =

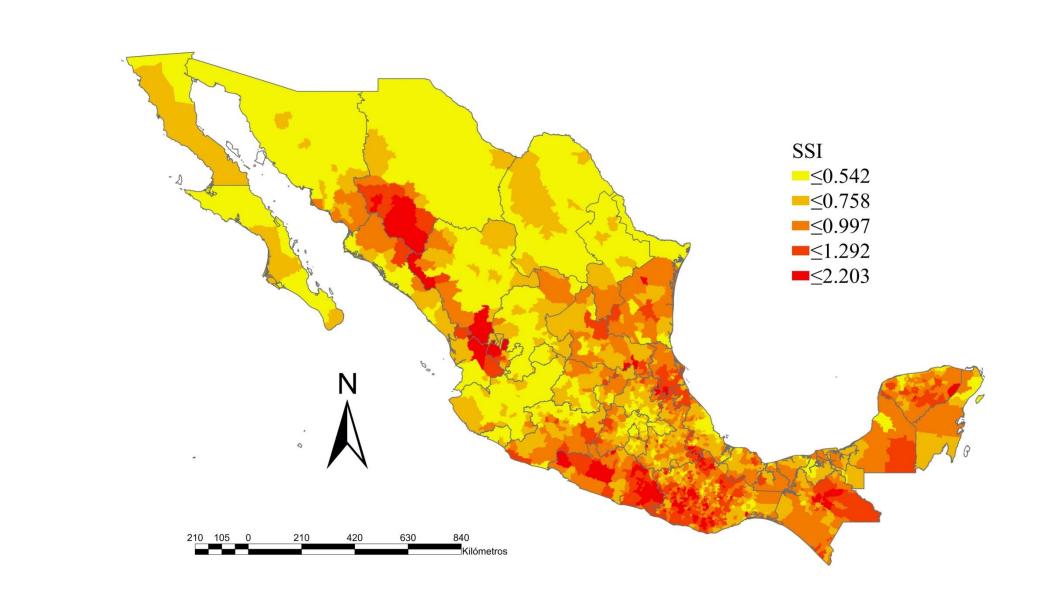
 #postiive / #negative
- » It only reflects the emotion of people actually tweeting
 - Restrictions from the Public API

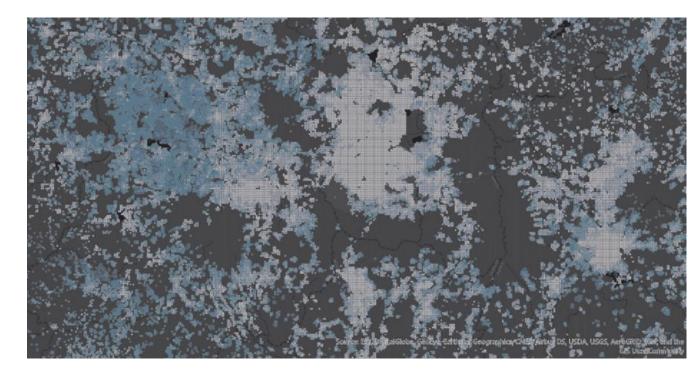
https://www.inegi.org.mx/app/animotuitero/#/app/multiline

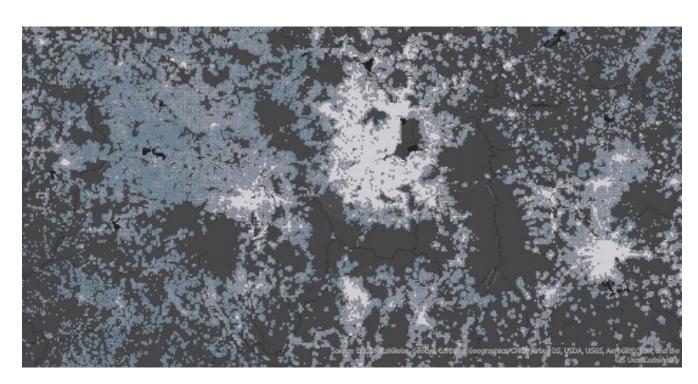


ML + Satellite imagery to identify vulnerable settlements

- » Vulnerable settlements to be identified at the microscale and in a more timely manner
- » 600x600 meters grid, matching a vulnerability index to an image
 - > DL algorithm to train a model
- » Multidisciplinary team
- » Challenge:
 - to improve accuracy metrics







Real

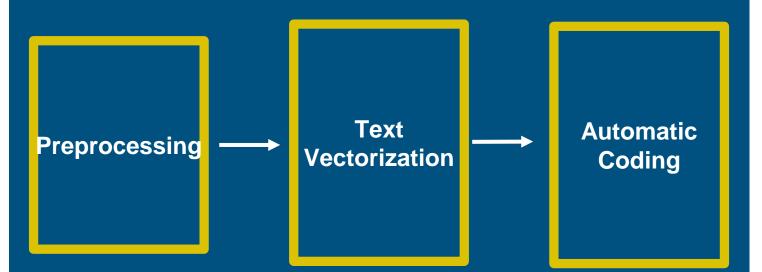
Predicted



Coding Occupation and Economic activity variables

70% 30% Deterministic Coding Entry batch (some Assited Coding = in the form of: records with Human-made standarized-text items) coding this text is equal to... to be coded then ..." Subsample Subsample Subprocess to ensure Human-coding made by an expert ("guru") a quality measure **Contrasting Process:** In the case of deterministic coding, accuracy must be at least 95%, otherwise adjustments are made to the algorithms. In the case of assisted coding, the sampling design ensures 90% accuracy.

- » Our goals:
 - Reduce the percentage human-made coding
 - Verify the quality of deterministic coding



- We can reduce the human workload up to 40%
- » Challenges:
 - To adjust the current process to add this new block based on NLP & ML





ML at INEGI

jose.ruizs@inegi.org.mx