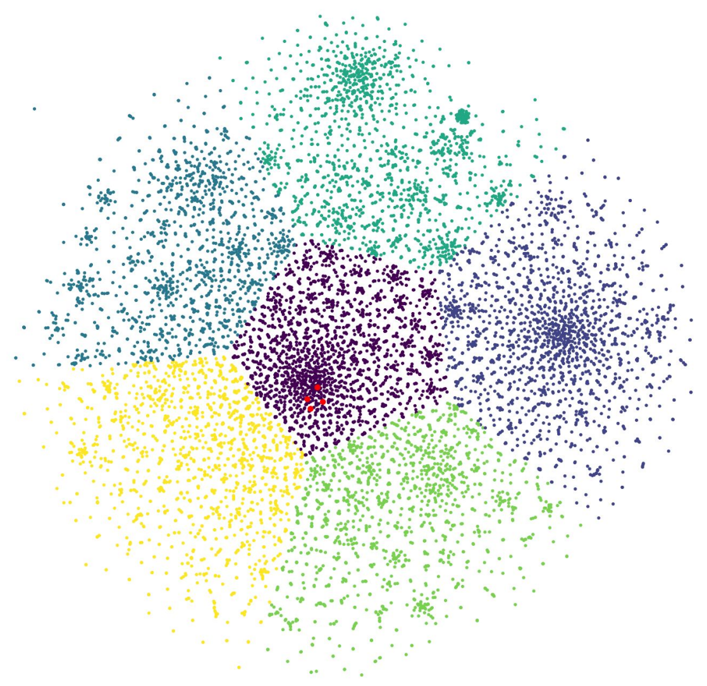


Improving Context-Aware Semantic Relationships in Sparse Mobile Datasets

Peter Hansel, Nik Marda, William Yin
{pwhansel, nmarda, wyin}@stanford.edu

Overview

- Cutting-edge NLP techniques often fail to capture semantic context
- Microblogging (and many other types of mobile datasets) have inputs other than text
- How do we make relationships between sentences more semantically salient using multimodal data?

Data and Features

- Politician Tweets Dataset [1]
  - Tweets associated with user locations
  - Coordinates collected using GeoPy Nominatim API
  - Date/time encoded as cyclical continuous feature
  - Data stripped of URLs and NLTK toolkit stopwords
- Tweet similarity data labeled by political science students and averaged

Existing Methods

- **Doc2Vec** generates a sentence embedding space allowing for comparison [2]
- **CoSal** uses contextually significant words in weighted BoW embeddings [3]
- Neither incorporates non-textual data

Models

- Iterative Minimization - Given embeddings *a*, *b*, and multimodal features $m_{a,i}$, $m_{b,i}$, iteratively optimized various distance functions d_i for various multimodal features:

$$f(a, b, (m_{a1}, m_{b1}), (m_{a2}, m_{b2}), ...) = a \cdot b + d_1(m_{a1}, m_{b1}) + d_2(m_{a2}, m_{b2}) + ...$$

- PCA for dimensionality reduction of sentence embedding space
- t-Distributed Stochastic Neighbor Embedding (t-SNE) for constructing visualizations and determining relative similarity [4]

Iterative Minimization

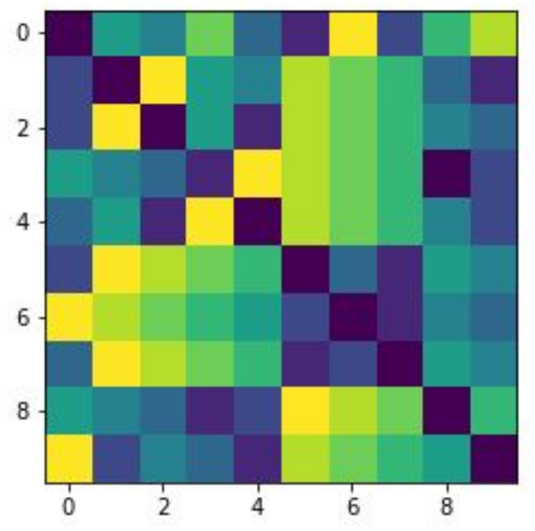
- Manually-annotated comparisons
- Distance function

$$d_i(m_{aj}, m_{bj}) = e^{-|m_{aj} - m_{bj}|}$$

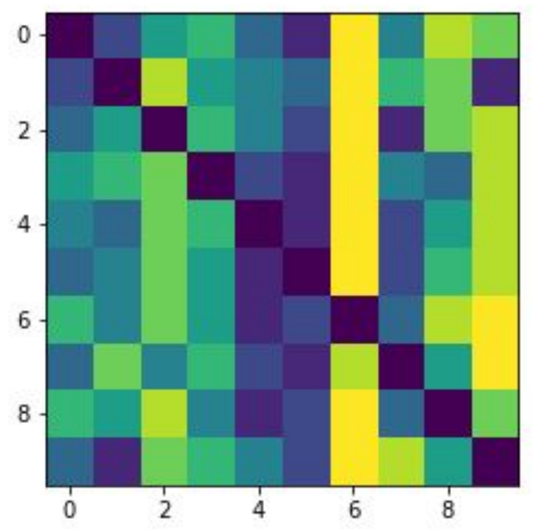
- Iteratively optimizing objective
  - Discrete ranking system means no continuous gradient
  - Minimizing this function:

$$L(\alpha_1, \alpha_2, ...) = \sum_{(i,j)} [\hat{y}(i,j) - y(i,j)]^2$$

- Scaling outputs of distance functions / integrating into f above

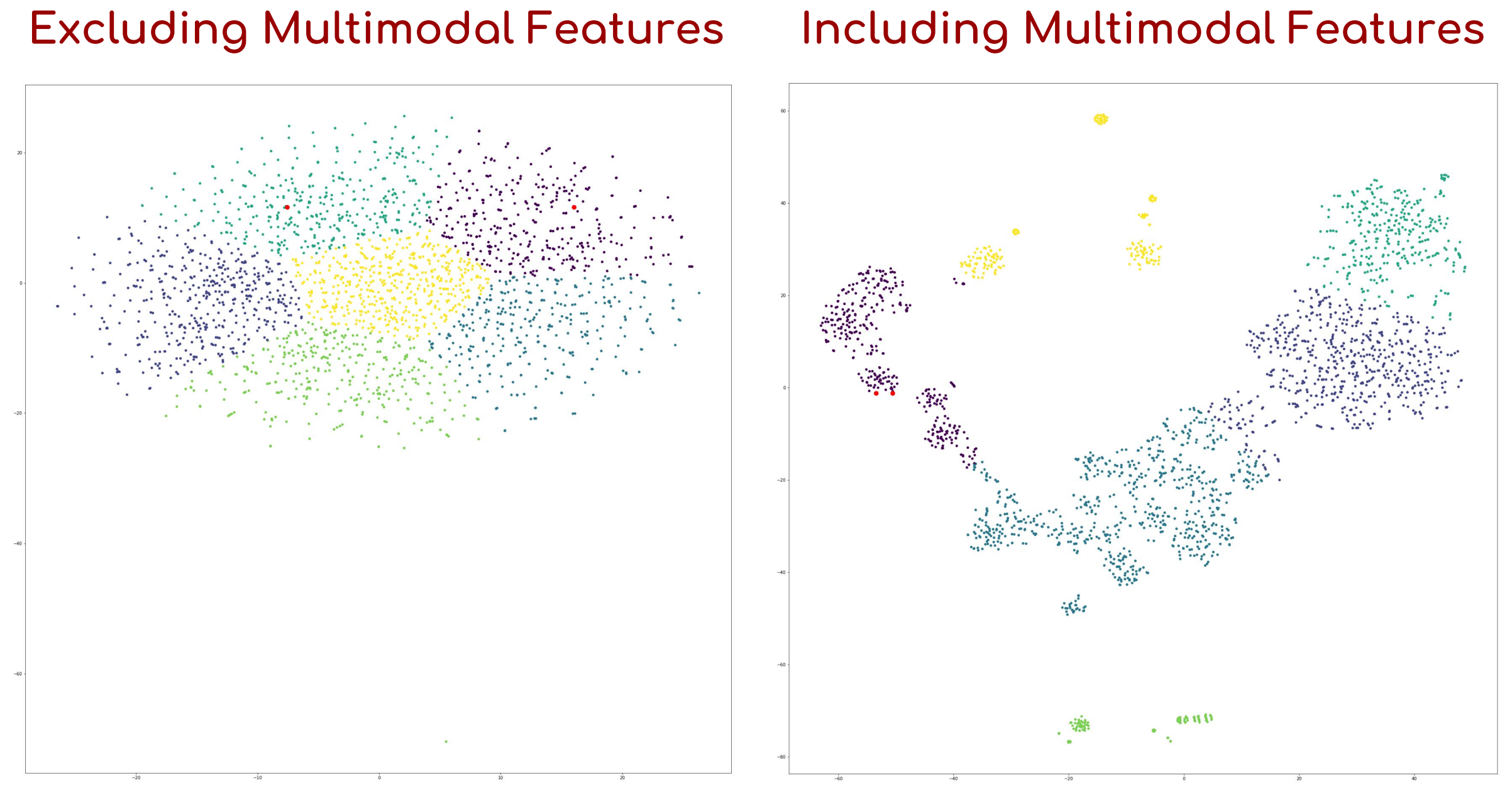


Labeled data Rankings



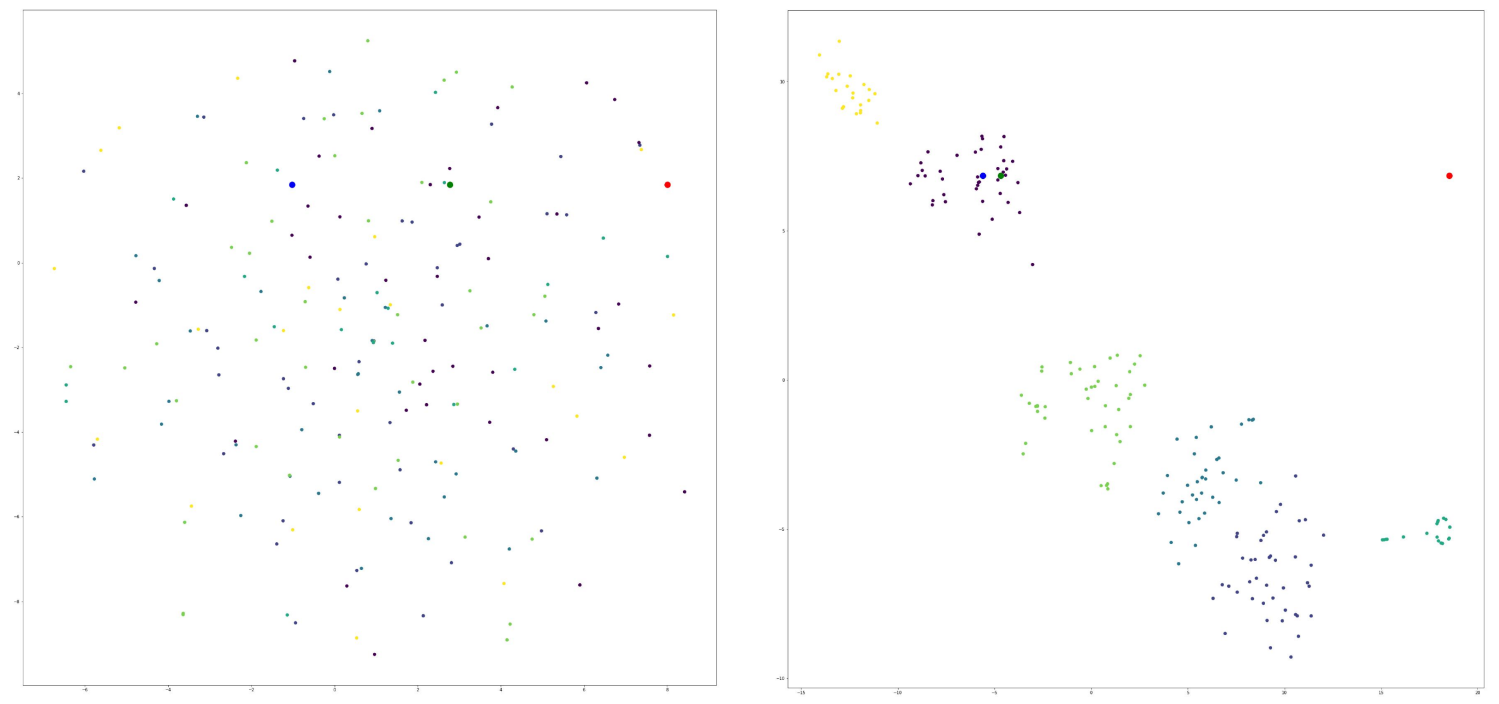
Iterative Minimization Rankings

PCA and t-SNE



“RT @CantorPress: House Republicans Unveil Debt Plan via @NROCorner #tcot #GOP #2Futures” - Peoria, AZ, 2011-07-25 18:58:42

“I'll be going on @foxnews at 11:20 (ET) to discuss the current negotiations of the #debtceiling. Check it out!” - Arizona, 2011-07-13 14:29:17



Green: “Good news- The House passed a bill to exempt those who lost coverage due to the failure of #Obamacare's co-ops from the individual mandate.” - Janesville, WI, 2016-10-03 20:39:43

Blue: “RT @DrPhilRoe: Bottom line: Obamacare is NOT working, especially not in Tennessee. Tennesseans deserve a #BetterWay.” - Jefferson, LA, 2016-10-06 19:34:56

Red: “It is too soon to rule out impacts to Florida. Please visit so that you and your family can get prepared.” - The Sunshine State, 2016-10-01 21:16:00

Discussion

- Multimodal data improves recognition of semantic relationships
- Especially valuable when tweets are about the same event but lack textual similarity
- Iterative Minimization has an upper bound on performance

Future Directions

- Test on tweets from local politicians and see if they differ from national politicians (controlling for location)
- Distort the word embedding to directly incorporate information from multimodal features
- Beyond Twitter and microblogging: other extended data

References

[1] B. Kay. “Politician Tweets.” *data.world*. 2018.

[2] Q. Le and T. Mikolov. "Distributed representations of sentences and documents." *ICML*. 2014.

[3] E. Zelikman. "Context is Everything: Finding Meaning Statistically in Semantic Spaces." *arXiv (2018).*

[4] L. Maaten and G. Hinton. "Visualizing data using t-SNE." *JMLR*. 2008.