The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions

Salvatore Giorgi[1], Daniel Preotiuc-Pietro[1], Anneke Buffone[1], Daniel Rieman[1], Lyle H. Ungar[1] & H. Andrew Schwartz[2]

[1] University of Pennsylvania [2] Stony Brook University

Motivation

How does one properly compute community-level lexical features?

- *Documents* can contain both location and user information
- *Users* produce text at various locations in time and space
- *Communities* can be considered a collection of words, documents or people

Data and Prediction Task

Twitter Data: a 10% Twitter sample from 2009-2015, over 30 billion tweets [1]

- Tweets are mapped to U.S. Counties (1.5 billion) [2]
- Users with less than 30 tweets are removed (over 5 million users in final data set)
- Counties with less than 100 users are removed (2041 U.S. Counties meet this threshold)

Community Level Data:

- *Income and Education* Median household income and percentage of people with a Bachelor's degree.
- *Life Satisfaction* Average response to the question "In general, how satisfied are you with in your life?" [3]
- *Mortality* Age-adjusted heart disease

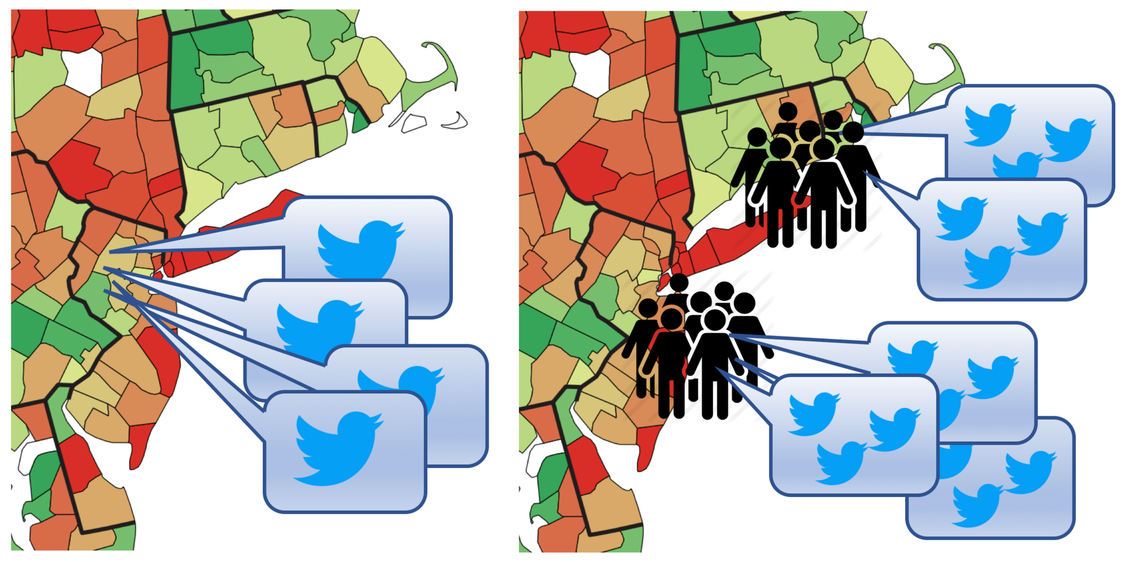
Prediction Task:

- Ten fold cross validation
- Ridge regression
- Randomized PCA for feature selection
- Feature sets: unigrams, topics and unigrams + topics

Contact Information

- http://wwbp.org
- github.com/wwbp/county_tweet_lexical_bank
- sgiorgi@seas.upenn.edu,has@cs.stonybrook.edu

Data Aggregation Methods and Main Result



- **Tweet to County**

$$feat_{i,j} = \frac{\text{number of tweets containing feature } i}{\text{number of users in county } j}$$

- **County**

$$feat_{i,j} = \frac{\text{number of times feature } i \text{ was used}}{\text{number of features used by county } j}$$

- **User to County**

$$feat_{i,j} = \frac{1}{N_j} \sum_{k \in U_j} \frac{\text{num. of times user } k \text{ used feature } i}{\text{number of features used by user } k}$$

where U_j is the set of users in county j and N_j is the total number of Twitter users in county j.

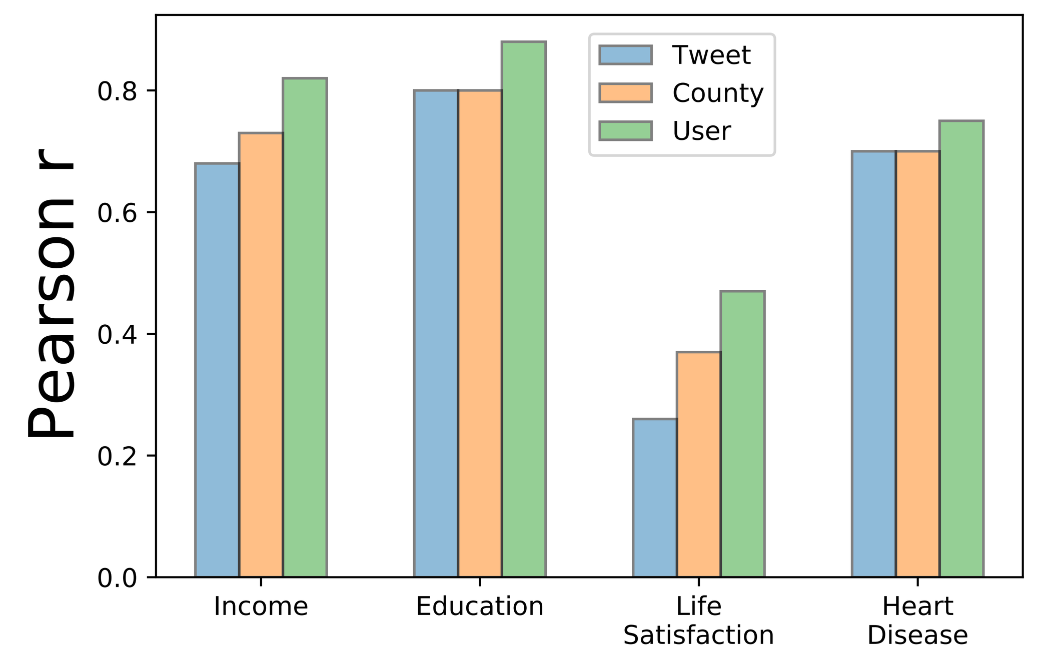
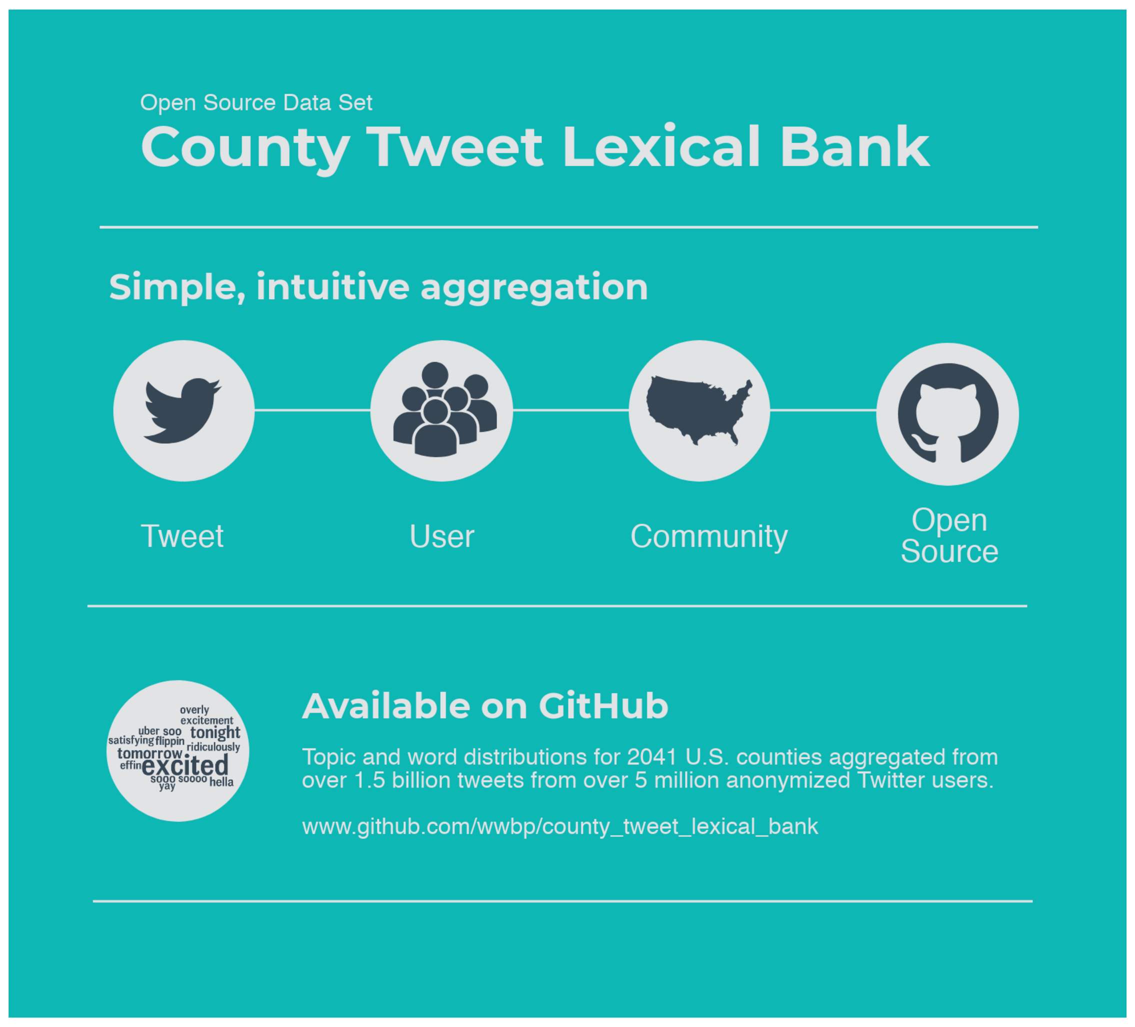
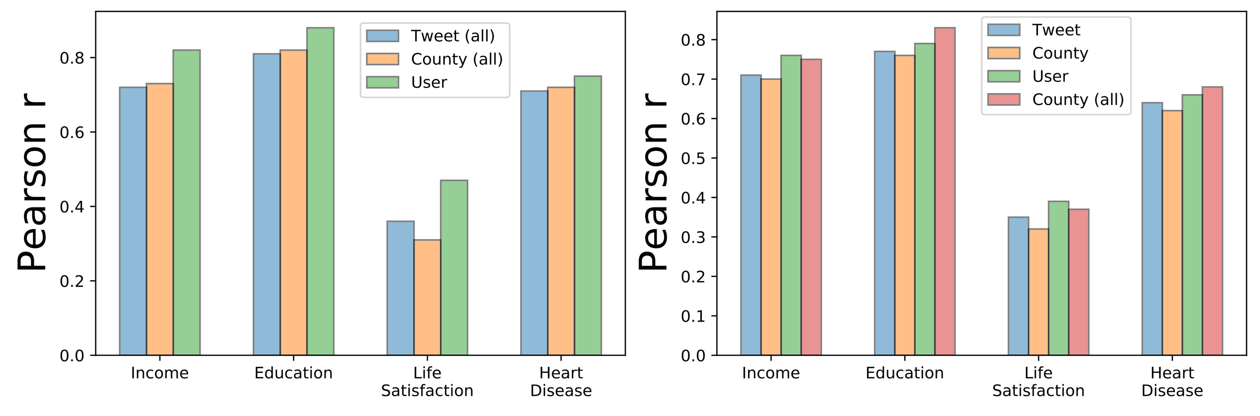


Figure: Predictive accuracy of unigrams + topics for 10% Twitter sample.



Changes In Data Size



(a) (b)

Figure: (a) Predictive accuracy of three aggregation methods without removing data. "All" methods do not throw away users with less than 30 tweets. "All" methods use approximately 1.6 billion tweets, "User to County" uses 1.3 billion., (b) Predictive accuracy using a 1% sample of random Twitter data.

Super Users

	Max Tweets	Income	Educat.	Life Satis.	Heart Disease	Num. Users Removed
County (all)	50	.73	.84	.34	.68	4,665,114
	500	.81	.87	.44	.75	611,661
	1000	.81	.87	.41	.75	217,517
	No Max	.73	.82	.31	.72	-
User to County	50	.68	.80	.34	.64	4,665,114
	500	.80	.87	.47	.76	611,661
	1000	.81	.87	.47	.76	217,517
	No Max	.81	.87	.48	.76	-

Table: Prediction results (Pearson r) using topics + unigrams. Users with more than "Max Tweets" number of tweets are removed from the sample.

Acknowledgements

This work was supported, in part, by the Templeton Religion Trust (grant TRT-0048).

References

[1] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan.
Trendminer: An architecture for real time analysis of social media text.
In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Workshop on Real-Time Analysis and Mining of Social Streams, ICWSM*, 2012.

[2] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin E P Seligman, and Lyle H Ungar.
Characterizing geographic variation in well-being using tweets.
In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2013.

[3] Nicole M Lawless and Richard E Lucas.
Predictors of regional well-being: A county level analysis.
Social Indicators Research, 101(3):341–357, 2011.

[4] H Andrew Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. H. Ungar.
Personality, gender, and age in the language of social media: The Open-Vocabulary approach.
PLoS ONE, 2013.