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

Salvatore Giorgi[1] **Daniel Preoţiuc-Pietro**[2] **Anneke Buffone**[1]
Daniel Rieman[1] **Lyle H. Ungar**[2] and **H. Andrew Schwartz**[3]
[1]Department of Psychology, University of Pennsylvania
[2]Computer and Information Science, University of Pennsylvania
[3]Computer Science, Stony Brook University
sgiorgi@sas.upenn.edu

Abstract

Nowcasting based on social media text promises to provide unobtrusive and near real-time predictions of community-level outcomes. These outcomes are typically regarding *people*, but the data is often aggregated without regard to users in the Twitter populations of each community. This paper describes a simple yet effective method for building community-level models using Twitter language aggregated by user. Results on four different U.S. county-level tasks, spanning demographic, health, and psychological outcomes show large and consistent improvements in prediction accuracies (e.g. from Pearson $r = .73$ to $.82$ for median income prediction or $r = .37$ to $.47$ for life satisfaction prediction) over the standard approach of aggregating all tweets. We make our aggregated and anonymized community-level data, derived from 37 billion tweets – over 1 billion of which were mapped to counties, available for research.

1 Introduction

Social media is an increasingly popular resource for large-scale population assessment which promises a cheap and non-intrusive complement to standard surveys with finer spatio-temporal scales (Coppersmith et al., 2015; Mowery et al., 2016; Wang et al., 2017). Twitter has been used — among other things — to measure community health (Paul and Dredze, 2011; Mowery et al., 2016; Eichstaedt et al., 2015), well-being (Schwartz et al., 2013), and public opinion on politics (O'Connor et al., 2010; Miranda Filho et al., 2015). By having access to measurements from multiple locations or communities, models trained on text data from social media can be used both to predict future measurements and to provide community estimates where these are lacking or are not robust. Such research is made possible by the massive amount of easily accessible user-generated data from public social media.

However, there has been little research on the way in which such data should be aggregated in order to compute community-level lexical feature estimates. Typically, data are aggregated in a "bag of words" style, disregarding tweets and authors (Culotta, 2014a; Schwartz et al., 2013; Eichstaedt et al., 2015; Curtis et al., 2018). We find, however, that giving equal weight to each user, rather than to each word or tweet, yields much more accurate community-level predictions.

In this paper, we conduct a series of experiments testing various simple yet intuitive aggregation methods. We show that choice of aggregation methods can result in substantial (one might even say "remarkable") boosts in accuracy when predicting U.S. county level outcomes (e.g. user-to-county aggregation yields a 7% to 27% increase in Pearson correlation). Contributions include (a) validation of aggregation approaches across four outcomes related to health, psychology, and demographics, (b) validation that aggregation has some effect on smaller sample of Twitter data, (c) show the effect of power tweeters (or "super users") and (d) release of resource-intensive community aggregated lexical data.

Related work. This is the first work we know of to explore simple aggregation techniques for population-level prediction tasks from language. Previous work has explored more sophisticated adjustments, such as addressing demographic-self selection bias in Twitter community predictions by re-weighting messages, finding small improvements (a 4.5% reduction in symmetric mean absolute percentage error) (Culotta, 2014b). In a political voting intention prediction application, (Lampos et al., 2013) modeled users and words jointly by learning separate regression weights for the two dimensions based on the intuition that each user contributes differently towards the outcome. However, their model was specifically adapted to problems that use time-series outcomes, rather than community-level aggregation. Distributions of lexical features are considered at multiple levels of analysis (message, user and community) in (Almodaresi et al., 2017) though each level considers one type of aggregation. Similar aggregation methods have been used in the context of topic modeling (Latent Dirichlet Allocation (Blei et al., 2003) and Author Topic Model (Rosen-Zvi et al., 2004)) by considering user, hashtag and conversation level aggregations (Alvarez-Melis and Saveski, 2016; Hong and Davison, 2010) but, again, community level aggregation and prediction tasks were not considered.

2 Data

Research was reviewed by an academic institutional review board and deemed exempt.

2.1 Twitter Data Collection

Twitter Sample A random 10% sample of the entire Twitter stream ('GardenHose') was collected between July 2009 and April 2014, which was then supplemented with a random 1% sample from May 2014 to February 2015. The total sample contains approximately 37.6 billion tweets (Preotiuc-Pietro et al., 2012).

County Mapping In order to map each tweet to a location within a county in the United States, we use both self-reported location information in user profiles and latitude/longitude coordinates associated with a tweet. If latitude/longitude coordinates are present then we trivially map the tweet to a county. The self-reported location information is a free text field and we use a cascading set of rules to map this field to a county. The rules are designed to avoid false positives (incorrect mappings) at the expense of fewer mappings. The full details of this process can be found in (Schwartz et al., 2013). Note that the latitude/longitude coordinates are a tweet attribute whereas the self-reported location is a user attribute yet both are used to map tweets to counties. Users are assigned a county by considering their earliest county mapped tweet.

In total, we are able to map 1.78 billion of the 37.6 billion tweets to a US county using the above-mentioned method. The county mapped data set was then filtered to contain only English tweets using the popular langid.py method (Lui and Baldwin, 2012), further reducing our tweet set to 1.64 billion tweets. For experiments with user-level data aggregation, we removed users who made relatively few (less than 30) posts in our data set.

	Number of Tweets				Number of Users
	Full Sample	County Mapped	English*	User Level**	
10%	37.6B	1.78B	1.64B	1.53B	5.25M
1%	-	-	199M	131M	1.57M

Table 1: Number of tweets in each section of the resource, including the total number of users. (*) The number of tweets used in the "all" experiments; (**) the number of tweets in the remaining experiments.

Publicly Available Stream The standard publicly available Twitter stream outputs approximately 1% of the public Tweet volume. Since a 10% sample is not available to most researchers, we replicated a 1% sample by taking a random 10% of our county mapped, English filtered 10% sample. The same process of county mapping, language filtering and user selection was applied to this data resulting in 131 million county mapped English tweets from 1.57 millions users. Table 1 presents the data set statistics.

The County Tweet Lexical Bank The County Tweet Lexical Bank is a U.S. County level data set comprised of two feature sets[1] :

- an aggregated "bag-of-words" count vector across all the county's messages in order to preserve anonymity. The unigrams represent the most frequent words in the data set;[2] ;
- a "bag-of-topics" representation for each county, with 2000 social media-derived topics described in (Schwartz et al., 2013).

Both feature sets will be releases across the 2009-2015 time span as well as individual years. Yearly updates will be included as they become available. As we are only releasing aggregated word-level features, as opposed to raw Tweets, this data release is within Twitter's Terms of Service.

[1] Available at `https://github.com/wwbp`

[2] While 25,000 features were used in the predictive tasks we removed some features (@-mentions, URLs, etc.) from the data release to preserve anonymity.

	N	Mean	Std Dev	Min	Max	Skew
Income	1750	4.66	0.11	4.33	5.07	0.47
Educat.	1750	21.57	9.46	5.70	70.30	1.20
Life Satis.	1952	3.39	0.03	3.26	3.51	0.02
Heart Dis.	2041	186.66	45.59	54.82	412.32	0.66

Table 2: Descriptives of U.S. County data used in the four prediction tasks.

2.2 Outcomes

The following U.S. county demographic, psychological and health variables were used in our prediction tasks. Table 2 gives statistics for each county variable.

Income and Education The census data for county median household income (log-transformed to reduce skewness; N=1,750) and percentage of people with a Bachelor's degree (N=1,750) were obtained from the 2010 U.S. Census Bureau's American Community Survey (ACS).

Life Satisfaction To assess subjective well-being we used the average response to the question "In general, how satisfied are you with in your life?" (1 = very dissatisfied and 5 = very satisfied) (Lawless and Lucas, 2011). Estimates are averaged across 2009 and 2010 (N=1,952).

Mortality Rates From the Centers of Disease Control and Prevention (CDC) we collected age-adjusted mortality rates for heart disease (N=2,041). Rates are averaged across 2010-2015.

3 Methods

3.1 Aggregation

Our aim is to use the user-level information based on the assumption that aggregating data first at the user-level would remove biases introduced by non-standard users of the platform. To this end, we explore three types of aggregation: (1) tweet to county, (2) county "bag of words" and (3) user to county.

Tweet to County Here we compute

$$feat_{i,j} = \frac{1}{N_j} \sum_{k} \mathbb{1}_i(unigram_k), \qquad (1)$$

where $\mathbb{1}_i$ denotes the indicator function for unigram$_i$. Here the ith feature for the jth unit of analysis (a U.S. county) is equal to the relative frequency of the unigram: the number of times each unigram was mentioned divided by N_j the total number of tweets from county j.

County Next, we use a method which was generally used in past research, which aggregates all messages to a community disregarding any meta-data, including tweet or user information. Previous state-of-the-art results using this method include life satisfaction r = .31 (Schwartz et al., 2013), atheroclerotic heart disease r = .42 (Eichstaedt et al., 2015) and education r = 0.15 (Culotta, 2014b). We therefore consider each county a "bag of words" using (1) with N_j equal to the number of unigrams from county j.

User to County The third method treats the unit of analysis (U.S. county) as a community of users. Therefore, feature weights are extracted at the user level, normalized and then averaged to communities:

$$feat_{i,j} = \frac{1}{N_j} \sum_{k \in U_j} r_k(unigram_i), \qquad (2)$$

where U_j is the set of users in county j, N_j is the total number of Twitter users in county j and $r_k(x)$ is the relative frequency of feature x for user k with $i \in$ {all unigrams} and $j \in$ {all counties}.

Features

We use as features a list of 2,000 social media-derived topics generated from Latent Dirichlet Allocation (Blei et al., 2003) using the complete MyPersonality Facebook data set consisting of approximately 15 million posts (Schwartz et al., 2013). The topic loadings are computed from the most frequent 25,000 unigrams in our data set. We also use a subset of these unigrams as additional features in our models (25,000 reduced to 10,000).

Experimental setup

For each of the four county level Census and health variables we built three models using 10-fold cross validation with the following features: (1) unigrams, (2) topics and (3) unigrams + topics. For consistency across tasks we only considered counties with 100 or more 30+ tweet users (N=2,041).

We used a feature selection pipeline which first removed all low variance features and then features that were not correlated with our census and health data. Principal component analysis was

	Income	Educat.	Life Satis.	Heart Disease
Tweet to County	.68	.80	.26	.70
County	.73	.80	.37	.70
User to County	**.82**	**.88**	**.47**	**.75**

(a) Unigrams + Topics, Pearson r

	Income	Educat.	Life Satis.	Heart Disease
Tweet to County	.67	.79	.22	.65
County	.72	.78	.37	.64
User to County	**.79**	**.87**	**.44**	**.73**

(b) Unigrams, Pearson r

	Income	Educat.	Life Satis.	Heart Disease
Tweet to County	.65	.77	.31	.71
County	.68	.80	.34	.71
User to County	**.81**	**.87**	**.47**	**.76**

(c) Topics, Pearson r

Table 3: Prediction results (Pearson r) for direct aggregation comparison on the 10% sample.

then applied to the reduced feature set for further dimensionality reduction. This preprocessing was used to avoid overfitting, since our model included more independent variables (2000 topic frequencies and/or 10k unigrams) than observations (at most 2,041 counties). For the prediction task we used linear regression with ℓ_2 regularization (Ridge regression) (Eichstaedt et al., 2015). The regression regularization parameter α was set to 1000 using grid search.

Because our initial dataset consisted of 37.6 billion tweets, using distributed IO was crucial for data aggregation and feature extraction. We used a Hadoop-style cluster consisting of 64 disks and 64 CPU cores across 8 physical machines. Over this cluster, we used Hadoop MapReduce for the county mapping step (taking approximately 1 week of wall clock runtime) and Spark for the feature aggregations (taking approximately 1 day of wall clock runtime). The entire pipeline of county mapping, English language filtering, feature extraction and prediction used the DLATK Python package (Schwartz et al., 2017)[3].

Experiments

Using the above setup we perform 3 experiments in order to explore the effects of data aggregation. We 1) directly compare aggregation methods using our 10% data; 2) compare aggregation methods using a 1% sample and, finally, 3) explore the effect of choosing an upper bound on the number of posts per Twitter users, looking at users with less than 50, 500, 1000 posts. This allows us to exclude frequent posters who are potentially organizations or bots.

	Income	Educat.	Life Satis.	Heart Disease
User to County	**.82**	**.88**	**.47**	**.75**
$N_{user-tweets}$	1.350B	1.350B	1.356B	1.360B
Tweet to County (all)	.72	.81	.36	.71
County (all)	.73	.82	.31	.72
$N_{all-tweets}$	1.621B	1.621B	1.628B	1.634B

Table 4: Prediction results (Pearson r, using unigrams + topics) using full 10% data vs. users with 30+ tweets. The number of tweets used in each task is listed to highlight the fact that the "User to County" tasks use less tweets than the "all" tasks.

4 Results and Discussion

Direct aggregation comparison. The results of our predictive experiments on the 10% data can be found in Table 3. Across all four tasks we see that the "User to County" approach outperforms the other aggregation methods, giving a higher Pearson r and obtaining state-of-the-art results for community-level predictions.

We see the largest gains for the "User to County" aggregation for the income outcome, with a 13 point increase in Pearson r for topics alone and a 9 point increase for unigrams + topics.

In Table 4 we remove the 30+ tweet requirement from the "Tweet to County" and "County" methods and compare against the "User to County" method (with the 30+ tweet requirement). Again we see the "User to County" method outperforms all others in spite of the fact that the "User to County" approach uses less data than both "all" approaches, which contains 108 million more tweets.

1% data. In Table 5 we repeat the above experiment on a 1% Twitter sample. Here we see that the "User to County" method outperforms both the "Tweet to County" and "County" methods (with all three tasks using the same number of tweets). When we compare the "User to County" and "County (all)" methods we see the "User to county" outperforming on two out of four tasks (Income and Life Satisfaction). Again, we note

[3] Available at https://github.com/dlatk

	Income		Educat.		Life Satis.		Heart Disease	
Tweet to County	.71	.62	.77	.71	.35	.32	.64	.63
County	.70	.60	.76	.67	.32	.28	.62	.62
User to County	.76	.70	.79	.74	.39	.28	.66	.66
$N_{user-tweets}$	127M	130M	127M	130M	127M	130M	127M	131M
County (all)	.75	.67	.83	.77	.37	.34	.68	.66
$N_{all-tweets}$	191M	195M	191M	195M	191M	197M	191M	198M
$N_{counties}$	949	1750*	949	1750*	954	1952*	960	2041*

Table 5: 1% sample prediction results (Pearson r) using topics + unigrams. $*$ same counties as the 10% prediction task.

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

that the "User to County" is using less data than the "County (all)". While, across the board, the performance increase is not as substantial as in the 10% results, we see comparable performance between "User to County" and "County (all)" methods despite the difference in the number of tweets.

Super users. One theory why we see such large gains depending on aggregation technique is that aggregating through users negates the effects of super users – those who post an extraordinary amount (such as organizations or bots). We implemented a maximum tweet requirement in order to remove these users and see if that accounts for the difference. Here we use both the "User to County" and "County (all)" samples and report results in Table 6. These results demonstrate that by keeping only users with less than 500 tweets we get results close to our "User to County (No Max)" method using the user-naive "County (all)" scheme. This shows that relatively few users (in this case 611k) can significantly decrease performance, though still leaves a small gain from the user to county approach. As seen in the lower half of the table, this thresholding does not increase performance when using the "User to County" method, which suggests such users can still be beneficial if they are just treated such that they can't dominate a community. This highlights the benefit of our simple method: we do not need to consider optimizations which may not generalize across data, such as upper-bound thresholds on the number of tweets per user. Further, the user-to-county aggregation seems to provide at least a small benefit beyond removal of super users.

5 Conclusion

This study explored the benefit of aggregation techniques for streaming user-generated data from individual messages to community level data, the typical setting for nowcasting. We showed that by simply aggregating to users first and then taking the mean within a county, we can obtain large gains (remarkably, up to a 13 point increase in Pearson correlation) over typical aggregation methods common in past work. In order to foster nowcasting research utilizing this more ideal aggregation, we will release the County Tweet Lexical Bank – a large aggregated and anonymized county-level data set, and computed on more than 1.6 billion tweets posted over 5 years.

Future work in this area can look at adjusting models to account for other meta-data such as temporal variation and diversity and to adjust for selection biases present in social media, where the user base on social media is not representative of the population of the community (Greenwood et al., 2016).

Acknowledgments

This work was supported by a grant from the Templeton Religion Trust (ID TRT0048). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Fatemeh Almodaresi, Lyle H. Ungar, Vivek Kulkarni, Mohsen Zakeri, Salvatore Giorgi, and H. Andrew Schwartz. 2017. On the Distribution of Lexical Features in Social Media. In *Annual Meeting of the Association for Computational Linguistics*, ACL.

David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. *ICWSM*, 2016:519–522.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses.

Aron Culotta. 2014a. Estimating county health statistics with twitter. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1335–1344. ACM.

Aron Culotta. 2014b. Reducing Sampling Bias in Social Media Data for County Health Inference. In *Joint Statistical Meetings Proceedings*.

Brenda Curtis, Salvatore Giorgi, Anneke E. K. Buffone, Lyle H. Ungar, Robert D. Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H. Andrew Schwartz. 2018. Can twitter be used to predict county excessive alcohol consumption rates? *PLOS ONE*, 13(4):1–16.

Johannes C Eichstaedt, H Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, Christopher Weeg, Emily E Larson, Lyle H Ungar, and Martin EP Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26:159–169.

Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. *Social Media Update*. Pew Research Center.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM.

Vasileios Lampos, Danie Preoţiuc-Pietro, and Trevor Cohn. 2013. A User-Centric Model of Voting Intention from Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL.

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

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Renato Miranda Filho, Jussara M Almeida, and Gisele L Pappa. 2015. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, pages 1254–1261. IEEE.

Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. 2016. Towards automatically classifying depressive symptoms from twitter data for population health.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 265–272.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. 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*.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

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. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 91–100. ACM.