# How Social Media Data Helps Predict Innovation

*Aaron Wang | Bureau of Business Research | University of Nebraska-Lincoln*

## Objective

Innovation leads to the creation of smarter and more efficient processes and is a key contributor to economic growth. Yet while the benefits of innovation are clear, measurements of innovation are both hard to obtain and difficult to quantify. Therefore, my research aims to answer the question –

Can social media data help predict county-level innovation?

**Literature –** Previous research by Eichstaedt (2015) shows the value of using Twitter data to predict heart disease mortality. Since then, several publications have utilized social media language data to better model socioeconomic factors. Methods proven effective in previous research have been used throughout this analyses.

## Data

**Twitter –** 1.1M geo-tagged Tweets from 2018, covering 282 counties with over 10,000 Tweeted words.

**Innovation 2.0 –** index of 56 county-level census measures of entrepreneurship, productivity, and socioeconomic wellbeing gathered by the Indiana Business Research Center (IBRC)

## Methods

Tweets were first cleaned to remove emoji, numbers, punctuation, and stopwords, then tokenized into individual words and analyzed by the following two methods. All scores have been user-weighted to account for differences in Tweet frequency.

**Sentiment Analysis –** Through the *sentimentr* package, each Tweet received a score for trust, joy, fear, anger, anticipation, surprise, disgust, and sadness. Scores are calculated from the NRC word-emotion lexicon and account for tweet length and valence shifters.

**Natural Language Processing –** Each tweet was assigned 2,000 topic probabilities, created through open-vocabulary approach Latent Dirichlet Allocation (LDA) and calculated by grouping together words that often appear together. Topics were then dimension-reduced, and all features not correlated with innovation at an adjusted p-value < $2.5 \times 10^{-5}$ were removed.

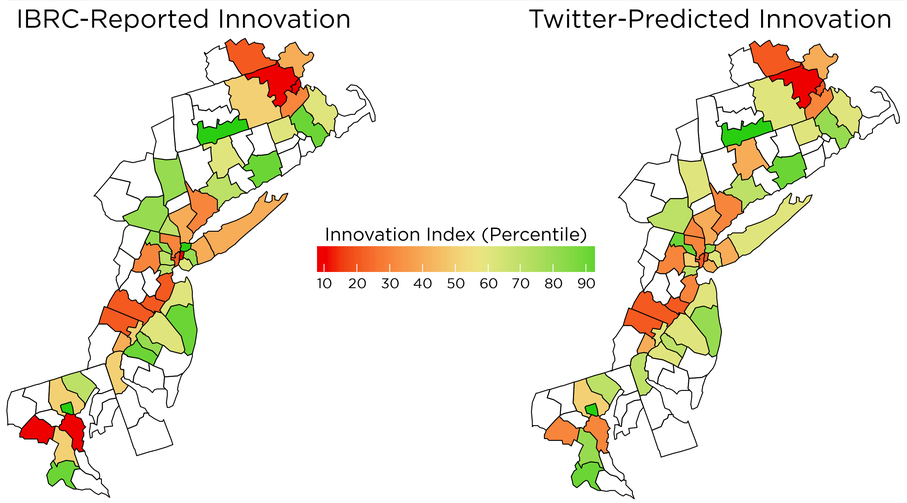
All Twitter-based models used the following ridge regression equation, where the regularization term λ was chosen via cross validation to reduce overfitting.

$$\beta = (XT\mathrm{X} + \lambda I)\, X^T y$$

## Topic Correlations

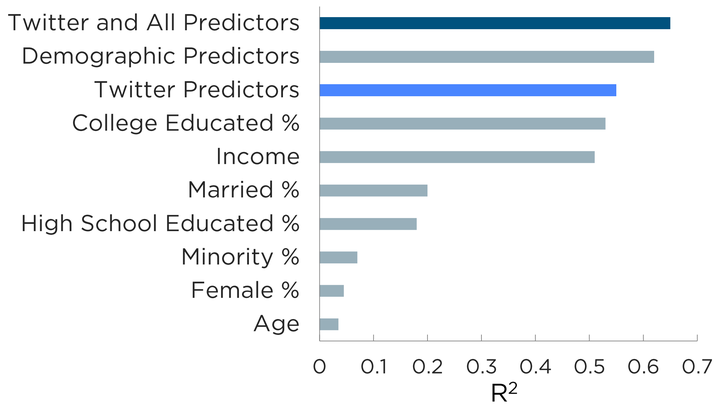

Topics most correlated with county-level innovation, significant at a Benjamini-Hochberg corrected p-value < $2.5 \times 10^{-5}$.
Word size represents its prevalence relative to all words within the topic.

## Prediction Map




Innovation prediction results for 50 counties in the Northeastern U.S.
Out-of-sample predictions were obtained through a cross-validation process.

## Model Comparison




Model performance was measured between predicted and actual county-level innovation. $R^2$ values were averaged from a 10-fold cross-validated process to avoid distortion of accuracy due to chance. Aggregate models have significant differences in performance at a p-value < .05.

## Limitations and Sources

The 2018 Twitter dataset temporally contrasts with the 2016 Innovation Index. The index is updated annually but 2018 values were not available during this analyses. However, previous research shows that up to 95% of a community's sentiment and language topics are maintained over a period of several months.

– World Well-Being Project and StatsAmerica, IBRC

## Conclusions

Results illustrate the value of incorporating Twitter data into existing models, as it presents a richer analyses of characteristics associated with county-level innovation. Several significant community language patterns were also revealed that provide insight beyond simple demographic factors. Twitter data is both easier to gather and faster to analyze and allows for on-demand predictions through real-time Tweets. Future research can investigate the precise relationships between Twitter language, index variables, and innovation, and may benefit from a more comprehensive social media dataset.

Aaron Wang | BBR Scholar | aaronwang@unl.edu