# Spatial Analysis of Emotional Language on Twitter in the U.S.

*Mohammad Atari*

*5/6/2018*

## Introduction

Understanding the relationship between language use and thought have long been an important and vibrant area of research within psychology. Yet, studying language can require time-intensive qualitative approaches, often with only a handful of respondents. Computational linguistics offers techniques to study language use at scale, requiring considerably less time and resources (Iliev et al., 2015). No longer constrained by results based on small samples of people, language offers many opportunities to directly study people's thoughts and emotions (Lazer et al., 2009). Yet as data move from gigabytes to terabytes to petabytes, finding an interpretable signal becomes a process of hunting for a needle in a hay field. Theories are needed to interpret data, and psychologists have developed such theories across hundreds of years. Additional benefit can come by collaborating with experts from multiple fields, including quantitative psychologists, statisticians, methodologists, economists, political scientists, health professionals, and educators.

One of the most important lines of research in psychology and language has focused on representation of human emotions in language. An emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response. Only very recently, computational social scientists have started to look at the representations of different emotions in language. Computer scientists have also started to examine affective processes in textual data, often labeled as sentiment analysis, defined as the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral. To do such analyses, scholars from different fields have relied on large-scale corpora (sometimes referred to as big data) from different sources. Before any analyses can be performed, data must be obtained. Careful consideration should be given toward which data will be most appropriate for the question at hand, and whether informative data are available and accessible (Kern et al., 2016). As Borgman (2015) noted, "having the right data is usually better than having more data" (p. 4). Some shared data resources exist. Social media platforms have provided psychologists and computer scientists with a rich, naturally occurring environment to investigate human behavior at scale (Serrano-Guerrero et al., 2015).

Twitter as an important social media platform for people across cultures to publically express their opinions and feelings about their community and their private life, has attracted increasing attention with millions of members in different countries (Li et al., 2013). A number of empirical studies have been published on sentiment analysis on Twitter (Poria et al., 2014). Methods for sentiment classification have been developed ranging from simple text mining to advanced symbol and feature recognition to a sentiment and subjective analysis (Oscar et al., 2017) and machine learning or lexicon-based approaches to more advanced hybrid methods (Serrano-Guerrero et al., 2015) and from sentiment orientation with only two modes (i.e., negative versus positive) to coarse measurement scale and more nuanced classifications (Fink et al., 2011). Yet, few sentiment analysis studies have utilized a geospatial and temporal computational methods.

A growing number of language studies on Twitter are trying to incorporate geospatial methods into their sentiment analyses. Medhat et al. (2014) categorized sentiment analyses into two groups: machine learning and lexicon-based approaches. Broadly, machine-learning methods are used to automatically identify sentiment polarity patterns in large corpora in order to learn mass opinions or emotions in textual data. A variety of algorithms have been developed (e.g., Rushdi Saleh et al., 2011) for this purpose. Most algorithms fall into the category of supervised machine learning. Lexicon-based approaches focus on measuring subjectivity and

opinions in texts using Semantic orientation (SO), which capture orientations of sentiments (i.e., negative or positive) and strengths or degrees of orientation (Taboada et al. 2011). Sentiment lexicons are the key for this type of methods. For example, Paltoglou and Thelwall (2012) proposed a lexicon-based approach to identify whether a text conveys negative or positive attitudes and to estimate the level of emotional intensity of a text in social media and microblogging environments.

More recently, Eichstaedt and colleagues (2015) collected tweets from across the United States, determined their counties of origin, and derived values for language variables (e.g., the relative frequencies with which people expressed anger or engagement) for each county. These authors correlated these county-level language measures with county-level age-adjusted heart disease mortality rates obtained from the independent health organizations. All procedures of these authors were approved by the University of Pennsylvania Institutional Review Board. In addition, they provided access to their data on Open Science Framework.

As mentioned before, Tweets messages on Twitter containing information about emotions, thoughts, behaviors, and other personally salient information. In 2009 and 2010, Twitter made a 10% random sample of tweets (the "Garden Hose") available for researchers through direct access to its servers. Eichstaedt et al. (2015) obtained a sample of 826 million tweets collected between June 2009 and March 2010. Many Twitter users self-reported their locations in their user profiles, and these authors used this information to map tweets to counties. This resulted in 148 million county-mapped tweets across 1,347 counties. An automatic process was used to extract the relative frequency of words and n-grams for every county. For example, the relative frequency of the word hate ranged from 0.009% to 0.139% across counties. Then they derived two types of language-use variables from counties' relative word-usage frequencies: variables based on (a) dictionaries and (b) topics. Dictionary-based variables were relative frequencies of psychologically related words from predetermined dictionaries (e.g., positive-emotion words accounted for 4.6% of all words in a county on average). Topic-based variables were the relative usage of 2,000 automatically created topics, which are clusters of semantically related words that can be thought of as latent factors. These authors used pre-established dictionaries for anger, anxiety, positive and negative emotions, positive and negative social relationships, and engagement and disengagement (Pennebaker et al., 2007; Schwartz et al., 2013). Topics had previously been automatically derived (Schwartz et al., 2013).

In this project, I used the data provided by Eichstaedt et al. (2015) as outlined above. The lexicon-based frequencies were used in this project. Latent factors (i.e., topic models) were not used for simplicity. I was particularly interested in emotion words and their geospatial distribution in the United States in over 1000 counties. The dictionaries that were used in the published paper are presented in Table 1. I used a number of Tidyverse packages to compute relative emotions in language use, merge the datasets from the paper and the ggplot2 package, and visualize the distrivbition of eight psychologically important emotional constructs (i.e., *anger*, *anxiety*, *negative relationships*, *positive relationships*, *negative emotion words*, *positive emotions words*, *engagement*, and *disengagement*).

```
## Warning: package 'knitr' was built under R version 3.4.3
```

```
## Parsed with column specification:
## cols(
##   Dictionary = col_character(),
##   `Top Ten Dictionary Words by Frequency` = col_character()
## )
```

Table 1: Emotion Categories and Their Sample Words

| Dictionary | Top Ten Dictionary Words by Frequency |
|---|---|
| Anger | shit f*** hate damn b*tch hell f**ing mad stupid b*tches |
| Negative Relationships | hate alone jealous blame evil rude lonely independent hated ban |
| Negative Emotion | sorry mad sad scared p*ssed crying horrible afraid terrible upset |
| Disengagement | tired bored sleepy lazy blah meh exhausted yawn distracted boredom |
| Anxiety | crazy pressure worry scared awkward scary fear doubt horrible afraid |
| Positive Relationships | love home friends friend team social welcome together kind dear |

| Dictionary | Top Ten Dictionary Words by Frequency |
|---|---|
| Positive Emotion | great happy cool awesome amazing glad excited super enjoy wonderful |
| Engagement | learn interesting awake interested alive learning creative alert involved careful |

## Data Manipulation

```
##loading libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(ggplot2)
library(ggmap)
##reading the dataset
cat = read_csv("~/Desktop/categorized dictionary with city name.csv")
```

```
## Parsed with column specification:
## cols(
##   group_id = col_integer(),
##   feat = col_character(),
##   city_name = col_character(),
##   lex_cat = col_character()
## )
```

```
##writing the data on a CSV format
write_csv(cat, "cat.csv")

#city and category
citycat = cat%>%
  group_by(city_name, lex_cat)%>%
  summarize(n())

#changing column names
citycat$city = citycat$city_name
citycat$emotion = citycat$lex_cat
citycat$word = citycat$`n()`

citycat = citycat%>%
  dplyr::select(c("city", "emotion", "word"))
```

```
## Adding missing grouping variables: `city_name`
```

```
city_words = citycat%>%
  group_by(city)%>%
  summarise(sum(word))
```

```
city_words_emotions = full_join(citycat, city_words, by = "city")

city_words_emotions = mutate(city_words_emotions, rel.emotion = word/sum(word))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
##different emotions
#anger
city_words_emotions_ANGER = filter(city_words_emotions, emotion == "ANGER")
#anxiety
city_words_emotions_ANXIETY = filter(city_words_emotions, emotion == "ANX")
#positive emotions
city_words_emotions_POSEMO = filter(city_words_emotions, emotion == "E+")
#negative emotions
city_words_emotions_NEGEMO = filter(city_words_emotions, emotion == "E-")
#engagment
city_words_emotions_ENGAGE = filter(city_words_emotions, emotion == "P+")
#disengagement
city_words_emotions_DISENGAGE = filter(city_words_emotions, emotion == "P-")
#positive relationships
city_words_emotions_POSREL = filter(city_words_emotions, emotion == "R+")
#negative relationships
city_words_emotions_NEGREL = filter(city_words_emotions, emotion == "R-")
```

## Plots

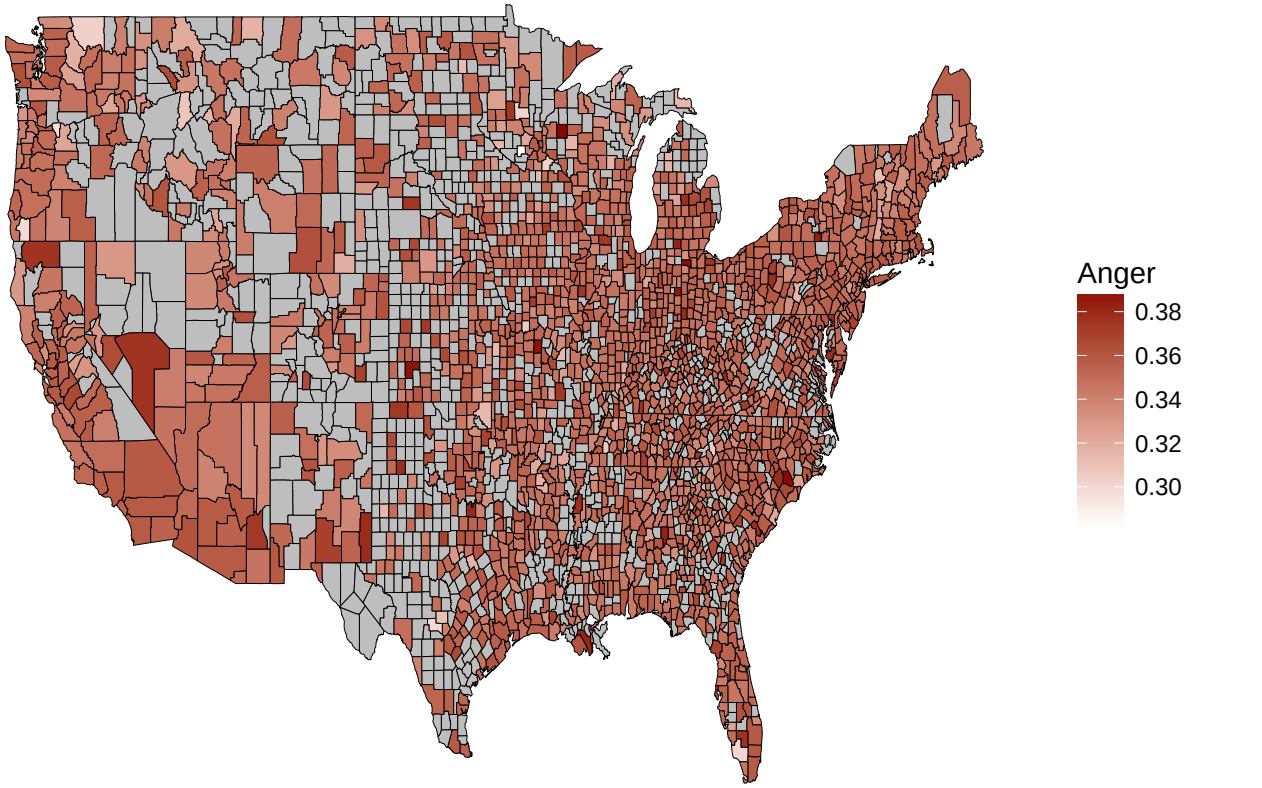
The following code provides merging the datasets and visualizations for the 8 emotional constructs.

```
#getting map data

USmap = map_data("county")
```

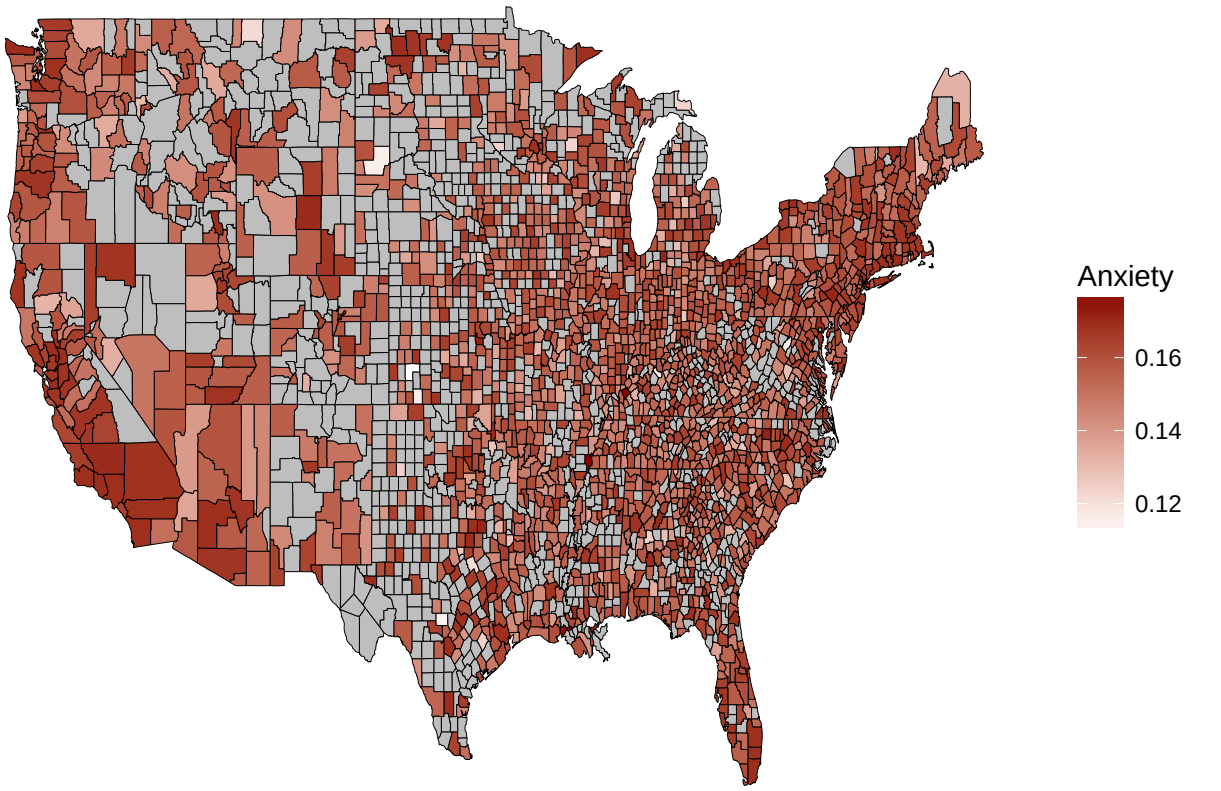
```
## Warning: package 'maps' was built under R version 3.4.4
```

```
USmap$city = USmap$subregion

#merging

data.ANGER = full_join(USmap, city_words_emotions_ANGER, by = "city")
data.ANX = full_join(USmap, city_words_emotions_ANXIETY, by = "city")
data.POSEMO = full_join(USmap, city_words_emotions_POSEMO, by = "city")
data.NEGEMO = full_join(USmap, city_words_emotions_NEGEMO, by = "city")
data.ENGAGE = full_join(USmap, city_words_emotions_ENGAGE, by = "city")
data.DISENGAGE = full_join(USmap, city_words_emotions_DISENGAGE, by = "city")
data.POSREL = full_join(USmap, city_words_emotions_POSREL, by = "city")
data.NEGREL = full_join(USmap, city_words_emotions_NEGREL, by = "city")

## Plots
#anger
ggplot(data.ANGER, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Anger"))+
```

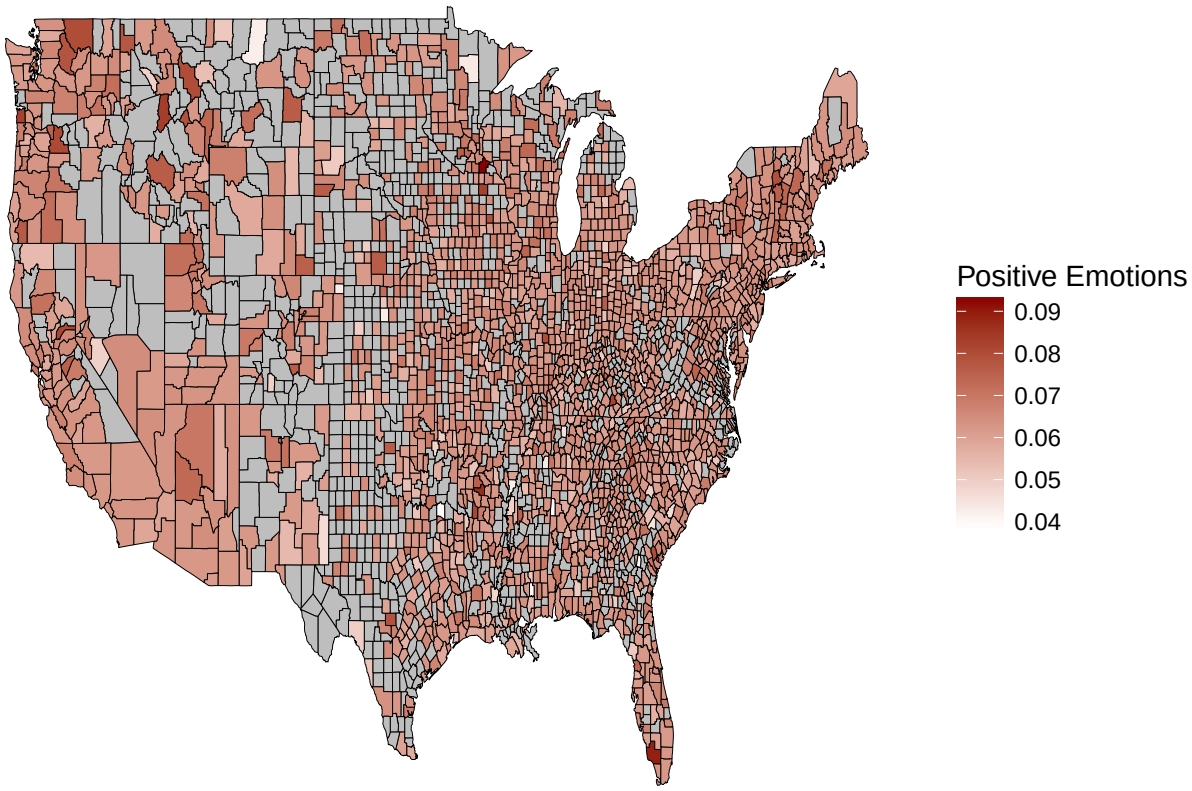
```
  theme_void()
```



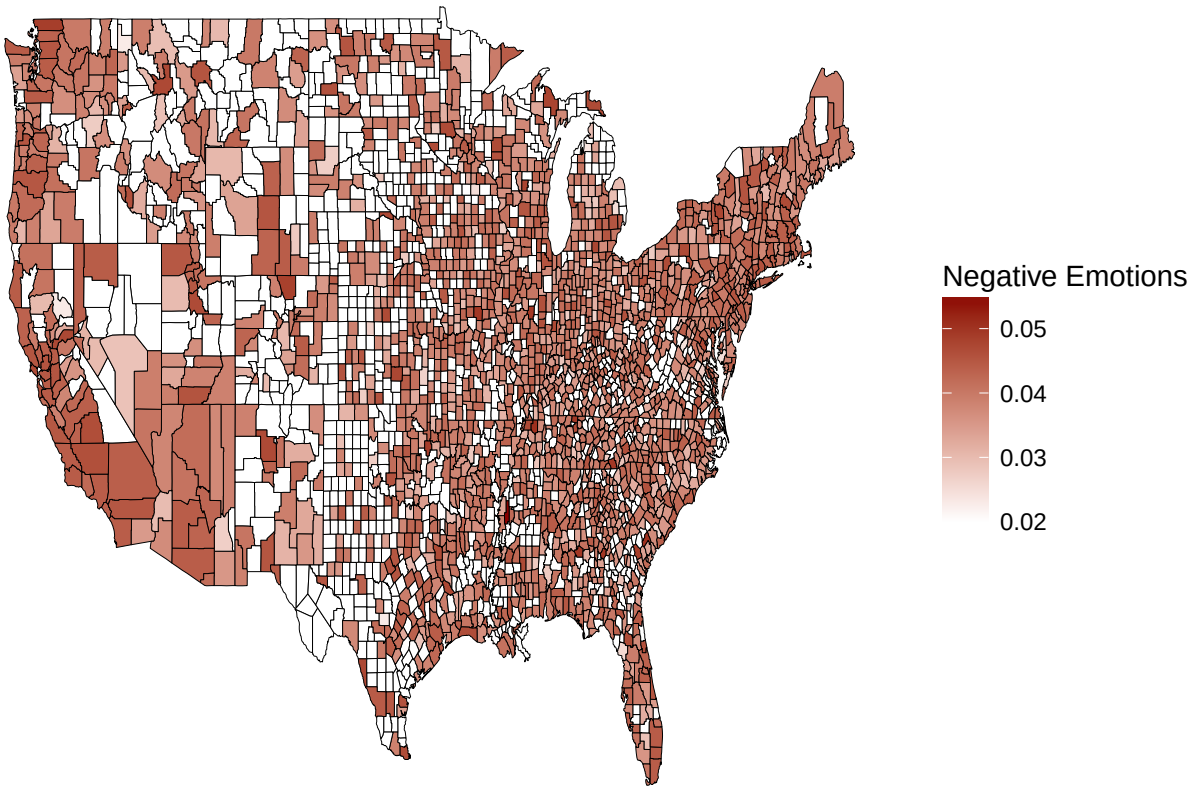

```
#anxiety
ggplot(data.ANX, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Anxiety"))+
  theme_void()
```



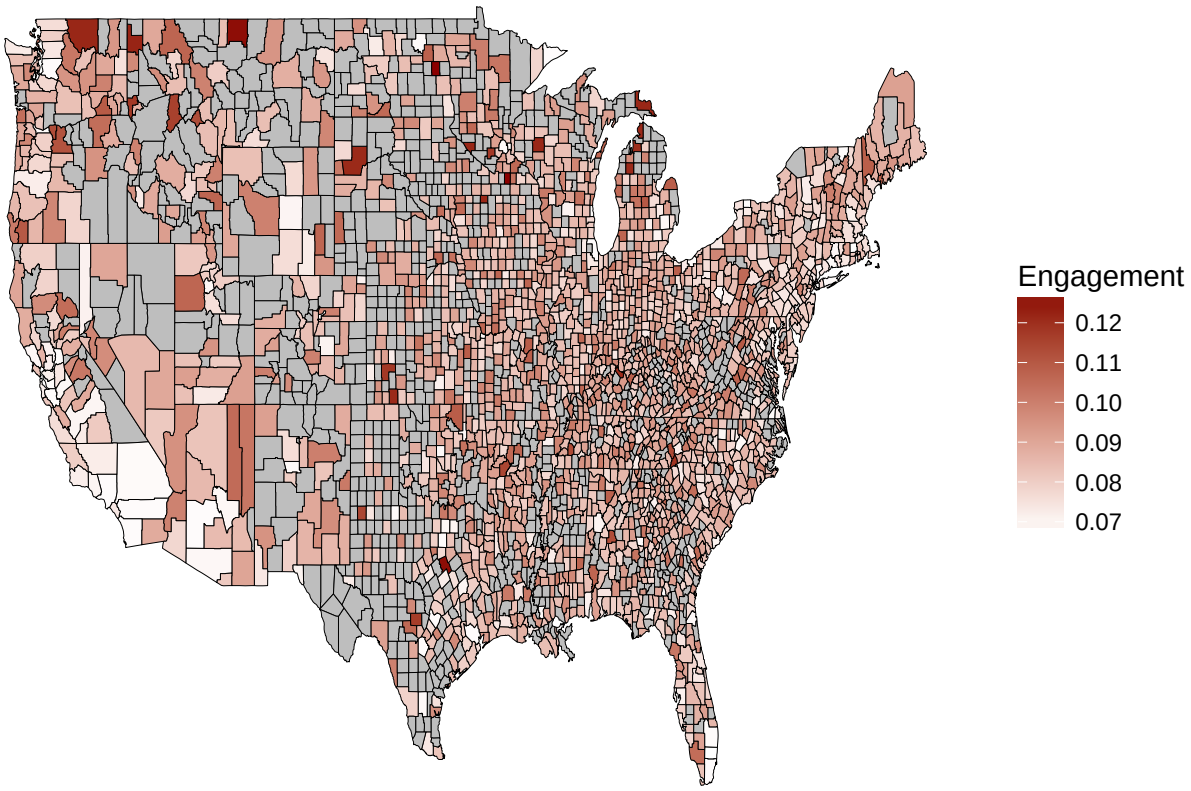

```
#positive emotions
ggplot(data.POSEMO, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Positive Emotions"))+
  theme_void()
```



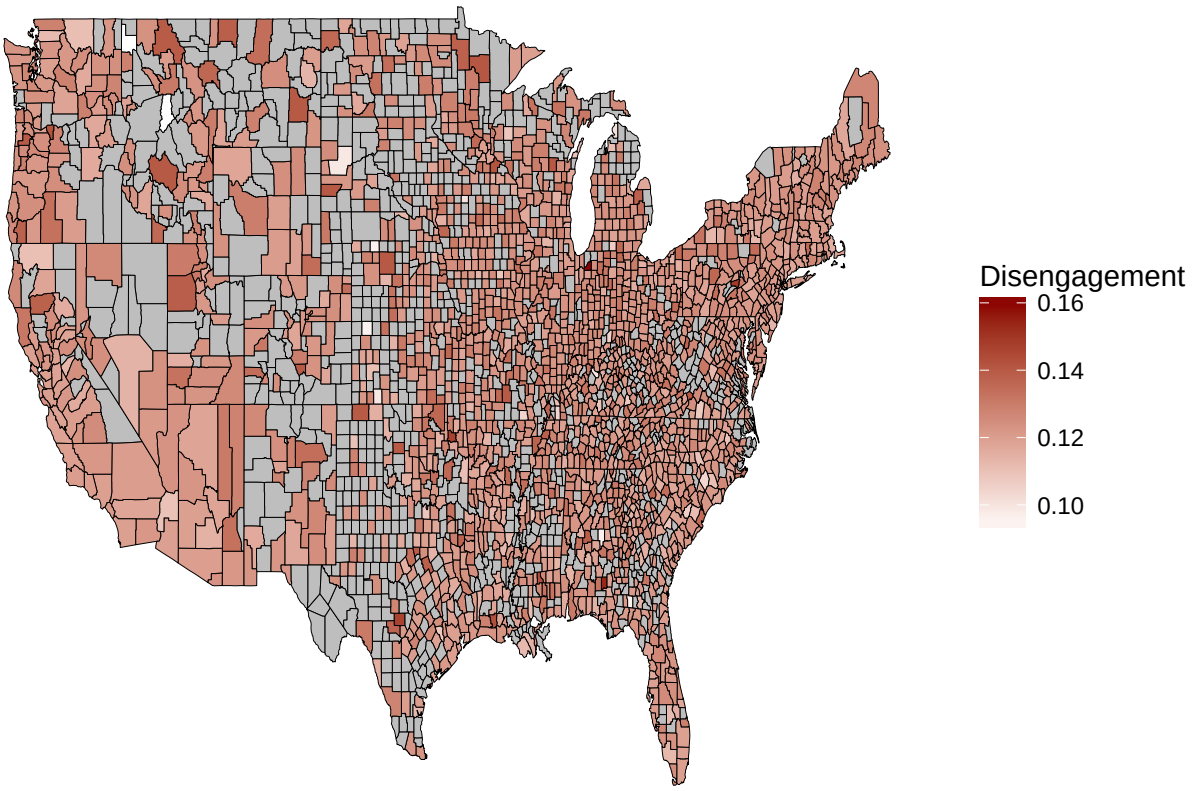

```
#negative emotions
ggplot(data.NEGEMO, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "white",
                      guide_legend("Negative Emotions"))+
  theme_void()
```



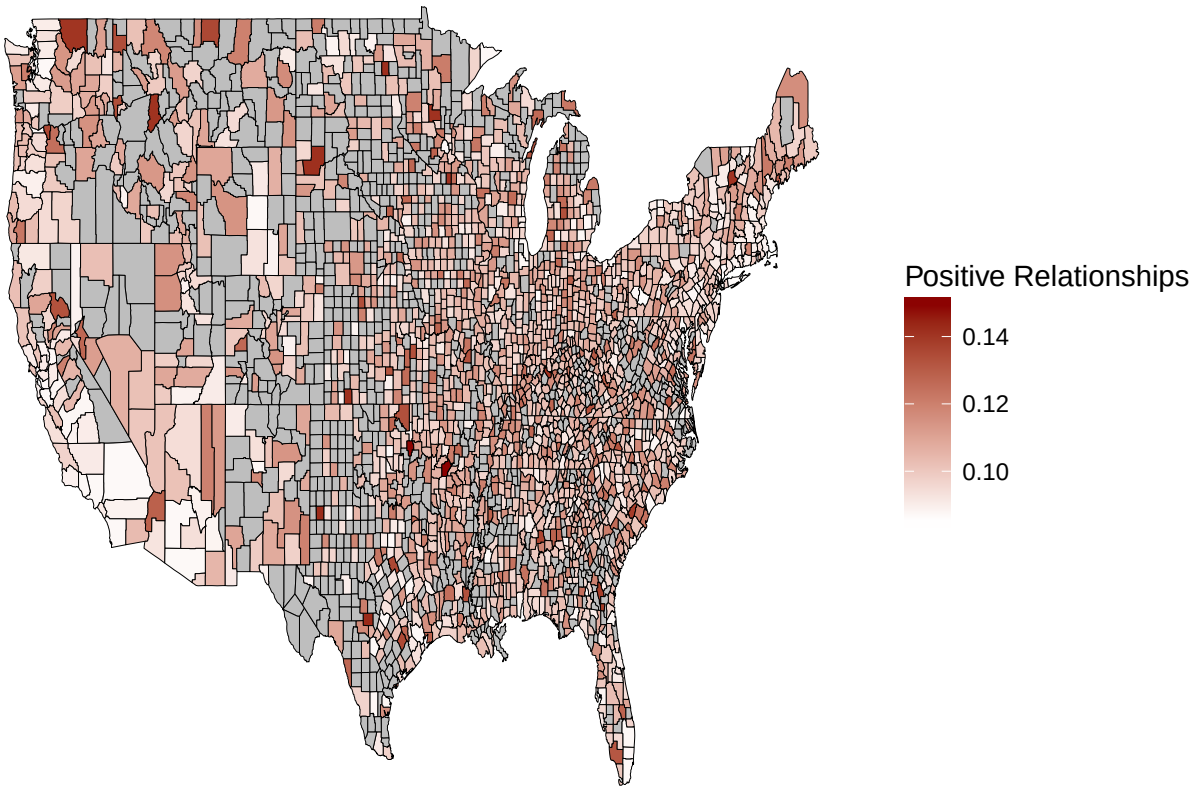

```
#engagment
ggplot(data.ENGAGE, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Engagement"))+
  theme_void()
```



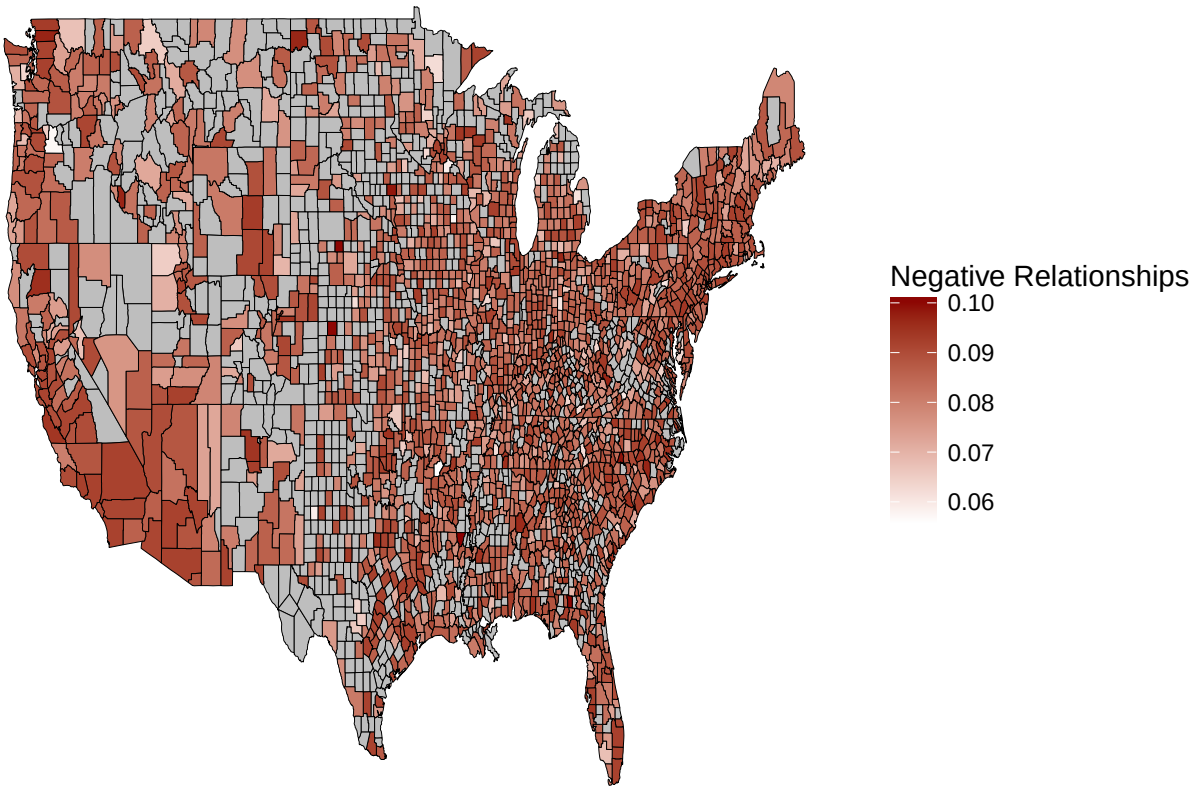

```
#disengagement
ggplot(data.DISENGAGE, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Disengagement"))+
  theme_void()
```




```
#positive relationships
ggplot(data.POSREL, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Positive Relationships"))+
  theme_void()
```




```
#negative relationships
ggplot(data.NEGREL, aes(x = long, y = lat, group = group, fill = rel.emotion))+
  geom_polygon(color = "black", size = .06)+
  scale_fill_gradient(low = "white", high = "darkred", na.value = "gray",
                      guide_legend("Negative Relationships"))+
  theme_void()
```




## References


- Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press.

- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological science, 26, 159-169.

- Fink, C. R., Chou, D. S., Kopecky, J. J., & Llorens, A. J. (2011). Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. Johns Hopkins APL Technical Digest, 30, 22-30.
  Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. Language and Cognition, 7, 265-290.

- Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. Psychological methods, 21, 507.

- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. Science, 323, 721.

- Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and geographic information science, 40, 61-77.

- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5, 1093-1113.

- Oscar, N., Fox, P. A., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 72, 742-751.

- Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology (TIST), 3, 66.

- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

- Poria, S., Gelbukh, A., Cambria, E., Hussain, A., & Huang, G. B. (2014). EmoSenticSpace: A novel framework for affective common-sense reasoning. Knowledge-Based Systems, 69, 108-123.

- Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38, 14799-14804.

- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8, e73791.

- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38.

- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37, 267-307.