RSG
the science of insight

Tolls and Trolls: Analyzing Sentiments from 17 Years of Toll Road Survey Comments

Rachel Schmidt and Tristan Cherry, RSG

June 26, 2018

Qualitative Data Wrangling

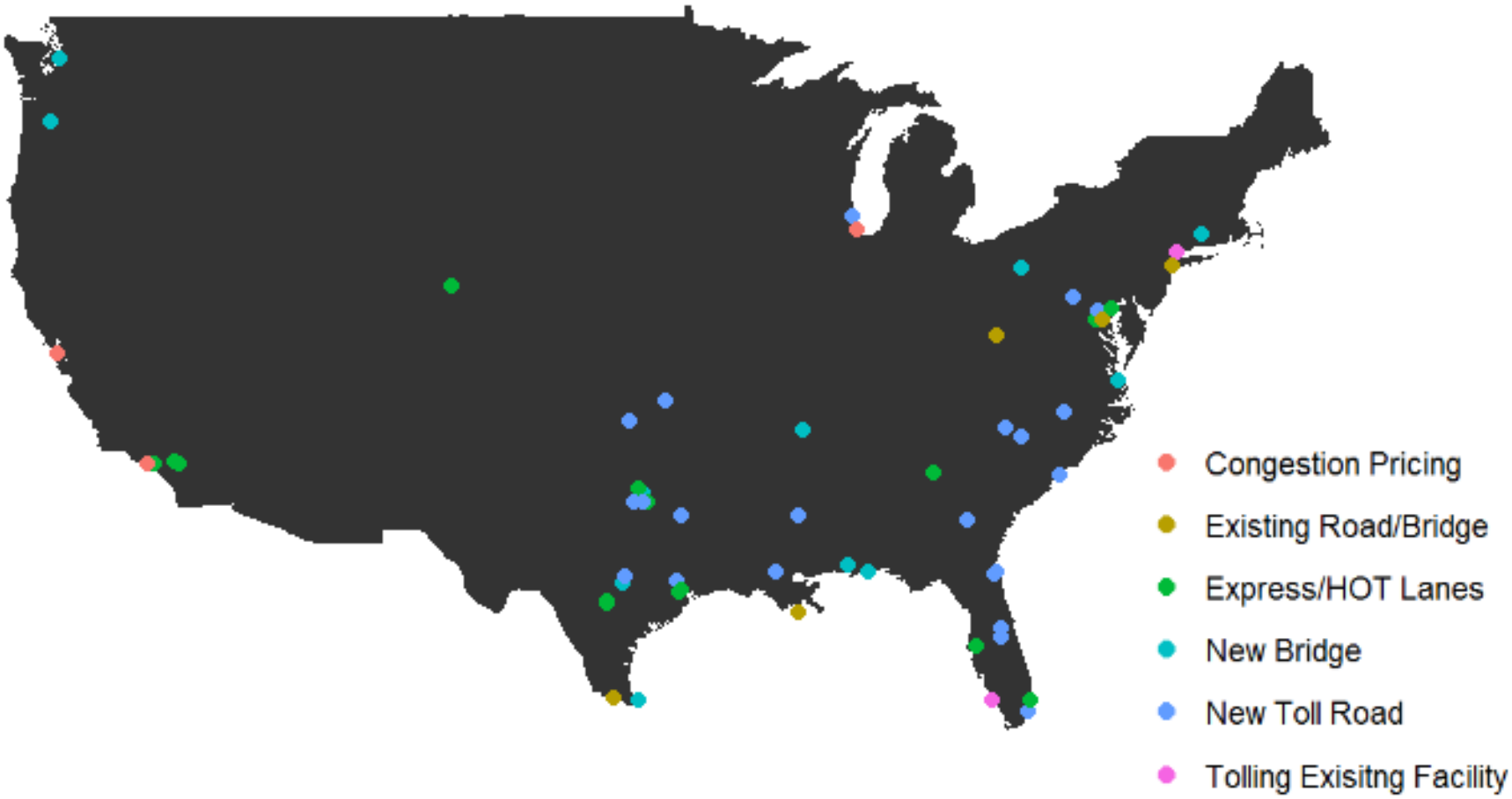
PROBLEM

- Many market research surveys and planning studies produce large volumes of text from open-ended comments
- Assessing and categorizing these data is time consuming and often unproductive

POTENTIAL SOLUTION

- Increasingly accessible, open source (free!) text mining applications and language processing tools have the potential to simplify and automate exploratory analysis

Road Pricing Surveys Since 1999



Data Set Properties

"If you have additional comments or suggestions about the survey, please enter them in the box below."

FUN FACTS

RSG Projects: 94

US States: 22

Number of Comments: 52,591

Words: 2,238,818

Average comment length: 42.6 words

Shortest comments: Tennessee (avg. 32.6 words)

Longest comments: Connecticut (avg. 58.8 words)

Analyzing Sentiment

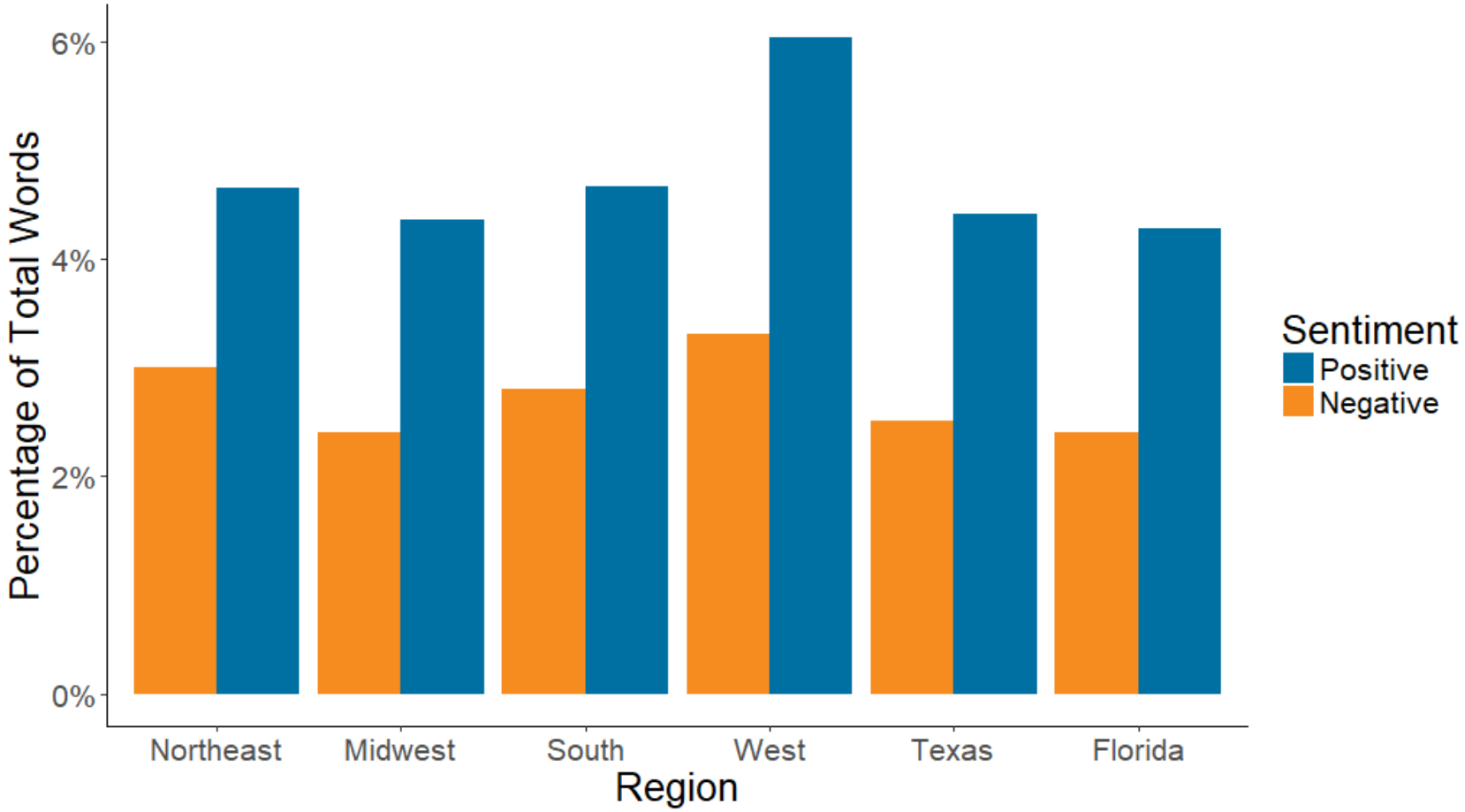
THE GIST

- Extracting emotional intent from text
- Often uses a lexicon to parse words into positive and negative scales, or group words into emotional categories

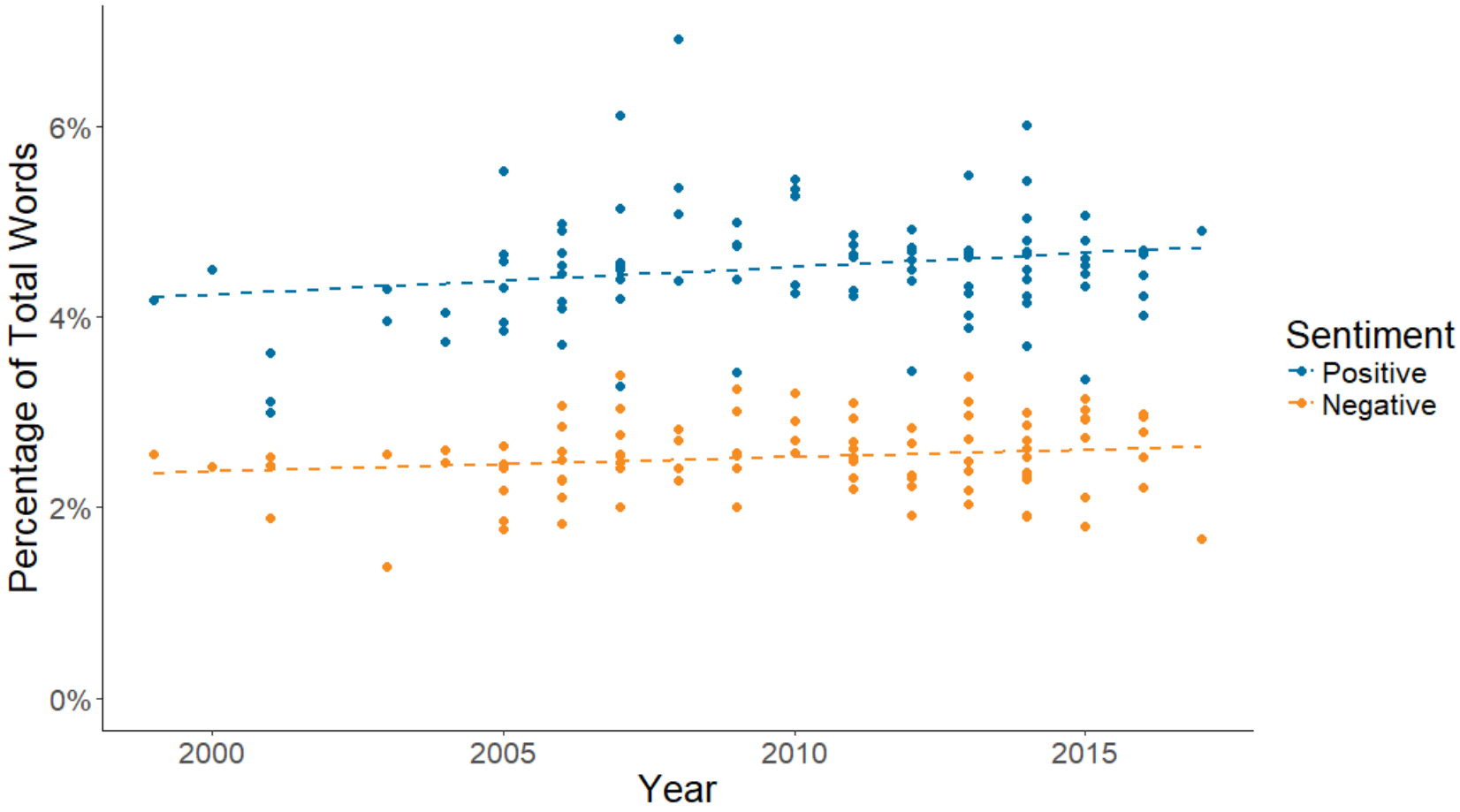
R PACKAGES

- **Quantitative Discourse Analysis Package (qdap):** Suite of functions centered around corpora and document-term matrix data structures
- **tidytext:** Applies tidy data principles and relational joins

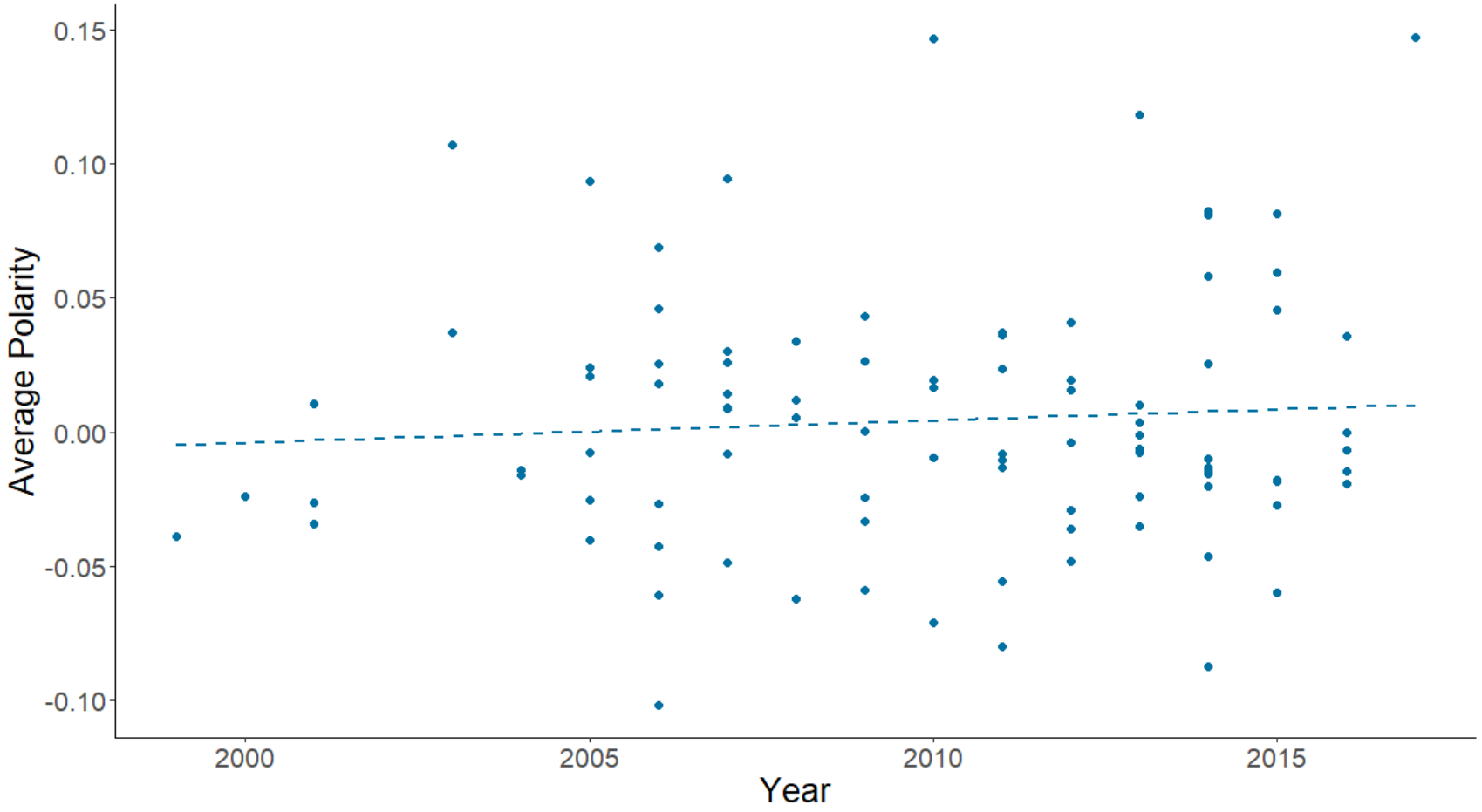
Sentiment by Region



Sentiment over Time: tidytext



Sentiment over Time: qdap



RSG Analyst v. Machine Classification

ASSESSMENT OF ACCURACY

- A sample of 2,175 comments from 5 projects with comments categorized into positive, negative, or neutral
- Compared to qdap and tidytext

Method	% of Classifications Consistent with RSG Analyst
tidytext: Bing lexicon	42%
tidytext: Afinn lexicon	39%
tidytext: NRC lexicon	30%
qdap's polarity function	41%

Conclusions

OPEN SOURCE LANGUAGE TOOLS

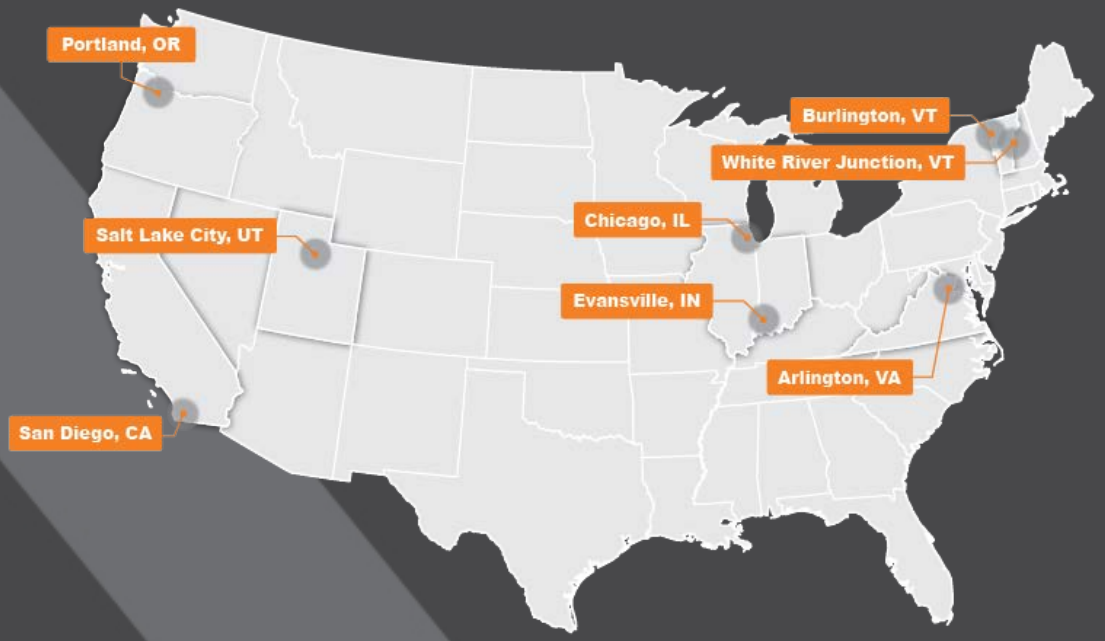
- tidytext has a simpler overall approach, and a familiarity with tidy principles and dplyr goes a long way
- qdap has steeper data processing and computing requirements

QUESTIONABLE ACCURACY

- Neither approach does a great job classifying tokens or assessing sentiment

BUT…

- Overall patterns through time seem intuitively correct
- How you pose the questions influences the types of responses you get (duh)



Contacts

RACHEL SCHMIDT
ANALYST

Rachel.Schmidt@rsginc.com

TRISTAN CHERRY
CONSULTANT

Tristan.Cherry@rsginc.com

www.rsginc.com