# Aggregating Information Based on Geolocated Twitter Data

Brian Mitchell

April 30, 2016

University of Minnesota, Morris

# Introduction

# Introduction

Surveys

- ☺ Representative
- ☺ Accurate demographics
- ☹ Costly
- ☹ Difficult to scale
- ☹ Lower response rate
- ☹ Infrequent
- ☹ Active participation

Twitter

- ☹ Non-representative
- ☹ Inaccurate demographics
- ☺ Inexpensive
- ☺ Easy to scale
- ☺ High-value
- ☺ Real-time
- ☺ Unobtrusive

# Outline

# Background

Twitter

# Twitter

- Microblogging social network
- 320 million monthly active users
- 80% of users active on mobile
- 140 characters
- Mentions, retweets, location, timestamp, images, polls, and links

# Twitter

Location in Twitter

- Opt-in feature
- 3-5% adoption
- `place_id`
  - Bounding box of coordinates
  - Precise coordinate if given
  - Neighborhood, city, point of interest
- User defined location on profile
  - Not validated
  - String of text

# Background

Bias

# Bias

*Twitter population does not match the general population*

- Higher rates of usage in some demographic groups
  - Young in age
  - Urban and suburban
  - African-American
- 75% “local”
- Well-educated people in occupations of management, business, science, and arts are more likely to include location

# Applications

Disaster Management

Tracking keywords and creating stories for social media curation

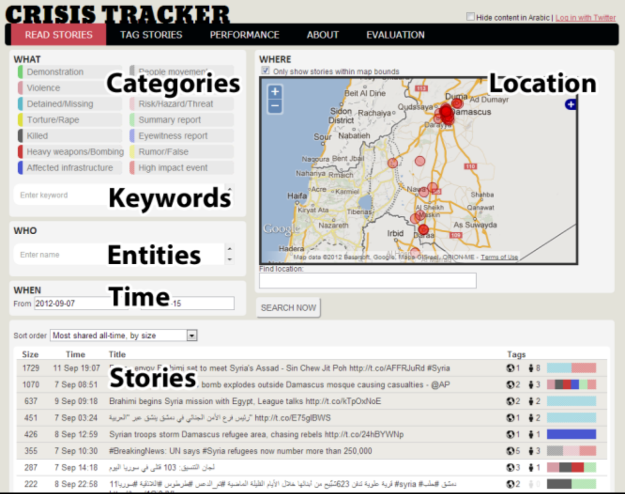



Figure 1: Rogstadius et al. and Ikawa et al.

# CrisisTracker: Rogstadius et al.

- Collection of tweets inside bounding box
- Some tweets filtered out (for example, “@username thanks!”)
- New tweets compared as a weighted set of words
- Fed through a similarity metric and locality-sensitive hashing
    - Hashes documents into “buckets” to be made into stories
    - Adapted by Petrović et al. for constant time searching
    - Adapted version can scale to huge numbers of documents (over 160 million)
- Stories
    - Timestamps, keywords, and number of users
    - 5,000 users who tweet most frequently to omit jokes, opinions, and summary articles

# CrisisTracker: Ikawa et al.

Adaptation by Ikawa et al. to infer locations from similar messages and classify messages based on the location

Four location types:

1. Locations in text
2. Focused locations
3. User's current location
4. User's location profile (home location)

**GeoNames**: geographical database with over 8 million names and coordinates

Location Name Recognition for finding locations in the text

Toponym Resolution assigns locations a coordinate

## CrisisTracker: Ikawa et al.

Confidence score: location popularity $\times$ region context

Location popularity: population of the location

Region context: focused locations included in the text

Highest confidence score $\rightarrow$ toponym resolution

*Evaluating of Location Name Recognition and Toponym Resolution*

- Subset of cities in Syria with a population over 15,000 from GeoNames
- Place names extracted by hand for a gold standard

# CrisisTracker: Ikawa et al. Evaluation

| | Country | State | City/ Town | Village | Total |
|---|---|---|---|---|---|
| #appearance | 250 | 39 | 41 | 12 | 342 |
| #unique | 20 | 7 | 11 | 8 | 46 |
| Precision | 0.996 | 1.000 | 1.000 | 0.917 | 0.994 |
| Recall | 0.992 | 1.000 | 0.927 | 0.750 | 0.977 |

#appearance: the total number of locations
#unique: number of locations after the removal of duplicate elements
Precision: how successful the technique is at finding known relevant data
Recall: how completely the technique finds relevant data

# CrisisTracker: Ikawa et al. Evaluation

- Accurate
- Faster than finding by hand
- To imporve performace:
    - Better geo-inferencing
    - Additional data sources

# Applications

## Migration Trends

# Migration Trends

Measuring migration flows

- Inconsistent
- Outdated
- Sometimes nonexistent
- Often limited to census years
- Needs to be normalized across data sources

# Migration Trends: Data and Pre-Processing

- Zagheni et al.
- Tweets from 500,000 users in OECD countries
- May 2011 to April 2013
- Oversampling in countries with low mobility
- Undersampling in countries with high mobility
- Fraction of users with geolocated tweets outside of their home country
- Users sampled with a probability inverse to the fraction
  - Country A: 50% of users posted tweets from a foreign country
    Country B: 5% of users posted tweets from a foreign country
    B would need a sample about 10 times larger than A
- Age and gender estimated with Face++

# Migration Trends: Difference-in-Differences

$m_c^t$: out-migration rate from country $c$ to all other countries at time $t$

$m_{oecd}^t$: average migration rate at time $t$ for all considered OECD countries

Estimator allows for change in the Twitter users' population if it is similar to the population change in OECD countries:

$$\hat{\delta} = (m_c^t - m_{oecd}^t) - (m_c^{t-1} - m_{oecd}^{t-1})$$

# Migration Trends: Results

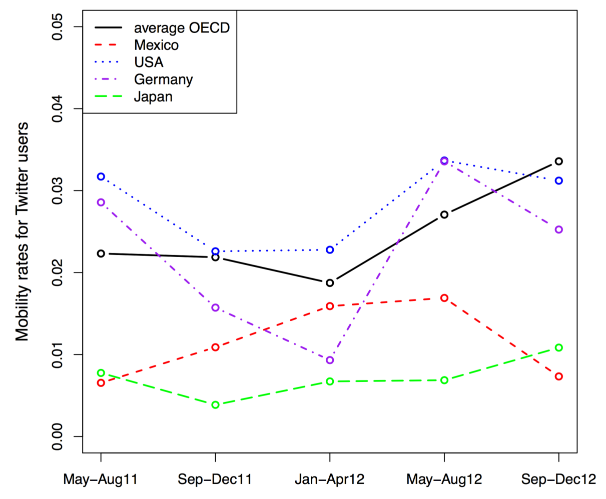



Figure 2: Zagheni et al.

- Estimating recent trends
- Some randomness or noise
- No available official data for training

# Applications

## Societal Happiness

# Societal Happiness

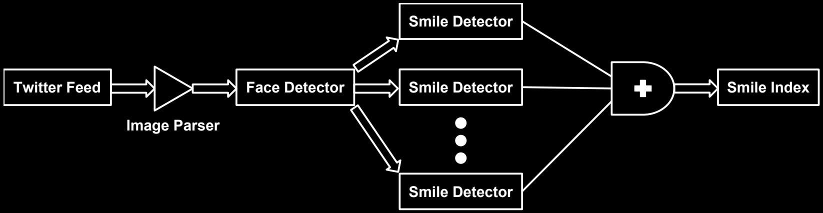
Measuring well-being

- Satisfaction with life (SWL) score
- Gross National Happiness (GNH)
- World Happiness Report

Data from polls, surveys, and other self-reporting

- Abdullah et al.
- 9 million tweets from Twitter's "garden hose" from January 1, 2012 to April 30, 2013
- Tweets with images uploaded via Twitter's official photo-sharing service
- Location is from the tweet, not photo

# Societal Happiness: Smile Index Framework



# Societal Happiness: Smile Index Framework

$$\text{Ratio} = \frac{\text{raw smile count at given location}}{\text{total number of images at given location}}$$

# Societal Happiness: Results

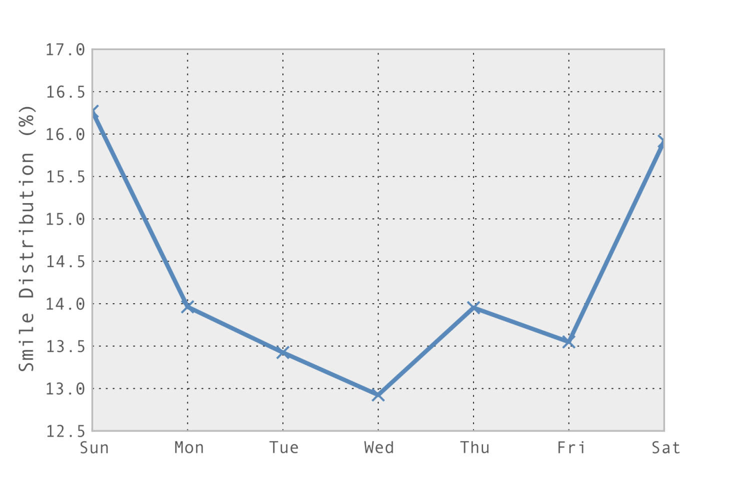



Figure 3: Abdullah et al.

# Societal Happiness: Results

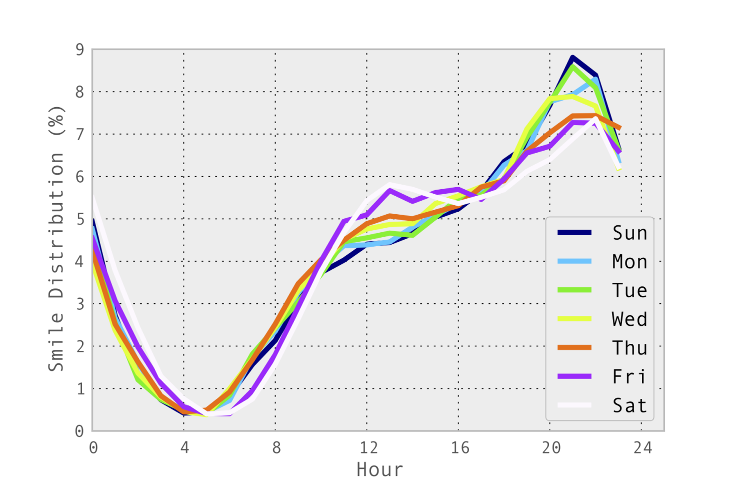



Figure 4: Abdullah et al.

- Daily and hourly results in line with prior research
- Increased response for celebratory events and holidays
- Decreased response for tragedies and disasters
- Cultural variance was not a significant hindrance

Future work

- Additional data sources
- Further investigation of using images for sentiment analysis

# Discussions & Conclusion

# Discussions & Conclusions

Twitter as a data source...

- High volume
- Immediate
- Biased
- Assumptions for demographic information
- Bad for small scale
- Useful for large-scale patterns and trends

# Questions?


S. Abdullah, E. L. Murnane, J. M. Costa, and T. Choudhury. Collective smile: Measuring societal happiness from geolocated images. 2015.

Y. Ikawa, M. Vukovic, J. Rogstadius, and A. Murakami. Location-based insights from the social web. 2013.

J. Rogstadius, M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo. Crisistracker: Crowdsourced social media curation for disaster awareness. 2013.

E. Zagheni, V. R. K. Garimella, I. Weber, and B. State. Inferring international and internal migration patterns from twitter data. 2014.

# License

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

---

This presentation is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.