PREDICTING WELL-BEING WITH GEO-REFERENCED DATA COLLECTED FROM SOCIAL MEDIA PLATFORMS

JOÃO LOFF, MANUEL REIS AND *BRUNO MARTINS*

UNIVERSITY OF LISBON AND INESC-ID

ACM SAC 2015

TÉCNICO LISBOA

MOTIVATION

How can we know how happy people are?

- Subjective well-being has been widely studied in psychology and related disciplines
- Measuring well-being can help individuals, organizations, and governments choose policies that are not just in the best economic interest, but which truly improve well-being
- Well-being is being tracked by governmental agencies and by private surveying organizations, such as Gallup-Healthways
    - Traditional surveying methodologies (i.e., expensive, coarse-grained temporal and spatial resolutions, …)

OUR WORK

Measuring well-being based on language in social media

- We report on a initial study leverage existing resources:
  - Large amounts of geo-referenced Twitter data
  - Existing lexicons associating words to emotion ratings
  - Data from previous well-being surveys (Gallup-Healthways)
- Learn predictive models for estimating well-being with basis on variables (i.e., word counts) derived from textual contents

OVERVIEW

- Motivation
- The Proposed Work
- *Related Work*
- Estimating Well-Being
- Experiments and Results
- Conclusions
- Future Work

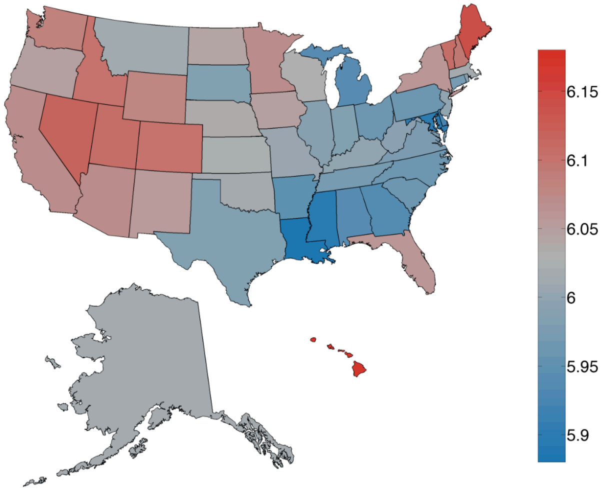
RELATED WORK (1)

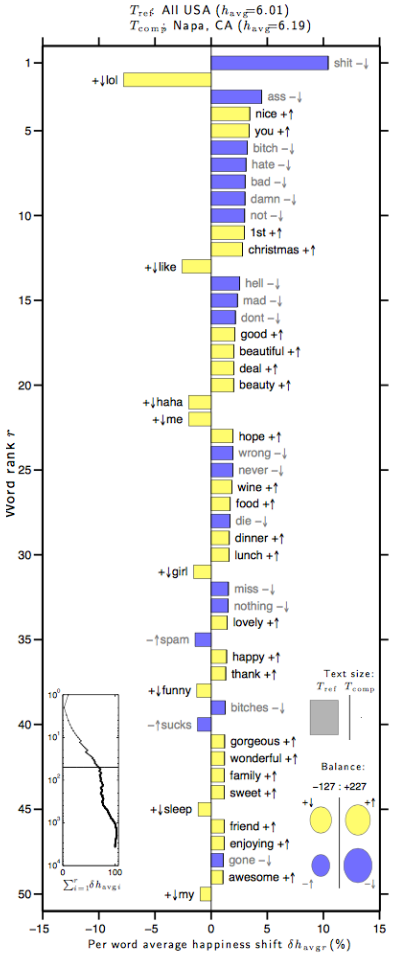
The Hedonometer Project - ***http://hedonometer.org/***

Crowdsourcing methodology to collect human evaluations on the "*happiness*" of words.

Simple procedure for extrapolating ratings into textual corpora (e.g., tweets from a given month/region).

…An instrument that measures the happiness of large populations in real time!





RELATED WORK (2)

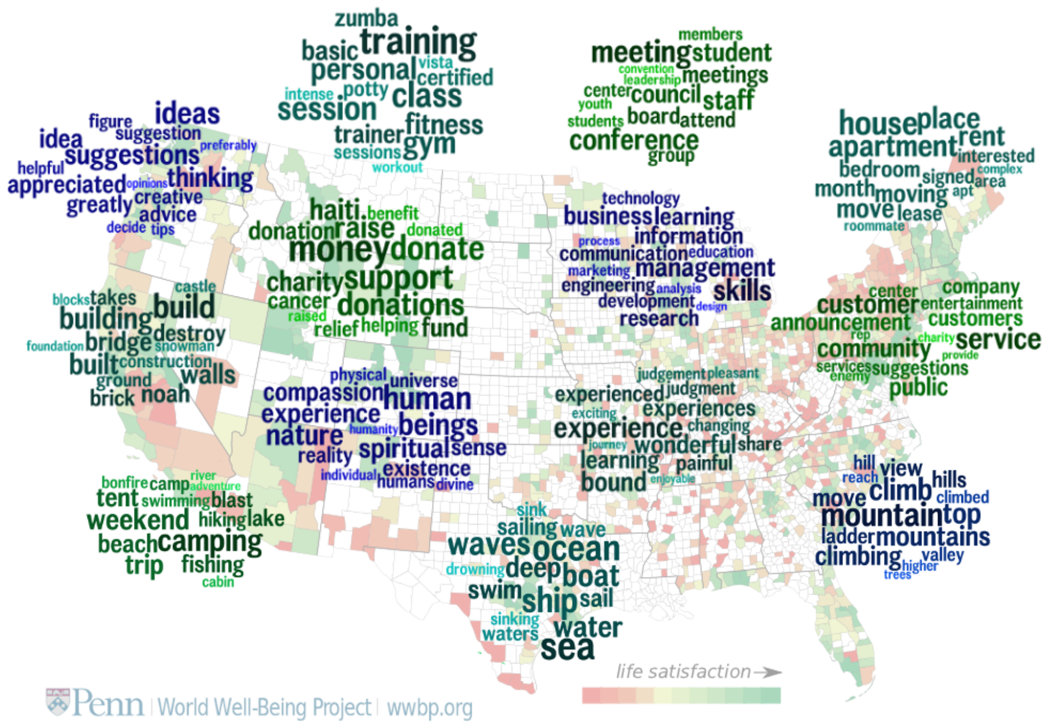
The World Well Being Project - http://wwbp.org

Language used in tweets from different U.S. counties was able to predict the results from a well-being survey.

Other studies within wwbp:

- ***Psychological Language on Twitter Predicts County-Level Heart Disease Mortality***
- ***Towards Assessing Changes in Degree of Depression through Facebook***

Characterizing Happy Communities:



OVERVIEW

- Motivation
- The Proposed Work
- Related Work
- ***Estimating Well-Being***
  - Lexicons considered in this study
  - Features representing well-being within particular geo-spatial regions
  - Predictive modeling
- **Experiments and Results**
- **Conclusions**
- **Future Work**

EMOTION LEXICONS

Our approach is based on counting words over tweets…

- **Affective Norms for English Words (ANEW) Lexicon**
  - A total of 1,034 English words rated by humans according to:
    - **Valence**, pleasantness of the stimulus (i.e., from *happy* to *unhappy*)
    - **Arousal**, intensity of feeling (i.e., from *excited* to *relaxed*)
    - **Dominance**, how much the reader feels *in control*
  - Adapted into other languages (e.g., Spanish)

Behavior
Research
Methods

BRM

- **LabMT Lexicon from the Hedonometer**
  - A total of 10,222 English words rated according to happiness (i.e., valence in the ANEW study) through crowdsourcing
  - Consistent with ratings from the ANEW study

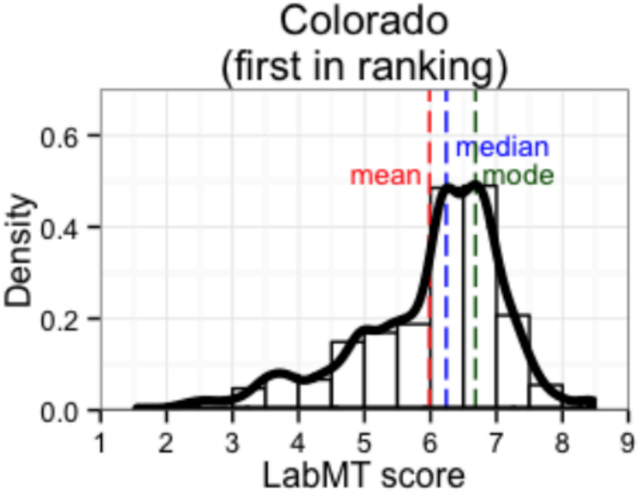
FEATURES LEVERAGING THE LEXICONS

Extrapolate word ratings into tweets… for each dimension:

$$v_{tweet} = \frac{\sum_{i=1}^{n} v_i \times f_i}{\sum_{i=1}^{n} f_i}$$

Compute features for geo-spatial regions, with basis on the corresponding geo-referenced tweets

- Features capturing the distribution of the tweet ratings
  - Mean, median, mode, standard deviation, …
  - Also capturing num. of tweets
- A total of 46 features



PREDICTIVE MODELING

Regions are represented as 46-dimensional feature vectors…

Regions are associated to well-being scores, as obtained through traditional surveys…

Regression modeling for estimating well-being:

- Linear least-squares regression modeling

$$y = Xb + e$$

- Model training with Elastic Net regularization approach

$$b = \arg\min_{b} ||y - Xb||^2 + \lambda_1 ||b||_1 + \lambda_2 ||b||_2^2$$

EXPERIMENTAL METHODOLOGY

Large collection of Twitter data geo-referenced to the U.S. territory

- Approximately 500,000 tweets from the year of 2012
- Tweets containing words from the lexica used in our study
- 48 states in continental U.S. (i.e., except Hawaii and Alaska)

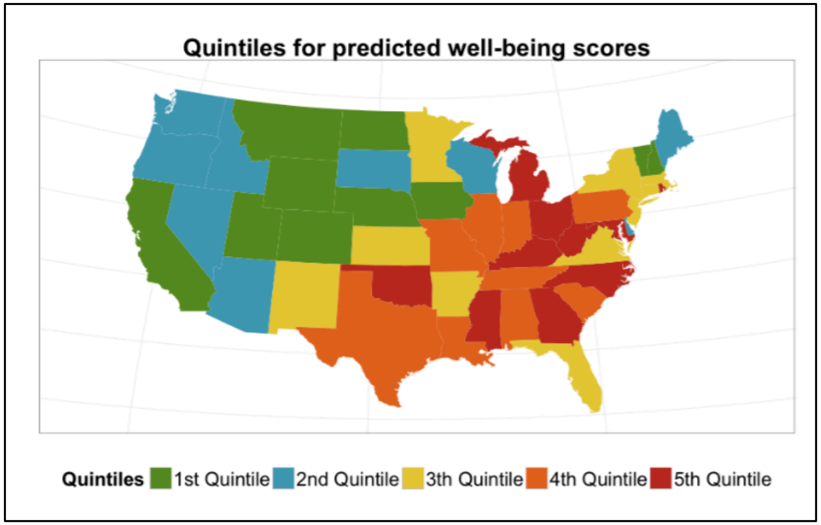
Gallup-Healthways Well-Being Index for 2012

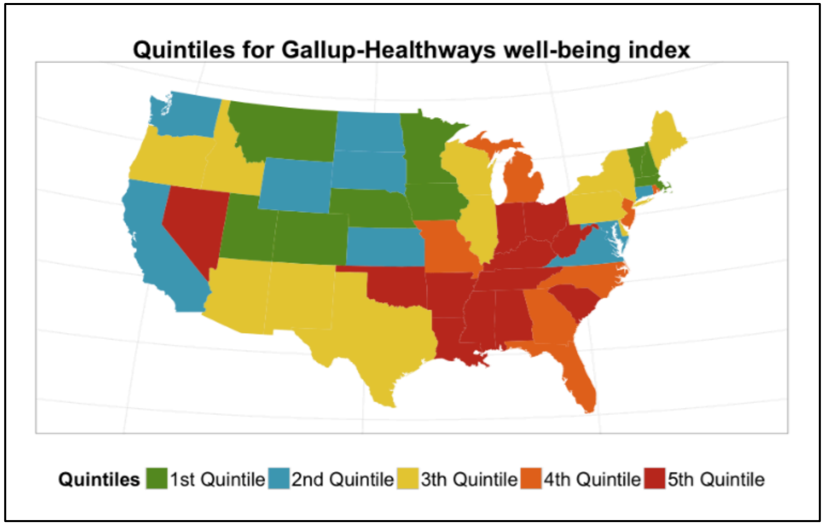
- Phone interviews with approx. 1,000 individuals (7 days/week)
- National average of 66.5 in 100.0 (61.3 in West Virginia ; 69.4 in Colorado)

Evaluation through leave-one-out cross-validation

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Pearson's correlation (ρ)
- Kendall's correlation (τ)

EXPERIMENTAL RESULTS (1)


Quintiles for predicted well-being scores

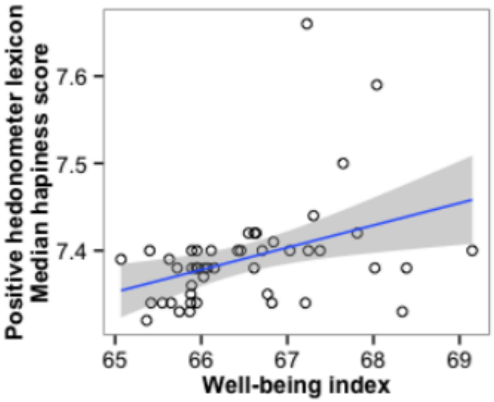

Quintiles for Gallup-Healthways well-being index

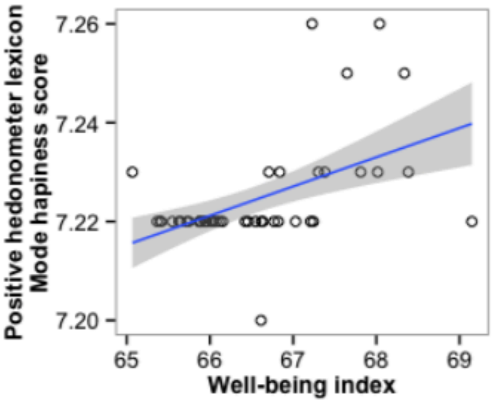
- Correlations of $\rho = 0.7441$ and $\tau = 0.5862$
  - Study from the *wwbp* reported on slightly inferior values on data from 2010
- Low errors of MAE=0.92 and a RMSE=1.22
  - Assigning average corresponds to MAE=1.40 and RMSE=1.73
- Errors in ranking states like Maryland, Minnesota and Nevada…

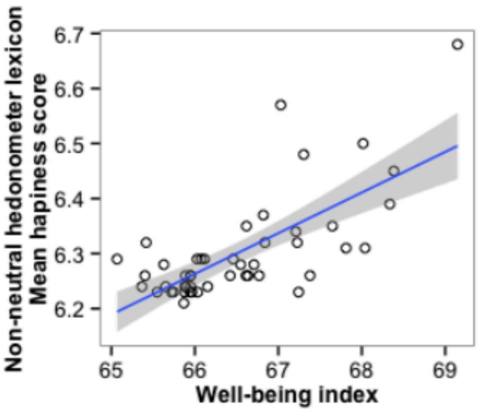
EXPERIMENTAL RESULTS (2)

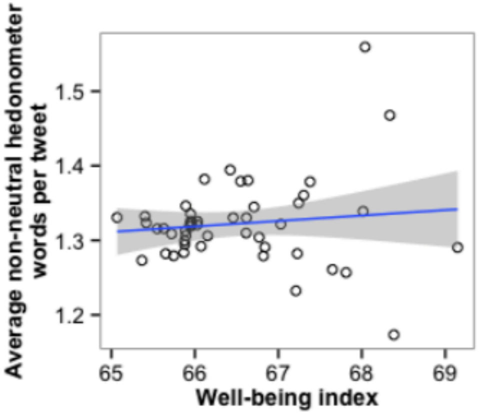
Elastic Net regularization:

- Only 28 features were informative for the regression model (of 46)
- Most of the features with positive values in the estimated regression coefficients were obtained from the hedonometer lexicon.
  - mode of the happiness score obtained from filtered version of lexicon, only considering non-neutral words









MAIN CONCLUSIONS

- **We evaluated a simple approach for estimating well-being through predictive models leveraging Twitter data**

- **Promising results in terms of correlations towards existing well-being surveys, although much remains to be done:**
    - Check if our predictive models generalize well to other years and/or across geographic regions
    - Additional variables for accounting with Twitter's demographics

FUTURE WORK

Increasing the number of geo-referenced tweets

- Explore automated geo-coding methods
- Important for thin-grained spatial-temporal resolutions

Estimate happiness ratings for more tweets

- Use distributional representations for words / documents
- Unsupervised embeddings (e.g., word2vec)
- Leverage ANEW-like lexicons for building predictive models for rating new words and/or documents

Other application domains besides tracking well-being

- Public health surveillance, public opinion, political pools, etc.
- See for instance www.popstar.pt

THANKS FOR YOUR ATTENTION!

Predicting Well-Being With Geo-Referenced Data Collected from Social Media Platforms

Research work financed by FCT through projects

KD-LBSN, DATASTORM and through INESC-ID's multi-annual funding

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Bruno Emanuel Martins

bruno.g.martins@ist.utl.pt