

The World Well-Being Project: Psychological Insight through Language Analysis

Gregory Park[‡] Johannes C. Eichstaedt[‡] H. Andrew Schwartz[†]

Margaret L. Kern[‡] Martin Seligman[‡] Lyle H. Ungar[†]

[†]Computer & Information Science, [‡]Psychology

University of Pennsylvania

{gregpark, jeich, hansens, mkern, seligman, ungar}@sas.upenn.edu

The World Well-Being Project (WWBP) is an interdisciplinary team of psychologists, computer scientists, and statisticians developing techniques to measure well-being through the language of social media. These techniques provide a new way to study a wide range of phenomena across individuals and communities. Ultimately, we hope that our data-driven approach will help individuals, organizations, and governments choose actions and policies that will improve psychological, social, physical, spiritual, and economic well-being for people worldwide.

Social scientists typically use questionnaires to study the well-being of people and their communities. However, questionnaires and other traditional survey methods are costly and time-consuming. We suggest the huge volume of language produced on social media, like Facebook and Twitter, is rich with psychological information and can provide an alternative survey methodology.

Across individuals and communities, we have found that social media language can yield accurate prediction models as well as new psychological insights. In the largest study of personality and language use to date [5], we used natural language processing techniques to learn the language of personality traits, gender, and age from a massive social media data set: language from over 15.4 million Facebook messages collected from over 70,000 volunteers (Stillwell, Kosinski, & Graepel, 2013). Predictive models of personality based on this language produced state-of-the-art accuracy when compared to users' responses to gold standard personality questionnaires. In addition, we used differential language analysis (DLA) to identify the words, phrases, and language topics that distinguished several groups of individuals, confirming many results from previous psychological research but also making several new discoveries [1, 2, 3].

For example, Figure 1 shows how the language of *emotional stability*—low levels of neuroticism and less proneness toward depression and anxiety—is characterized by several activities and mindsets that may foster greater emotional stability, including athletics (e.g., *workout*, *volleyball*, *basketball*), gratitude (e.g., *blessed*, *life is good*, *praise*), and

Emotional stability

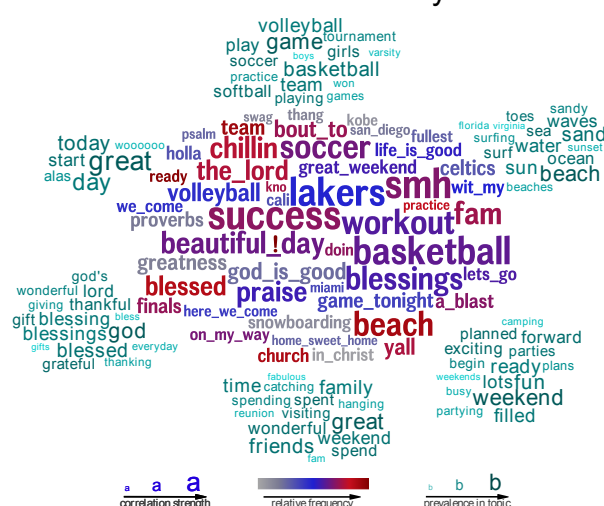


Figure 1: The language of emotional stability—characterized by words about athletics, gratitude, and nature—revealed by applying differential language analysis to the language written by over 70,000 Facebook users.

nature (e.g., *beach*, *sun*, *beautiful day*). The potential to discover new connections between language and underlying traits makes DLA a powerful tool for generating insight and hypotheses.

At the community level, we have used language from Twitter to predict the average life satisfaction of U.S. counties [4]. By combining county-level representative surveys with over 800 million tweets, we built predictive model of life satisfaction based on the word and phrase use across 1,300 counties. Resulting predictions from this model were accurate, outperforming models based on traditional predictors of county well-being (e.g., demographic and socioeconomic factors). Combining traditional predictors with language resulted in the even higher accuracy (see Figure 2). Similar approaches can potentially complement and extend the traditional survey methods currently used to monitor regional well-being.

Using DLA to identify the language most associated with county-level life satisfaction brought to the foreground many processes and correlates that were expected in light of the existing psychological literature, and some that are less fre-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGKDD 2014 New York, New York USA

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

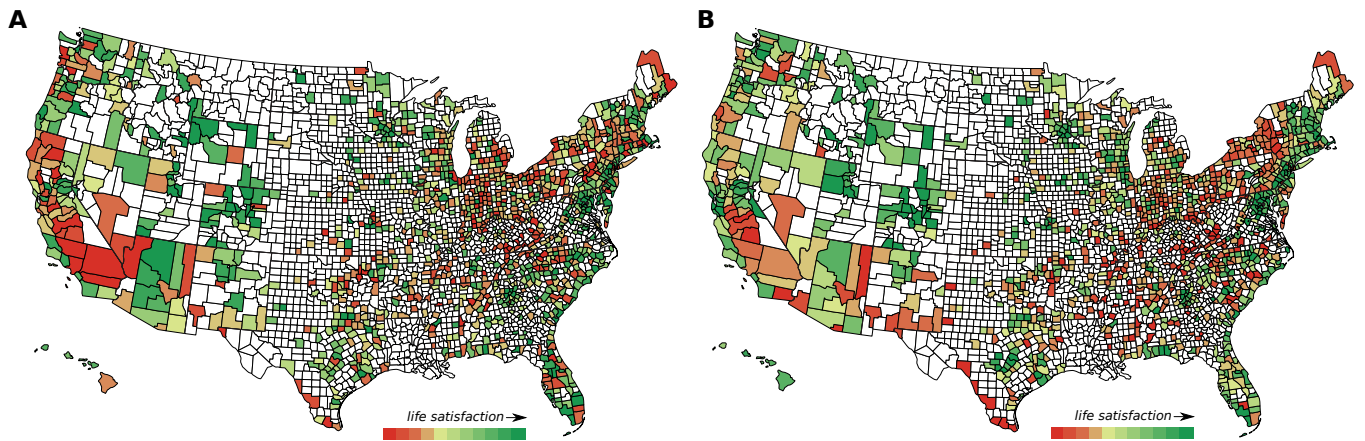


Figure 2: Map of county life satisfaction as measured (A) using survey data and (B) as predicted using a model combining language features from Twitter and demographic factors. Green regions are higher satisfaction, while red are lower. White regions are those for which the survey sample size is too small to have valid measurements. ($r = 0.535$, $p < 0.001$)

quently discussed. The language of high life satisfaction reflected several distinct themes, including physical activity (*training, class, session, gym*), activities that are likely associated with high socioeconomic status (*money, ideas, meetings, management*) and engagement (*experience, bound, wonderful*). Interestingly, the themes of money that emerged often placed it a pro-social and philanthropic context (*donate, charity*), highlighting how language use not only captures traditional predictors of well-being such as socioeconomic status, but also characterizing what it is about these other predictors that uniquely connects them to life satisfaction. In the case of money, our analyses suggests that how one spends it—and not simply having lots of it—is particularly important to life satisfaction.

Through collaborations with leading experts, we plan to expand this approach to study other aspects of psychological and physical well-being: depression and anxiety, disease, and healthy behaviors. In addition to providing a new means to track these factors, we hope our language analyses will generate additional insight into pathways to greater well-being for individuals and their communities.

Lastly, a major goal of WWBP is to create tools and datasets for the broader research community. We plan to build user-friendly research tools and tutorials to allow social scientists to apply these methods to their own language data. In addition, our language-based models and regional predictions (e.g., county-level life satisfaction and other aspects of well-being) will be made available to researchers who may be interested in including these indicators in their own sociological, economic, or policy-related research.

WWBP aims to improve our understanding of the psychological determinants, correlates, and outcomes associated with psychological, physical, and social well-being, as revealed through social media. By sharing our work with the research community, policymakers, and the general public, we hope to broaden the public discourse on what constitutes a good life, and how one should achieve it.

- [1] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell, M. Kosinski, S. M. Ramones, and M. E. Seligman. The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169, 2014.
- [2] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, M. Kosinski, L. Dziurzynski, and M. E. Seligman. From “sooo excited!!!” to “so proud”: Using language to study development. *Developmental psychology*, 50(1):178, 2014.
- [3] H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, E. Blanco, M. L. Kern, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar. Toward personality insights from language exploration in social media. In *Proceedings of the AAAI Workshop on Analyzing Microtext*, 2013.
- [4] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman, et al. Characterizing geographic variation in well-being using tweets. In *ICWSM*, 2013.
- [5] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.