Research Statement: H. Andrew Schwartz

Postdoctoral Research Fellow, University of Pennsylvania

www.seas.upenn.edu/~hansens/
hansens@seas.upenn.edu
215-746-5085

What can language analyses reveal about human health and well-being?

Upon proudly introducing her Ph.D. relative, a friend once added, "yes, a doctor, but not the kind that helps people."

With the growth of social media, now used regularly by more than one-seventh of the world's population, scientists are presented with an unprecedented resource of objective and quantifiable behavioral and textual data. People freely post their daily activities, feelings, and thoughts in the form of status updates, tweets, and other personal online discourse. This "big data", unlike that from questionnaires or controlled studies, is unprompted and requires no *a priori* theories to collect, and thus opens the door for enormous data-driven discovery in the health and social sciences.

Researchers have begun leveraging natural language data for applications such as monitoring influenza outbreaks, predicting the stock market, and personally targeting advertisements. While such pursuits are exciting and worthwhile, I believe the biggest application of such natural language analyses are toward science itself: discovery of new behavioral and psychological drivers of health and well-being.

Recently, along with colleagues from psychology, medicine, and computer science, I've begun exploring how scalable natural language processing algorithms can be developed to mine social media for scientific discovery about people. Language use features, in the form of *words*, *phrases*, and *topics*,[1] are extracted from Facebook status updates or Tweets, sometimes dealing with billions of messages using distributed computing.[2] Once language use features are extracted, we perform correlational analyses and machine learning for *(a)* prediction of outcomes (e.g. personality, heart disease, life satisfaction) from social media messages and *(b)* insights from language about the behavioral and psychological factors of well-being and health.

Prediction. My research shows that *words, phrases,* and *topics* in Tweets and Facebook posts are predictive of many health and psychological outcomes. Language from Twitter significantly improved accuracy in a predictive model of county life satisfaction, as measured by the CDC, over and above standard demographic and socio-economic controls (age, gender, ethnicity, income, and education; see fig. 1) [1]. Similarly, atherosclerotic coronary heart disease mortality was predicted significantly higher by a twitter language model than by models based on SES, smoking, hypertension, and a combined model with ten socioeconomic demographic and health variables (see fig. 2) [2]. On the individual level, we found personality, gender, age, and life satisfaction were strongly linked with language in status updates from Facebook, and that an *open-vocabulary* analysis[3] resulted in significantly higher predictive accuracies than those from traditional closed-vocabulary approaches [3, 4, 5]

Insight. Results from our predictive models of health, well-being, and personality speak to the potential of language to predict many health and psychologically-relevent outcomes. However, language can be used for more than prediction; A parallel goal is to gain a greater understanding of the nature of people. The problem changes from obtaining a single high prediction accuracy to that of describing the many links between language and outcomes. The *topics* used to predict county *life satisfaction* reveal insight into what makes a community thrive beyond broad socio-economic and demographic variables. For example, topics relating to outdoor activities, spiritual meaning, philanthropy, exercise, and work correlate with increased life satisfaction (see fig. 3) [1]. In the personality study over Facebook, topics related to laughter, college, getting a job, and having children were characteristic of the different age groups (see fig. 4). We found

[1] ***phrases***: words more likely to occur together than by chance (sometimes called *collocations*; e.g. "sick of", "a beautiful day").
topics: groups of semantically-related words automatically derived via Bayesian *Latent Dirichlet Allocation* (e.g. see fig. 3).

[2] Utilizing a *Hadoop*-style cluster to distribute feature extraction over many machines.

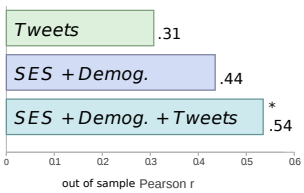

fig 1: Accuracy of predictive models on county life-satisfaction. *Tweets*: topic usage from tweets. *SES + Demog.*: age, gender, ethnicity, income, and education. *SES + Dmog. + Tweets*: SES and topics from tweets. * = sig. ($p < .01$) improvement over *SES + Demog.* [1]

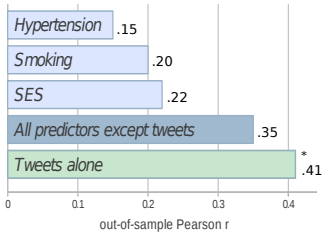

fig 2: Accuracy of predictive models on county heart disease mortality. *Hypertension, smoking:* county rates; *SES:* income and education. *All predictors except tweets*: rates of diabetes, obesity, smoking, hypertension; % black, Hispanic, female, married; income + education. *Tweets alone*: words, phrases, topics. * = sig. ($p < .02$) improvement over *all predictors*. [2]

[3] ***open-vocabulary***: data-driven approaches which discover predictive words rather than relying on *a priori* word or category judgments [3].

emotionally stable individuals mention sports and social activities more, and *conscientious* (organized, responsible, thorough) people do not just mentioned work and planning more but also vacations and weekends [3, 4]. In a followup, we found language-based evidence for the aging positivity hypothesis: that low-arousal positive emotion increases (e.g. 'proud', 'grateful') while negative emotion and high-arousal positive emotion (e.g. 'hate', 'bored', 'excited') decrease across age[4] [6]. Most recently, in the case of heart disease, counties using more angry (e.g. 'hate', 'f*ck') and disengaged (e.g. 'bored', 'tired') language showed significantly higher rates of mortality [2].

Computational Linguistics about People. Natural language data in its raw digital form, a pattern of characters, is not very meaningful. Techniques from computational linguistics (e.g. part-of-speech tagging, syntactic parsing, semantic interpretation, and sentiment analysis) help to make sense of such patterns. However, traditional computational linguistics is primarily concerned with understanding *language* itself rather than understanding the *people* behind the language. Understanding people based on language warrants building new techniques. A new area of information science is emerging at the intersection of computational linguistics, health, and psychology, which blends the development of social science theory with novel algorithms. The challenge is creating methods that are interpretable enough to be trusted while also being sophisticated enough to fully leverage the information available in billions of messages.[5]

In order to gain deeper psychological and behavioral insights, more algorithms and methods are needed to mine the wealth of natural language data. Below, I outline select information science research problems which I plan to tackle.

- **Semantic interpretation for human insights.** Much of my core computational linguistics research has focused on lexical semantics [8, 9]. I plan to explore the use of word sense disambiguation and verb argument structure and roles to garner more precise insights about people. Most words are ambiguous given context and meaning emerges out of patterns of language rather than single tokens; current work which simply looks at words or categorical language is just scratching the surface of data-driven discovery. As as simple example, while we found "sick" to be highly predictive of neurotic (depressive) personality, further investigation revealed it was the sense as in "sick of; hate" that was most predictive. Work would build on existing semantic analysis techniques to more appropriately handle social media messages (which are brief, informal, and contain novel terms) and integrating such information into predictors of psychological and health outcomes.
- **Joint and multi-level language modeling.** Individual messages such as status updates or tweets are the units of expression, but health, well-being, and psychological states are attributes of people or groups. I plan to explore multi-level models covering message to user (human) levels and user to group levels. Potentially useful techniques include cascaded models, where one level of predictions feed into the next, or Bayesian joint models and adaptating *hierarchical Diriclet processes* for output of user-level variables. Similar approaches could also be used to expand typical analysis for latent personality traits to handle textual data. Because standard *factor analysis* (used for personality trait discovery in psychology) does not naturally handle the sparse and ambiguous nature of words, it seems promising to adapt techniques such as *latent Dirichlet allocation*, for the purpose of distinguishing people rather than topics.
- **Integrating discourse relation detection and sentiment analysis for capturing optimistic and pessimistic explanatory style.** Psychological theory of optimism and pessimism consists of more than positive and negative sentiment; is requires considering how people *explain* good and bad events on their lives. For example, "I always failed my tests" would be considered pessimistic while "My teacher unfairly failed me on my stats quiz" would be optimistic. I am developing a method to capture optimistic and pessimistic explanatory style in messages, which requires the identification

[4] more at wwbp.org

fig 3: Topics most characteristic of county-level *life satisfaction* (significantly correlated at $p < 0.001$). Word size is scaled by prevalence in topic [1].

fig 4: Topics most characteristic of individuals, from top to bottom, aged 13-18, 19 to 22, 23 to 29, and 30 plus, (significantly correlated at $p < 0.001$). [3]

[5] While data-driven analysis is my primary focus, there is also room to improve computational tools already accepted in social science. For example, I developed a novel technique for refining closed-vocabulary lexica such as those from *Linguistic Inquiry and Word Count* by automatically removing the most ambiguous words [7].

of causal *discourse relations* to distinguish events from their causal explanations, and sentiment analysis to find the valence of the event.

- **Detecting temporal-orientation and conditional statements.** People's "temporal-orientation" (focus on the past, present, or future) and "prospection" (thoughts and forecasts about the future, such as if-then thinking) have been linked with health and achievement. I am starting to use temporal orientation as revealed by social media to predict many health and well-being outcomes. Beyond the tense of verbs, people use many temporal indicators (e.g. 'tomorrow', '3 weeks ago', 'next year') that can be leveraged for identifying higher resolution temporal orientation (e.g. *4 days in the future* rather than simply *future*). Further, I seek to automatically identify conditional statements with respect to evaluation of future events (e.g. "if I go out drinking tonight, I may not make it to my son's soccer game tomorrow morning."). This research will combine syntactic parsing to capture conditionals with semantic analysis for temporal orientation.

Scientific Applications. Serious behavioral, psychological, and health questions motivate my creation of novel language analyses. Over time, I hope to have solid answers to the questions outlined below.

- **Can we predict individual disease recovery or longevity based on language use?** I am collaborating with researchers in medicine to write and distribute a cell-phone "app" to consenting hospital or clinical trial patients in order to match language with individual health outcomes. I am also examining childhood writings from a cohort of 10,000 citizens of the United Kingdom with biomedical outcomes tracked across their lifetime so see which behaviors, as revealed through language, are most predictive of health outcomes.
- **Can language-based prediction of personality partially replace questionnaire-based assessments?** Questionnaires have been primary tool of psychology for over a century, but they are time-consuming to administer and they are subject to self-report biases. Along with a quantitative psychologist, I am evaluating our language-based models for psychometric validity (e.g. construct, internal, and external validity and reliability).
- **Which diseases are most easily predicted at the community level from social media language?** Expanding on the heart-disease work, I am now comparing which community health-outcomes are best predicted by Twitter language, and what their characteristics are. Controlling for socio-economic factors, this may give insight into those diseases that are most influenced by the psychology or behavior of a community.
- **How does the temporal-orientation of individuals and communities influence their health, economics, and well-being?** Early results suggest, for example, use of more future-oriented words is predictive of lower spread of HIV among counties with larger amounts of risky behavior.

In recent decades, the biomedical sciences have been transformed based on large scale computational analyses such as the microarray and gene sequencers. Rather than testing specific hypotheses about genes, the process changed to that of *data-driven discovery.* Exploratory computational linguistics and information science over the new massive datasets of personal discourse (social media) are bringing behavioral and psychological factors into the reach of data science. Given the numerous findings revealed by language thus far, it is perhaps not far-fetched that the next major break-through in heart disease or cancer treatment will derive from the data-driven insights of computational linguistics – novel scientific discoveries of behavioral and psychological factors that "help people".

fig 5: Words and phrases most characteristic of extraversion (top) and introversion (bottom). (significantly correlated at $p < 0.001$). Word size is scaled by correlation strength. Color indicates relative frequency.) [3].

[1] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. P. Seligman, and L. H. Ungar, "Characterizing geographic variation in well-being using tweets," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.

[2] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. J. Park, S. Jha, M. Agrawal, L. Dziurzynski, M. Sap, C. Weeg, R. Merchant, D. Labarthe, M. E. P. Seligman, and L. H. Ungar, "Twitter correlates with atherosclerotic heart disease mortality across us counties.," *In Preparation*, 2013.

[3] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLOS ONE 8(9)*, 2013.

[4] H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, 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 Spring Symposium Series*, 2013.

[5] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, M. Agrawal, A. Kapelner, M. E. P. Seligman, and L. H. Ungar, "Predicting well-being through language use: Multi-level modeling messages and users," *In Preparation*, 2013.

[6] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, M. Kosinski, L. Dziurzynski, and M. E. P. Seligman, "From "sooo excited!!!" to "so proud": Using language to study development," *Developmental Psychology*, 2013.

[7] H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, S. Ramones, M. E. P. Seligman, and L. H. Ungar, "Choosing the right words: Characterizing and reducing error of the word count approach," in *Proceedings of *SEM-2013: Second Joint Conference on Lexical and Computational Semantics*, 2013.

[8] H. A. Schwartz and F. Gomez, "Acquiring knowledge from the web to be used as selectors for noun sense disambiguation," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, (Manchester, England), pp. 105–112, August 2008.

[9] H. A. Schwartz, F. Gomez, and L. Ungar, "Improving supervised sense disambiguation with web-scale selectors," in *COLing-2012: the 24th International Conference on Computational Linguistics*, (Mumbai, India), pp. 105–112, August 2012.

(This is a partial list of most relevant publications. The complete list is available in the curriculum vitae.)