Toward Personality Insights from Language Exploration in Social Media

H. Andrew Schwartz, Johannes C. Eichstaedt, Lukasz Dziurzynski,
Margaret L. Kern, Martin E. P. Seligman and **Lyle H. Ungar**
University of Pennsylvania

Eduardo Blanco
Lymba Corporation

Michal Kosinski and **David Stillwell**
University of Cambridge

Abstract

Language in social media reveals a lot about people's personality and mood as they discuss the activities and relationships that constitute their everyday lives. Although social media are widely studied, researchers in computational linguistics have mostly focused on prediction tasks such as sentiment analysis and authorship attribution. In this paper, we show how social media can also be used to gain psychological insights. We demonstrate an exploration of language use as a function of age, gender, and personality from a dataset of Facebook posts from 75,000 people who have also taken personality tests, and we suggest how more sophisticated tools could be brought to bear on such data.

Introduction

With the growth of social media such as Twitter and Facebook, researchers are being presented with an unprecedented resource of personal discourse. Computational linguists have taken advantage of these data, mostly addressing prediction tasks such as sentiment analysis, authorship attribution, emotion detection, and stylometrics. A few works have also been devoted to predicting personality (i.e, stable unique individual differences). Prediction tasks have many useful applications ranging from tracking opinions about products to identifying messages by terrorists. However, for social sciences such as psychology, gaining insight is at least as important as making accurate predictions.

In this paper, we explore the use of various language features in social media as a function of gender, age, and personality to support research in psychology. Some psychologists study the words people use to better understand human psychology (Pennebaker, Mehl, and Niederhoffer 2003), but they often lack the sophisticated NLP and big data techniques needed to fully exploit what language can reveal about people. Here, we analyze 14.3 million Facebook messages collected from approximately 75,000 volunteers, totaling 452 million instances of n-grams and topics. This data set is an order-of-magnitude larger than previous studies of language and personality, and allows qualitatively different analysis. To examine the thousands of statistically significant correlations that emerge from this analysis, we employ a *differential word cloud* visualization which displays words or n-grams sized by relationship strength rather than the standard, word frequency. We also use Latent Dirichlet Allocation (LDA) to find sets of related words, and plot word and topic use as a function of Facebook user age.

Background

Psychologists have long sought to gain insight into human psychology by exploring the words people use (Stone, Dunphy, and Smith 1966; Pennebaker, Mehl, and Niederhoffer 2003). Recently, such studies have become more structured as researchers leverage growing language datasets to look at what categories of words correspond with human traits or states. The most common approach is to count words from a pre-compiled word-category list, such as Linguistic Inquiry and Word Count or *LIWC* (Pennebaker et al. 2007). For example, researchers have recently used *LIWC* to find that males talk more about occupation and money (Newman et al. 2008); that females mention more social and emotional words (Newman et al. 2008; Mulac, Studley, and Blau 1990); that conscientious (i.e. efficient, organized, and planful) people mention more positive emotion words and filler plus talk about family (Mehl, Gosling, and Pennebaker 2006; Sumner, Byers, and Shearing 2011); that people low in agreeableness (i.e. appreciative, forgiving, and generous) use more anger or swear words (Mehl, Gosling, and Pennebaker 2006; Yarkoni 2010; Sumner, Byers, and Shearing 2011); or that most categories of function words (articles, prepositions, pronouns, auxiliaries) vary with age, gender, and personality (Chung and Pennebaker 2007).

Such studies rarely look beyond *a priori* categorical language (one exception, (Yarkoni 2010), is discussed below). One reason is that studies are limited to relatively small sample sizes (typically a few hundred authors). Given the sparse nature of words, it is more efficient to group words into categories, such as those expressing positive or negative emotion. In this paper, we use an open-vocabulary approach, where the vocabulary being examined is based on the actual text, allowing discovery of unanticipated language.

On the other hand, open-vocabulary or data-driven ap-

proaches are commonplace in computational linguistics, but rarely for the purpose of gaining insights. Rather, open-vocabulary features are used in predictive models for many tasks such as authorship attribution / stylistics (Holmes 1994; Argamon, Šarić, and Stein 2003; Stamatatos 2009), emotion and interaction style detection(Alm, Roth, and Sproat 2005; Jurafsky, Ranganath, and McFarland 2009), or sentiment analysis (Pang, Lee, and Vaithyanathan 2002; Kim and Hovy 2004).

Personality refers to biopsychosocial characteristics that uniquely define a person (Friedman 2007). A commonly accepted framework for organizing traits, which we use in this paper, is the Big Five model (McCrae and John 1992). The model organizes personality traits into five continuous dimensions:

- *extraversion*: active, assertive, energetic, enthusiastic, outgoing
- *agreeableness*: appreciative, forgiving, generous, kind
- *conscientiousness*: efficient, organized, planful, reliable
- *neuroticism*: anxious, self-pittying, tense, touchy, unstable
- *openness*: artistic, curious, imaginative, insightful, original

A few researchers have looked particularly at personality for their predictive models. Argamon et al. (2005) noted that personality was a key component of identifying authors and examined function words and various taxonomies in relation to two personality traits, neuroticism and extraversion over approximately 2200 student essays. They later examined predicting gender while emphasizing function words (Argamon et al. 2009). Mairesse and Walker; Mairesse et al. (2006; 2007) examined all five personality traits over approx. 2500 essays and 90 individuals' spoken language data. Bridging the gap with Psychology, they used *LIWC* as well as other dictionary based features rather than an open-vocabulary approach. Similarly, Golbeck et al. (2011) used *LIWC* features to predict personality of a sample of 279 Facebook users. Lastly, Iacobelli et al. (2011) examined around 3,000 bloggers, the largest previous study of language and personality, for the predictive application of content customization. Bigrams were among the best predictive features, motivating the idea that words with context add information linked to personality. Most of these works include some discussion on the best language features (i.e. according to information gain) within their models, but they are focused on producing a single number: an accurate personality score, rather than a comprehensive list of language links for exploration.

To date, we are only aware of one other study which explores open-vocabulary word-use for the purpose of gaining personality insights. Yarkoni (2010) examined both words and manually-created lexicon categories in connection with personality of 694 bloggers. They found between 13 and 393 significant correlations depending on the personality trait. To contrast with our approach, we examined an orders-of-magnitude larger sample size (75,000 volunteers) and a more extensive set of open-vocabulary language: multi-word n-grams and topics. The larger sample size allows a more comprehensive, and less fitted results (i.e. we find thousands of significant correlations for each personality trait, even when adjusting significance for the fact that we look at tens of thousands of features). Outside of the Big 5 personality construct, works have used language processing techniques to link language with psychosocial variables. Select examples include link language with happiness (Mihalcea and Liu 2006; Dodds et al. 2011), location (Eisenstein et al. 2010), or over decades in books (Michel et al. 2011).

Data Set

For the experiments in this paper , we used the status updates of approximately 75,000 volunteers who also took a standard personality questionnaire and reported their gender and age (Kosinski and Stillwell 2012). In order to insure a decent sample of language use per volunteer, we restricted the analyses to those who wrote at least 1,000 words across their status updates. 74,941 met this requirement and also reported their gender and age. Out of those, we had 72,791 individuals with *extraversion* ratings, 72,853 with *agreeableness* ratings, 72,863 with *conscientiousness* ratings, 72,047 with *neuroticism* ratings, and 72,891 with *openness* ratings.

Differential Language Analysis: A General Framework for Insights

Our approach follows a general framework for insights consisting of the three steps depicted in Figure 1:

1. **Linguistic Feature Extraction**: Extract the units of language that we wish to correlate with (i.e. n-grams, topics, etc...).
2. **Correlation Analysis**: Find the relationships between language use and psychological variables.
3. **Visualization**: Represent the output of correlation analysis in an easily digestible form.

Linguistic Feature Extraction

Although there are many possibilities, as initial results we focus on two types of linguistic features:

N-Grams: *sequences of one to three tokens.*
We break text into tokens utilizing an emoticon-aware tokenizer built on top of Christopher Pott's "happyfuntokenizing" [1]. For sequences of multiple words, we apply a collocation filter based on point-wise mutual information (*PMI*)(Church and Hanks 1990; Lin 1998) which quantifies the difference between the independent probability and joint-probability of observing an n-gram (given below). We eliminated uninformative ngrams which we defined as those with a $pmi < 2 * len(ngram)$ where $len(ngram)$ is the number of tokens (tok). In practice, we record the relative frequency of an n-gram ($\frac{freq(ngram)}{total\_word\_usage}$) and apply the Anscombe transformation (Anscombe 1948) to stabilize variance between volunteers' relative usages.

$$pmi(ngram) = \log \frac{p(ngram)}{\prod_{token \in ngram} p(token)}$$

[1] http://sentiment.christopherpotts.net/code-data/

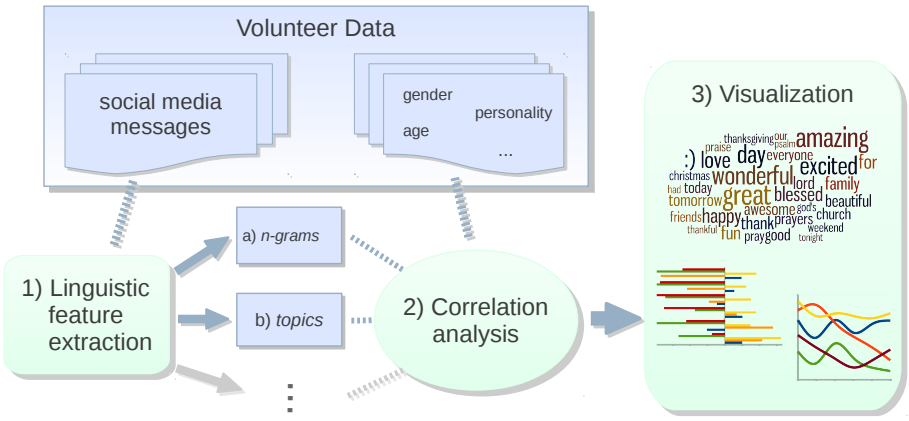

Figure 1: The differential language analysis framework used to explore connections between language and psychological variables.

Topics: *semantically related words derived via LDA.*
LDA (Latent Dirichlet Allocation) is a generative process in which documents are defined as a distribution of topics, and each topic in turn is a distribution of tokens. Gibbs sampling is then used to determine the latent combination of topics present in each document (i.e. Facebook messages), and the words in each topic (Blei, Ng, and Jordan 2003). We use the default parameters within an implementation of *LDA* provided by the Mallet package (McCallum 2002), except that we adjust alpha to 0:30 to favor fewer topics per document, as status updates are shorter than the news or encyclopedia articles which were used to establish the parameters. One can also specify the number of topics to generate, giving a knob to the specificity of clusters (less topics implies more general clusters of words). We chose 2,000 topics as an appropriate level of granularity after examining results of *LDA* for 100, 500, 2000, and 5000 topics. To record a person's use of a topic we compute the probability of their mentioning the topic ($p(topic, person)$ – defined below) derived from their probability of mentioning tokens ($p(tok|person)$) and the probability of tokens being in given topics ($p(topic|tok)$). While n-grams are fairly straight-forward, topics demonstrate use of a higher-order language feature for the application of gaining insight.

$$p(topic, person) = \sum_{tok \in topic} p(topic|tok) * p(tok|person)$$

Across all features, we restrict analysis to those in the vocabulary of at least 1% of our volunteers in order to eliminate obscure language which is not likely to correlate. This results in 24,530 unique n-grams and 2,000 topics.

Correlation Analysis

After extracting features, we find the correlations between variables using ordinary least squares linear regression over standardized (mean centered and normalized by the standard deviation) variables. We use language features (n-grams or topics) as the explanatory variables – the features in the regression, and a given psychological outcome (such as introversion/extraversion) as the dependent variable. Linear regression, rather than a straight Pearson correlation, allows us to include additional explanatory variables, such as gender or age in order get the unique effect of the linguistic feature (adjusted for effects from gender or age) on the psychological outcome. The coefficient of the target explanatory variable[2] is taken as the strength of the relationship. Since the data is standardized, 1 indicates maximal covariance, 0 is no relationship, and -1 is maximal covariance in opposite directions. A separate regression is run for each language feature.

To limit ourselves to meaningful relationships, two-tailed significance values are computed for each coefficient, and since we explore thousands of features at once, a Bonferonni-correction is applied (Dunn 1961). For all results discussed, a Bonferonni-corrected p must have been below 0.001 to be considered significant[3]

Visualization

Hundreds of thousands of correlations result from comparing tens of thousands of language features with multiple dimensions of psychological variables. Visualization is thus crucial for efficiently gaining insights from the results. In this work, we employ two visualizations: *differential word clouds* and *standardized frequency plots*.

Differential word clouds: *comprehensive display of the most distinguishing features.* When drawing word clouds, we make the size of the n-grams be proportional to the correlation strength and we select their color according to their frequency. Note that unlike standard word clouds which simply show the frequency of words, we emphasize what differentiates the variable. We use word cloud software provided by Wordle[4] as well as that of the D3 data-driven visualization package[5]. In order to provide the most compre-

[2] often referred to as β in statistics or simply a "weight" in machine learning

[3] A passing p when examining 10,000 features would be below 10^{-7} (or $\frac{.001}{10000}$).

[4] http://wordle.net/advanced

[5] http://d3js.org

Figure 2: N-grams most correlated with females (top) and males (bottom), adjusted for age ($N = 74,941$: $46,572$ females and $28,369$ males; Bonferroni-corrected $p < 0.001$). Size of words indicates the strength of the correlation; color indicates relative frequency of usage. Underscores (_) connect words in multiword phrases.

hensive view, we prune features from the word cloud which contain overlap in information so that other significant features may fit. Specifically, using inverse-frequency as proxy for information content (Resnik 1999), we only include an ngram if it contains at least one word which is more informative than previously seen words. For example, if 'day' correlates most highly but 'beautiful day' and 'the day' also correlate but less significantly, then 'beautiful day' would remain because 'beautiful' is adding information while 'the day' would be dropped because 'the' is less informative than 'day'. We believe a differential word cloud representation is helpful to get an overall view of a given variable, functioning as a supplement to a definition (i.e. what does it mean to be neurotic in Figure 3).

Standardized frequency plot: *standardized relative frequency of a feature over a continuum.* It is often useful to track language features across a sequential variable such as age. We plot the standardized relative frequency of a language feature as a function of the outcome variable. In this case, we group age data in to bins of equal size and fit second-order LOESS regression lines (Cleveland 1979) to the age and language frequency data over all users. We adjust for gender by averaging male and female results.

While we believe these visualizations are useful to demonstrate the insights one can gain from differential language analysis, the possibilities for other visualization are quite large. We discuss a few other visualization options we are also working on in the final section of this paper.

Results

We first present the n-grams that distinguish *gender*, then proceed to the more subtle task of examining the traits of *personality*, and last to exploring variations in topic use with *age*.

Gender Figure 2 presents age-adjusted differential word clouds for females and males. Since gender is a familiar variable, it functions as a nice proof of concept for the analysis. In agreement with past studies (Mulac, Studley, and Blau 1990; Thomson and Murachver 2001; Newman et al. 2008), we see many n-grams related to emotional and social processes for females (e.g. 'excited', 'love you', 'best friend') while males mention more swear words and object references (e.g. 'shit', 'Xbox', 'Windows 7'). We also contradict past studies, finding, for example, that males use fewer emoticons than females, contrary to a previous study of 100 bloggers (Huffaker and Calvert 2005). Also worth noting is that while 'husband' and 'boyfriend' are most distinguishing for females, males prefer to attach the possessive modifier to those they are in relationships with: 'my wife' or 'my girlfriend'.

Personality Figure 3 shows the most distinguishing n-grams for *extraverts* versus *introverts*, as well as *neurotic* versus *emotionally stable* (word clouds for the other personality factors are in the appendix). Consistent with the definition of the personality traits (McCrae and John 1992), extraverts mention social n-grams such as 'love you', 'party', 'boys', and 'ladies', while introverts mention solitary activities such as 'Internet', 'read', and 'computer'. Moving beyond expected results, we also see a few novel insights,

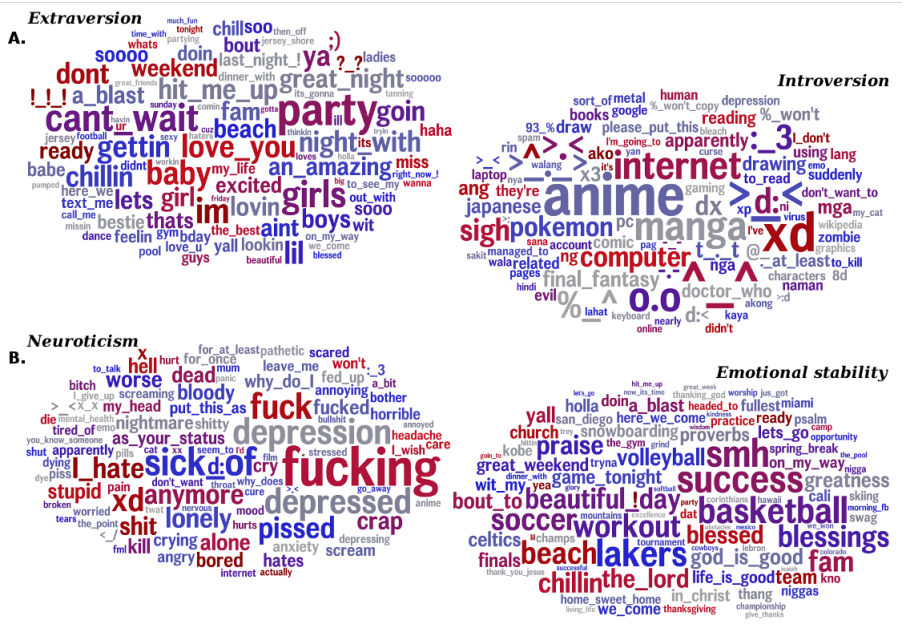

Figure 3: **A.** N-grams most distinguishing extraversion (top, e.g., 'party') from introversion (bottom, e.g., 'computer'). **B.** N-grams most distinguishing neuroticism (top, e.g. 'hate') from emotional stability (bottom, e.g., 'blessed') ($N = 72,791$ for extraversion; $N = 72,047$ for neuroticism; adjusted for age and gender; Bonferroni-corrected $p < 0.001$). Results for *openness*, *conscientiousness*, and *agreeableness* can be found on our website, wwbp.org.

such as the preference of introverts for Japanese culture (e.g. 'anime', 'pokemon', and eastern emoticons '> . <' and '^_^'). A similar story can be found for neuroticism with expected results of 'depression', 'sick of', and 'I hate' versus 'success', 'a blast', and 'beautiful day'. [6] More surprisingly, sports and other activities are frequently mentioned by those low in neuroticism: 'backetball', 'snowboarding', 'church', 'vacation', 'spring break'. While a link between a variety of life activities and emotional stability seems reasonable, to the best of our knowledge such a relationship has never been explored (i.e. does participating in more activities lead to a more emotionally stable life, or is it only that those who are more emotionally stable like to participate in more activities?). This demonstrates how open-vocabulary exploratory analysis can reveal unknown links between language and personality, suggesting novel hypotheses about behavior; it is plausible that people who talk about activities more also participate more in those activities.

Age We use age results to demonstrate use of higher-order language features (*topics*). Figure 4 shows the n-grams and topics most correlated with two age groups (*13 to 18* and *23 to 29* years old). The differential word cloud of n-grams is shown in the center, while the most distinguishing topics, represented by their 15 most prevalent words, surround. For 13 to 18 year olds, we see topics related to *Web shorthand*, *classes*, *going back to school*, *laughing*, and *young relationships* while 23 to 29 year olds mention topics related to *job search, work, drinking, household chores,* and *time management*. Additionally, we show n-gram and topic use across age in standardized frequency plots of Figure 5. One can follow peaks for the predominant topics of *school, college, work,* and *family* across the age groups. We also see more psychologically oriented features, such as 'I' and 'we' decreasing until the early twenties and then 'we' monotonically increasing from that point forward. One might expect 'we' to increase as people marry, but it continues increasing across the whole lifespan even as weddings flatten out. A similar result is seen in the social topics of Figure 5B.

Toward Greater Insights

While the results presented here provide some new insight into gender, age, and personality they mostly confirm what is already known or obvious. At a minimum, our results serve as a foundation to establish *face validity* – confirmation that the method works as expected. Future analyses, as described below, will delve deeper into relationships between language and psychosocial variables.

Language Features The linguistic features discussed so far are relatively simple, especially n-grams. It is well-known that individual words (unigrams) and words in context (bigrams, trigrams) are useful to model language; in our previous analysis we exploited this fact for modeling personality types. However, n-grams ignore all links but the ones between words within a small window, and do not provide

[6] Finding 'sick of' rather than simply 'sick' shows the benefit of looking at n-grams in addition to single words ('sick' has quite different meanings than 'sick of').

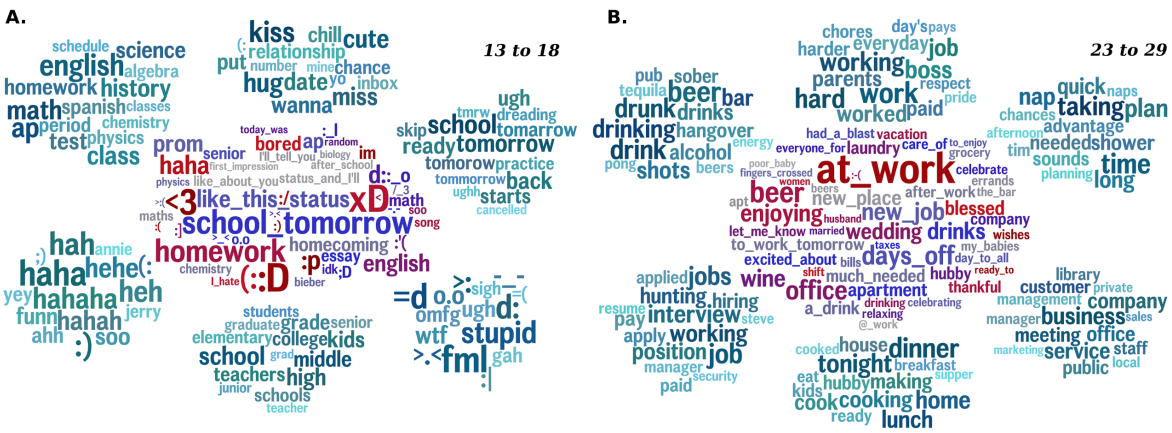

Figure 4: **A.** N-grams and topics most distinguishing volunteers aged *13 to 18*. **B.** N-grams and topics most distinguishing volunteers aged *23 to 29*. *N-grams* are in the center; *topics*, represented as the 15 most prevalent words, surround. ($N = 74,941$; correlations adjusted for gender; Bonferroni-corrected $p < 0.001$). Results for *19 to 22* and *30+* can be found on our website, wwbp.org.

any information about the words or link. That is, they only capture that two words co-occur.

Computational linguistics has witnessed a growing interest in automatic methods to represent the meaning of text. These efforts include named entity recognition (Finkel, Grenager, and Manning 2005; Sekine, Sudo, and Nobata 2002) (identifying words that belong to specific classes, e.g., diseases, cities, organizations) and semantic relation extraction (Carreras and Màrquez 2005) (labeling links between semantically related words, e.g., in "John moved to Florida because of the nice weather", the *weather* is the CAUSE of *moving*, *Florida* the DESTINATION and *John* the AGENT; *nice* is VALUE of *weather*).

We plan to incorporate features derived from the output of the above tools into our personality analyses. Thus, the correlation analysis and visualization steps will consider the meaning behind words. To demonstrate how these features may help, consider the word *cancer*. It is useful to identify the instances that are a named entity *disease*, e.g., compare (1) "Pray for my grandpa who was just diagnosed with terminal *cancer*." and (2) "I have been described as a *cancer* or a virgo, what do you guys think?". Second, it seems useful to analyze the semantic relations in which *cancer* is involved. For example, compare (3) "Scan showed cancer is gone!" and (4) "My brother in law passed away after a seven month battle with a particularly aggressive cancer". In (3), *cancer* is the EXPERIENCER of *gone*; in (4) *cancer* CAUSE *passed away*. In turn, this information could be used to determine which personality types are prone to express positive and negative feelings about diseases.

Correlation Analysis and Visualization We have so far looked at differential use in one dimension (e.g. extraversion), controlling for variations in other dimensions (e.g. age and gender). Preliminary explorations suggest that "interactions" provide useful insights e.g., noting how the language of female extroverts differs from that of male extroverts, or how young introverts use language differently from old introverts. These interactions can be computed in multiple ways. The simplest is to use our existing methods, as described above, to do differential language analysis between these subsets of the population. More insight, however, is derived by creating different plots that show, for example, each (significant) word in a two dimensional space, where the one dimension is given by, for example, that word's coefficient on age (controlling for gender) and the other dimension is given by the same word's coefficient on introversion (again controlling for gender). Even more subtle effects can be found by including interaction terms in the regression model, such as a cross-product of age and introversion.

Conclusion

We presented a case study on the analysis of language in social media for the purpose of gaining psychological insights. We examined n-grams and *LDA* topics as a function of gender, personality, and age in a sample of 14.3 million Facebook messages. To take advantage of the large number of statistically significant words and topics that emerge from such a large dataset, we crafted visualization tools that allow for easy exploration and discovery of insights. For example, we present results as word clouds based on regression coefficients rather than the standard word frequencies. The word clouds mostly show known or obvious findings (i.e. *extraverts* mention 'party'; *introverts* mention 'internet'), but also offer broad insights (i.e. *emotionally stable* individuals mention more sports and life activities plus older individuals mention more social topics and less anti-social topics), as well as fine-grained insights (i.e. men prefer to preface 'wife' or 'girlfriend' with the possessive 'my'). We envision this work as a foundation to establishing more sophisticated language features, analyses and visualizations.

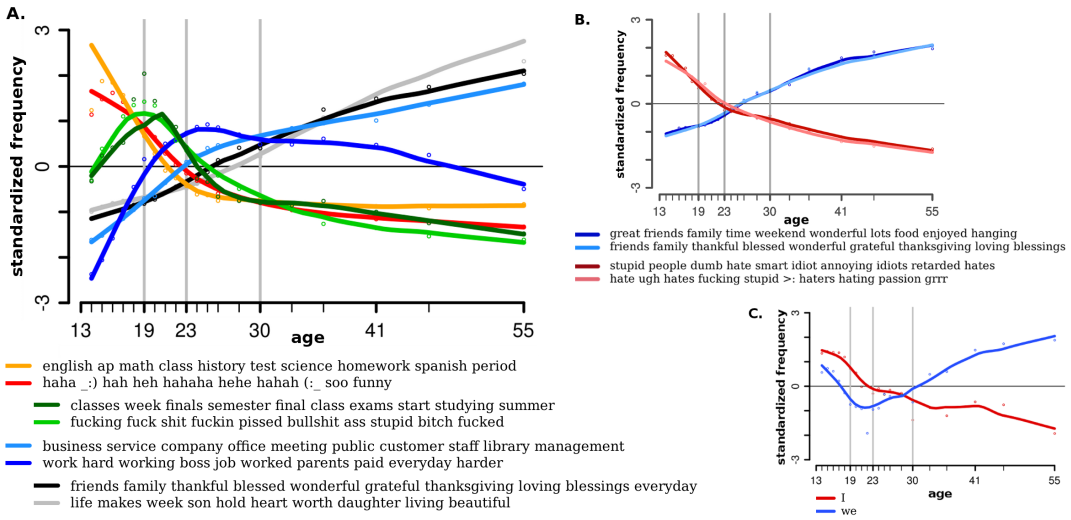

Figure 5: **A.** Standardized frequency for the top 2 topics for each of 4 bins across age. Grey vertical lines divide bins: 13 to 18 (red: $n = 25,496$ out of $N = 75,036$), 19 to 22 (green: $n = 21,715$), 23 to 29 (blue: $n = 14,677$), and 30+ (black: $n = 13,148$). **B.** Standardized frequency of social topic use across age. **C.** Standardized 'I', 'we' frequencies across age. (Lines are fit from second-order LOESS regression (Cleveland 1979) controlled for gender).

Acknowledgements

Support for this research was provided by the Robert Wood Johnson Foundations Pioneer Portfolio, through a grant to Martin Seligman, Exploring Concept of Positive Health.

References

Alm, C.; Roth, D.; and Sproat, R. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Empirical Methods in Natural Language Processing*, 579–586.

Anscombe, F. J. 1948. The transformation of poisson, binomial and negative-binomial data. *Biometrika* 35(3/4):246–254.

Argamon, S.; Dhawle, S.; Koppel, M.; and Pennebaker, J. W. 2005. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society*.

Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52(2):119–123.

Argamon, S.; Šarić, M.; and Stein, S. S. 2003. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth international conference on Knowledge discovery and data mining*, 475–480.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J of Machine Learning Research* 3:993–1022.

Carreras, X., and Màrquez, L. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 152–164.

Chung, C., and Pennebaker, J. 2007. The psychological function of function words. *Social communication: Frontiers of social psychology* 343–359.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the Am Stati Assoc* 74:829–836.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE* 6(12):26.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56(293):52–64.

Eisenstein, J.; O'Connor, B.; Smith, N.; and Xing, E. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Finkel, J. R.; Grenager, T.; and Manning, C. D. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.

Friedman, H. 2007. Personality, disease, and self-healing. *Foundations of Health Psychology*.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting personality from twitter. In *Proc of the 3rd IEEE Int Conf on Soc Comput*, 149–156.

Holmes, D. 1994. Authorship attribution. *Computers and the Humanities* 28(2):87–106.

Huffaker, D. A., and Calvert, S. L. 2005. Gender, Identity, and Language Use in Teenage Blogs. *J of Computer-Mediated Communication* 10(2):1–10.

Iacobelli, F.; Gill, A. J.; Nowson, S.; and Oberlander, J. 2011. Large scale personality classification of bloggers. In *Proceedings of the Int Conf on Affect Comput and Intel Interaction*, 568–577.

Jurafsky, D.; Ranganath, R.; and McFarland, D. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technology Conference of the NAACL*, 638–646.

Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*.

Kosinski, M., and Stillwell, D. 2012. mypersonality project. In *http://www.mypersonality.org/wiki/*.

Lin, D. 1998. Extracting collocations from text corpora. In *Knowledge Creation Diffusion Utilization*. 57–63.

Mairesse, F., and Walker, M. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL*, 85–88.

Mairesse, F.; Walker, M.; Mehl, M.; and Moore, R. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30(1):457–500.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. In *http://mallet.cs.umass.edu*.

McCrae, R. R., and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60(2):175–215.

Mehl, M.; Gosling, S.; and Pennebaker, J. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology* 90(5):862.

Michel, J.-B.; Shen, Y. K.; Aiden, A. P.; Veres, A.; Gray, M. K.; Team, T. G. B.; Pickett, J. P.; Hoiberg, D.; Clancy, D.; Norvig, P.; Orwant, J.; Pinker, S.; Nowak, M.; and Lieberman-Aiden, E. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.

Mihalcea, R., and Liu, H. 2006. A corpus-based approach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*.

Mulac, A.; Studley, L. B.; and Blau, S. 1990. The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles* 23:439–470.

Newman, M.; Groom, C.; Handelman, L.; and Pennebaker, J. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3):211–236.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 79–86.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of liwc2007 the university of texas at austin. *LIWC.NET* 1:1–22.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology* 54(1):547–77.

Resnik, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11:95–130.

Sekine, S.; Sudo, K.; and Nobata, C. 2002. Extended named entity hierarchy. In *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, 1818–1824.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Stone, P.; Dunphy, D.; and Smith, M. 1966. The general inquirer: A computer approach to content analysis.

Sumner, C.; Byers, A.; and Shearing, M. 2011. Determining personality traits & privacy concerns from facebook activity. In *Black Hat Briefings*, 1 – 29.

Thomson, R., and Murachver, T. 2001. Predicting gender from electronic discourse. *Brit J of Soc Psychol* 40(Pt 2):193–208.

Yarkoni, T. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44(3):363–373.