DLATK: Differential Language Analysis ToolKit

H. Andrew Schwartz[†] **Salvatore Giorgi**[‡] **Maarten Sap**[§]
Patrick Crutchley[‖] **Johannes C. Eichstaedt**[‡] **Lyle Ungar**[‡]
† Stony Brook University ‡ University of Pennsylvania
§ University of Washington ‖ Qntfy
has@cs.stonybrook.edu, sgiorgi@sas.upenn.edu

Abstract

We present *Differential Language Analysis Toolkit* (DLATK), an open-source python package and command-line tool developed for conducting social-scientific language analyses. While DLATK provides standard NLP pipeline steps such as tokenization or SVM-classification, its novel strengths lie in analyses useful for psychological, health, and social science: (1) incorporation of extra-linguistic structured information, (2) specified levels and units of analysis (e.g. document, user, community), (3) statistical metrics for continuous outcomes, and (4) robust, proven, and accurate pipelines for social-scientific prediction problems. DLATK integrates multiple popular packages (SKLearn, Mallet), enables interactive usage (Jupyter Notebooks), and generally follows object oriented principles to make it easy to tie in additional libraries or storage technologies.

1 Introduction

The growth of NLP for social and medical sciences has shifted attention in NLP research from understanding language itself (e.g. syntactic parsing or characterizing morphology) to understanding how language use characterizes people (e.g. by correlating language use characteristics with traits of the person producing the language). Much of this work has been done using Facebook and Twitter (Coppersmith et al., 2014).

Analyzing language for social science applications requires different tools and techniques than conventional NLP. Structured data are often beneficial to facilitate the use of the extensive extra-linguistic information such as the time and location of the post and author demographics (or even health or school records). Models can be made at multiple levels of analysis: documents, users, and different geographic (zip code, state or country) or temporal resolutions. Many of the outcomes (or dependent variables) are continuous (e.g. scores on personality tests), and researchers are often as interested in interpretable insights as they are with predictive accuracy (Kern et al., 2014a).

There are small "tricks" to obtain accurate predictive models or high correlations between language features and outcomes. Emoticon-aware tokenizers are needed, robust methods for creating LDA topics (different packages produce clusters of strikingly different quality), and subtle issues of regularization arise when combining demographic and language features in models. When these choices are combined with the complexity of the structured data, even NLP and data scientists can fail to produce high quality models. We therefore built a platform that integrates a variety of open-sourced tools, alongside our "tricks" and optimizations, to provide a well-documented, easy-to-use program for undertaking reproducible research in the area of NLP for the social sciences.

This software, which has now been used for the data analysis behind 32 papers in psychology, health care, and NLP, is now available under a GPLv3 software license.[1]

2 Overall Framework

The core of DLATK is a Python library depicted in Figure 1. The base class, DLAWorker, sits on top of a data engine (e.g. MySQL, HDFS/Spark) and is used to track corpus basics (corpus location, unit-of-analysis). The next level of classes acts on either: messages, features or outcomes. MessageAnnotator filters messages (removing du-

[1] http://dlatk.wwbp.org or http://github.com/dlatk/

Proceedings of the 2017 EMNLP System Demonstrations, pages 55–60
Copenhagen, Denmark, September 7–11, 2017.

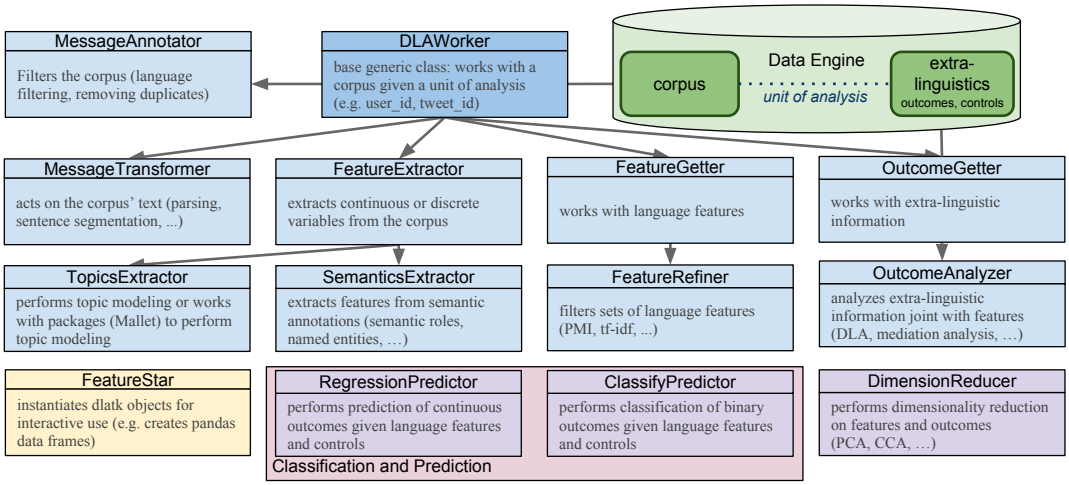


Figure 1: Basic DLATK package class structure.

plicate tweets, language filtering, etc.) while MessageTransformer acts on message text (tokenizing, part of speech tagging, etc.). FeatureExtractor converts document text to features (ngrams, character ngrams, etc.) and is responsible for writing while FeatureGetters read for downstream analysis or further refinement via FeatureRefiner. OutcomeGetter reads outcome tables (i.e., extra-linguistic information). Its child class OutcomeAnalyzer works with both linguistic and extra-linguistic information for statistical analyses (correlation, logistic regression, etc.) and various outputs (wordclouds, correlation matrices, etc.).

The bottom classes do not inherit but are utilizers of FeatureGetters and OutcomeGetters. This includes two classes for prediction, RegressionPredictor and ClassifyPredictor which carry out machine learning tasks: cross validation, feature selection, training models, building data-driven lexica, etc, while DimensionReducer provides unsupervised transformations on language and outcomes Finally, the FeatureStar class ("star" for wildcard) is used to interact with the other classes and transform important information into convenient data structures (e.g. Pandas dataframes).

3 Differential Language Analyses

The prototypical use of DLATK is to perform *differential language analysis* – the identification of linguistic features which either (a) independently explain the most variance for *continuous outcomes* or (b) are individually most predictive of *discrete outcomes* (Schwartz et al., 2013b). Unlike predictive techniques where one seeks to *produce outcome(s)* given language (discussed next), here, the goal is to *produce language* that is most related to or independently discriminant of outcomes.[2] DLATK supports several metrics for performing differential language analysis.

Continuous DLA Metrics. We support a variety of metrics for comparing language to continuous outcomes (e.g. age, degree of depression, personality factor scores, income). Primary metrics are based on **Pearson Product-Moment Correlation Coefficient** (Agresti and Finlay, 2008). When one requests control variables (e.g. finding the relationship with degree of depression, controlling for age and gender) then **ordinary least squares linear regression** is used (Rao, 2009) wherein the control variables are included alongside the linguistic variable as covariates and the outcome is the dependent variable.

Discrete DLA Metrics. While linear regression produces meaningful results for most situations, it is often ideal to use other metrics for discrete or Bernoulli outcomes. **Logistic regression** can be used in place of linear regression where, by assuming a dichotomous outcome, statistical significance tests are usually more accurate (Menard, 2002). Where controls are not needed, there are many other options, often less computationally complex, such as **TF-IDF**, **Informative Dirichlet Prior**,[3] or classification accuracy metrics like

[2]Even in basic prediction methods, like linear regression, the relationship between each linguistic feature and the outcome is complex – dependent on the covariance structure between all the variables. DLA works in a univariate, per-feature fashion or with a limited set of control variables (e.g. age and gender when discriminating personality).

[3]Bayesian approach to log-odds (Monroe et al., 2008).

Area Under the ROC Curve (Fawcett, 2006).

Multiple Hypothesis Testing. Most of the metrics have a corresponding standard significance test (e.g. Student's t-test for Pearson correlation and OLS regression), and most output confidence intervals by default. Permutation testing has been implemented for many of metrics without standard significance tests, such as AUC-ROC, with the linguistic feature vector shuffled relative to outcome (and controls, if applicable) multiple times to create a null distribution. Standard practice in differential language analysis (Schwartz et al., 2013b) is to correlate each of potentially thousands of single features (e.g., normalized usage of one single- or multi-word expression) with a given outcome. Thus, correcting for multiple comparisons is critical. When used through the interface script, DLATK by default corrects for multiple comparisons using the Benjamini-Hochberg method of FDR correction (Benjamini and Hochberg, 1995). Other options, such as the more conservative Bonferroni correction (Dunn, 1961) are also available.

4 Predictive Methods

As with traditional NLP, many social-scientific research objectives can be framed as *prediction* tasks, in which a model is fit to language features to predict an outcome. DLATK implements many available regression and classification tools, supplemented with feature selection functions for refining the feature space. A wide range of feature selection techniques have been empirically refined for accurate use in regression problems.

Feature selection. DLATK's ClassifyPredictor and RegressionPredictor classes include methods for feature selection, which is critical given what may be a very large space of linguistic features, e.g., 100s of thousands of 1- to 3-grams in a corpus. Both classes allow for pass-through of scikit-learn Pipelines (e.g. univariate feature selection based on feature correlation with outcome and family-wise error) and dimensionality reduction methods (e.g., PCA on feature matrix), including combination methods where FS and DR steps are applied to the original data in a serial manner.

Regression Models. DLATK supports a variety of regression models in order to take in features as well as extra-linguistic information and output a continuous value predictions. These include variants on penalized linear regression: **Ridge**, **Lasso**, **Elastic-Net**, as well as non-linear techniques such as **Extremely Random Forests**. A common pipeline, referred to as "magic sauce" applies univariate feature selection and PCA to linguistic features independent of controls, and then uses ridge to fit a linear model from a combined reduced space to the outcomes.

Classification Models. DLATK implements a rich variety of classifiers, including **Logistic Regression** and **Support Vector Classifiers** with L_1 and L_2 regularization, as well ensemble and gradient boosting techniques such as **Extremely Randomized Trees**. As with regression, techniques have been setup so as to leverage extra-linguistic information effectively either as additional predictors or controls to try to "-out-predict".

5 Notable Functionality

Linguistic information Because DLATK was designed to exploit the full power of social media, a special emoticon-aware tokenizer is used while also leveraging Python's unicode capabilities. Though not specifically designed to be language independent, DLATK has been used in one non-English study (Smith et al., 2016).

Extra-linguistic information. Most functionality in DLATK is designed with extra-linguistic, also referred to as "outcomes", in mind. Such information ranges from meta-information of social media posts, such as time or location, to user attributes such as demographics or strong baselines one may wish to out-predict. For DLA, this means that one not only distinguishes target extra-linguistic information, but that controls are available. For prediction, extra-linguistic information can be incorporated as input to a model, taking into account the fact that such features are often less sparse and more reliable features of people than individual linguistic features.

Multiple Levels of Analysis. DLATK allows one to work with a single corpus at multiple levels of analysis, simply as a parameter to any action. For example, one may choose to analyze tweets themselves or group them by user_id, location, or even a combination of user and date. Extra-linguistic information often dictates particular levels of analyses (e.g. community level mortality rates or user-level personality questionnaire responses). Analysis setups are flexible for levels of analysis – for example, one can dynamically

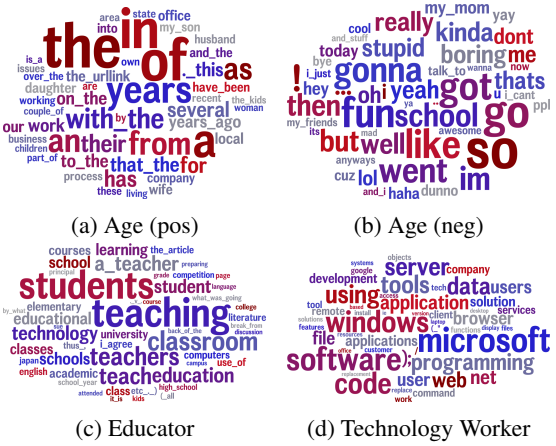
threshold which of the units of analyses are available (e.g. only include users with at least 1000 words or counties with 50,000 words).

Integration of Popular Packages. DLATK sits on top of many popular open source packages used for data analysis and machine learning (scikit-learn (Pedregosa et al., 2011) and statsmodels (Seabold and Perktold, 2010)) as well as NLP specific packages (Stanford parser (Chen and Manning, 2014), TweetNLP (Gimpel et al., 2011) and NLTK (Loper and Bird, 2002)). LDA topics can be created with the Mallet (McCallum, 2002) interface. After creation these topics can then be used downstream in any standard DLATK analysis pipeline. The *pip* and *conda* package management systems control python library dependencies.

Interactive Usage. The standard way to interact with DLATK is with the interface script through the command line. Often users will only see the two end points (the document input and the analysis output) and as a result this package is used as a "black box". In order to encourage data exploration the FeatureStar class converts the language features and extra-linguistic information into Pandas dataframes (McKinney, 2011) allowing users to import our methods into existing code. Sample use cases include opening up predictive models to explore feature coefficients and easily reading linguistic data into standard data visualization tools.

Visualization. When running DLA we often run separate correlations over tens of thousands of language features. While a single word might not give us considerable insight into our extra-linguistic information groups of words taken together can often tell a compelling story. To this end DLATK offers wordcloud output in the form of n-gram and topic clouds images. Figure 2 shows 1- to 3-grams significantly correlated with (a) age (positive; higher age), (b) age (negative; lower age), (c) educator occupation and (d) technology occupation. This was run over the Blog Authorship Corpus (Schler et al., 2006) packaged with DLATK. Here color represents the words frequency in the corpus (grey to red for infrequent to frequent) and size represents correlation strength.

Comparison to social-scientific tools. Traditional programs for text analysis in the social sciences are based on dictionaries (list of words associated with a particular psychological 'construct'



(a) Age (pos) (b) Age (neg)

(c) Educator (d) Technology Worker

Figure 2: 1- to 3-grams correlated with age and occupation class.

or language categories, such as 'positive emotion' or references to work and occupational terms). According to citations, the most popular tool is Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), followed by DICTION (Hart, 1984) and the General Inquirer (Stone et al., 1966). For a given document, these programs provide the relative frequency of occurrence of terms from the dictionaries. The use of dictionaries has the advantage that they provide relatively parsimonious language in a given text sample, and that the results are in principle comparable across studies. DLATK also reproduces the functionality of these dictionary-based approaches. Dictionaries, however, are often opaque units of analysis, as their overall frequency counts are determined by a few highly frequent words. If these words are ambiguous, interpretations of dictionary-based results can be misleading (Schwartz et al., 2013a). DLATK allows for the determination of words which drive a given dictionary category, and it can also produce data-driven lexica based on predictive models over ngrams, or even find, within a given dictionary category, the words most associated with.

DLATK can provide researchers with enough information to generate hypotheses and clarify the "nomological net" of a construct (Cronbach and Meehl, 1955); That is, help identify the psychological and social processes and constructs that relate to (are sufficiently correlated with) the outcome under investigation. Further, the fact that DLATK incorporates language features and controls in prediction tests allows the researcher to gauge how much construct-related variance is captured in language compared to meaningful demo-

graphic or socioeconomic baselines.

6 Evaluations

DLATK has been used as a data analysis platform in over 30 peer-reviewed publications, with venues ranging from general-interest (PLoS ONE: Schwartz et al., 2013b) to computer science methods proceedings (EMNLP: Sap et al., 2014) to psychology journals (JPSP: Park et al., 2015).

The most straightforward use for DLATK is to provide insight on linguistic features associated with a given outcome, the *differential language analyses* presented in Schwartz et al. (2013b). Other works to primarily use DLATK for correlation-type analyses examine outcomes like age (Kern et al., 2014b), gendered language and stereotypes (Park et al., 2016; Carpenter et al., 2016b), and efficacy of app-based well-being interventions (Carpenter et al., 2016a).

Another area one can evaluate the utility of DLATK is in building predictive models. Table 1 summarizes some predictive models reported in peer-reviewed publications. DLATK works to create models at multiple scales, i.e., for predicting aspects of single messages (e.g., tweet-wise temporal orientation; Schwartz et al., 2015), or predicting user-level attributes (e.g., severity of depression; Schwartz et al., 2014), or predicting community-level health outcomes (e.g., heart disease mortality; Eichstaedt et al., 2015).

Outcome	Score	Source
Demographic (user-level)		
Age	$R = 0.83$	Sap et al. (2014)
Gender	Acc $= 0.92$	
Big-Five Personality (user-level)		
Openness	$R = 0.43$	Park et al. (2015)
Conscientiousness	$R = 0.37$	
Extraversion	$R = 0.42$	
Agreeableness	$R = 0.35$	
Neuroticism	$R = 0.35$	
Temporal orientation (message-level)		
3-way classif	Acc $= 0.72$	Schwartz et al. (2015)
Intensity & affect (message-level)		
Intensity	$R = 0.85$	Preoţiuc-Pietro et al. (2016)
Affect	$R = 0.65$	
Mental health (user-level)		
PTSD	AUC $= 0.86$	Preoţiuc-Pietro et al. (2015)
Depression	AUC $= 0.87$	
Degree of dprssn	$R = 0.39$	Schwartz et al. (2014)
Physical health (US county-level)		
Heart disease mortality	$R = 0.42$	Eichstaedt et al. (2015)

Table 1: Survey of predictive model scores trained using DLATK in peer-reviewed publications. Scores reported are: R: Pearson correlation; Acc: accuracy; AUC: area under the ROC curve.

7 Conclusion

DLATK has been under development for over five years. We have discussed some of its core functionality, including support for extra-linguistic features, multiple levels of analysis, and continuous variables. However, its biggest benefits may be flexibility and reliability due to many years of refinement over dozens of projects. We aspire for DLATK to serve as *a multipurpose Swiss Army Knife* for the researcher who is trying to understand the manifestations of social, psychological and health factors in the lives of language users.

Acknowledgments

This work was supported, in part, by the Templeton Religion Trust (grant TRT-0048). DLATK is an open-source project out of the University of Pennsylvania and Stony Brook University. We wish to thank all those who have contributed to its development, including, but not limited to: Youngseo Son, Mohammadzaman Zamani, Sneha Jha, Megha Agrawal, Margaret Kern, Gregory Park, Lukasz Dziuzinski, Phillip Lu, Thomas Apicella, Masoud Rouhizadeh, Daniel Rieman, Selah Lynch and Daniel Preoţiuc-Pietro.

References

Alan Agresti and Barbara Finlay. 2008. *Statistical Methods for the Social Sciences*. Allyn & Bacon, Incorporated.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Jordan Carpenter, P. Crutchley, R. D. Zilca, H. A. Schwartz, L. K. Smith, A. M. Cobb, and A. C. Parks. 2016a. Seeing the "big" picture: Big data methods for exploring relationships between usage, language, and outcome in internet intervention data. *Journal of Medical Internet Research*, 18(8).

Jordan Carpenter, D. Preoţiuc-Pietro, L. Flekova, S. Giorgi, C. Hagan, M. Kern, A. Buffone, L. Ungar, and M. Seligman. 2016b. Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychological and Personality Science*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. *ACL 2014*, 51.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Olive Jean Dunn. 1961. Multiple comparisons among

means. *Journal of the American Statistical Association*, 56(293):52–64.

Johannes C Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, C. Weeg, E. E. Larson, L. H. Ungar, and M. Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26:159–169.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL*, pages 42–47.

Roderick P Hart. 1984. *Verbal style and the presidency: A computer-based analysis*. Academic Pr.

Margaret L Kern, J. C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell, M. Kosinski, S. M. Ramones, and M. Seligman. 2014a. The online social self: An open vocabulary approach to personality. *Assessment*, 21:158–169.

Margaret L Kern, J. C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, M. Kosinski, L. Dziurzynski, and M. Seligman. 2014b. From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology*, 50:178–188.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proc. of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Wes McKinney. 2011. pandas: a foundational python library for data analysis and statistics.

Scott Menard. 2002. *Applied logistic regression analysis*. 106. Sage.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Polit. Anal.*, 16(4):372–403.

Greg Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, D. J. Stillwell, M. Kosinski, L. H. Ungar, and M. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108:934–952.

Gregory Park, D. B. Yaden, H. A. Schwartz, M. L. Kern, J. C. Eichstaedt, M. Kosinski, D. Stillwell, L. H. Ungar, and M. Seligman. 2016. Women are warmer but no less assertive than men: Gender and language on facebook. *PloS one*, 11(5):e0155885.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. P. Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proc. of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, NAACL.

Daniel Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. H. Ungar. 2015. Mental illness detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proc. of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL.

C Radhakrishna Rao. 2009. *Linear statistical inference and its applications*, volume 22. John Wiley & Sons.

Maarten Sap, G. Park, J. C. Eichstaedt, M. L. Kern, D. J. Stillwell, M. Kosinski, L. H. Ungar, and H. A. Schwartz. 2014. Developing age and gender predictive lexica over social media. In *EMNLP*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium*.

H Andrew Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *Proc. of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, ACL, pages 118–125.

H Andrew Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, S. Ramones, M. Seligman, and L. H. Ungar. 2013a. Choosing the right words: Characterizing and reducing error of the word count approach. In **SEM: Conf on Lex and Comp Semantics*, pages 296–305.

H Andrew Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. H. Ungar. 2013b. Personality, gender, and age in the language of social media: The Open-Vocabulary approach. *PLoS ONE*.

H Andrew Schwartz, G. Park, M. Sap, E. Weingarten, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, J. Berger, M. Seligman, and L. Ungar. 2015. Extracting human temporal orientation from Facebook language. In *NAACL*.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Laura Smith, S. Giorgi, R. Solanki, J. C. Eichstaedt, H. A. Schwartz, M. Abdul-Mageed, A. Buffone, and Lyle H. Ungar. 2016. Does 'well-being' translate on twitter? In *EMNLP*.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.