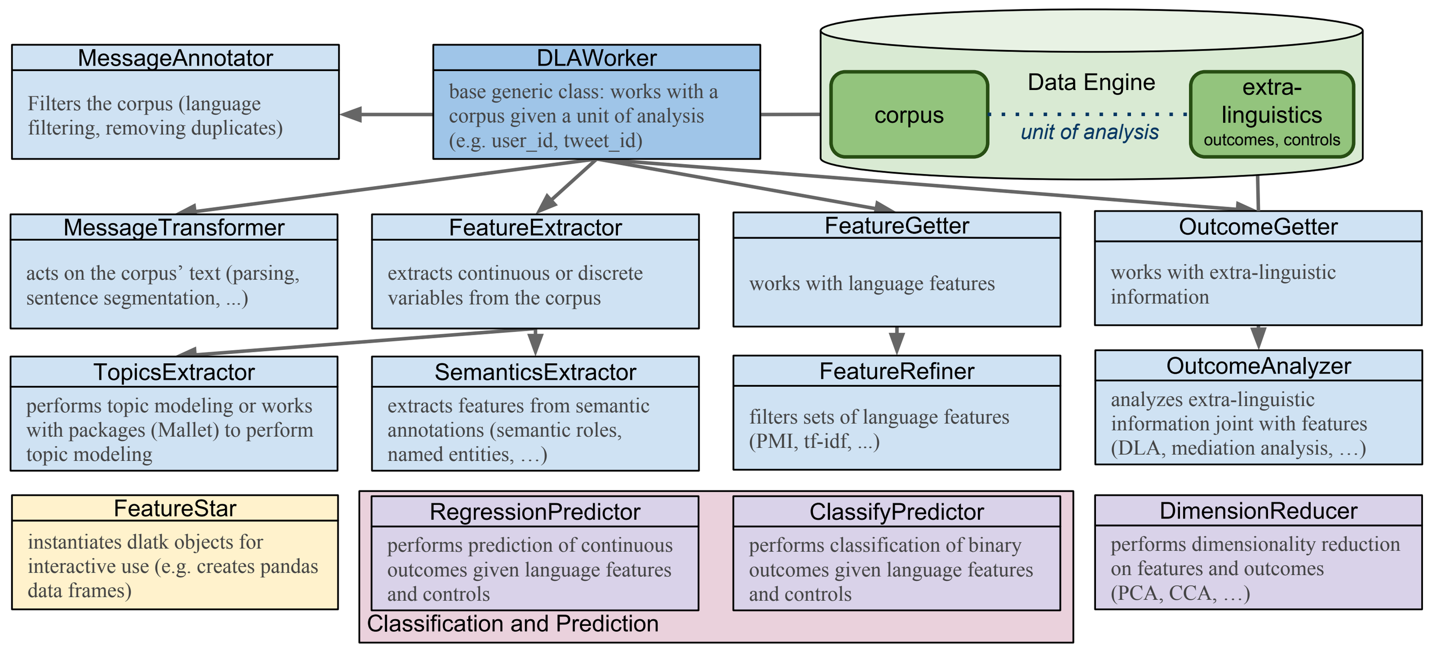
DLATK: Differential Language Analysis ToolKit

H. Andrew Schwartz†, Salvatore Giorgi‡, Maarten Sap§, Patrick Crutchley‖, Johannes C. Eichstaedt‡, Lyle Ungar‡

† Stony Brook University ‡ University of Pennsylvania § University of Washington ‖ Qntfy

Natural Language Processing for Social Scientific Applications



Key Functionality 1: Multiple Levels of Analysis

DLATK allows one to work with a single corpus at multiple levels of analysis (document, user, date, community). At each level one can incorporate extra-linguistic information

- *Document*: time, location, likes
- *User*: demographics, medical records, questionnaire responses
- *Community*: Census or CDC data

Key Functionality 2: Extra-linguistic information

DLATK enables incorporation of "extra-linguistic" or human-/community-level attributes (e.g. examples of such information: age, gender, personality, health, income, education-level.)

- Differential language analysis can utilize as either an 'outcome' or 'control' to reveal distinguishing language.
- Prediction can incorporate alongside linguistic features and has functionality to handle the heterogeneity of including both linguistic and human/community features.

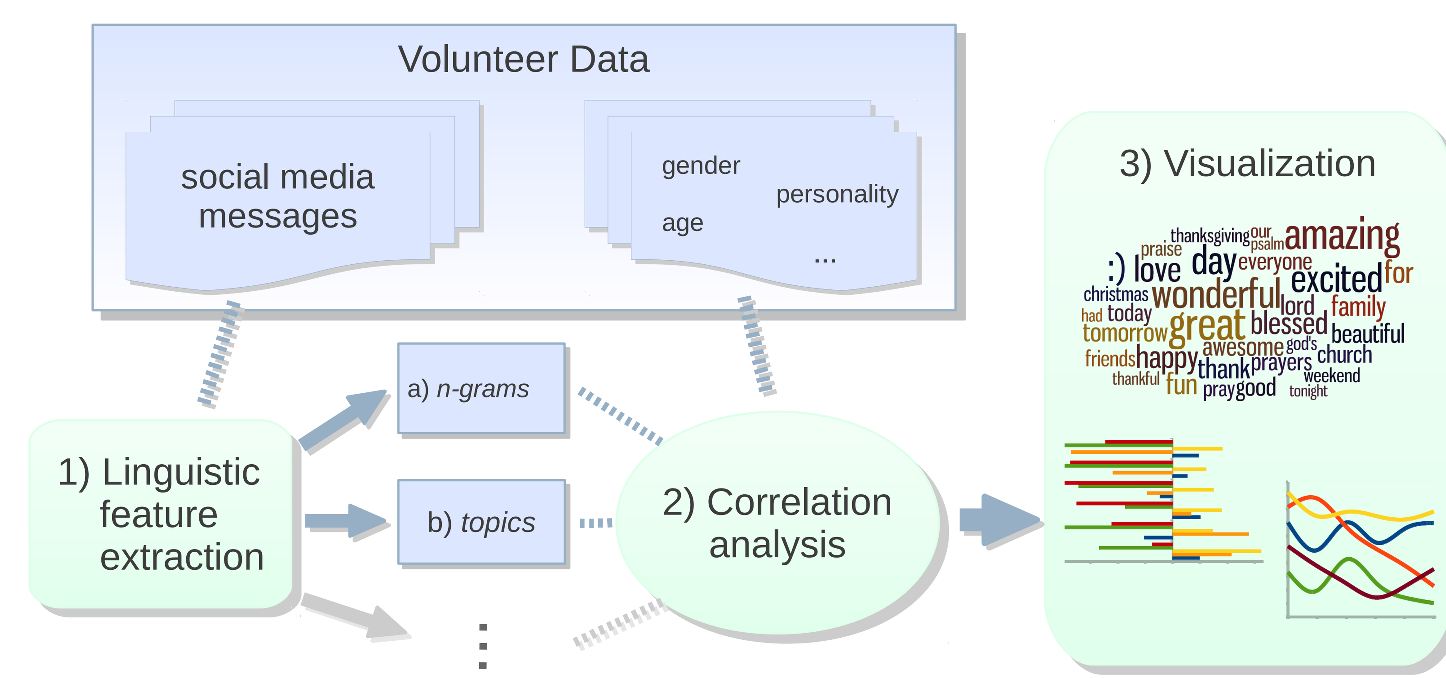
Key Functionality 3: Integration of Popular Packages

- *Python*: numpy, scikit-learn, statsmodels, pandas
- *NLP*: Stanford parser, TweetNLP, NLTK
- *Other*: Mallet, IBM wordcloud
- *Install*: pip, conda, GitHub

Analysis Pipelines

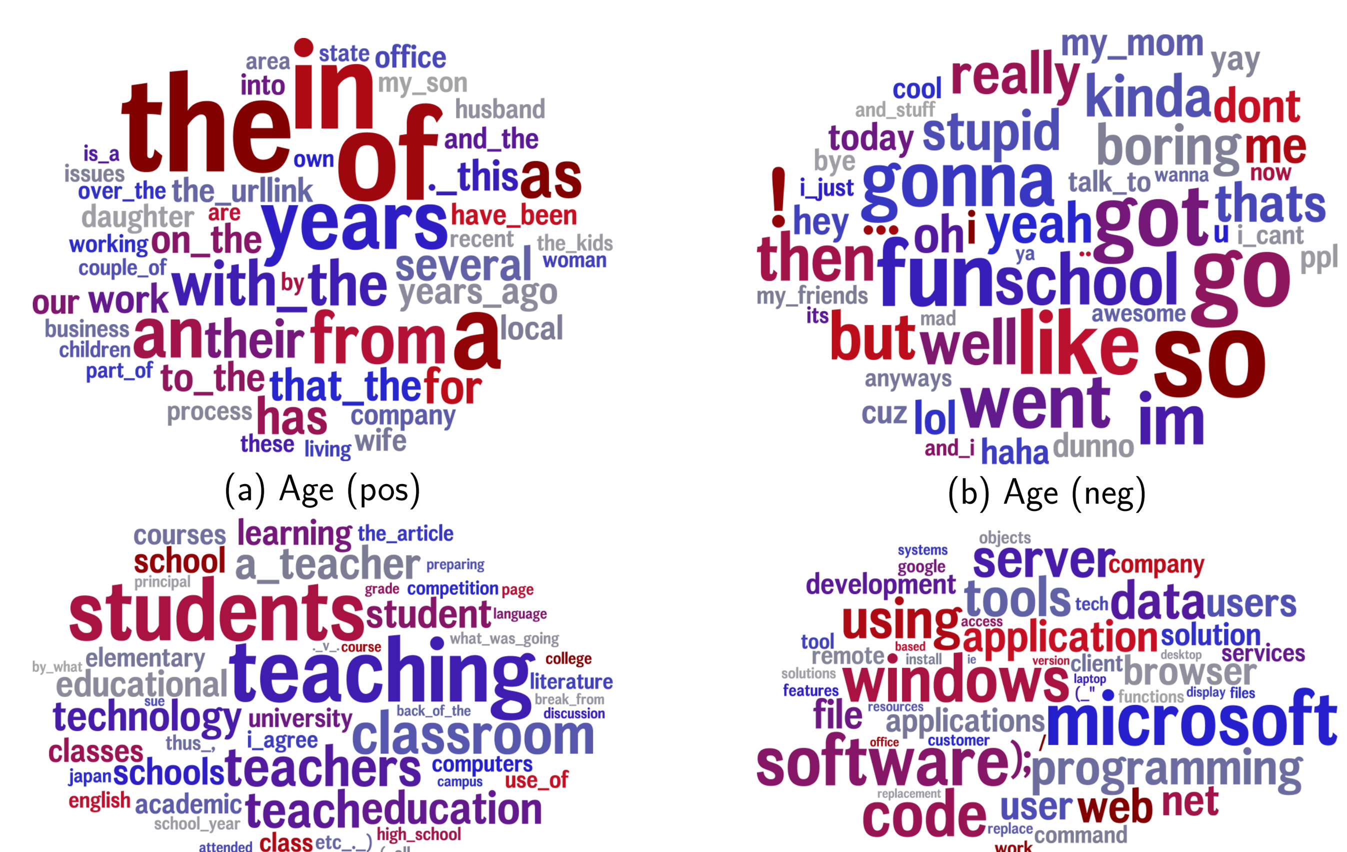
- Feature Extraction (n-grams, part of speech, topics / lexica)
- Correlation (Differential Language Analysis)
- Prediction and Classification
- Dimensionality reduction and clustering
- Mediation
- Wordcloud visualization

Use Case 1: Differential Language Analysis



Differential Language Analysis (DLA): 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* [1].

- Prototypical use of DLATK is to perform DLA
- Goal is to *produce language* that is most related to or independently discriminant of outcomes
- Univariate, per-feature fashion or with a limited set of control variables
- Corrects for multiple comparisons using the Benjamini-Hochberg method of FDR correction



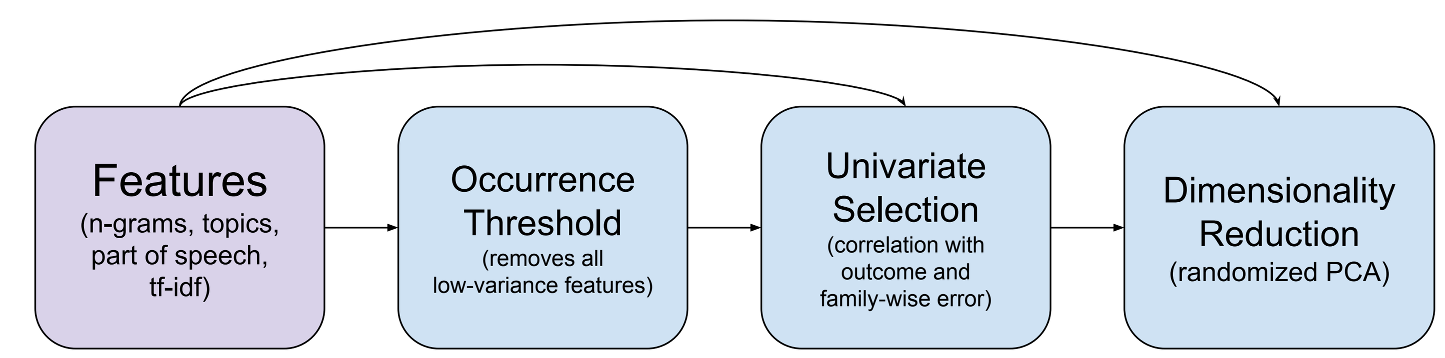
(a) Age (pos) (b) Age (neg) (c) Educator (d) Technology Worker

Figure: 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 [2] packaged with DLATK. Here color represents the word's frequency in the corpus (grey to red for infrequent to frequent) and size represents correlation strength.

Contact Information

- Web: http://dlatk.wwbp.org
- GitHub: http://github.com/dlatk/
- Email: has@cs.stonybrook.edu, sgiorgi@seas.upenn.edu

Use Case 2: Prediction / Classification



Outcome	Score	Source
Demographic (user-level)		
Age	$R = 0.83$	[3]
Gender	Acc $= 0.92$	
Big-Five Personality (user-level)		
Openness	$R = 0.43$	[4]
Conscientiousness	$R = 0.37$	
Extraversion	$R = 0.42$	
Agreeableness	$R = 0.35$	
Neuroticism	$R = 0.35$	
Temporal orientation (message-level)		
3-way classif	Acc $= 0.72$	[5]
Intensity & affect (message-level)		
Intensity	$R = 0.85$	[6]
Affect	$R = 0.65$	
Mental health (user-level)		
PTSD	AUC $= 0.86$	[7]
Depression	AUC $= 0.87$	
Degree of dprssn	$R = 0.39$	[8]
Physical health (US county-level)		
Heart disease mortality	$R = 0.42$	[9]

Table: 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.

Acknowledgements

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

[1] 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.
Personality, gender, and age in the language of social media: The Open-Vocabulary approach.
PLoS ONE, 2013.

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

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

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

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

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

[7] Daniel Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. H. Ungar.
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, 2015.

[8] H Andrew Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar.
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, 2014.

[9] 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.
Psychological language on Twitter predicts county-level heart disease mortality.
Psychological Science, 26:159–169, 2015.