Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students

Philip Resnik
University of Maryland
College Park, MD 20742
resnik@umd.edu

Anderson Garron
University of Maryland
College Park, MD 20742
agarron@cs.umd.edu

Rebecca Resnik
Mindwell Psychology Bethesda
5602 Shields Drive
Bethesda, MD 20817
drrebeccaresnik@gmail.com

Abstract

We investigate the value-add of topic modeling in text analysis for depression, and for neuroticism as a strongly associated personality measure. Using Pennebaker's Linguistic Inquiry and Word Count (LIWC) lexicon to provide baseline features, we show that straightforward topic modeling using Latent Dirichlet Allocation (LDA) yields interpretable, psychologically relevant "themes" that add value in prediction of clinical assessments.

1 Introduction

In the United States, where 25 million adults per year suffer a major depressive episode (NAMI, 2013), identifying people with mental health problems is a key challenge. For clinical psychologists, language plays a central role in diagnosis: many clinical instruments fundamentally rely on what is, in effect, manual coding of patient language. Automating language assessment in this domain potentially has enormous impact, for two reasons. First, conventional clinical assessments for affective disorders (e.g. The Minnesota Multiphasic Personality Inventory, MMPI) are based on norm-referenced self-report, and therefore depend on patients' willingness and ability to report symptoms. However, some individuals are motivated to underreport symptoms to avoid negative consequences (e.g. active duty soldiers, parents undergoing child custody evaluations), and others lack the self awareness to report accurately.[1] Second, many people – e.g. those without adequate insurance or in rural areas – cannot access a clinician qualified to perform a psychological evaluation (Sibelius, 2013; APA, 2013). There is enormous value in inexpensive screening measures that could be administered in primary care and by social workers and other providers.

We take as a starting point the well known lexicon-driven methods of Pennebaker and colleagues (LIWC, Pennebaker and King (1999)), which relate language use to psychological variables, and improve on them straightforwardly using topic modeling (LDA, Blei et al. (2003)). First, we show that taking automatically derived topics into account improves prediction of neuroticism (emotional instability, John and Srivastava (1999)), as measured by correlation with widely used clinical instruments, when compared with lexically-based prediction alone. Neuroticism is of particular interest as a personality measure because higher scores on neuroticism scales are consistent with increased distress and more difficulty coping; individuals with high levels of neuroticism may also be at higher risk of psychiatric problems categorized as Axis I in the DSM-IV (Association, 2000), including the internalizing disorders (depression, anxiety).[2] Second, we show a similar correlation improvement result for prediction of depression, adding improvement

[1] As evidenced by the fact that assessments such as the MMPI-2-RF include validity scales to detect, e.g., defensiveness, atypical responses, and overly positive self-portrayals (Tellegen et al., 2003).

[2] The Diagnostic and Statistical Manual of Mental Disorders is a widely used organization of mental health conditions; it served as the standard for diagnosis from 1994 until the release of the (quite controversial) DSM-5 in May, 2013. Axis I in the DSM-IV includes all the major diagnostic categories, excluding mental retardation and personality disorders.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1348–1353,
Seattle, Washington, USA, 18-21 October 2013.

on precision (with no decrease in recall), as well as comparison with human performance by clinical psychologists. Third, we show that LDA has identified meaningful, population-specific themes that go beyond the pre-defined LIWC categories and are psychologically relevant.

2 Predicting neuroticism

2.1 Experimental framework

Data. We utilize a collection of 6,459 stream-of-consciousness essays collected from college students by Pennebaker and King (1999) between 1997 and 2008, averaging approximately 780 words each. Students were asked to think about their thoughts, sensations, and feelings in the moment and "write your thoughts as they come to you".

Each essay is accompanied by metadata for its author, which includes scores for the Big-5 personality traits (John and Srivastava, 1999): agreeableness, conscientiousness, extraversion, neuroticism, and openness. Because Big-5 assessment can be done using a variety of different survey instruments (John et al., 2008), and different instruments were used from year to year, we treat the data from each year as an independent dataset.

Any author missing any Big-5 attribute was excluded. Essays were tokenized using NLTK (Bird et al., 2009) and lowercased, eliminating those that failed tokenization because of encoding issues. This resulted in a dataset containing 4,777 essays with associated Big-5 metadata.

LIWC features. For each document, we calculate the number of observed words in each of Pennebaker and King's 64 LIWC categories. These include, among others, syntactic categories (e.g. pronouns, verbs), affect categories (e.g. negative emotion words, anger words), semantic categories (e.g. causation words, cognitive words), and topical categories (e.g. health, religion). For instance, the anger category contains 190 word patterns specifying, for example, words descriptive of contexts involving anger (e.g. *brutal*, *hostile*, *shoot*) and words that would be used by someone when angry (e.g. *bullshit*, *hate*, *moron*). We also explored including essays' average sentence length and total word-count, as an initial proxy for language complexity, which often figures into psychological assessments. However, results adding these features to LIWC did not differ significantly from LIWC alone, and for brevity we do not report them.[3]

LDA features. We use vanilla LDA as implemented in the Mallet toolkit (McCallum, 2002), developing a k-topic model on just training documents, and using the posterior topic distribution for each training and test document as a set of k features. Mallet's stoplist and default parameters were used for burn-in, lag, number of iterations, priors, etc. Details on train/test splits and number k of topics appear below.

LIWC+LDA features. The union of the LIWC features (one feature per category) and LDA features (one feature per topic).

Prediction. We utilize linear regression in the WEKA toolkit (Hall et al., 2009), estimated on training documents, to predict the neuroticism score associated with the author of each test document.[4]

2.2 Results

Table 1 shows the quality of prediction via linear regression, averaged over the eleven datasets, 1997 through 2008, using Pearson correlation (r) as the evaluation metric. For each year, we used 10-fold cross-validation to ensure proper separation of training and test data. We experimented with LDA using 20, 30, 40, and 50 topics.[5]

A first thing to observe is that the multiple regressions using all LIWC categories produce much stronger correlations with neuroticism than the individual category correlations reported by Pennebaker and King.[6] There the strongest individual correlations with neuroticism for any LIWC categories are .16 (negative emotion words) and -.13 (positive emotion words), though it should be noted that their goal was to validate their categories as a meaningful

[3] Using richer measures of complexity, e.g. Pakhomov et al. (2011), is a topic for future work.

[4] In previous work we have found that multiple linear regression is competitive with more complicated techniques such as SVM regression, though we plan to explore the latter in future work.

[5] Full year-by-year data appears in supplemental materials at `http://umiacs.umd.edu/~resnik/papers/emnlp2013-supplemental/`.

[6] The comparison is not perfect, since they used Big-5 data collected between 1993 and 1998, and we also eliminated some files during preprocessing.

Feature set	LIWC	LDA20	LIWC+LDA20	LDA30	LIWC+LDA30	LDA40	LIWC+LDA40	LDA50	LIWC+LDA50
Average r	0.413	0.384	0.430	0.407	0.442*	0.420	0.459**	0.440	0.443*

Table 1: Prediction quality for neuroticism, for alternative feature sets (Pearson's r). *p$<$ .03, **$p < .02$

way to explore personality differences, not prediction.

As noted in Table 1, paired t-tests (df=10, $\alpha = .05$) establish that, in comparing the cross-year averages, augmenting LIWC with topic features improves average correlation significantly over using LIWC features alone for 30, 40, and 50 topics. LDA features alone do not improve significantly over LIWC.

3 Predicting Depression

3.1 Experimental framework

Data. We use essays collected by Rude et al. (2004) similarly to §2.1; in this case, students were asked to "describe your deepest thoughts and feelings about being in college". Each essay is accompanied by the author's Beck Depression Inventory score (BDI). BDI (Beck et al., 1961) is a widely used 21-question instrument that correlates strongly with ratings of depression by psychiatrists. Following Rude et al., we treat BDI ≥ 14 as the threshold for a positive instance of depression.[7] Text preprocessing was done as in §2.1, with 124 documents in total averaging around 390 words each.

Training/test split. Because only 12 of 124 authors met the BDI ≥ 14 threshold, we did not split randomly, lest the test sample include too few positive instances to be useful. Instead we included a random 6 of the 12 above-threshold cases, plus 24 more items sampled at random, to create a 30-item test set. To form the training set from the complementary items, we added two more copies of each positive instance to help address class imbalance (Batista et al., 2004) and, following Rude et al., we excluded items with with BDI of 0 or 1 as potentially invalid.

Human comparison. We created a set of expert human results for comparison by asking three practicing clinical psychologists to review the test documents and rate whether or not the author is suffering depression.[8] They were asked to "decide how at-risk this person might be for depression", assigning 0 (no significant concerns), 1 (mild concerns, but does not require futher evaluation), or 2 (requires attention, refer for further evaluation). Following recommended practice for cases where different labels are not equally distinct from each other (Artstein and Poesio, 2008), we evaluate inter-coder agreement using Krippendorff's α; our α, computed for ordinal data, is 0.722.

Features. We ran 50-topic LDA on the 4,777 essays from §2.1 plus the BDI training items, using the posterior topic distributions as features as in §2.1. As in §2.1, the LIWC features comprised one count per LIWC category, and LIWC+LDA features were the union of the two.

3.2 Results

Regression on LIWC features alone achieved $r = .288$, and adding topic features improved this substantially to $r = .416$. Treating BDI ≥ 14 as the threshold for positive instances (i.e. that an author is depressed), Table 2 shows that adding topic features improves precision without harming recall. Automatic prediction is more conservative than human ratings, trading recall for precision to achieve comparable F-measure on this test set.[9]

[7] Each question contributes a score value from 0 to 3, so BDI scores range from 0 to 63.

[8] These psychologists all have doctoral degrees, are licensed, and spend significant time primarily in assessment and diagnosis of psychological disorders. None were familiar with the specifics of this study.

[9] A reviewer observes, correctly, that in a scenario where a system is providing preliminary screenings to aid psychologists, the precision/recall tradeoff demonstrated here would potentially be undesirable, since a presumed goal would be to not miss any cases, even at the risk of some false positives. We note, however, that the real world is unfortunately replete with situations where there is significant cost or social/professional stigma associated with interventions or follow-up testing; in such situations it might be high precision that is desirable. These are challenging questions, and the ability to trade off precision versus recall more flexibly is a topic we are interested in investigating in future work.

	P	R	F1
LIWC	.43	.50	.46
LIWC+LDA	.50	.50	.50
Rater 1	.38	.83	.52
Rater 2	.33	.83	.47
Rater 3	.33	.66	.44

Table 2: Prediction quality for depression.

4 Qualitative themes

In order to explore the relevance of the themes uncovered by LDA, the third author, a practicing clinical psychologist, reviewed the 50 LDA categories created in §3.1. Each category, represented by its 20 highest-probability words, was given a readable description. Then, for each category, she was asked: "If you were conducting a clinical interview, would observing these themes in a patient's responses make you more (less) likely to feel that the patient merited further evaluation for depression?"

Table 3 shows the seven topics selected as particularly indicative of further evaluation.[10] These capture population-specific properties in ways that LIWC cannot — for example, although LIWC does have a *body* category, it does not have a category that corresponds to somatic complaints, which often co-occur with depression. Similarly, some words related to energy level, e.g. *tired*, would be captured in LIWC's *body*, *bio*, and/or *health* category, but the LDA theme corresponding to low energy or lack of sleep, another potential depression cue, contains words that make sense there only in context (e.g. *tomorrow*, *late*). Other themes, such as the one labeled HOMESICKNESS, are clearly relevant (potentially indicative of an adjustment disorder), but even more specific to the population and context.

5 Related Work

The application of NLP to psychological variables has seen a recent uptick in community activity. One recent shared task brings together research on the Big-5 personality traits (Celli et al., 2013; Kosinski et al., 2013), and another involved research on identification of emotion in suicide notes (Pestian et al., 2012). Other examples include NLP research on autistic spectrum disorders (Van Santen et al., 2010; Prudhommeaux et al., 2011; Lehr et al., 2013) and dementia (Pakhomov et al., 2011; Lehr et al., 2012; Roark et al., 2011).

With regard to depression, Neuman et al. (2012) develop a corpus-based "depression lexicon" and produce promising screening results, and De Choudhury et al. (2013) predict social network behavior changes related to post-partum depression. Neither, however, evaluates using formal instruments for clinical assessment.

Related investigations involving LDA include Zhai et al. (2012), who use LIWC to provide priors for corpus-specific emotion categories; Stark et al. (2012), who combine LIWC and LDA-based features in classification of social relationships; and Schwartz et al. (2013), who use lexical and topic-based features in Twitter to predict life satisfaction.

6 Conclusions

In this paper, we have aimed for a small, focused contribution, investigating the value-add of topic modeling in text analysis for depression, and for neuroticism as a strongly associated personality measure. Our contribution here is not technical: corpus-specific topics/themes are anticipated by Zhai et al. (2012), and Stark et al. (2012) employ topic-based features for prediction in a supervised setting. Rather, our contribution here has been to show that topic models can get us beyond the LIWC categories to relevant, population-specific themes related to neuroticism and depression, and to support that claim using evaluation against formal clinical assessments. More data (e.g. Kosinski et al. (2013)) and more sophisticated models (e.g. supervised LDA, Blei and McAuliffe (2008), and extensions such as Nguyen et al. (2013)) will be the key to further progress.

Acknowledgments

We are grateful to Jamie Pennebaker for the LIWC lexicon and for allowing us to use data from Pennebaker and King (1999) and Rude et al. (2004), to the three psychologists who kindly took the time to provide human ratings, and to our reviewers for helpful comments. This work has been supported in part by NSF grant IIS-1211153.

[10] All 50 can be found in the supplemental materials at `http://umiacs.umd.edu/~resnik/papers/emnlp2013-supplemental/`.

VEGETATIVE/ENERGY LEVEL	sleep tired night bed morning class early tomorrow wake late asleep long hours day sleeping nap today fall stay time
SOMATIC	hurts sick eyes hurt cold head tired back nose itches hate stop starting water neck hand stomach feels kind sore
NEGATIVE/TROUBLE COPING	don('t) hate doesn care didn('t) understand anymore feel isn('t) stupid make won('t) wouldn talk scared wanted wrong mad stop shouldn('t)
ANGER/FRUSTRATION	hate damn stupid sucks hell shit crap man ass god don blah thing bad suck doesn fucking fuck freaking real
HOMESICKNESS	home miss friends back school family weekend austin parents college mom lot boyfriend left houston visit weeks wait high homesick
EMOTIONAL STRESS	feel feeling thinking makes make felt feels things nervous scared lonely feelings afraid moment happy worry comfortable stress excited guilty
ANXIETY	feel happy things lot sad good makes bad make hard mind happen crazy cry day worry times talk great wanted

Table 3: LDA-induced themes related to depression.

References

[APA2013] APA. 2013. The critical need for psychologists in rural america. http://www.apa.org/about/gr/education/rural-need.aspx, Downloaded September 16, 2013.

[Artstein and Poesio2008] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

[Association2000] American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR)*. American Psychiatric Association, 4th edition, July.

[Batista et al.2004] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June.

[Beck et al.1961] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and J1 Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561.

[Bird et al.2009] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly.

[Blei and McAuliffe2008] David Blei and Jon McAuliffe. 2008. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Celli et al.2013] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Computational personality recognition (shared task) workshop. In *International Conference on Weblogs and Social Media*. AAAI.

[De Choudhury et al.2013] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 3267–3276. ACM.

[Hall et al.2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[John and Srivastava1999] Oliver P John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2:102–138.

[John et al.2008] O. P. John, L. P. Naumann, and C. J. Soto. 2008. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of personality: Theory and research*, pages 114–158.

[Kosinski et al.2013] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

[Lehr et al.2012] Maider Lehr, Emily Tucker Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *INTERSPEECH*.

[Lehr et al.2013] Maider Lehr, Izhak Shafran, Emily Prudhommeaux, and Brian Roark. 2013. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *Proceedings of NAACL-HLT*, pages 211–220.

[McCallum2002] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[NAMI2013] NAMI. 2013. Major depression fact sheet, April. http://www.nami.org/Template.cfm?Section=depression.

[Neuman et al.2012] Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. 2012. Proactive screening for depression through metaphorical and automatic

text analysis. *Artif. Intell. Med.*, 56(1):19–25, September.

[Nguyen et al.2013] Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Neural Information Processing Systems*.

[Pakhomov et al.2011] Serguei Pakhomov, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. Computerized assessment of syntactic complexity in alzheimers disease: a case study of iris murdochs writing. *Behavior Research Methods*, 43(1):136–144.

[Pennebaker and King1999] James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

[Pestian et al.2012] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, Christopher Brew, et al. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl. 1):3.

[Prudhommeaux et al.2011] Emily T Prudhommeaux, Brian Roark, Lois M Black, and Jan van Santen. 2011. Classification of atypical language in autism. *ACL HLT 2011*, page 88.

[Roark et al.2011] Brian Roark, Margaret Mitchell, J Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090.

[Rude et al.2004] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

[Schwartz et al.2013] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle Ungar. 2013. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

[Sibelius2013] Kathleen Sibelius. 2013. Increasing access to mental health services, April. http://www.whitehouse.gov/blog/2013/04/10/increasing-access-mental-health-services.

[Stark et al.2012] Anthony Stark, Izhak Shafran, and Jeffrey Kaye. 2012. Hello, who is calling?: can words reveal the social nature of conversations? In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 112–119. Association for Computational Linguistics.

[Tellegen et al.2003] A. Tellegen, Y.S. Ben-Porath, J.L. McNulty, P.A. Arbisi, and B. Graham, J.R.and Kaemmer. 2003. *The MMPI-2 Restructured Clinical Scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.

[Van Santen et al.2010] Jan PH Van Santen, Emily T Prud'hommeaux, Lois M Black, and Margaret Mitchell. 2010. Computational prosodic markers for autism. *Autism*, 14(3):215–236.

[Zhai et al.2012] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. lda: a flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 879–888, New York, NY, USA. ACM.