Dictionaries and Decision Trees for the 2019 CLPsych Shared Task

Micah Iserman, Taleen Nalabandian, and **Molly E. Ireland**
Department of Psychological Sciences, Texas Tech University, Lubbock, Texas
first.last@ttu.edu

Abstract

In this summary, we discuss our approach to the CLPsych Shared Task and its initial results. For our predictions in each task, we used a recursive partitioning algorithm (decision trees) to select from our set of features, which were primarily dictionary scores and counts of individual words. We focused primarily on Task A, which aimed to predict suicide risk, as rated by a team of expert clinicians (Shing et al., 2018), based on language used in SuicideWatch posts on Reddit. Category-level findings highlight the potential importance of social and moral language categories. Word-level correlates of risk levels underline the value of fine-grained data-driven approaches, revealing both theory-consistent and potentially novel correlates of suicide risk that may motivate future research.

1 Introduction

The shared task for this year's CLPsych workshop focused on predicting Reddit users' risk for suicide (none, low, moderate, and severe, as coded by clinical psychologists with suicide expertise) based on language used in their posts (Shing et al. 2018; for a review, see Zirikly et al. 2019). Reddit is a social media website that hosts over 138,000 active forums (or subreddits; as of 2017[1]) in which users can post on any topics of interest.

Social media sites like Reddit, Facebook, and Twitter have increasingly become an important source of data for researchers. Studies have demonstrated how language use in social media posts reflects various psychological processes, ranging from personality (Youyou et al., 2017) to mental health (e.g., postpartum depression; De Choudhury et al., 2014). For instance, Eichstaedt et al. (2018) were able to accurately distinguish depressed patients from non-depressed controls based on Facebook statuses posted before the date of their diagnosis.

Certain language categories have been implicated as markers of mental health conditions (such as anxiety; Dirkse et al., 2015). Relevant to this shared task, suicidal ideation tends to be positively correlated with rates of first-person singular pronoun use (Stirman and Pennebaker, 2001) and negative emotion word use (e.g., anger, sadness; Coppersmith et al., 2016). Self-focused and negative language appear to be associated with psychological distress in general, relating to a variety of mental health issues, such as psychosis (Fineberg et al., 2016), neuroticism (Tackman et al., 2018), and depression (Rude et al., 2004). Notably, self-focused language correlates with psychological distress across a variety of contexts (such as across public Facebook posts; De Choudhury et al., 2014), whereas the use of negative emotional language tends to be limited to more private or intimate contexts (such as in conversations with romantic partners; Baddeley et al., 2012).

Based on previous research, we went into this year's shared task with a particular interest in first-person singular pronouns and overtly negative content words. Although our models cast a wide net, making use of all available lexicons, we expected categories relating to negative affect, self-focus, and social distance to be most predictive of suicide risk, as rated by expert coders.

2 Method

Preprocessing. We first removed any entries not from users in the task A or B sets, or with only "nan" as the post body. This left 11,856 posts from 329 users, which we cleaned automatically in order to (a) standardize encoding, such as for quotation or apostrophe marks; (b) remove some code elements, such as HTML tags or characters; (c)

[1] https://www.redditinc.com/

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 188–194
Minneapolis, Minnesota, June 6, 2019.

remove some formatting that could make identifying word or sentence boundaries more difficult, such as periods within word; (d) standardize some common typing-related practices, such as repeating characters within some words for emphasis (e.g., "reeeeeaallllly"); and (e) replace some standard formatted elements with tags, such as URLs, references to subreddits, and simple emojis.

After cleaning and tokenizing texts, we applied a spelling correction processes in two phases: First, we applied a more generic version of the process (to be described), and checked its output for (a) miscorrections (such as specialized terminology like "reddit", "macbook", and "moba"), which we added to the list defining correctly spelled words, and (b) frequent misspellings not caught by the process, which we added to a map between correctly spelled words and their misspelled instances. This caught some of the most frequent miscorrections and missed misspellings, but was limited by available time. We applied the process again with these refinements and allowed it to correct the misspellings it identified.

The spelling correction process used the hunspell package (Ooms, 2018) and its US English dictionary to mark words as misspelled (on its own at first, then manually supplemented; only considering words over 3 characters long). The process then measured edit distance (optimal string alignment, calculated with the stringdist package; van der Loo, 2014) between each marked and unmarked (correctly spelled) word found in the text. If a misspelled word was within 2 edit distance of one and only one correctly spelled word, it was considered a matched to that word. If a word was within 1 edit distance of multiple words, these were considered potential matches, and the qgram and soundex distance were calculated between them and the original misspelling—a combination of these new distances and the frequency of the potential matches determined which of these would claim the misspelling (as shown in equation 1, where a is the misspelling, and b is each word in the set of words within 1 edit distance; document frequency is the number of posts in which the word appears).

$$\underset{b \in matches}{\arg\min} \frac{qgram(a,b) + soundex(a,b)}{documentfrequency(b)} \quad (1)$$

If a misspelled word did not meet the edit distance criteria, corrections suggested by hunspell were considered: If any of these were more frequent than the misspelling, the most frequent of them was considered its correction. Otherwise, if any suggested corrections contained spaces (i.e., the misspelling was suggested to be a combination of words), and if the individual suggested words were all found in the texts, the most frequent combination was taken to be its correction.

AFINN	Nielsen (2011)
Hu & Liu	Hu and Liu (2004)
General Inquirer	Stone and Hunt (1963)
labMT	Dodds et al. (2011)
LIWC	Pennebaker et al. (2015)
Lusi	Ireland and Iserman (2018)
Moral Foundations	Frimer et al. (2018)
Netspeak	Ireland and Iserman (2019)
NRC	Mohammad (2017)
Senticnet	Cambria et al. (2010)
SentimentDictionaries	Pröllochs et al. (2018)
SentiWordNet	Baccianella et al. (2010)
Slangsd	Wu et al. (2016)
Vader	Hutto and Gilbert (2014)
Whissell	Whissell (1989)
Age and Gender	Sap et al. (2014)
PERMA	Schwartz et al. (2016)

Table 1: Dictionaries/Lexicons.

Most of the genuine spelling errors appeared to be typing related (e.g., *ddin't*, *favirite*), with other common errors seeming to be formatting related (such as words being combined, or parts of words being appended to others). Other corrections effectively standardized across certain word variants (e.g., forms of *highschool* to *high-school*, words with commonly omitted apostrophes to have apostrophes, or British to English spellings) or casual language (e.g., *wana*, *coulda*).

Features. Table 1 lists the dictionaries we used to score the texts. Those with multiple words or parts of words in single entries had each term searched for exactly in the raw text. Otherwise, terms were searched for in the tokens extracted from all texts, allowing for partial matches when words were marked at the beginning or end with an asterisk (as in the case of dictionaries intended for Linguistic Inquiry and Word Count; LIWC; Pennebaker et al., 2015). We also used LIWC to process its internal 2015 dictionary, prior to which we trimmed 3 or more sequential PERSON tags to 1, as some posts with many tags (such as posts con-

taining code examples) caused entries to overflow.

Many individual categories were nearly identical, so we removed those correlating over .9 with any other category (done iteratively, such that only one of each similar category was retained, preferring to retain LIWC categories). In addition to these pre-built dictionaries, we considered each manually replaced tag (such as those for proper names, subreddits, and emojis) to be its own category, counting their instances up and including them as features. The final set of features included dictionary categories and counts of each token, as well as Language Style Matching (Ireland and Pennebaker, 2010) between (a) each post and the posting user's average language style across all of their posts, and (b) each post and the average language style of the subreddit in which it was posted.

Model. We ended up using a simple recursive partitioning model (as calculated by the rpart package; Therneau and Atkinson, 2018), with all features predicting the ratings for each task (with tasks simply defining the particular posts to be included). For final predictions, we trained each model on the full task specific training data (though the submitted task C model was accidentally trained on the task B data), then aggregated within user, assigning each user the rating that had the largest average probability across their posts (as depicted in Figure 2).

We also briefly considered other models (with a small set of features, selected by their correlations with any rating or the continuous rating scale), such as linear regressions predicting a numeric version of the ratings (with their predictions being binned), separate logistic regressions predicting each category, and multinomial logistic regressions (both with the subset of features, and an elastic net regularized version with all features), as well as a random forest model, but these all either performed worse than our final model in our own testing splits, or seemed to overly capitalize on priors (tending to predict only the most common ratings, even more than our final models). Of course, there are many strategies that might be explored to address the uneven distribution of ratings, but our first step in this brief analysis was to compare the performance of a few different models. We also considered mixed-effects models estimating a per-user intercept adjustment, but these did not work well, at least for task A, since most users had only 1 r/SuicideWatch post.

Task	*a*	*b*	*c*	*d*	mean	rank
A	.667	.200	.140	.600	.402	6th
B	.000	.000	.000	.591	.148	11th
C		.000	.000	.353	.118	8th

Table 2: F1 scores for each rating level in each task. Rankings out of 12, 11, and 8 for each task respectively.

3 Results and Discussion

As the results on the official test sample depict (Table 2), our models tended to only predict extreme ratings, capitalizing on the prior ratings distributions. Because the model could perform well in each task by identifying features that marked *a*- or *d*-rated users (with *d* being the most common rating; as ratings applied per user, across posts), trees in tasks B and C in particular tended to be very simple. This tendency was exacerbated by the fact that some users had multiple posts, which meant any idiosyncrasies in word use or topics of discussion among prolific posters could be used as a cue for their entire rating level.

In terms of differences in higher-level language dimensions, posts in r/SuicideWatch were more likely to be coded as high risk (category *d*) if they had higher Clout scores (used *I* more, and *we* and *you* less), talked about family (e.g., *dad*, *grandma*) at relatively low rates, and used less positive affective language (as indexed by sentimentr). With respect to moral language, higher-risk posts referred more often to care (e.g., *help*, *pity*; Moral Foundations Dictionary, Frimer et al. 2018) as well as both vice and virtue, as measured by the General Inquirer lexicons (e.g., *ability*, *burn*). In terms of sentence structure and punctuation, higher risk posts used more periods, fewer parentheses, and more hyperbolic or extreme statements (e.g., *quite*, *extreme*; overstatement, General Inquirer), and fewer third-person singular pronouns (e.g., *him*, *she*; LIWC shehe), relative to lower-risk posts.

At the word level, lower-risk posts (ratings *a* and *b*) seem to be more social, including more communicative words (like *called*, *said*, and *told*) and words connoting warmth (such as *comfortable*), more *we*, and specific family references (such as *brother*, *cousin*, and *mom*). Higher-risk posts (ratings *c* and *d*) seem to reflect more certainty, finality, or black-and-white thinking (*every*, *anymore*, *anything*, *end*), more focus on physical harm (*knife*, *hurts*) and life or death (*alive*,

die). Higher risk posts also included a number of negations (*don't*, *can't*, *no*; see Weintraub 1989). Swearing (e.g., *fucking*) was indicative of the highest risk level as well, perhaps reflecting intense negative affect or disregard for social norms. Perhaps the most notable and theory-consistent word-level correlates of the highest risk level were self-focused pronouns, including *I*, *me*, and *myself*. Self-focused pronouns are commonly associated with depression (Rude et al., 2004), suicidality (Stirman and Pennebaker, 2001), or, more broadly, vulnerability to stress (Tackman et al., 2018). See Figure 1 for additional word-level correlates of risk-level ratings.

Some of the linguistic correlates of risk categorization are consistent with our prediction that posts would be viewed as indicating higher suicide risk to the degree that they used more negative and socially distant language. The interpersonal theory of suicide (Van Orden et al., 2010) is a leading psychological model of suicide risk. The theory proposes that people are more likely to attempt or die by suicide to the degree that they feel a thwarted desire to belong, believe they are a burden on their loved ones, and have acquired the capability to die (or no longer fear death). Talking infrequently about family and using fewer third-person singular references that might refer to other people in their lives could reflect social isolation.

Although not predicted *a priori*, the moral language correlates seem to be relatively face valid. People using care-related words from the revised Moral Foundations Dictionary (Frimer et al., 2018) may have simply been requesting help more explicitly than people who did not use words such as *help*, *mercy*, or *comfort* (Graham et al., 2009; Sagi and Dehghani, 2014). The General Inquirer *vice* and *virtue* categories (Stone and Hunt, 1963) are less intuitive, but discussing basic moral questions of good and evil may reflect the thwarted belonging dimension of the interpersonal theory of suicide (e.g., discussing wanting to be good but disappointing loved ones; Van Orden et al. 2010).

The punctuation categories are less straightforward to interpret. Using more periods and fewer parentheses seems to indicate simpler writing. Others have observed that writing about serious trauma is often better quality than writing about more mundane or lighter-hearted topics, partly due to its less convoluted sentence structures and more straightforward style (Pennebaker, 1997). Perhaps that is some of what experts were decoding in the severe-risk posts: Posts using simpler punctuation may have indicated a more urgent or certain desire to die, and thus were coded as high risk.

4 Conclusion

It is important to remember that the expert coders in Shing et al. (2018) had no more information than we do about these users. We do not know whether the people whose r/SuicideWatch posts comprised this sample have died by suicide since posting, either immediately following an expressed intention to die or later on, related to long-term complications of problems mentioned in their posts. Thus, there are bound to be some false positives in every risk category.

In lieu of additional information, it may be most productive to view these expert ratings as accurate. It could be the case that the main value of tasks like this—where teams aim to find specific linguistic features that correlate with holistic risk annotations—is to find variables that expert clinicians have procedural but not declarative access to in memory or everyday experiences with clients (Schneider et al., 1990). Clinical psychologists often note that they intuit someone's diagnosis or risk at a glance, without being able to easily verbalize what it is about that client that places them in a certain diagnostic category (Hamm, 1988). To the degree that those intuitions are accurate, it would benefit both computational linguists (to bolster the accuracy of predictive models) and clinicians (to improve treatment and diagnosis) if we could determine what behavioral variables are influencing those perceptions—perhaps particularly in the context of noisy, relatively low-fidelity samples of behavior, such as posts in mental heath forums on Reddit.

Figure 1: Word cloud based on posts in r/SuicideWatch, aggregated within user. Words are colored by the rating they most correlate with (a = green, b = yellow, c = orange, d = red), sized by correlation size, and shaded by document frequency (lighter words being used by more users).

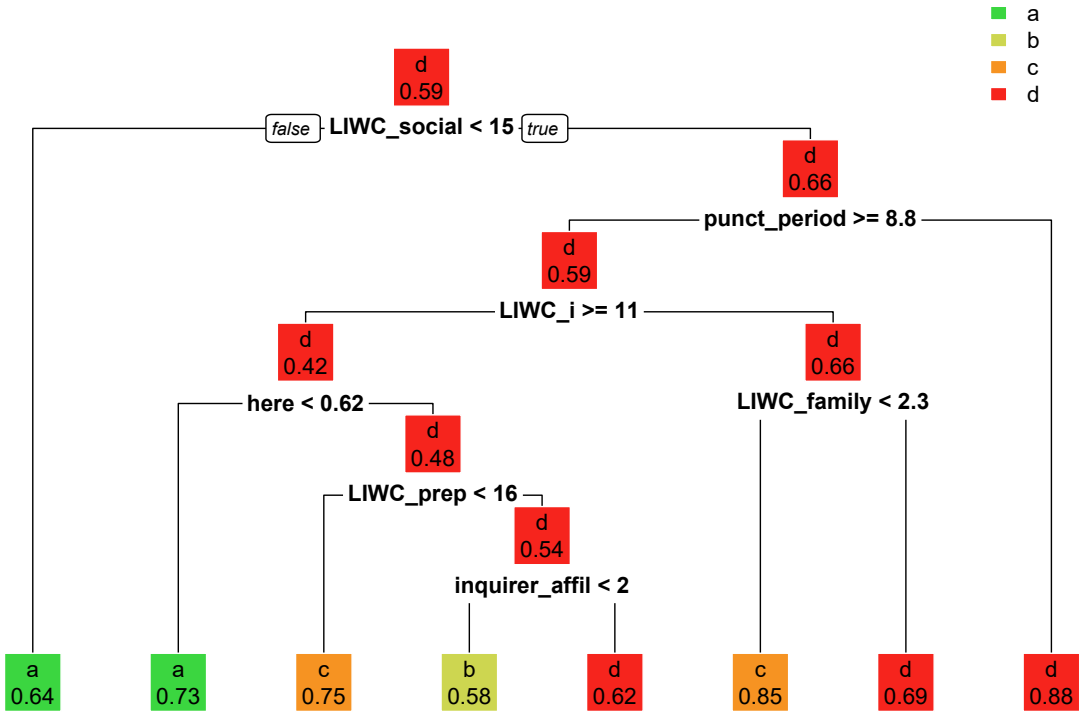

Figure 2: Decision tree fit to the full training set of posts in r/SuicideWatch (task A; in-sample, overall accuracy = 72%, macro F1 = .524). Variables are single words or dictionary categories, and values are percentages of total word count in each post. Each node is colored and labeled by the dominant rating, and displays the probability of that rating in the subset.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Jenna L Baddeley, James W Pennebaker, and Christopher G Beevers. 2012. Everyday social behavior during a major depressive episode. *Social Psychological and Personality Science*, 4(4):445–452.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium: Commonsense Knowledge*, volume FS-10-02 of *AAAI Technical Report*. AAAI.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638. ACM.

Dale Dirkse, Heather D Hadjistavropoulos, Hugo Hesser, and Azy Barak. 2015. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. *Cognitive behaviour therapy*, 44(1):21–32.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):1–1.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

SK Fineberg, J Leavitt, S Deutsch-Link, S Dealy, CD Landry, K Pirruccio, S Shea, S Trent, G Cecchi, and PR Corlett. 2016. Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*, 46(12):2605–2615.

Jeremy Frimer, Jonathan Haidt, Jesse Graham, Morteza Dehgani, and Reihane Boghrati. 2018. *Moral Foundations Dictionary 2.0*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Robert M Hamm. 1988. Clinical intuition and clinical analysis: expertise and the cognitive continuum. *Professional judgment: A reader in clinical decision making*, pages 78–105.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Molly E. Ireland and Micah Iserman. 2018. *LUSI Lab Development Dictionaries*.

Molly E. Ireland and Micah Iserman. 2019. *LUSI Lab Revised Netspeak Dictionary*.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Social Psychological and Personality Science*, 99(3):549–571.

Saif M. Mohammad. 2017. Word affect intensities. *CoRR*, abs/1704.08798.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Jeroen Ooms. 2018. *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 3.0.

James W Pennebaker. 1997. *Opening up: The healing power of expressing emotions*. Guilford Press.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. 2018. Statistical inferences for polarity identification in natural language. *PLOS ONE*, 13(12):1–21.

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

Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Social science computer review*, 32(2):132–144.

Maarten Sap, Greg Park, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, Lyle H Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social

media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wolfgang Schneider, Joachim Körkel, and Franz E. Weinert. 1990. Expert knowledge, general abilities, and text processing. In Wolfgang Schneider and Franz E. Weinert, editors, *Interactions Among Aptitudes, Strategies, and Knowledge in Cognitive Performance*, pages 235–251. Springer New York, New York, NY.

H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, Michal Kosinski, Martin E P Seligman, and Lyle H Ungar. 2016. Predicting individual well-being through the language of social media. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:516—527.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Shannon Wiltsey Stirman and James W Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2018. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*.

Terry Therneau and Beth Atkinson. 2018. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Mark P.J. van der Loo. 2014. The stringdist package for approximate string matching. *The R Journal*, 6:111–122. Version 0.9.4.7.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review*, 117(2):575.

Walter Weintraub. 1989. *Verbal behavior in everyday life*. Springer Publishing Co.

Cynthia M. Whissell. 1989. Chapter 5 - the dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *The Measurement of Emotions*, pages 113 – 131. Academic Press.

Liang Wu, Fred Morstatter, and Huan Liu. 2016. Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *CoRR*, abs/1608.05129.

Wu Youyou, David Stillwell, H Andrew Schwartz, and Michal Kosinski. 2017. Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological science*, 28(3):276–284.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*.