Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health

Danielle Mowery, Albert Park, Mike Conway
Biomedical Informatics
University of Utah
421 Wakara Way, Ste 140
Salt Lake City, Utah, 84108
firstname.lastname@utah.edu

Craig Bryan
Psychology
University of Utah
380 S 1530 E BEH S 502
Salt Lake City, Utah, 84112
firstname.lastname@utah.edu

Abstract

Major depressive disorder, a debilitating and burdensome disease experienced by individuals worldwide, can be defined by several *depressive symptoms* (e.g., *anhedonia* (inability to feel pleasure), *depressed mood*, *difficulty concentrating*, etc.). Individuals often discuss their experiences with depression symptoms on public social media platforms like Twitter, providing a potentially useful data source for monitoring population-level mental health risk factors. In a step towards developing an automated method to estimate the prevalence of symptoms associated with major depressive disorder over time in the United States using Twitter, we developed classifiers for discerning whether a Twitter tweet represents *no evidence of depression* or *evidence of depression*. If there was evidence of depression, we then classified whether the tweet contained a *depressive symptom* and if so, which of three subtypes: *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*. We observed that the most accurate classifiers could predict classes with high-to-moderate F1-score performances for *no evidence of depression* (85), *evidence of depression* (52), and *depressive symptoms* (49). We report moderate F1-scores for depressive symptoms ranging from 75 (*fatigue or loss of energy*) to 43 (*disturbed sleep*) to 35 (*depressed mood*). Our work demonstrates baseline approaches for automatically encoding Twitter data with granular depressive symptoms associated with major depressive disorder.

1 Introduction

Major depressive disorder is one of the most debilitating diseases experienced by individuals worldwide according to the World Health Organization (Mathers and Loncar, 2006; Centers for Disease Control and Prevention, 2012). Major depressive disorder is clinically defined as experiencing one or more of the following symptoms: *fatigue*, *inappropriate guilt*, *difficulty concentrating*, *psychomotor agitation or retardation*, or *weight loss or gain*, as well as continuously experiencing 2 weeks or more of *depressed mood* and *anhedonia* (American Psychiatric Association, 2000; American Psychiatric Association, 2013). For individuals experiencing major depressive disorder, these symptoms often create both personal and interpersonal burdens e.g., reduced productivity at work, hindered interactions with others, and disrupted eating and sleeping behaviors (National Institute of Mental Health, 2016).

1.1 Social Media and Mental Health

In the United States, the traditional means of estimating the prevalence and burden of depression symptoms has involved national face-to-face and telephone interview-based surveys. However, these surveys are both expensive to conduct and typically administered only once per year. Social media platforms like Twitter, in conjunction with natural language processing and machine learning, can be leveraged to support the analysis of very large data sets for population-level mental health research (Conway and O'Connor, 2016). For example, using social media data, researchers have characterized smoking and drinking problems (Tamersoy et al., 2015; Myslín et al., 2013), classified phases of substance addiction (MacLean et al., 2015), predicted the likelihood of recovering from an eating disorder (Chancellor et al., 2016), and identified individuals at risk of committing suicide (De Choudhury et al., 2016).

Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 182–191, Osaka, Japan, December 12 2016.

1.2 Social Media and Depression

For major depressive disorder or depression, researchers have found that individuals discuss their mental health issues on social media (De Choudhury et al., 2014; Park et al., 2013) and that social media data can predict individuals at risk for depression (De Choudhury et al., 2013; Park et al., 2012) as well as specific subtypes e.g., postpartum depression (De Choudhury et al., 2014; De Choudhury et al., 2013). However, the majority of these studies do not explicitly analyze symptoms and risk factors (e.g. *disturbed sleep*, *fatigue or loss of energy*) associated with depression that could be useful in creating population-level mental health monitoring systems.

1.3 Populations and Depression

Depression experiences and risk factors vary widely by population. It has been shown that depression can affect individuals of different ethnicities (Oquendo et al., 2004) and ages (Pratt and Brody, 2008) at different rates. Moreover, depression can initiate at widely different ages (Kessler et al., 2009) and depressive symptoms can vary based on life stage. For example, children may experience depression intermittently or persistently into adulthood demonstrating episodes of *irritability*, *negativity*, and *sulking*; whereas, older adults may experience depression following bereavement or while suffering from a chronic disease, and are less likely to admit sadness, making it hard to diagnose depressive disorder (National Institute of Mental Health, 2015). Although depression affects both genders; women experience a significantly greater percentage of lifetime major depression (11.7%) compared to men (5.6%) (Ford and Erlinger, 2004). When depressed, women tend to experience *depressed mood*, *inappropriate guilt*, and *worthlessness*; in contrast to, men who tend to experience *difficulty sleeping*, *irritability*, *fatigue*, and *anhedonia* (National Institute of Mental Health, 2015). Additionally, some personality traits (e.g., neurotism) are strongly correlated with depressive disorders (Kotov et al., 2010) as well as with subjective well-being (Lucas and Diener, 2009).

1.4 Natural Language Processing and Depression

Despite the progress toward understanding how depression is expressed in social media, relatively little work has been addressed at the detection of specific depressive symptoms and risk factors associated with depression from Twitter data. Exceptions include Cavazos-Rehg et al. (2016) and some of our previous works (Mowery et al., 2016; Mowery et al., 2015). Cavazos-Rehg et al. (2016) applied a qualitative technique to study 2,000 randomly selected tweets containing one or more depression-related keywords (depressed, #depressed, depression, #depression), finding that two-thirds of the tweets described depressive symptoms of *depressed mood or irritable most of the day*, *guilt or worthlessness*, *self harm*, and *contemplating suicide or desires death*. In our previous work, we created a schema based on 9 DSM-5 (American Psychiatric Association, 2013) depressive symptoms and 12 DSM-IV (American Psychiatric Association, 2000) psychosocial stressors and classified the most prevalent symptoms (*depressed mood* and *fatigue or loss of energy*) and stressors (*problems with social environment*) (Mowery et al., 2015). This paper builds upon these works toward encoding Twitter tweets representing depressive symptoms of major depressive disorder by (1) accounting for basic demographic information (i.e., age, and gender) and personality traits (i.e., neuroticism and openness) as features, (2) developing supervised classifiers for automatically classifying not only whether a tweet is depressive-related or not, but classifying it as a depressive symptom of one or more subtypes, and (3) assessing whether machine learning-based classification can detect depression-related symptom and specific symptom subtype-related Twitter tweets more precisely than keywords alone.

2 Methods

Specifically, we conducted a quantitative study to train and test a variety of machine learning classifiers to discern whether or not a tweet contains *no evidence of depression* or *evidence of depression*. If there was evidence of depression, then whether the tweet contained one or more *depressive symptoms* and further classified the symptom subtype of *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*.

2.1 Dataset

We leveraged an existing dataset annotated for depressive stressors and psychosocial stressors that we developed called the Depressive **S**ymptoms and Psychosocial **S**tressors **A**ssociated with **D**epression (SAD) dataset (Mowery et al., 2016) . The SAD dataset was annotated with high reliability (overall pairwise F1-score of >0.76%) by three annotators - two psychology undergraduates and a postdoctoral biomedical informatics researcher. The SAD dataset contains 9,300 tweets queried using a subset of the Linguistic Inquiry Word Count (LIWC) lexicon[1] (Pennebaker et al., 2001). Specifically, the "SAD" category lexicon of LIWC was supplemented with depression-indicative keywords selected by a clinical psychologist (author CB). Each tweet was annotated with one or more classes from a linguistic annotation scheme based on DSM-5 (American Psychiatric Association, 2013) and DSM-IV (American Psychiatric Association, 2000) depression criteria resulting in 9,473 annotations. The full schema includes 9 depressive stressors and 12 psychosocial stressors classes. However, for this study, we focused our attention to the three most prevalent depressive symptoms subtypes: *depressed mood* (n=1,010 tweets, e.g., "Feeling so defeated today"), *disturbed sleep* (n=98 tweets, e.g., "Living a never-ending life of insomnia"), *fatigue or loss of energy* (n= 427 tweets, e.g., "I am so tireeeeeed!!") (see Figure 1). In an attempt to classify whether a tweet represented *no evidence of depression* (n=6,829 tweets) or *evidence of depression* (n=2,644 tweets), specifically, *depressive symptoms* (n=1,656 tweets) and one or more of these three *subtypes*, we encoded the following feature groups described in **Features** below.

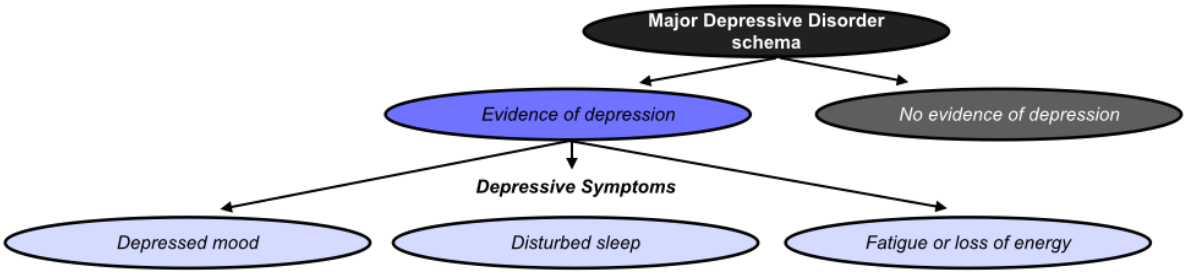

Figure 1: Major depressive disorder schema. Light purple boxes are depressive symptom subtypes. *No evidence of depression* and *evidence of depression* are mutually exclusive classes.

2.2 Features

We included a variety of binary features (present: 1 or absent: 0), including many subsets designed to collapse similar features into a smaller set of semantically similar values to reduce the feature space.

- **N-grams** may provide meaningful, highly predictive terms indicative of a particular symptom (Mowery et al., 2015) e.g., "tired" may indicate *fatigue or loss of energy*. We encoded unigrams (n=16,773 unigrams) using the Twokenizer[2].
- **Syntax** has been shown to be useful for discerning whether a person is depressed or not e.g., usage of first person vs third person pronouns (Coppersmith et al., 2014; Coppersmith et al., 2015). We encoded parts of speech using ARK (Gimpel et al., 2011; Owoputi et al., 2012).
- **Emoticons** can be used to demonstrate positive or negative emotion, which could be an indicator of whether an individual is experiencing a depressive mood. We encoded whether the tweet contained emoticons representing four values: happy, sad, both, or neither.
- **Age/Gender** have been correlated with some depressive symptoms (Pratt and Brody, 2008; Ford and Erlinger, 2004; National Institute of Mental Health, 2016). Because age and gender information is

[1] http://liwc.wpengine.com/
[2] http://www.cs.cmu.edu/~ark/TweetNLP/

not readily available with tweets, we applied age and gender lexicons to predict the age and gender for each tweet (Sap et al., 2014).

- **Sentiment** subjectivity terms (e.g., 5 point-scale from strongly subjective to strongly objective) and polarity terms (e.g., 5 point-scale from strongly positive to strongly negative) may indicate a person's sentiment and its strength toward people, events, and things. We leveraged the Multi-Perspective Question Answering lexicons to encode these subjectivity and polarity scales (Wilson et al., 2005).

- **Personality traits** have been useful predictors of depressive states (Kotov et al., 2010) e.g., depressed individuals exhibit more inward-looking behavior. We encoded personality traits of openness, conscientiousness, extraversion/introversion, agreeableness/antagonism, neuroticism.

- **Linguistic Inquiry Word Counts** terms e.g., words associated with negative emotion including **anxiety** and **anger**, biological state such as **health** and **death**, cognitive mechanisms including **cause** and **tentativeness** have been used to accurately distinguish a depressed from a non-depressed individual (Coppersmith et al., 2014; Coppersmith et al., 2015). Preotiuc-Pietro et al. (Preotiuc-Pietro et al., 2015) also observed terms associated with **illness management** (e.g., "meds", "pills", "therapy") associated with depressed individuals. We encoded each tweet with terms indicative with several linguistic topics including: *syntactic terms*: **function**, **personal pronoun**, **I**, **we**, **she/he**, **they**, **I pronouns**, **articles**, **verbs**, **auxillary verb**, **past**, **present**, **future**, **adverbs**, **prepositions**, **conjugates**; *qualifier terms*: **negation**, **quantifiers**, **numbers**; *semantic terms*: **swearing**, **social**, **family**, **friends**, **humans**, *emotion terms*: **affect**, **positive emotion**, **negative emotion**, **anxiety**, **anger**, **sadness**; *mental postulation terms*: **cognitive mechanism**, **insight**, **cause**, **discrepancy**, **tentativeness**, **assent**, **filler**, **certainty**, **inhibitory**, **inclusive**, **exclusive**, **perception**, **hearing**, **seeing**; *health-related terms:* **biology**, **body**, **health**, **sexual**, **ingest**, **non-FLU**; *temporal/spatial terms:* **relative**, **motion**, **space**, **time**; *life terms:* **work**, **achievement**, **leisure**, **home**, **money**, **religion**, **death**.

Age/gender and **personality traits** lexicons can be found at the World Well-Being Project website[3]. **Sentiment** lexicons can be found at the Multi-Perspective Question Answering Subjectivity website[4].

2.3 Classifiers

We trained and tested supervised machine learning classifiers for predicting depression-related classes: 1) whether a tweet represents *no evidence of depression* or *evidence of depression* and 2) if the tweet is depression-related, whether it is classed as a *depressive symptom* and specifically by subtypes of *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*. We trained each classifier using scikit learn[5] with 5-fold cross validation using all features (described in **Experiments** below) and then reported performances using average recall and average precision (all classifiers) as well as average F1-scores (most accurate classifiers only) for each class level. We assessed six supervised machine learners – decision tree, random forest, logistic regression, support vector machine, linear perceptron, and naïve Bayes.

- **Decision Tree** learns a prediction model by determining a sequence of the most informative features that maximize the split distinguishing one output class label from another by leveraging recursive partitioning and measuring the information gained for each split using entropy. We chose decision trees because of their simple representation of tree structures for interpretation. We tested models produced with both depth restriction of 5 and no depth restriction by applying an optimised version of the CART algorithm.

[3] http://wwbp.org/lexica.html
[4] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.
[5] http://scikit-learn.org/

- **Random Forests** learn many decision trees during its training and classifying a predicted class label based on the mode of the classes or the mean of the prediction of the aggregate individual trees; thus, reducing the likelihood of overfitting by a single decision tree model. Similar to the decision trees experiment, we also tested models produced with both depth restriction of 5 and no depth restriction.

- **Logistic Regression** learns a logit regression model in which the dependent variable is the class label. Logistic regression models that leverage regularization avoid over-fitting particularly when the dataset contains only a few number of training examples for a class label, many irrelevant features for classification, and a large number of parameters that must be learned. We tested models with both L_1 and L_2 regularization.

- **Support Vector Machine** learns a model that linearly separates two classes in a high dimensional space. We chose to train classifiers using support vector machines because of their ability to tolerate a large number of features while maintaining high performance, to minimize the likelihood of overfitting by using support vectors for classification, and to withstand sparse data vectors that could be produced by encoding a high number of features. We trained the model using a linear kernel.

- **Linear Perceptron** learns a prediction model based on a linear predictor function leveraging a set of weights from a feature vector. We chose linear perceptron because of their efficiency and ability to be easily trained with large datasets.

- **Naïve Bayes** learns a prediction model that leverages posterior probabilities of each class and conditional probabilities of the class for each individual feature. We chose naïve Bayes because a naive assumption of independence between features can prove effective for many similar text classification problems.

2.4 Experiments

We performed the following two experiments leveraging the aforementioned features and classifiers.

2.4.1 Most Accurate Classifiers

For predicting each class label, we leveraged all features sets to train and test each classifier, then compared the output of each classifier against the manual reference standard. We report the best performing classifier for each label according to average F1-score and average precision.

2.4.2 Most Precise Classifiers

Searching for relevant data from the Twitter API[6] requires the identification of keywords appropriate for the task at hand. In the case of identifying depression-related tweets, the limitation of a purely keyword-based (e.g., "depression") approach are obvious (e.g.,"Brexit may cause worldwide economic depression!"). A key aim of our work is understanding the extent to which machine learning methods improve precision compared to keyword-based methods alone. Therefore, we aimed to determine how much more precise the outputs of machine learning classifiers could be compared to a simple keyword query. Specifically, we aimed to determine whether the LIWC keywords used to query the Twitter tweets (Table 1) provide greater precision than the most precise machine learning algorithm for discerning whether a tweet contained an expression of *depressive symptoms* and, if so, by subtypes of *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*.

3 Results

We assessed the performance of six supervised machine learners – decision tree, random forest, logistic regression, support vector machine, linear perceptron, and naïve Bayes – and a variety of features for classifying whether or not a tweet contains *no evidence of depression* or *evidence of depression*. If there was evidence of depression, then we determined whether the tweet contained one or more *depressive*

[6] https://dev.twitter.com/overview/documentation

Depression Categories	Linguistic Inquiry Word Count keywords
Depressive symptoms	**all keywords for subtypes below*
Depressed mood	abandon*, ache*, aching, agoni*,
	alone, broke*, cried, cries, crushed, cry, damag*, defeat*, depress*, depriv*, despair*,
	devastat*, disadvantage*, disappoint*, discourag*, dishearten*, disillusion*, dissatisf*,
	doom*, dull*, empt*, gloom*, grave*, grief, griev*, grim*,fail*, flunk*,
	heartbr*, helpless*, homesick*, hopeless*, hurt*, inadequa*, inferior*, isolat*, lame*, lone*,
	longing*, lose, loser*, loses, losing, loss*, lost, melanchol*, miser*,mourn*, neglect*,
	overwhelm*, pain, pathetic*, pessimis*, piti*, pity* , regret*, reject*, remorse*, resign*, ruin*,
	sad, sobbed, sobbing, sobs, solemn*, sorrow*, suffer*, tears*, traged*, tragic* , unhapp*,
	unimportant, unsuccessful*, useless*, weep*, wept, whine*, whining, woe*, worthless*, yearn*
Disturbed sleep	insomnia
Fatigue or loss of energy	fatigu*, tired

Table 1: Linguistic Inquiry Word Count keywords used for query by depression-related tweets from Twitter API (Mowery et al., 2015).

symptoms and classified the tweet by subtype as *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*.

3.1 Most Accurate Classifiers

Overall, we observed that support vector machines were able to produce the highest F1-scores for most (4/6) of the classifications (Figure 2). In terms of the binary classification, a tweet could be classified into the majority class of *no evidence of depression* (logistic regression$_{L_1 \text{ regularization}}$) with an F1-score of 85 and into the minority class of *evidence of depression* (support vector machine) with an F1-score of 52. For tweets representing evidence of depression, *depressive symptoms* could be predicted with an F1-score of 49 (support vector machine). F1-scores for depressive symptoms ranged from 35 (*depressed mood*: support vector machine) to 43 (*disturbed sleep*: support vector machine) to 75 (*fatigue or loss of energy*: decision tree$_{\text{restriction depth of 5}}$).

For most classes, the performance differences for the most accurate classifier in terms of precision and recall scores were most often not more than 5 points from each other. A notable exception with higher recall (82) than precision (70) was *fatigue or loss of energy*. In contrast, *disturbed sleep* demonstrated higher precision (58) over recall (36).

3.2 Most Precise Classifiers

In Figure 3, half of the classes were precisely classified using decision trees with a depth restriction of 5. Compared to the most precise classifier for each class, LIWC keyword terms produced lower precision for the class of *depressive symptoms* (-49 points), *depressed mood* (-34 points), and *fatigue or loss of energy* (-28 points). We only observed higher precision leveraging the original LIWC keywords compared to the machine learning classifier for *disturbed sleep* (+11 points).

4 Discussion

In this study, we evaluated several supervised classifiers for accurately classifying whether a tweet expressed *evidence of depression* or not, *depressive symptoms* and their subtypes. Furthermore, we assessed whether rich features i.e., demographic and personality features, with machine learning approaches improved upon precision of simple keywords for precisely detecting *depressive symptoms* and subtypes of *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy* from Twitter tweets.

4.1 Most Accurate Classifiers

Overall, we observed that support vector machines were able to produce the highest F1-scores for most of the classifications (Figure 2). We hypothesize that the support vector machine produced superior results due to its ability to tolerate a large number of features while maintaining high performance and to withstand sparse data vectors produced by encoding the large number of features. In terms of the binary classification, we could discern a tweet containing *evidence of depression* with moderate performance

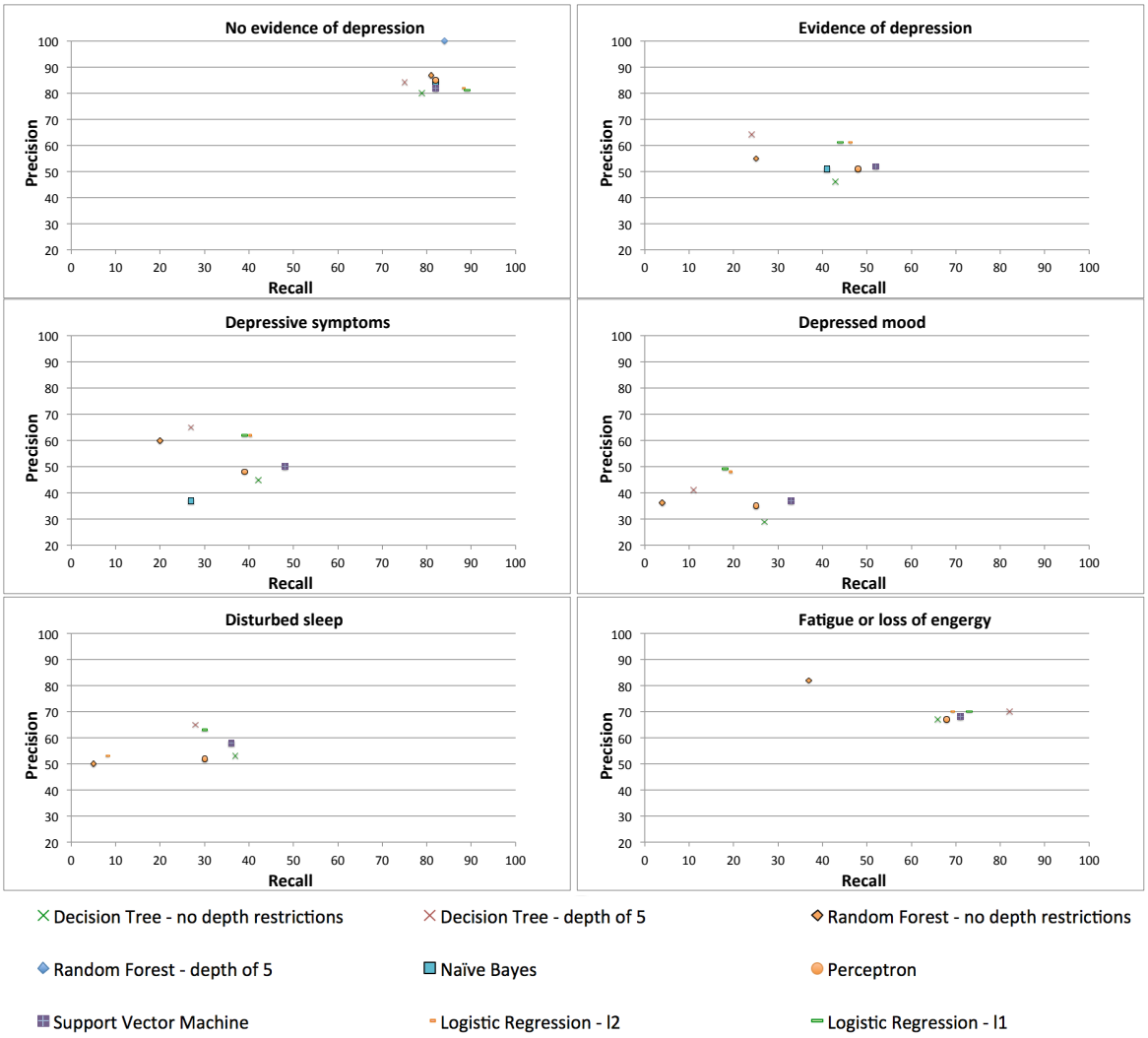

Figure 2: Classifier performances for each class. Reported recall and precision values are averages over 5 fold cross-validation. Only classifiers with precision values greater than 20 are shown. $_a$ = no depth restriction. $_b$ = restriction depth of 5. $_c$ = L_1 regularization.

(F1-score: 52) and even precision and recall suggesting that a machine learning approach will identify a little over half of the depression-related tweets with a similar portion of which are a true signal of *evidence of depression*. We observed similar results with identifying *depressive symptoms*. In terms of particular subtypes, *fatigue or loss of energy* could be most reliably classified – we suspect this is due to the high, unambiguous usage of the words like "tired" and "fatigue" and other features e.g., SAD emoticon :(. In practical use of these classifiers, we would expect lower recall, but more precise classification which is important for reducing the likelihood of producing inflated prevalence estimates of depression risk factors at a population level.

4.2 Most Precise Classifiers

Furthermore, in Figure 3, we observed that a range of learning classifiers are needed to most precisely classify depressive symptoms and subtypes. Decision trees (*depressed mood* and *depressive symptoms*) and random forests (*fatigue or loss of energy*) produced substantially higher precision than the set of LIWC query keywords. The only exception was observed for *disturbed sleep* which might be explained by again the low ambiguity of "insomnia". This finding suggests that for some symptoms machine learning algorithms can reduce the likelihood of sampling noisy tweets that do not indicate one or more depressive symptoms. A practical implication of this finding could be developing a highly sensitive

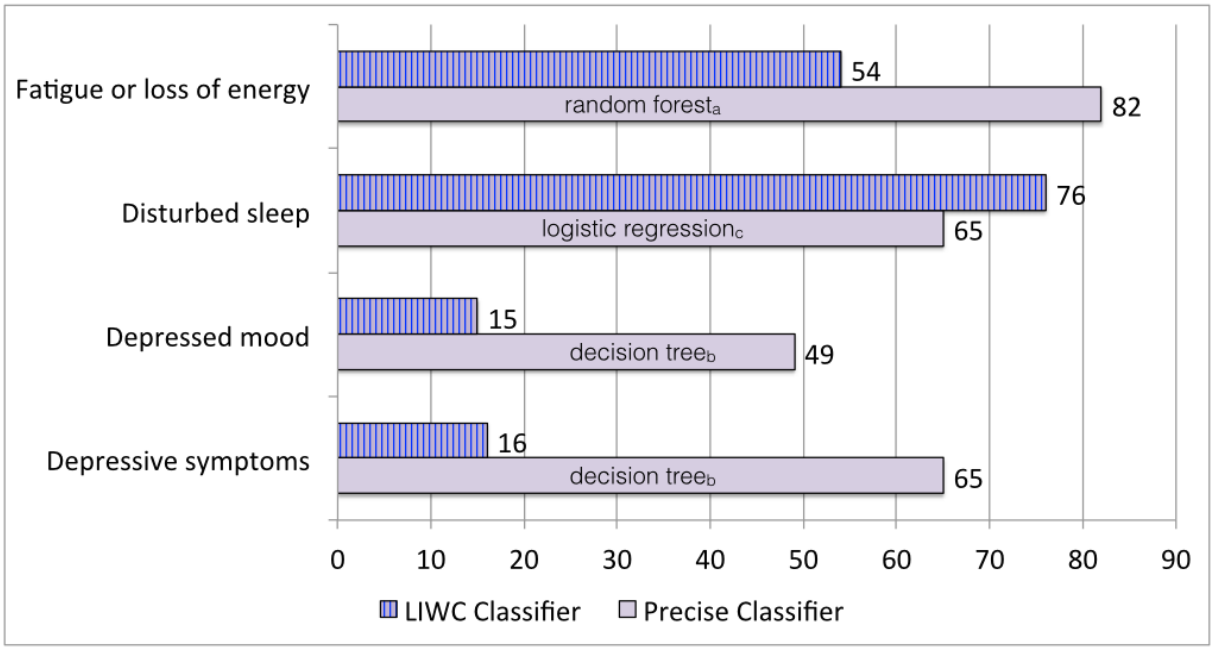

Figure 3: Performance of best average precision classifier for *depressive symptoms* and for each *subtype*. $_a$ = no depth restriction, $_b$ = restriction depth of 5. $_c$ = L_1 regularization.

lexicon for querying the Twitter API for depressive-related tweets, then applying highly precise filtering to identify tweets more likely to contain depressive symptoms and particular subtypes.

4.3 Comparison to Related Work

In comparison to our previous work (Mowery et al., 2015), we observed a very similar classification trend of high performance for *no evidence of depression* and *fatigue or loss of energy* as well as moderate performance for *depressed mood*. When comparing particular classifier performances between studies, most classifiers performed with equal or slightly lower recall and precision suggesting the addition of demographics and personality features may not greatly improve performance compared to simple unigrams for this dataset on an atomic tweet-level (in contrast to a user-level with many tweets over time (Sap et al., 2014)). These consistent findings suggest we can reach the state-of-the-art performance for detecting these subtypes with perhaps a rather simple unigram model. However, in future work, we will experiment with larger n-grams, network-based features, and feature selection approaches to develop more precise classifiers for these subtypes and other depressive symptom subtypes not addressed in this study e.g., *anhedonia*, *inappropriate guilt*, *worthlessness*, and *irritability*, etc. We will also conduct a feature ablation study to better understand the contribution of features with respect to classifier performance.

5 Conclusion

In conclusion, we developed classifiers for discerning whether a tweet contained *evidence of depression* and if so, we encoded whether it was a *depressive symptom*, in addition to encoding the subtypes *depressed mood*, *disturbed sleep*, or *fatigue or loss of energy*. We showed that in most cases the use of machine learning classifiers improve precision in identifying *depression symptom* and subtype-related tweets compared to the use of keywords alone.

6 Acknowledgements

This work was funded by grants from the National Library of Medicine of the National Institutes of Health (K99LM011393/R00LM011393), and was granted a review exemption by the University of Utah Institutional Review Board (IRB 00076188). To protect tweeter anonymity, we have not reproduced tweets verbatim. Example tweets shown were generated by the researchers as exemplars only.

References

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR)*. Author, Washington, DC.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. American Psychiatric Association, Washington, DC.

Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosa, Meghana Bharadwaj, and Laura J. Bierut. 2016. A Content Analysis of Depression-related Tweets. *Computers in Human Behavior.*, 54:351–357.

Centers for Disease Control and Prevention. 2012. Behavioral Risk Factor Surveillance System Survey Data. http://www.cdc.gov/brfss/.

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery Amid Pro-Anorexia. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 2111–2123, New York, New York, USA. ACM Press.

Mike Conway and Daniel O'Connor. 2016. Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications. *Current Opinion in Psychology*, 9:77–82.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June 27th 2014. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, CO, USA, June 5th 2015.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, pages 3267–3276, Paris, France, April 27 - May 2, 2013. ACM Press.

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 - CSCW '14*, pages 626–638, New York, New York, USA. ACM Press.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110, San Jose, CA, USA. ACM Press.

Daniel E. Ford and Thomas P. Erlinger. 2004. Depression and C-Reactive Protein in US Adults Data From the Third National Health and Nutrition Examination Survey. *Archives of Internal Medicine*, 164(9):1010–1014.

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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ronald C Kessler, Sergio Aguilar-gaxiola, Jordi Alonso, Somnath Chatterji, Sing Lee, Johan Ormel, T Bedirhan Üstün, and Philip S Wang. 2009. The Global Burden of Mental Disorders: An Update from the WHO World Mental Health (WMH) Surveys. *Epidemiologia e Psichiatria Sociale*, 18(01):23–33.

Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. 2010. Linking "Big" Personality Traits to Anxiety, Depressive, and Substance Use Disorders: A Meta-Analysis. *Psychological Bulletin*, 135(5):768–821.

Richard Lucas and Ed Diener. 2009. Personality and Subjective Well-Being. In *The Science of Well-Being*, volume 37, pages 75–102. Springer Netherlands.

Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1511–1526, New York, NY, USA. ACM.

Colin D. Mathers and Dejan Loncar. 2006. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Med*, 3(11):e442.

Danielle L. Mowery, Craig Bryan, and Mike Conway. 2015. Toward Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In *Proceeding of 2nd Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality*, pages 89–98. Association for Computational Linguistics.

Danielle L. Mowery, Hilary A. Smith, Tyler Cheney, Craig Bryan, and Mike Conway. 2016. Identifying Depression-related Tweets from Twitter Public Health Monitoring. In *Online Journal of Public Health Informatics*, volume 8, page e144.

Mark Myslín, Shu-Hong Zhu, Wendy W Chapman, and Mike Conway. 2013. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research*, 15(8):e174.

National Institute of Mental Health. 2015. Depression (NIH Publication No. 15-3561). http://www.nimh.nih.gov/health/publications/depression-what-you-need-to-know-12-2015/depression-what-you-need-to-know-pdf_151827.pdf.

National Institute of Mental Health. 2016. National Institute of Mental Health. Mental Health Information: Depression. https://www.nimh.nih.gov/health/topics/depression/index.shtml.

Maria A. Oquendo, Dana Lizardi, Steven Greenwald, Myrna M. Weissman, and J. John Mann. 2004. Rates of Lifetime Suicide Attempt and Rates of Lifetime Major Depression in Different Ethnic Groups in the United States. *Acta Psychiatrica Scandinavica*, 110(6):446–451.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical report, Carnegie Mellon University.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive Moods of Users Portrayed in Twitter. In *ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, pages 1–8.

Minsu Park, David W. Mcdonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 476–485.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count [computer software]*. Mahwah, NJ: Erlbaum Publishers.

Laura A. Pratt and Debra J. Brody. 2008. Depression in the United States Household Population, 2005-2006. Technical report, US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. https://www.cdc.gov/nchs/products/databriefs/db07.htm.

Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Ungar Lyle. 2015. The Role of Personality, Age, and Gender in Tweeting about Mental Illness. In *e 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,*, pages 21–23. Association for Computational Linguistics.

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)*, pages 1146–1151, Dona, Qatar, October 25-29th 2014.

Acar Tamersoy, Munmun De Choudhury, and Duen Horng Chau. 2015. Characterizing Smoking and Drinking Abstinence from Social Media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 139–148, Guzelyurt, TRNC, Cyprus, September 1-4th 2015. ACM.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.