# Are You Really Okay? A Transfer Learning-based Approach for Identification of Underlying Mental Illnesses

**Ankit Aich** and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{aaich2, parde}@uic.edu

## Abstract

Evidence has demonstrated the presence of similarities in language use across people with various mental health conditions. In this work we investigate these relationships both as described in literature and as a data analysis problem. We also introduce a novel transfer learning based approach that learns from linguistic feature spaces of previous conditions and predicts unknown ones. Our model achieves strong performance, with $F_1$ scores of 0.75, 0.80, and 0.76 at detecting depression, stress, and suicidal ideation in a first-of-its-kind transfer task and offering promising evidence that language models can harness learned patterns from known mental health conditions to aid in their prediction of others that may lie latent.

## 1 Introduction

Mental health conditions are a pervasive but historically often overlooked societal and individual concern (Bertolote, 2008). In recent decades their study has gained increasing priority, and within the past decade this study has extended to techniques for automated analysis and detection of mental health conditions, including through patterns detected in written and spoken language (Resnik et al., 2014). Most work on automated assessment of mental health seeks to identify and possibly alleviate specific mental health conditions. Researchers have focused on a myriad of target illnesses and diagnoses such as depression (Schwartz et al., 2014a), schizophrenia (Gutiérrez et al., 2017), or even suicideal ideation[1] (Homan et al., 2014). However, to date they have not yet examined the overlap or interplay between these target illnesses. This overlap may present a valuable source of information, particularly in the resource-poor settings common in mental health and healthcare applications more generally.

In this paper we ask three important research questions centered on the interplay between the linguistic footprints of known and latent mental health conditions (MHCs),[2] and present answers to them with evidence.

- **RQ1:** *How do features relate across multiple MHCs?*
- **RQ2:** *Can we represent different MHCs under the same feature spaces and find relations?*
- **RQ3:** *Can we identify underlying MHCs using the language of known ones?*

Our first question relates to the linguistic markers of MHCs. We comprehensively examine existing psycholinguistic and mental health research to search for common underlying threads (§3). To answer our second question, we investigate the relation between the identified features using well defined and trusted NLP baselines (§4). Finally, we answer our last question by experimentally determining the success with which we can use similar and dissimilar linguistic feature spaces to predict the presence of latent MHCs (§5). To do so, we leverage transfer learning to achieve a strong benchmark accuracy of 85%.

## 2 Background

According to the National Institute of Mental Health, 43.6 million adults (nearly 18.1% of the

[1] Presence of *Suicidal Ideation* (SI) is not an illness, but a diagnosis which encompasses thoughts ranging from contemplation to preoccupations with death via suicide (Harmer et al., 2022).

[2] We define MHCs as any condition ranging along the spectrum from issues causing mental health concerns such as stress, to actual defined illnesses such as depression, or diagnoses such as SI.

*Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 89 - 104
July 15, 2022 ©2022 Association for Computational Linguistics

U.S. population) experience mental health conditions in a given year.[3] Oftentimes, symptoms may be recognizable when interacting with close relations (Insel, 2008) or even on social media (Berry et al., 2017). Berry et al. (2017) investigate the popularity of social media as an outlet for mental health discussion at length, finding reasons including anonymity, sense of empowerment, sense of community, and perceptions of the internet as a safe space. A growing number of approaches have sought to leverage social media data to aid in the automated identification of specific MHCs, with work to date including automated detection of depression (Yasaswini et al., 2021; Schwartz et al., 2014a; Tasnim and Stroulia, 2019; Rosenquist et al., 2010), post-traumatic stress disorder (Li et al., 2010), anxiety (Shen and Rudzicz, 2017), and stress (Naik et al., 2018). However, these approaches have lagged behind the state of the art in more fundamental NLP tasks. In particular, work harnessing high-powered transfer learning models has remained either scarce or singularly focused on one illness (Pegah et al., 2019; Howard et al., 2019).

We aim to fill this translational gap by synthesizing fundamental progress with the applied problem of detecting the presence of underlying MHCs. We follow Blodgett et al. (2020)'s lead and model our approach not only on existing NLP models, but on findings from pyscholinguistic and other domain-specific literature as well, including those correlating retention (Shen et al., 2009), cognitive attention and complexity (Vuilleumier, 2006; Tausczik and Pennebaker, 2010), reasoning (Jung et al., 2014), and problem-solving skills (Isen et al., 1987) with specific mental health conditions. Little has been done towards this problem with RQs of multi-task learning from social media being very recent (Benton et al., 2017b). This work, to the best of our knowledge, is the first of its kind to study correlation among diseases in both theory and practice. We examine prior literature to identify correlating themes across illnesses, analyze language data from individuals with different mental health conditions to find practical correlations and trends, and present transfer learning-based classification models to identify undiagnosed illnesses given known features grounded in mental health and psycholinguistic theory and NLP practice.

## 3 Feature Correlation Across Varying Mental Health Conditions

A natural question that arises when using social media data is its reliability as an information source. Social media is increasingly seen as a popular choice and acceptable platform for healthcare information exchange (Gkotsis et al., 2016a), and its use has been investigated in numerous predictive healthcare tasks. Classical models (e.g., support vector machines) trained on simple text-based features have reliably predicted mental health emergencies (Franco-Penya and Mamani Sanchez, 2016). Audio features have also been found to be excellent markers of mood or other prosodic signals, including for automated detection of depression (Lamers et al., 2014). Language models have demonstrated an ability to learn powerful, quantifiable signals from tweets to predict users' mental states (Coppersmith et al., 2014), and more clinically advanced mental health conditions such as psychosis have also been detected using short appraisals of social media posts (Birnbaum et al., 2017). Predicting depression on social media is a long standing research track (De Choudhury et al., 2021), and social media has also shown that signals to identify suicidal ideation can be traced with high efficacy (Choudhury et al., 2016). Platforms like Reddit[4] can be instrumental in terms of support, resources, and self-disclosure about mental health (Choudhury and De, 2014; Valizadeh et al., 2021).

One of the first traceable thematic identifications of correlated, quantifiable information regarding mental state and wellbeing was by Fleming et al. (1992), suggesting that a lack of social support combined with social isolation was present in patients showing signs of depression or post-partum depression. The same work also identified effects of psychological stress on attitude, emotion, and behavior. The relationship between social isolation, loneliness, and clinical depression was later also validated by MNSc et al. (1996), and the relationship between latent stress and surface depression has since persisted as a recurring theme across mental health literature (Scott et al., 2000).

Homan et al. (2014) found that high levels of stress or distress are related to higher levels of suicidal ideation. Schwartz et al. (2014b) also pointed to trepidation, frustration, annoyance, helplessness, and again stress as major themes correlating with expression of mental illness. Depression and stress

[3] www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-us-adults.shtml

[4] reddit.com

co-exist in latent forms for other surface illnesses such as schizophrenia as well, as demonstrated by Mitchell et al. (2015) who extracted LIWC (Tausczik and Pennebaker, 2010) features from social media data to detect advanced psychosis and schizophrenia in social media.

Perhaps one of the most interesting finds in translational mental health research is the direct relationship between depression, suicidal ideation, and stress (Preoţiuc-Pietro et al., 2015). Preoţiuc-Pietro et al. (2015) provide evidence that depressive language correlates with sustained periods of low sentiment and has similar topical themes to language produced by suicidal or dysphoric individuals.

Although NLP researchers have experimented with a wide range of linguistic features for mental health assessment and analysis, several have emerged as being particularly discriminating. Metadata such as hashtags or the name of a forum (Mills, 2017) can be powerful features to detect mental health conditions such as suicidal ideation (Gkotsis et al., 2016b). Specific words or hashtags can be used to identify personality profiles, as well as stigma or awareness of mental health conditions on social media (Hwang and Hollingshead, 2016). Degrading or negative n-grams (e.g., *crazy*, *mad*, or *nuts*) can distinguish personality types and mental health outlook (Hwang and Hollingshead, 2016), and part-of-speech (POS) tags can also be informative in social media data (Gkotsis et al., 2016b). Tausczik and Pennebaker (2010) characterize speech at a granular level with social and personal profiles, and present LIWC, a powerful tool to extract such features (Malmasi et al., 2016). N-grams have been powerful markers of depression or PTSD (Pedersen, 2015), and can be valuable tools for feature discovery (Tanana et al., 2016). Lexicon-based features, word embedding features, or annotated posts from social media are also informative (Shickel et al., 2016). Across this systematic review of mental health within NLP literature, the following key relations become evident:

- Stressful and emotional events affect measured cognitive complexity (Shen et al., 2009; Vuilleumier, 2006; Isen et al., 1987).
- Depression, stress, and suicide are related with often overlapping diagnoses, and have intersecting themes of general negativity and hopelessness (Fleming et al., 1992; MNSc et al., 1996; Scott et al., 2000; Schwartz et al., 2014b; Homan et al., 2014).
- N-grams, lexicon-based features, word embeddings, and POS features are powerful tools for social media analysis of mental health problems (Gkotsis et al., 2016b; Hwang and Hollingshead, 2016; Pedersen, 2015; Malmasi et al., 2016).

We experiment further with these features in the following subsections.

## 4 Feature Relationships in Mental Health Data

### 4.1 Data Sourcing and Ethical Guidelines

To fully understand the relationships among linguistic features in mental health contexts we explore datasets associated with three different MHCs. Gaining access to datasets in this area proved challenging, as also discussed by Harrigian et al. (2021), for numerous reasons including IRB restrictions, personal reluctance, or unresponsiveness to data access requests. We ultimately acquired datasets pertaining to *suicide* (Shing et al., 2018; Zirikly et al., 2019), *stress* (Turcan and McKeown, 2019), and *depression* (Losada and Crestani, 2016; Parapar et al., 2021).

In conducting our exploration, we followed the ethical and privacy guidelines defined by Benton et al. (2017a). No identifiable information is collected, and all data is stored on secured servers and obtained via written agreements from the creators. The institutional review board (IRB) at our institution declared our experiments on these datasets as exempt from further review.

### 4.2 Data Description

We studied and analyzed each dataset. All datasets were created with a mixed and randomized population of social media users; thus, the selection of participants was not constrained by gender, background, or other factors. Our *suicide* dataset is sourced from Reddit (Shing et al., 2018; Zirikly et al., 2019) and contains posts and labels for users diagnosed as having suicidal ideation or matched controls. Our *stress* dataset is the Dreaddit dataset published by Turcan and McKeown (2019). It is a publicly available dataset with binary labels indicating the presence of stress[5] (*stressed* and *not*

[5] The authors also ask annotators to indicate instances for which the label is unclear; instances for which this is the majority label are later dropped.

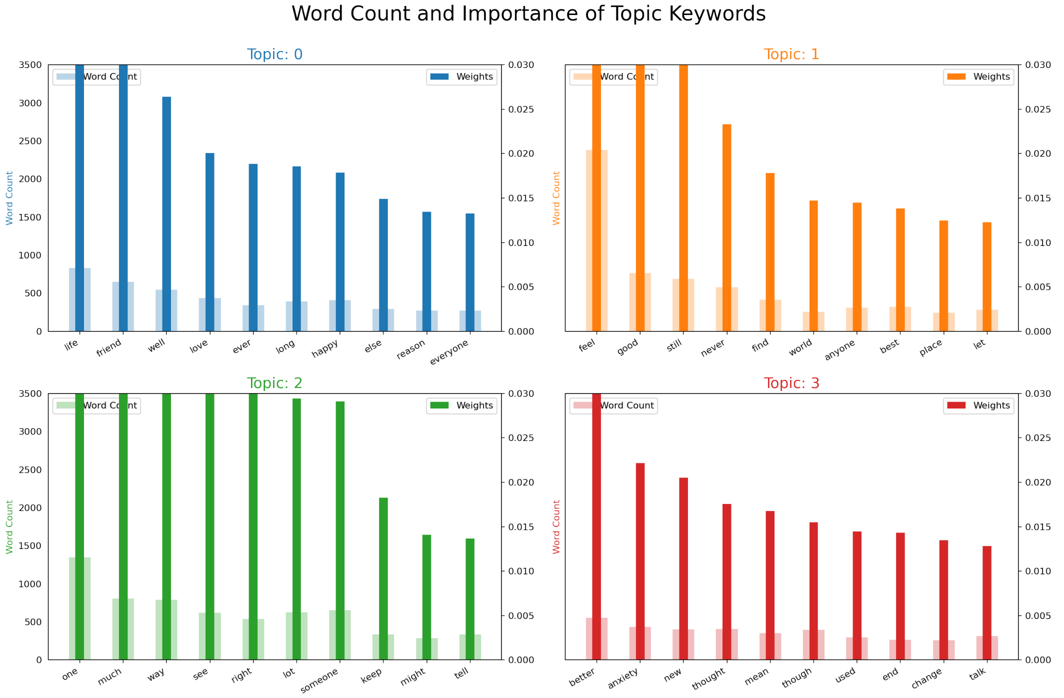

Figure 1: Top four topics identified when applying LDA to *depression*.

*stressed*) in individuals posting on Reddit. Our *depression* dataset is sourced from Twitter[6] and has binary labels (*depression* and *no depression/control*) and raw and pre-processed tweet text (Losada and Crestani, 2016; Parapar et al., 2021).

*Depression* contains 531,453 posts from 892 users, and *stress* contains 187,444 posts. Our *suicide* dataset samples 1097 users at random from a pool of 11,129 initial users, and picks 934 from among those to create a four-class dataset with risk assessment classes: *None*, *Low*, *Moderate*, and *Severe*. We aggregate these into binary labels of 0 (*None*, *Low*) and 1 (*Moderate*, *Severe*).

As per our agreements with the creators of these datasets we are unable to share data directly, but we provide a table in the Appendix to summarize dataset statistics. We encourage researchers to examine the data and related private datasets, and thank the respective authors as well as Harrigian et al. (2021) for creating a curated repository of mental health data and pointers facilitating data discovery.

### 4.3 Data Analysis

As noted in §3, trauma, stress, depression, and mental illness measurably impact reasoning, problem solving, and overall cognitive complexity. Tausczik and Pennebaker (2010) map these effects to psycholinguistic features including sentence complexity, words per sentence, and average word length on a scale of 0-100, where scores less than 50 denote lower cognitive reasoning and analysis.

We perform Latent Dirichlet Allocation (LDA) across the *depression* and *stress* data to identify topical themes. We present a graph in Figure 1 showing the four top themes identified for *depression*, visualized with frequency counts for the thematic terms (a similar graph for *stress* is provided in Figure 3 in the Appendix). To determine thematic titles, we apply Ryan and Bernard (2003)'s Keywords In Context (KWIC) approach, qualitatively examining context and finding the words that adhere to it. We detail our outcomes in Tables 1 and 2, considering the top words identified per theme using LDA and subsequently using KWIC to assign theme names. We find that social support, connections, and familial stress are common topical themes across both illnesses, validating our findings in §3 that similarities in language exist among people suffering from different MHCs. This manifests in our n-gram analyses as well (e.g., with terms such as *feel*, *don't know*, and *life*), further highlighting the intersection of themes across different MHCs.

We further assess the cognitive complexities of a random sample of 380 individuals from *depression* and *suicide*, measured as the average of (a)

[6] www.twitter.com

| **Identified Theme** | **Keywords in Context** |
|---|---|
| Social Support | Life, Friend, Love, Happy, Everyone, Reason |
| Feelings & Connections | Feel, Good, Anyone, Never, Find |
| Action Taken | One, Someone, Might, Tell |
| Therapeutic | Anxiety, Mean, End, Talk, Better |

Table 1: Identified themes applying KWIC to LDA topics for *depression*.

| **Identified Theme** | **Keywords in Context** |
|---|---|
| Failed Connections | Relationship, Didn't, Work, Someone, Need |
| Social and Familial Stress | Doesn't, Feel, Right, Dad, Girl, Kid |
| Pessimism | Don't, Can't, Family, Know, Good |
| Chronic Stress | Year, Still, Issue, Hard, Without |

Table 2: Identified themes applying KWIC to LDA topics for *stress*.

the ANALYTIC feature extracted by LIWC and (b) the average number of short (length $\leq 6$) words per sentence, mapped to a 0-100 scale. We plot the cognitive complexity scores (Y axis) in for each individual in the sample (X axis bars), and observe a slightly lower cognitive complexity for individuals in *suicide* (see Figures 4 and 5 in the Appendix). This is in line with our first finding in §3, and the complementary knowledge that suicidal ideation is often a more extreme expression of depression (Brådvik, 2018).

Finally, to examine the role of sentiment, negativity, and hopelessness (our second finding in §3), we also quantitatively analyze the most frequent trigrams associated with *depression*, *suicide*, and *stress* (see Figures 6, 7, and 8 in the Appendix). We similarly analyze bigrams and unigrams (see Figures 9 and 14 in the Appendix). We find that the top n-grams for all three illnesses are evocative of emotion, confirming substantial overlap across illnesses. N-grams associated with *depression* place additional emphasis on memories (e.g., "campsite tent fire") and specific mental health diagnoses (e.g., "major depressive disorder"), whereas n-grams associated with *suicide* place greater emphasis on confusion (e.g., "basically i'm wondering") and helplessness (e.g., "someone please help"). N-grams associated with *stress* echo many of these themes, with an additional emphasis on uncertainty (e.g., "don't really know").

## 5 Classification and Transfer Learning

### 5.1 Task Outline

We model the primary task as a binary classification problem to predict labels at the user level as 1 (*Diagnosed*) or 0 (*Undiagnosed*) for a mental illness or disease $D$. This can be formulated as:

$$Y_d = M(D)$$

where $Y$ is the label of a classification model $M$ on a domain $D$. This domain, $D$, can be defined as:

$$D = \{X, \mathrm{P}(X)\} \tag{1}$$

where $X$ is the feature space and $\mathrm{P}(X)$ is the marginal probability distribution for:

$$X = \{x_1, x_2, ..., x_n\}$$

For our MHC domain, we can define a task, $T$, as follows:

$$T = \{Y, f(\cdot)\} \tag{2}$$

Here, $Y$ is the label space. This is obtained from a classification function $f(\cdot)$, which learns from our data having features $X$ and labels $Y$ as follows:

$$\{(x_i, y_i) | i \in \{1, 2, ..., n\}, x_i \in X, y_i \in Y\} \tag{3}$$

In Equation 3, each data point in the task is represented by the subscript $i$, where $(x_i, y_i)$ corresponds to the feature vector and label for point $i$ in a dataset of length $n$. Represented mathematically, our function predicts a label $y_i = f(x_i)$ using the conditional probability distribution of $Y$ given $X$:

$$T = \{Y, P(Y|X)\} \tag{4}$$

Thus, given a transfer learning task with source (**S**) and target (**T**), there are four aspects of the task which might differ:

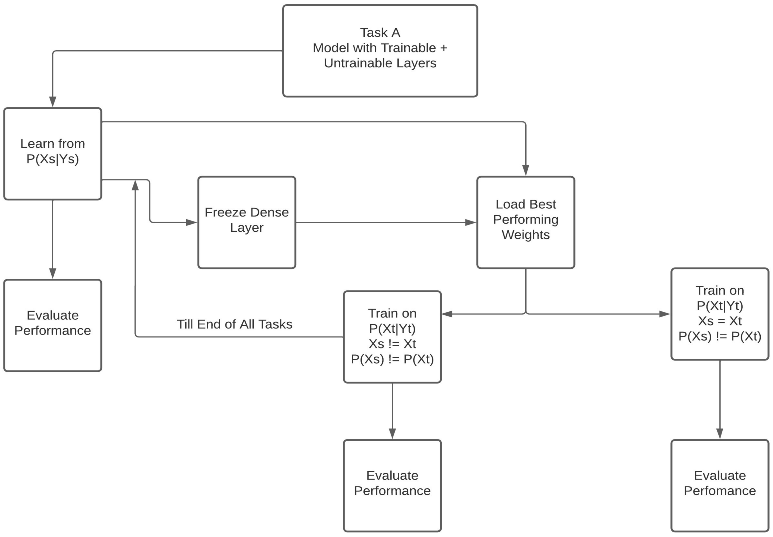

Figure 2: Model Architecture Flow. When $\mathrm{Xs} = \mathrm{Xt}$, both datasets use LIWC features. This keeps the feature space the same, with differing marginal distributions owing to separate datasets. When $\mathrm{Xs} \neq \mathrm{Xt}$, datasets have LIWC features in the source space and Word2Vec features in the target space.

- The feature space $\mathbf{X}$ of the source and target
- The marginal distribution $\mathbf{P}(\mathbf{X})$
- The label space $\mathbf{Y}$
- The conditional distribution $\mathbf{P}(\mathbf{Y}|\mathbf{X})$

We conduct our experiments under two variable conditions. In the first, we keep the feature space similar across transfer tasks, using LIWC (Tausczik and Pennebaker, 2010) features for both the source and target tasks. In the second we keep the feature space different between the two tasks, using LIWC features for the source task and Word2Vec (Mikolov et al., 2013) features for the target task. The label spaces are also the same, with binary classification labels across all tasks.

### 5.2 Feature Description

We extract both Word2Vec and LIWC features for each dataset. Word2Vec is a popular vector representation model that learns to predict words given their contexts from millions of online resources (Mikolov et al., 2013). Linguistic Inquiry and Word Count (LIWC) features are common in mental health tasks due to their demonstrated high performance for a wide range of applications including personality modeling, mental state assessment, affective analysis, and language understanding (Ludwig et al., 2013; Park et al., 2014; Schwartz et al., 2013; Pervin and Cervone, 2010; Coviello et al., 2014; Tumasjan et al., 2010; Riffe et al., 2019). They leverage syntactic patterns to provide feature representations that correlate with psycholinguistic characteristics (e.g., measuring cognitive complexity based on word length and words per sentence).

The creators of the *stress*, *depression*, and *suicide* datasets use a variety of features in their own work. We choose Word2Vec features and LIWC features since these exhibit the highest overlap across tasks in prior task-specific work. For instance, Losada and Crestani (2016) use TF-IDF vectorized text, and vectorized embeddings and LIWC features are also used by both Turcan and McKeown (2019) and Shing et al. (2018). The former use Word2Vec embeddings with BERT (Devlin et al., 2019) along with several attributes from LIWC including *clout*, *tone*, and *pronoun* features. The latter use domain-specific word embeddings from a SkipGram model trained on Reddit data, as well as bag-of-words features, topical features, readability scores, and features induced from LIWC, a mental health lexicon (Zirikly et al., 2016), and NRCLex (Mohammad and Turney, 2013).

### 5.3 Model Architecture and Training

Each model trains on $K$ tasks, where $K \in \{1, 2, .., N\}$, and is comprised of trainable and untrainable layers. Before all of our transfer tasks, each training dataset is padded to the same size (the vocabulary size from the largest training dataset). We consider a convolutional neural network (CNN), as well as to a lesser extent other models such as bidirectional long short-term memory (BiLSTM), LSTM, and RNN models.[7]

Each model in the input layer accepts the training data, consisting of the distribution of the feature space and labels to predict a classification label. Accuracy and $F_1$ scores are calculated for each task, and during transfer the dense trainable layer is frozen and the weights from the best performing epoch are loaded. Training then proceeds on the next task. This loop continues until all tasks have been learned and evaluation metrics have been calculated. Figure 2 illustrates this process.

Our best performing model is a CNN fine-tuned for transfer learning between datasets and a novel stress→depression→suicide prediction task. This model, as well as a BiLSTM alternative used in preliminary experiments, uses a one-dimensional max pooling layer with a poolsize of 2, flattening, a dropout of 0.5, and a frozen dense layer. The output layer has one node with a sigmoid activation.

## 6 Results and Discussion

Our experiments offer a first-of-its-kind examination of transfer learning across multiple MHCs. Since there are no directly comparable transfer learning models, we compare individual task performance to the respective benchmarks established by the dataset creators using task-specific models. These models leverage many architectures and feature types, intersecting in their use of vector representations and LIWC features. Specifically, we compare to the following:

- **Depression:** Losada and Crestani (2016) use TF-IDF vectorized embeddings with a logistic regression classifier.
- **Stress:** Turcan and McKeown (2019) use LIWC features and Word2Vec embeddings with a logistic regression classifier.
- **Suicide:** Shing et al. (2018) use LIWC features, Word2Vec embeddings, bag-of-words features, LDA features, and NRCLex features with a CNN classifier.

| Model | Depression | Stress | Suicide |
|---|---|---|---|
| Losada and Crestani (2016) | 0.66 | — | — |
| Turcan and McKeown (2019) | — | 0.79 | — |
| Shing et al. (2018) | — | — | 0.42 |
| **Ours** | **0.75** | **0.80** | **0.76** |

Table 3: Performance comparison between existing task-specific models (Losada and Crestani, 2016; Turcan and McKeown, 2019; Shing et al., 2018) and our transfer learning model reported here. Performance is measured using $F_1$.

For our own transfer CNN model (our highest-performing model), we train on: *stress* when using a target task of **depression**; *depression* when using a target task of **stress**; and *stress* and *depression* when using a target task of **suicide**, based on patterns of MHC expression identified in earlier reviewed literature. We report our findings in Table 3, using $F_1$ to measure performance. As shown, our model outperforms existing benchmarks with relative performance improvements of 13.64%, 1.27%, and 80.95% for *depression*, *stress*, and *suicide*, respectively and achieving a new state of the art with $F_1$ scores of 0.75, 0.80, and 0.76. We hope that these results will motivate other researchers to experiment with transfer learning across MHCs.

This answers one of our research questions: It is indeed possible to predict MHCs given information about existing ones, validating findings in mental health literature (Saini and Mandeep, 2020). However, the accuracy with which we can predict unseen mental health conditions depends on the feature space we use. LIWC features, which explicitly encode the psychological meaning of words, work better than Word2Vec features which rely purely on distributional semantics.

We also experiment with an alternative model grounded in psychological evidence that *suicide* may occur as a natural escalation from *stress* and then *depression*. We train our same core CNN model first on *stress*, then on *depression*, and then on *suicide* and achieve an 85% accuracy at the target task of **suicide**. Our BiLSTM model achieves an accuracy of 75% on **depression** when

[7] Preliminary experiments using RNN and LSTM achieved weaker performance than CNN and BiLSTM, so we did not pursue further experimentation with those models.

first trained on *stress*, and then an accuracy of 76% on **suicide** when subsequently trained on *depression*, echoing this trend albeit to a lesser degree. The strong performance of this technique further supports our finding that shared language characteristics across MHCs make this a promising and impactful sandbox for experiments with transfer learning.

## 7 Research Answers

In §1, we asked three important research questions. Following our analyses, we present concrete answers to them in this section.

**How do features relate across multiple MHCs?** Mental health conditions have similar manifestations in language, and correspondingly in their linguistic signatures. We provide evidence for this in our literature review (§3) and analyses (§4). Although we cannot through linguistic analysis conclusively measure the similarity of two MHCs, we can discern that the language usage and its features have significant overlap across MHCs (see Figure 1 and Tables 1 and 2, and other figures and tables in the Appendix).

**Can we represent different MHCs under the same feature spaces and find relations?** Yes, using semantically descriptive features such as LIWC it is possible to find relations (§4). We demonstrate that using standard NLP tools such as LDA or n-gram language modeling it is possible to see similar themes and topical relationships (§4).

**Can we identify underlying MHCs using the language of known ones?** Yes and No! While models trained on one task and transferred efficiently can predict unseen MHCs with a higher accuracy then when predicting them using only target domain data, these are linguistic classifications only (§5). AI models are still far from being able to conclusively identify MHCs, and should not be considered as replacements for professional mental health care.

Given these research answers, we close by discussing how we can carry this forward and what it means for NLP in mental health.

## 8 Conclusion and Future Directions

In this work, we examine the utility of transfer learning for the identification of three MHCs: depression, stress, and suicidal ideation. These MHCs vary in their clinical classification and severity. Depression is formally defined as a mental illness (Kanter et al., 2008), stress is a process which may ultimately result in mental illness (Salleh, 2008), and suicidal ideation is classified as a disorder (Fehling and Selby, 2021). Although we achieve promising performance in detecting these conditions, nothing—not even actual diagnosis by a human expert—can conclusively identify a mental illness with 100% certainty (Allsopp et al., 2019).

We presented a qualitative exploration of the overlap and interplay between language and mental health across multiple MHCs, and also presented quantitative correlations among words, tokens, themes, topics, and large feature space representations using well-known, established NLP methods. Finally, we introduced a transfer learning model to predict unseen mental health conditions using similar and dissimilar feature spaces, the first of its kind. Our model outperforms the baselines established by benchmark models for detecting depression, stress, and suicide with percent increases in measured performance of 13.64%, 1.27%, and 80.95%, respectively. The model also achieved an 85% accuracy at detecting suicidal ideation in a psychologically informed model that trains on datasets in an order established by clinical evidence, with *stress* followed by *depression*[8] and then ultimately *suicide* (Orsolini et al., 2020).

Although this paper demonstrated preliminary evidence that similarities in feature spaces can be leveraged to better predict unknown MHCs, in the future we wish to explore this further with a larger variety of models. We also plan to further examine the role that transfer learning order has in establishing performance.[9] Other work has found that social media-based models do not always generalize and may incur substantial performance losses (Harrigian et al., 2020), and other factors such as social concerns, self-disclosure bias, and temporal artifacts may also influence model performance (Harrigian et al., 2020). We hope that researchers will use our findings to explore new ways to increase the efficiency and usefulness of AI-supported treatment and diagnosis of MHCs (Allsopp et al., 2019).

[8] www.psychologytoday.com/us/blog/in-practice/201303/why-stress-turns-depression

[9] In some early experiments not reported here, reversing the transfer learning order of our model resulted in performance that peaked at an $F_1$=0.48.

## 9 Acknowledgements

This work was funded by a start-up grant from the University of Illinois at Chicago and by the National Institute Of Mental Health of the National Institutes of Health under Award Number R01MH116902. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the anonymous technical and clinical reviewers for their insightful comments. We are also very grateful to all of the teams who shared their valuable datasets with us.

## References

Kate Allsopp, John Read, Rhiannon Corcoran, and Peter Kinderman. 2019. Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research*, 279.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. #whywetweetmh: Understanding why people use twitter to discuss mental health problems. *Journal of Medical Internet Research*, 19.

José Bertolote. 2008. The roots of the concept of mental health. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, 7:113–6.

Michael Birnbaum, Sindhu Kiranmai Ernala, Asra Rizvi, Munmun Choudhury, and John Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research*, 19:e289.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Louise Brådvik. 2018. Suicide risk and mental disorders. *International Journal of Environmental Research and Public Health*, 15:2028.

M.D. Choudhury and S. De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 71–80.

Munmun Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems . CHI Conference*, volume 2016, pages 2098–2110.

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. Association for Computational Linguistics.

Lorenzo Coviello, Yunkyu Sohn, Adam Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas Christakis, and James Fowler. 2014. Detecting emotional contagion in massive social networks. *PloS one*, 9:e90315.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long* *and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kara B. Fehling and Edward A. Selby. 2021. Suicide in dsm-5: Current evidence for the proposed suicide behavior disorder and other possible improvements. *Frontiers in Psychiatry*, 11.

Alison Fleming, E Klein, and C Corter. 1992. The effects of a social support group on depression, maternal attitudes and behavior in new mothers. *Journal of child psychology and psychiatry, and allied disciplines*, 33:685–98.

Hector-Hugo Franco-Penya and Liliana Mamani Sanchez. 2016. Text-based experiments for predicting mental health emergencies in online web forum posts. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 193–197, San Diego, CA, USA. Association for Computational Linguistics.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016a. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 63–73, San Diego, CA, USA. Association for Computational Linguistics.

George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. 2016b. Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105, San Diego, CA, USA. Association for Computational Linguistics.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930, Copenhagen, Denmark. Association for Computational Linguistics.

Bonnie Harmer, Sarah Lee, Truc vi H Duong, and Abdolreza Saadabadi. 2022. Suicidal ideation. *StatPearls*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the 7th Workshop on Computational Linguistics and Clinical Psychology: Improving Access*.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, Baltimore, Maryland, USA. Association for Computational Linguistics.

Derek Howard, Marta Maslej, Justin Lee, Jacob Ritchie, Geoffrey Woollard, and Leon French. 2019. Transfer learning for risk classification of social media posts: Model evaluation study (preprint). *Journal of Medical Internet Research*, 22.

Jena D. Hwang and Kristy Hollingshead. 2016. Crazy mad nutters: The language of mental health. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 52–62, San Diego, CA, USA. Association for Computational Linguistics.

Thomas Insel. 2008. Assessing the economic costs of serious mental illness. *The American journal of psychiatry*, 165:663–5.

Alice Isen, Kimberly Daubman, and Gary Nowicki. 1987. Positive affect facilitates creative problem solving. *Journal of personality and social psychology*, 52:1122–31.

Nadine Jung, Christina Wranke, Kai Hamburger, and Markus Knauff. 2014. How emotions affect logical reasoning:evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Frontiers in psychology*, 5:570.

Jonathan Kanter, Andrew Busch, Cristal Weeks, and Sara Landes. 2008. The nature of clinical depression: Symptoms, syndromes, and behavior analysis. *The Behavior analyst / MABA*, 31:1–21.

Sanne M.A. Lamers, Khiet P. Truong, Bas Steunenberg, Franciska de Jong, and Gerben J. Westerhof. 2014. Applying prosodic speech features in mental health care: An exploratory study in a life-review intervention for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 61–68, Baltimore, Maryland, USA. Association for Computational Linguistics.

Huanhuan Li, Li Wang, Zhanbiao Shi, Yuching Zhang, wu Kankan, and Ping Liu. 2010. Diagnostic utility of the ptsd checklist in detecting ptsd in chinese earthquake victims. *Psychological reports*, 107:733–9.

David Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822, pages 28–39.

Stephan Ludwig, ko de ruyter, Mike Friedman, Elisabeth Brüggen, Martin Wetzels, and Gerard Pfann. 2013. More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77:87–103.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 133–137, San Diego, CA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Max Mills. 2017. Sharing privately: the effect publication on social media has on expectations of privacy. *Journal of Media Law*, 9:1–27.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Marja-Terttu MNSc, Marita PhD, Marja-Terttu Tarkka, and Marita Paunonen. 1996. Social support and its impact on mothers'experiences of childbirth. *Journal of Advanced Nursing*, 23:70 – 75.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Laura Orsolini, Roberto Latini, Maurizio Pompili, Gianluca Serafini, Umberto Volpe, Federica Vellante, Michele Fornaro, Alessandro Valchera, Carmine Tomasetti, Silvia Fraticelli, Marco Alessandrini, Raffaella Rovere, Sabatino Trotta, Giovanni Martinotti, Massimo di Giannantonio, and Domenico De Berardis. 2020. Understanding the complex of suicide in depression: from research to clinics. *Psychiatry investigation*, 17:207–221.

Javier Parapar, Patricia Martín-Rodilla, David Losada, and Fabio Crestani. 2021. *Overview of eRisk 2021: Early Risk Prediction on the Internet*, pages 324–344. Springer Link.

Gregory Park, H. Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2014. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108.

Ted Pedersen. 2015. Screening Twitter users for depression and PTSD with lexical decision lists. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages

46–53, Denver, Colorado. Association for Computational Linguistics.

Abed-Esfahani Pegah, Howard Derek, Maslej Marta, Sejal Patel, Vamika Mann, Sarah Goegan, and Leon French. 2019. Transfer learning for depression: Early detection and severity prediction from social media postings. *ERisk*, 2380.

L Pervin and D Cervone. 2010. *Personality. Theory and research*. Wiley.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Philip Resnik, Rebecca Resnik, and Margaret Mitchell, editors. 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Daniel Riffe, Stephen Lacy, Brendan Watson, and Frederick Fico. 2019. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Routledge.

(James) Rosenquist, J.H. Fowler, and Nicholas Christakis. 2010. Social network determinants of depression. *Molecular psychiatry*, 16:273–81.

Gery Ryan and H. Bernard. 2003. Techniques to identify themes. *Field Methods - FIELD METHOD*, 15:85–109.

Satvinder Saini and Mandeep. 2020. A study of perceived stress and loneliness in older people with depression. *International Journal on Aging Human Development*.

Mohd Salleh. 2008. Life event, stress and illness. *The Malaysian journal of medical sciences : MJMS*, 15:9–18.

H. Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin Seligman, and Lyle Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8:e73791.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014a. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014b. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kathryn Scott, Phyllis Klaus, and Marshall Klaus. 2000. The obstetrical and postpartum benefits of continuous support during childbirth. *Journal of women's health & gender-based medicine*, 8:1257–64.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Liping Shen, Minjuan Wang, and Ruimin Shen. 2009. Affective e-learning: Using ”emotional” data to improve learning in pervasive learning environment. *Educational Technology & Society*, 12:176–189.

Benjamin Shickel, Martin Heesacker, Sherry Benton, Ashkan Ebadi, Paul Nickerson, and Parisa Rashidi. 2016. Self-reflective sentiment analysis. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 23–32, San Diego, CA, USA. Association for Computational Linguistics.

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.

Michael Tanana, Aaron Dembe, Christina S. Soma, Zac Imel, David Atkins, and Vivek Srikumar. 2016. Is sentiment in movies the same as sentiment in psychotherapy? comparisons using a new psychotherapy sentiment database. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 33–41, San Diego, CA, USA. Association for Computational Linguistics.

Mashrura Tasnim and Eleni Stroulia. 2019. *Detecting Depression from Voice*, pages 472–478. Springer Link.

Yla Tausczik and James Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Social Science Computer Review*, volume 10. Sage Journals.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Patrik Vuilleumier. 2006. Vuilleumier p. how brains beware: neural mechanisms of emotional attention. trends cogn sci 9: 585-594. *Trends in cognitive sciences*, 9:585–94.

U. Yasaswini, Y. Sasidhar, P. Sai, P. Eswar, and V. Swathi. 2021. Detecting depression in tweets using distilbert. *International Journal of Innovative Research in Computer Science & Technology*, 9.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The GW/UMD CLPsych 2016 shared task system. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 166–170, San Diego, CA, USA. Association for Computational Linguistics.

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*.

## A Dataset Descriptions

In Table 4, we present dataset statistics for *depression*, *stress*, and *suicide*. Further details regarding these datasets can be found in the original papers (Losada and Crestani, 2016; Turcan and McKeown, 2019; Shing et al., 2018). We deeply thank all the authors and creators of these datasets.

## B Analytical Figures

In this section we include additional figures produced during data analysis. Figures 4 and 5 show cognitive complexity for individuals with depression and suicidal ideation, and Figure 3 shows graphical representations of LDA analyses on people with stress.

## C Extended Qualitative Analysis of N-Gram Frequency

In this section we include figures showing the most frequent n-grams associated with *depression*, *suicide*, and *stress*. Trigrams for *depression*, *suicidal ideation*, and *stress* are shown in Figures 6, 7, and 8, respectively. Bigrams for *depression*, *suicidal ideation*, and *stress* are shown in Figures 9, 10, and 11, and unigrams for the same three MHCs are shown in Figures 12, 13, and 14.

| **Dataset** | **Size** | **Labeling Scheme** | **Labels Used in Our Experiments** | **Baseline $F_1$** |
|---|---|---|---|---|
| *Depression* | 531,453 posts from 892 users | Binary | Binary | 0.66 |
| *Stress* | 187,444 posts | Binary | Binary | 0.79 |
| *Suicide* | 11,129 initial users, downsampled to 934 | Categorical (4 Categories) | Aggregated Binary | 0.42 |

Table 4: Additional descriptive statistics regarding *depression* (Losada and Crestani, 2016), *stress* (Turcan and McKeown, 2019), and *suicide* (Shing et al., 2018).

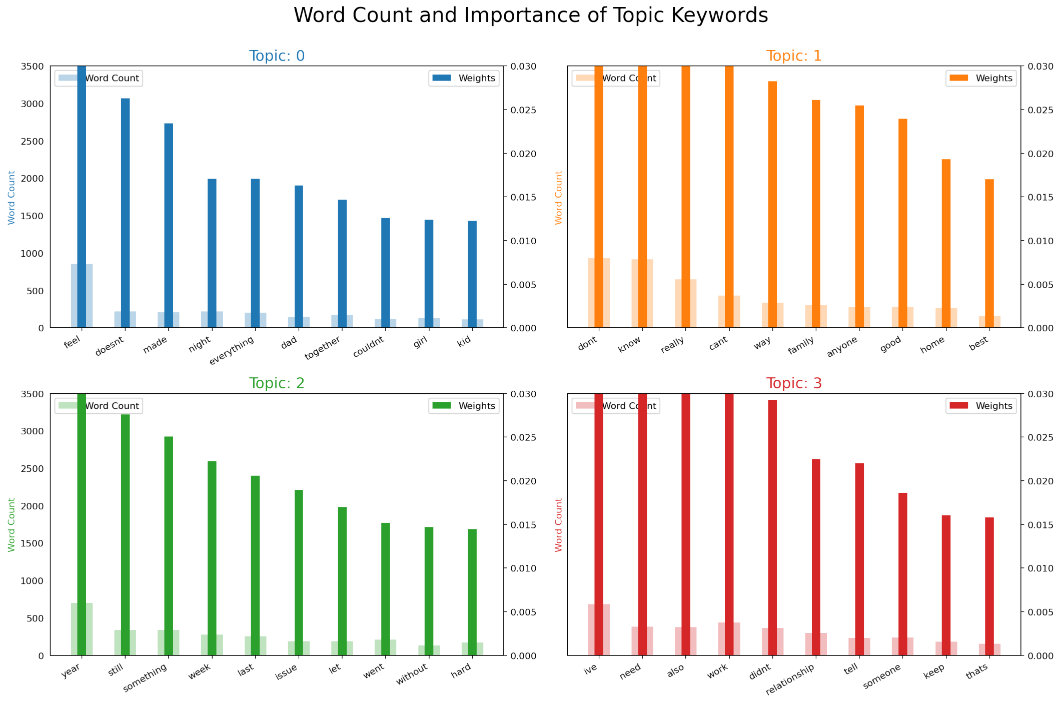

Figure 3: Top four topics identified when applying LDA to *stress*.

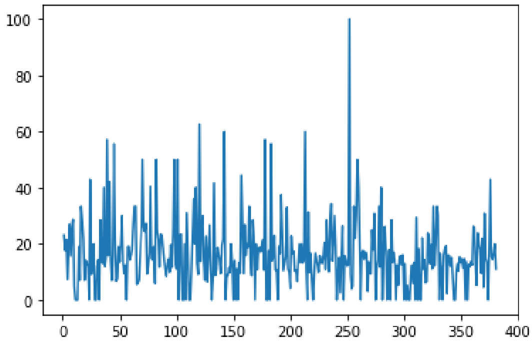

Figure 4: Cognitive complexity of a random subsample with depression.

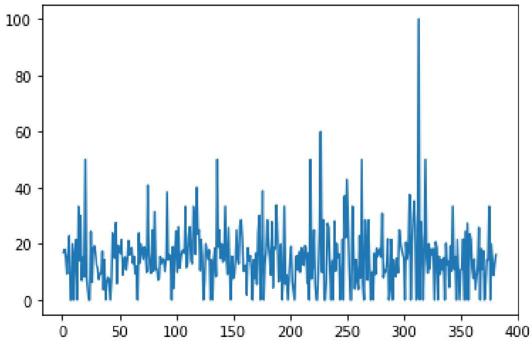

Figure 5: Cognitive complexity of a random subsample with suicidal ideation.

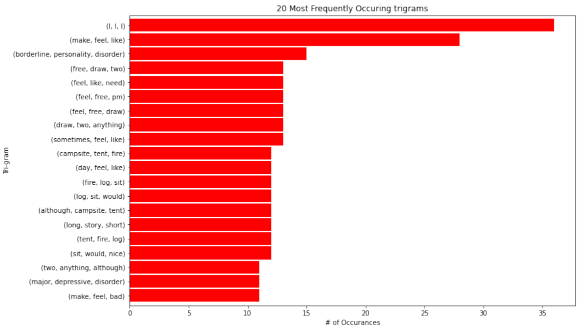

Figure 6: Most frequent trigrams in a random subsample (*depression*).

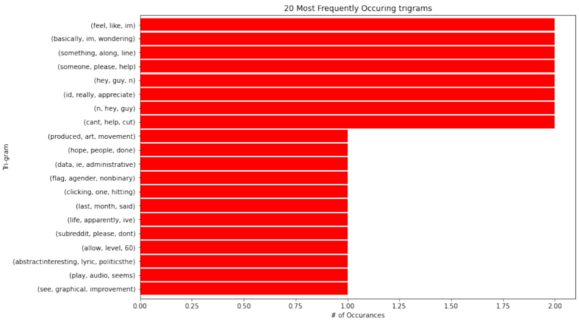

Figure 7: Most frequent trigrams in a random subsample (*suicide*).

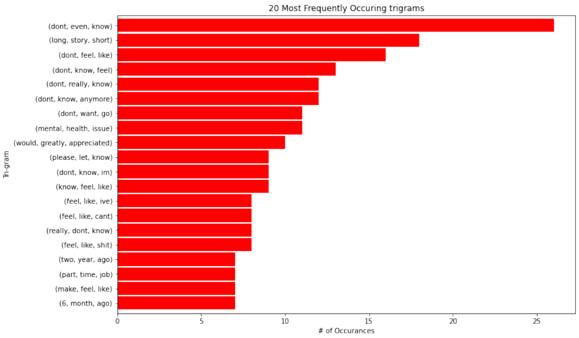

Figure 8: Most frequent trigrams in a random subsample (*stress*).

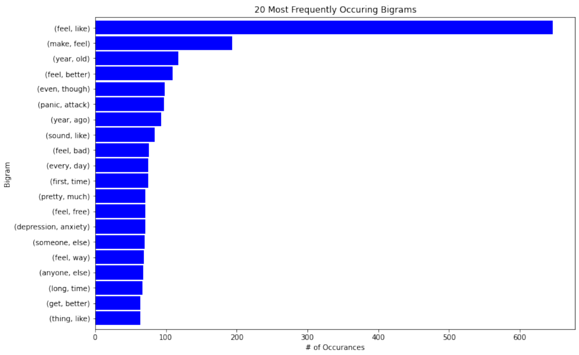

Figure 9: Most frequent bigrams in a random subsample (*depression*).

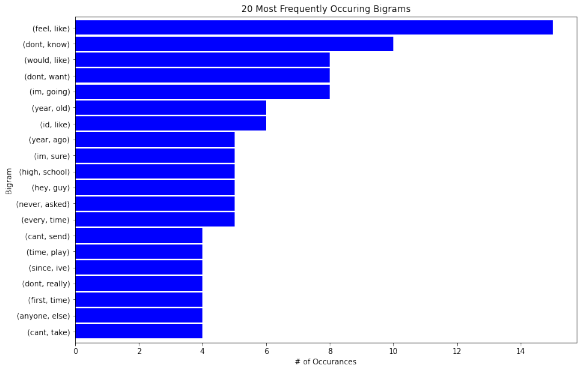

Figure 10: Most frequent bigrams in a random subsample (*suicide*).

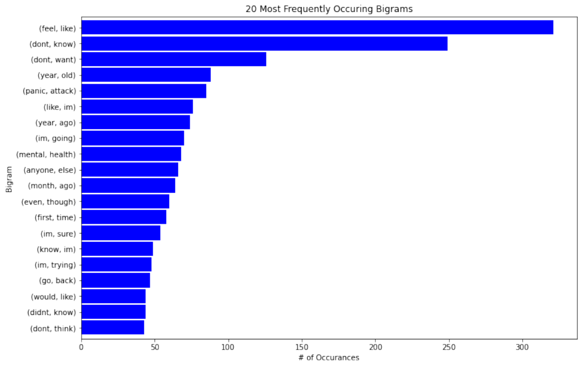

Figure 11: Most frequent bigrams in a random subsample (*stress*).

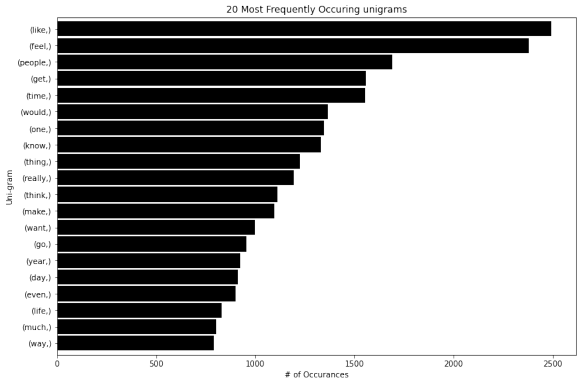

Figure 12: Most frequent unigrams in a random subsample (*depression*).

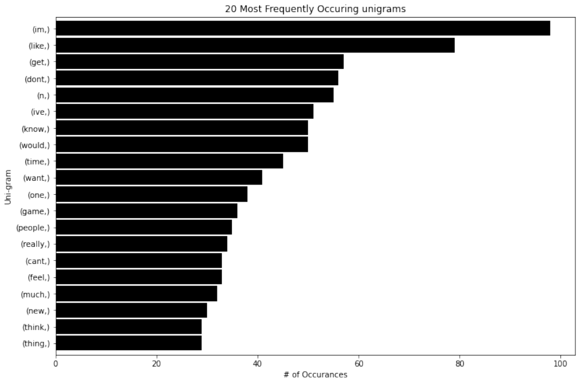

Figure 13: Most frequent unigrams in a random subsample (*suicide*).

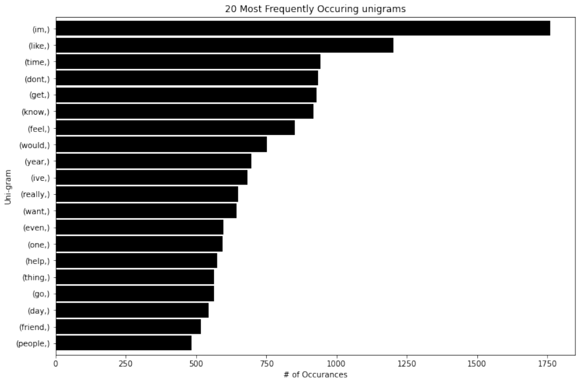

Figure 14: Most frequent unigrams in a random subsample (*stress*).