A Cross-modal Review of Indicators for Depression Detection Systems

Michelle Renee Morales
Linguistics Department
The Graduate Center, CUNY
New York, NY 10016
mmorales@gradcenter.cuny.edu

Stefan Scherer
USC Institute for
Creative Technologies
Los Angeles, CA 90094
scherer@ict.usc.edu

Rivka Levitan
Computer Science Department
Brooklyn College, CUNY
Brooklyn, NY 11210
levitan@sci.brooklyn.cuny.edu

Abstract

Automatic detection of depression has attracted increasing attention from researchers in psychology, computer science, linguistics, and related disciplines. As a result, promising depression detection systems have been reported. This paper surveys these efforts by presenting the first cross-modal review of depression detection systems and discusses best practices and most promising approaches to this task.

1 Introduction

Given advancements in hardware and software, coupled with the explosion of smartphone use, the forms of potential health care solutions have begun to change and interest in developing technologies to assess mental health has grown. Among the latest technologies are depression detection systems, which use indicators from an individual in combination with machine learning to make automated depression level assessments. Researchers have made significant progress, but challenges remain. One major challenge is the existing disconnect between language technology subfields: approaches to depression assessment from natural language processing (NLP), speech processing, and human-computer interaction (HCI) tend to silo by subfield, with little discussion about the utility of combining promising approaches. This existing disconnect necessitates a bridge to facilitate greater collaboration and cooperation across subfields and modalities.

Experts across several fields are attempting to build valid tools for depression assessment. Each subfield tends to approach the task from a unique perspective, with slightly different goals, and completely different data sources. Due to these experimental differences, it is difficult to compare approaches and even more difficult to combine promising approaches. For example, if we consider data sources alone, NLP research has aimed to detect depression from writing, both formal and informal (i.e. online text), speech processing research has aimed to assess depression level from audio while HCI and related fields try to assess depression level from video. Each data source is then labeled for depression through different approaches, including rating scales, self-report surveys, manual annotation, etc. As a result, we see various definitions of how depression is defined across studies. Regardless of the existing differences, every study and system share the common goal of discovering a way to use technology to help assess depression.

This survey paper aims to serve as a bridge between the subfields by providing the first review of depression detection systems across subfields and modalities. This paper focuses on the following research questions, how has depression been defined and annotated in detection systems? What kinds of depression data exists or could be obtained for depression detection systems? What (multimodal) indicators have been used for the automatic detection of depression? How do we evaluate depression detection systems? Each research question could serve as the main focus of an entire paper. Therefore, this review briefly touches upon each question and dedicates the most focus to reviewing indicators of depression and subsequently features for depression detection systems. We cover numerous features across modalities, including visual, acoustic, linguistic, and social. We briefly review approaches to defining and annotating depression, existing data sources, and how to evaluate depression detection systems. Lastly, we end our discussion with the practical or ethical issues that require attention when building systems for depression detection.

Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology, pages 1–12,
Vancouver, Canada, August 3, 2017. © 2017 Association for Computational Linguistics

2 Defining and Labeling Depression

2.1 Clinical Definition and Diagnostics

According to the Diagnostic and Statistical Manual of Mental Disorders (APA, 2013), the most widely used resource in diagnosing mental disorders in the United States, most people will experience some feelings of depression in their lifetime, although it does not meet the criteria of an illnesss until a person has experienced, for longer than a two-week period, a depressed mood and/or a markedly diminished interest/pleasure in combination with four or more of the following symptoms: significant unintentional weight loss or gain, insomnia or sleeping too much, agitation or psychomotor retardation noticed by others, fatigue or loss of energy, feelings of worthlessness or excessive guilt, diminished ability to think or concentrate, indecisiveness, or recurrent thoughts of death. In addition, diagnosis requires that the symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

Commonly used assessment tools for depression include clinical interviews or self-assessments. The Hamilton Rating Scale for Depression (HAM-D) (Hamilton, 1960) is a widely used assessment tool and is often regarded as the most standard assessment tool for depression for both diagnosis and research purposes (Cummins et al., 2015a). The HAM-D is clinician-administered, includes 21 questions, and takes 20 to 30 minutes to complete. The interview assesses the severity of symptoms associated with depression and gives a patient a score, which relates to their level of depression. Some symptoms included are depressed mood, insomnia, agitation, and anxiety. Each of the questions has 3 to 5 possible responses which range in severity, scored between 0-2, 2-3, or 4-5 depending on the importance of the symptom. All scores are then summed and the total is arranged into 5 categories (normal-severe).

There also exist commonly used self-report measures, including the the Beck Depression Inventory (BDI-II) (Beck et al., 1961). The BDI-II is a self-report questionnaire that consists of 21 items and takes 5 to 10 minutes to complete. The question items aim to cover important cognitive, affective, and somatic symptoms associated with depression. Each question receives a score on a scale from 0-3 depending on how severe the symptom was over the previous week. Similar to HAM-D, all scores are summed and the final score is categorized into 4 different levels (minimal-severe). Other diagnostic tools include the Montgomery-Åsberg Depression Rating Scale (Montgomery and Asberg, 1979), the Patient Health Questionnaire (Kroenke et al., 2001), and the Quick Inventory of Depressive Symptomology (Rush et al., 2003).

2.2 Scalable Approaches to Annotation

When working with datasets, it is not always feasible to acquire clinical ratings for depression level. As a result, researchers have come up with innovative ways of acquiring depression labels at scale, notably from social media sources. Given the explosion of social media, this domain is especially rich in data for mental health research. However, any research in this domain must take into account the ability of online users to be anonymous or even deceptive.

Coppersmith et al. (2015) looked for tweets that explicitly stated “I was just diagnosed with depression”. Moreno et al. (2011) evaluated Facebook status updates using references to depression symptoms such as “I feel hopeless” to ultimately determine depression label. Choudhury et al. (2013) used crowdsourcing, via the Amazon Mechanical Turk platform, to collect Twitter usernames as well as labels for depression. Reece and Danforth (2016) used a similar crowdsourcing approach to collect both depression labels and Instagram photo data. In some approaches to annotation, depression is subsumed into broader categories like distress, anxiety, or crisis. For example, Milne et al. (2016) used judges to manually annotate how urgently a blog post required attention, using a triage system of green/amber/red/crisis.

These innovative approaches to data annotation highlight the potential of social media data. This domain offers a very rich data source which can be used to build, train, and test models to automatically perform mental health assessments at a large scale.

3 Datasets

The task of depression detection is inherently interdisciplinary and all disciplines—psychology, computer science, linguistics—bring an essential set of skills and insight to the problem. However, it is not always the case that a team is fortunate enough to have collaborators from all disciplines. One way to promote collaboration is to

Dataset	Modality	Depression Label Annotation	Reference
AVEC 2013	Video/audio	Self-report survey (BDI-II)	Valstar et al. (2013)
AVEC 2014	Video/audio	Self-report survey (BDI-II)	Valstar et al. (2014)
Crisis Text Line	Text	Crisis counselor judgment	Lieberman and Meyer
DAIC	Video/audio/text	Self-report survey (PHQ-8)	Gratch et al. (2014)
DementiaBank Database	Video/audio/text	Clinical diagnosis of depression (HAM-D)	Becker et al. (1994)
ReachOut Triage Shared Task	Text	Expert judged for crisis/green/amber/red	Milne et al. (2016)
SemEval-2014 Task 7	Text	Hand labeled for depression	Pradhan et al. (2014)

Table 1: Datasets for depression detection systems.

organize challenges and publicly release data and code. Public datasets are invaluable resources that can give new researchers the ability to work on the task while connecting accomplished researchers across disciplines. The Computational Linguistics and Clinical Psychology (CLPsych) Shared Task (2013-2017) and the Audio/Visual Emotion Recognition (AVEC) Workshop Depression Subchallenge (2013-2016) are examples of depression detection system challenges that spurred interest, promoted research, and built connections across the research community. In this section, we describe the kinds of depression data that exist, listed in Table 1. We focus solely on datasets that are publicly available to download. For a detailed list of databases both private and public that have been used in speech processing studies see (Cummins et al., 2015a).

Both the AVEC 2013 and 2014 corpora are available to download[1]. The AVEC challenges are organized competitions aimed at comparing multimedia processing and machine learning methods for automatic audio, video and audiovisual emotion and depression analysis, with all participants competing under strictly the same conditions. The AVEC 2013 corpus (Valstar et al., 2013) includes 340 video clips in German of subjects performing a HCI task while being recorded by a webcam and a microphone. The video files each contain a range of vocal exercises, including free and read speech tasks. The level of depression is labeled with a single value per recording using the BDI-II. The AVEC 2014 corpus (Valstar et al., 2014) is a subset of the AVEC 2013 corpus. In total, the corpus includes 300 videos in German; the duration ranges from 6 seconds to 4 minutes. The files include a read speech passage (Die Sonne und der Wind) and an answer to a free response question.

The Crisis Text Line [2] is a free 24/7 crisis support texting hot line where live trained crisis counselors receive and respond quickly to texts. The main goal of the organization is to support people with mental health issues through texting. The organization includes an open data collaboration. In order to gain access, researchers must complete an Institutional Review Board application with their own university and an application with Crisis Text Line, which gives researchers access to a vast amount of text data annotated by conversation issue, including but not limited to depression, anger, sadness, body image, homelessness, self-harm, suicidal ideation, and more.

The Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014) contains clinical interviews in English designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The interviews were conducted by an animated virtual interviewer called Ellie. The DAIC interviews were meant to simulate the first step in identifying mental illness in health care settings, which is a semi-structured interview where health care providers ask a series of open-ended questions with the intent of identifying clinical symptoms. The corpus includes audio and video recordings and extensive questionnaire responses. Each interview includes a depression score from the PHQ-8 (Kroenke et al., 2009). A portion of the corpus was released during the AVEC 2016 Depression Sub-challenge and is available to download[3]. The publicly-available dataset also includes transcripts of the interview.

The DementiaBank Database[4] represents data collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh (Becker et al., 1994). DementiaBank is a shared database of multimedia interactions for the study of communication in dementia. A subset of the participants from the dataset also have HAM-D depression scores.

The ReachOut Triage Shared Task dataset[5] consists of 65,024 forum posts written between July 2012 and June 2015 (Milne et al., 2016). A subset

[1] https://avec2013-db.sspnet.eu/

[2] www.crisistextline.org

[3] http://dcapswoz.ict.usc.edu/

[4] http://dementia.talkbank.org/

[5] http://clpsych.org/shared-task-2016/

of the corpus (1,227 posts) is manually annotated by three separate expert judges indicating how urgently a post required a moderators attention. Labels included crisis, red, amber, and green.

The SemEval-2014 Task 7 (Pradhan et al., 2014) dataset[6] represents clinical notes which are annotated for disorder mentions, including mental disorders such as depression.

4 Indicators of Depression

Ideally, machine learning tools for depression detection should have access to the same streams of information that a clinician utilizes in the process of forming a diagnosis. Therefore, features used by such classifiers should represent each communicative modality: face and gesture, voice and speech, and language. This section provides a review of each modality highlighting markers that have had success in systems.

4.1 Visual Indicators

Visual indicators have been widely explored for depression analysis, including body movements, gestures, subtle expressions, and periodical muscular movements.

Girard et al. (2014) investigated whether a relationship existed between nonverbal behavior and depression severity. In order to measure nonverbal behavior they used the Facial Action Coding System (FACS) (Ekman et al., 1978). FACS is a system used to taxonomize human facial movements by their appearance on the face. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists. FACS is composed of facial Action Units (AUs), which represent the fundamental actions of individual muscles or groups of muscles. Girard et al. (2014) found that participants with high levels of depression made fewer affiliative facial expressions, more non-affiliative facial expressions, and diminished head motions. Scherer et al. (2013b) also investigated visual features using FACS and found that depression could be predicted by a more downward angle of the gaze, less intense smiles, shorter average durations of smile, longer self-touches, and fidgeting.

In addition to FACS features for video analysis, others have considered Space-Time Interest Points (STIP) features (Cummins et al., 2013; Joshi et al., 2013), which capture spatio-temporal changes including movements of the face, hands, shoulder, and head. Using STIP features, Joshi et al. (2013) found that they could detect depression with 76.7% accuracy. Their results showed that body expressions, gestures, and head movements can be significant visual cues for depression detection.

4.2 Speech Indicators

Recent research has shown the promise in using speech as a diagnostic and monitoring aid for depression (Cummins et al., 2015b,a, 2014; Scherer et al., 2014; Williamson et al., 2014a). The speech production system of a human is very complex and as a result slight cognitive or physiological changes can produce acoustic changes in speech. This idea has driven the research on using speech as an objective marker for depression. Depressed speech has consistently been associated with a wide range of prosodic, source, formant and spectral indicators. For a thorough review of speech processing research for depression detection see (Cummins et al., 2015a).

Many researchers have provided evidence for the robustness of prosodic indicators to capture depression level, specifically noting the promise of speech-rate (Mundt et al., 2012; Hönig et al., 2014). Cannizzaro et al. (2004) examined the relationship between depression and speech by performing statistical analyses of different acoustic measures, including speaking rate, percent pause time, and pitch variation. Their results demonstrated that speaking rate and pitch variation had a strong correlation with the depression rating scale. Moore et al. (2008) investigated the suitability for a classification system formed from the combination of prosodic, voice quality, spectral, and glottal features and reported maximum accuracy of 91% for male speakers and 96% accuracy for females speakers when classifying between absence/presence of depression.

Stassen et al. (1998) found for 60% of patients in their study that speech pause duration was significantly correlated with their HAM-D score. Alpert et al. (2001) also found significant differences in speech pause duration between spontaneous speech of their depressed group versus their control group. Cannizzaro et al. (2004) found a significant correlation between reduced speaking rate and HAM-D score. Mundt et al. (2012) found six prosodic timing measures to be significantly correlated with depression severity, includ-

[6] http://alt.qcri.org/semeval2014/task7/index.php?id=data-and-tools

ing total speech time, total pause time, percentage pause time, speech pause ratio, and speaking rate. Hönig et al. (2014) reported a positive correlation with increasing levels of speaker depression and average syllable duration. Trevino et al. (2011) found that changes in speech rate are stronger at the phoneme level, finding stronger relationships between speech rate and depression severity when using phone-duration and phone-specific measures instead of a global speech rate. Cohn et al. (2009) investigated vocal prosody and found that variation in fundamental frequency and latency of response to interviewer questions achieved 79% accuracy in distinguishing participants with moderate/severe depression from those with no depression.

Low et al. (2011) investigated various acoustic features, including spectral, cepstral, prosodic, glottal and a Teager energy operator based feature. In their best performing systems, using sex-dependent models, they achieved 87% accuracy for males and 79% for females. In Cummins et al. (2011) spectral features, particularly mel-frequency cepstral coefficients (MFCCs) were found to be useful, distinguishing 23 depressed participants from 24 controls with an accuracy of 80% in a speaker-dependent configuration. Scherer et al. (2013a) found glottal features (normalized amplitude quotient and quasi-open quotient) differed significantly between depressed and control groups. When used to detect depression they found glottal features to differentiate between the 2 groups with 75% accuracy. Alghowinem et al. (2013) investigated a number of feature sets for detecting depression from spontaneous speech and found loudness and intensity features to be the most discriminative.

4.3 Linguistic and Social Indicators

While most literature concerning depression detection systems has focused on the speech signal, there is a related body of work on detecting depression from writing using linguistic cues. For clinical psychologists, language plays a central role in diagnosis. Therefore, when building language technology in the domain of mental health it is essential to consider both the acoustic and linguistic signal. For an in-depth review of NLP applications for mental health assessment see Calvo et al. (2017).

Features derived from the speech signal are motivated by ways in which the cognitive and physical changes associated with depression can lead to differences in speech. Similarly, psychological and sociological theories suggest that depressed language can be characterized by specific linguistic features. Aaron Beck's (1967) cognitive theory of depression posits that people prone to depression possess a depressive schema, leading them to see themselves and the world in pervasively negative terms. When activated, these schema give rise to depressive thinking. A stressful event can then trigger these schema, leading an individual to perceive the event in a negative way and, as a result, cause an episode of depression. Pyszczynski and Greenberg (1987) speculated that depressed individuals think a great deal about themselves, stressing the role of self-focused attention and extreme self-criticism. Also related is the social integration model by Durkheim (1951), which posits that the perception of oneself as not integrated into society (detached from social life) is key to suicidality and is also relevant to the depressed persons' perceptions of self.

These theories have motivated empirical studies of depressed language which have in turn provided support for their validity. Stirman and Pennebaker (2001) provided evidence consistent with both the self-focus and social integration perspectives by studying the word usage of suicidal and non-suicidal poets. They conducted a comparison of 300 poems from the early, middle, and late periods of nine poets who committed suicide and nine who did not. They used the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007), which is a text analysis tool that can be used to count words in psychologically meaningful categories. Using LIWC, they found that suicidal poets used more first-person singular (I, me, my) words, and fewer first-person plural (we, us, our) words. In related work, Poulin et al. (2014) used medical records and a text analysis approach to predict suicide risk with an accuracy of 65%, finding that certain words were predictive of suicide.

Later work by Rude et al. (2004) analyzed narratives written by currently-depressed, formerly-depressed, and never-depressed college students. In the context of an essay task, they examined linguistic patterns using LIWC, including the use of first person singular, first person plural, social references, and negatively/positively valenced words. As hypothesized based on Pyszcynski and Greenberg's model of self-focus, depressed students used significantly more first person singular words than did never-depressed individuals. They also found that depressed students used more negatively valenced words and fewer positive emotion words, supporting both the negative focus predicted by Beck's cognitive theory of depression and the self-preoccupation predicted by Psyzcynski and Greenberg's control theory of depression. Given the success of LIWC in Rude et al.'s work, many other researchers have incorporated LIWC into depression detection systems with encouraging results. Nguyen et al. (2014) found LIWC to be useful in capturing topic and mood which showed good predictive validity in depression classification between clinical and control groups in blog post texts. Morales and Levitan (2016b) incorporated LIWC into a depression detection system and found certain LIWC categories to be useful in measuring specific depression symptoms, including sadness and fatigue.

Various approaches to modeling word usage have had much success in detecting depression. Coppersmith et al. (2015) accurately identified depression with high accuracies using n-gram models in Twitter text. Althoff et al. (2016) presented a large-scale quantitative study on the discourse of counseling conversations. They developed a set of discourse features to measure how correlated linguistic aspects of conversations were with outcomes. Features in their study included: sequence-based conversation models, language model comparisons, message clustering, and psycholinguistics-inspired word frequency analyses. Their results were also consistent with Psyzcynski and Greenberg's theory of depression, in that texters with a smaller amount of self-focus were associated with more successful conversations. In addition, Schwartz et al. (2014) showed that regression models based on Facebook language can be used to predict an individuals degree of depression.

In addition to considering word usage, researchers have also explored syntactic characteristics of depressed language. Zinken et al. (2010) investigated whether an analysis of a depressed patients' syntax could help predict improvement of symptoms. This work built upon previous findings that showed the health benefit of expressive writing (Pennebaker, 1997). Building upon this work, Zinken et al. considered the psychological relevance of syntactic structures of language use. Word use and syntactic structure were analyzed to explore whether the degrees to which a participant constructs relationships between events in a brief text can inform the likelihood of successful participation in depression treatment. They also used LIWC and targeted 2 categories: causation words and insight words. In addition, they manually coded eight different syntactic structures (ranging from simple to complex) in the patients' narratives. They found that certain structures were correlated with patients' potential to complete a self-help treatment. Zinken et al.'s findings demonstrate the promise in investigating syntactic characteristics of an individual's language use. Moreover, related work has found that differences in frequencies of part-of-speech (POS) tags were useful in detecting depression from writing (Morales and Levitan, 2016b).

Resnik et al. (2015) explored the use of supervised topic models in the analysis of detecting depression from Twitter. They use 3 million tweets from about 2,000 twitter users, of whom roughly 600 self-identify as having been diagnosed with depression. This work provided a more sophisticated model for text-based feature development for detecting depression, yielding promising results using supervised Latent Dirichlet Allocation (LDA). LDA uncovers underlying structure in a collection of documents by treating each document as if it were generated as a mixture of different topics. Qualitative examples confirmed that LDA models can uncover meaningful and potentially useful latent structure for the automatic identification of important topics for depression detection.

With the rise of social media, posts on sites such as Twitter and Facebook provide an interesting domain to investigate depression. Not only do these domains provide rich text data but also social metadata which captures important social behaviors and characteristics, like number of friends/followers, number of likes, retweets, etc. De Choudhury et al. (2014) studied Facebook data shared voluntarily by 165 new mothers. Their work aimed to detect and predict onset of postpartum depression (PPD). They considered multiple behavioral features including activity (frequency of status updates, media items, and wall posts), social capital (likes and comments on status updates or media), emotional expression and linguistic style measured through LIWC. They found that experiences of PPD were best predicted by increased isolation, which was modeled by reduced social activity and interaction on Facebook and decreased access to social capital.

Wang et al. (2013) constructed a model to detect depression from online blog posts. The features they extracted included first person singular and plural pronouns, polarity of each sentence using their polarity calculation algorithm, ratio of first person singular pronouns to first person plural pronouns, use of emoticons, user interactions with others (@username mentions), and number of posts. Using 180 users, the features given above, and three different kinds of classifiers Wang et al. (2013) report a a precision of 80% when classifying between depressed versus non-depressed users.

4.4 Multimodal Indicators

Researchers have also investigated multimodal indicators for depression detection. Scherer et al. (2013a), investigated visual signals and voice quality in a multimodal system, finding that they were able to distinguish interviewees with depression from those without depression with an accuracy of 75%.

Morales and Levitan (2016b) provided a comparative investigation of speech versus text-based features for depression detection systems, finding that a multimodal system leads to the best performing system. In addition, Morales and Levitan investigated using an automatic speech recognition system (ASR) to automatically transcribe speech and found that text-based features generated from ASR transcripts were useful for depression detection.

Fraser et al. (2016) extracted a large number of textual features and acoustic features. Textual features included POS tags, parse tree constituents, psycholinguistic measures, measures of complexity, vocabulary richness, and informativeness. Acoustic features include fluency measures, MFCCs, voice quality features, and measures of periodicity and symmetry. Using these multimodal features, Fraser et al. were able to detect depression with 65.8% accuracy. Related work on suicide risk assessment found that multimodal indicators were able to discriminate between suicidal and non-suicidal patients (Venek et al., 2016).

5 Evaluation

Depression detection can be divided into three different prediction tasks: presence (depressed vs. not depressed), severity (normal, mild, moderate, severe, and very severe), and score level prediction. With each task comes a set of evaluation metrics. In regards to the first two groups, performance is usually reported in terms of classification accuracy (Acc.). Given that accuracy is heavily affected by skewness in datasets, often times sensitivity (Sens.), specificity (Spec.), precision (Prec.), and F1-score (harmonic mean of precision and recall) are also reported. For score level prediction, performance is usually reported as a measure of differences between values predicted and the values actually observed, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). In Table 2 we report, to our knowledge, the best performing depression detection systems from 2016.

As Table 2 highlights, it is very difficult to make systematic comparisons across studies. Data, task, label, and experimental set-up tend to vary across study. Therefore, it is hard to understand which approach is most promising. However, in regards to features, it tends to be the case that combining features from multiple modalities leads to improvements (Morales and Levitan, 2016a; Scherer et al., 2013a; Fraser et al., 2016; Williamson et al., 2016; Valstar et al., 2016). In many cases, researchers may only have access to certain labels. However, when data sources do contain score labels reporting both error for regression as well as classification performance metrics will help facilitate comparisons across systems. Given that each feature or subset of features are meant to measure specific depression indicators or symptoms, it is also extremely important to understand how well each feature is performing. Therefore, it is best to always include correlation experiments, such as Pearson correlation tests, in order to make it transparent which features are important.

5.1 Confounding Factors

Specific variability factors have been shown to be strong confounding factors for depression detection systems (Cummins et al., 2015a, 2014, 2013, 2011; Sturim et al., 2011). Variability factors include traits like gender, age, emotion, or personality of the speaker. Therefore, it is important to keep these factors in mind when building a detection system. For example, in many studies systems have achieved better results using sex-dependent classifiers (Moore et al., 2008; Low et al., 2011; Yang et al., 2016; Scherer et al., 2014). Oth-

Reference	Task	Features	MAE	Acc.	Spec.	Sens.	Prec.	F1
Fraser et al. (2016)	Binary	MFCCs/lexical/syntax		0.66	0.61	0.71		
Milne et al. (2016)	4 classes	N-grams		0.78				
Kim et al. (2016)	4 classes	TF-IDF n-gram/post embedding		0.85				
Malmasi et al. (2016)	4 classes	Lexical/syntax/metadata		0.83				
Brew (2016)	4 classes	TF-IDF unigrams/metadata		0.79				
Valstar et al. (2016)	Binary	Visual				0.78	0.47	0.58
		Acoustic				0.89	0.27	0.41
		All				0.78	0.47	0.58
	PHQ-8	Visual	6.12					
		Acoustic	5.72					
		All	5.66					
Williamson et al. (2016)	PHQ-8	Visual	5.33					0.53
		Acoustic	5.32					0.57
		Semantic	3.34					0.84
		All	4.18					0.81
Yang et al. (2016)	PHQ-8	Visual/acoustic	6.70			0.67	0.50	0.57

Table 2: Best performing depression detection systems. F1 score, precision, and sensitivity are reported for the *depressed* class.

ers (Morales and Levitan, 2016a) have used unsupervised clustering prior to depression detection, finding that this approach could tease out participant differences and in turn lead to performance improvements. However, these approaches to dealing with variability factors usually mean a reduction in training data, which at times can be a substantial trade-off.

Another factor to consider, is comorbidity. Comorbidity refers to the simultaneous presence of two chronic diseases or conditions. For example, Alzheimer's disease (AD) and depression frequently co-occur. Fraser et al. (2016) found that their depression detection system performed considerably lower on patients with comorbid depression and AD than on those patients with only depression. Therefore, comorbidity can lead to a more difficult task given the wide overlap of symptoms in the two conditions. Factors such as gender, age, and comorbidity, can have substantial effects on system performance. In order to better understand performance across studies and the effect of variability factors more transparency is necessary, in regards to dataset details and descriptions. In addition, researchers should begin to consider more diverse populations in their studies. Thus far, most research and data collection efforts have focused on detecting depression from young and otherwise healthy participants. In order to generalize detection systems, datasets representing other populations need to be considered.

6 Discussion

As with any technology or tool there is always risk of misuse and therefore it is important to discuss general ethical considerations with pursuing this line of research. It is especially important to define and outline appropriate use of these systems. Mental health professionals should view language technology for depression detection as a mechanism to complement current diagnoses by giving them access to a novel and rich non-intrusive data source. It is understandable that mental health professionals as well as the general population may be uncomfortable with the possibility that technologies might have to predict psychological states, especially when relatively accurate predictions can be made. To be clear, these systems are not proposed as standalone diagnostic tools that could replace current approaches to diagnosing mental health issues, but instead proposed as part of a broader awareness, detection, and support system. These technologies provide numerous advantages, including large-scale and remote assessment, which in turn could help a broader population. These methods could also provide a lower cost complement to traditional depression assessments. In addition, these tools could help health professionals manage current patients more efficiently, allowing clinicians to monitor their patients continuously. Determining how machines should augment and assist in diagnosis is a complicated issue. However, there exists evidence that mechanical prediction (statistical, algorithmic, etc.) is typically as accurate or more accurate than clinical prediction (Grove et al., 2000). Moreover, mechanical predictions do not require an expert judgment and are completely reproducible. Although there are general ethical considerations, it is important to highlight the potential of mental health assessment tools to enhance the quality of life for society.

7 Conclusion

In this paper, we present a review of the latest work on depression detection systems. We provide a cross-modal review of indicators for depression detection systems, covering visual, acoustic, linguistic, and social features. We also outline approaches to defining and annotating depression, existing data sources, and how to evaluate depression detection systems. This paper serves as a bridge between the subfields by providing the first review across subfields and modalities. Given that depression detection is inherently a multimodal problem, this paper is an important contribution to the research community as it serves as a great resource for understanding multimodal features as well as what factors to consider when designing a depression detection system. Lastly, in order for the research community to progress together researchers should begin to follow the best practices (Stodden and Miguez, 2013). Best practices lead to communication standards, which will help disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of computational research. Without the data and code that underlie scientific discoveries, is is all but impossible to verify published findings. We urge researchers to focus on reproducible research, through the dissemination, availability, and accessibility of data and code.

References

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. 2013. A comparative study of different classifiers for detecting depression from spontaneous speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, pages 8022–8026.

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patients speech. *Journal of Affective Disorders* 66(1):59–69. https://doi.org/10.1016/S0165-0327(00)00335-9.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *arXiv preprint arXiv:1605.04462* .

APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Aaron T Beck. 1967. *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press.

Aaron T Beck, C Ward, M Mendelson, et al. 1961. Beck depression inventory (bdi). *Arch Gen Psychiatry* 4(6):561–571.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6):585–594.

Chris Brew. 2016. Classifying reachout posts with a radial basis function svm. *red* 14(23):27.

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using nonclinical texts. *Natural Language Engineering* page 137. https://doi.org/10.1017/S1351324916000383.

Michael Cannizzaro, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter J Snyder. 2004. Voice acoustical measurement of the severity of major depression. *Brain and cognition* 56(1):30–35.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre. 2009. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, pages 1–7.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 31–39.

Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Interspeech*. pages 2997–3000.

Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, and Jarek Krajewski. 2014. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 970–974.

Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. 2013. Diagnosis of depression by behavioural signals: a multimodal approach. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*. ACM, pages 11–20.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015a. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71:10–49.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, Sebastian Schnieder, and Jarek Krajewski. 2015b. Analysis of acoustic space variability in speech affected by depression. *Speech Communication* 75:27–49.

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*. ACM, pages 626–638.

Emil Durkheim. 1951. Suicide (g. simpson, trans.).

Paul Ekman, Wallace V Friesen, and Joseph C Hager. 1978. Facial action coding system (facs). *A technique for the measurement of facial action. Consulting, Palo Alto* 22.

Kathleen C Fraser, Frank Rudzicz, and Graeme Hirst. 2016. Detecting late-life depression in alzheimers disease through analysis of speech and language. *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology* .

Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing* 32(10):641–647.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*. pages 3123–3128.

William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12(1):19.

Max Hamilton. 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry* 23(1):56.

Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. 2014. Automatic modelling of depressed speech: relevant features and relevance of gender. In *INTERSPEECH*. pages 1248–1252.

Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can body expressions contribute to automatic depression analysis? In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, pages 1–7.

Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cécile Paris. 2016. Data61-csiro systems at the clpsych 2016 shared task. In *CLPsych@HLT-NAACL*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9. *Journal of general internal medicine* 16(9):606–613.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders* 114(1):163–173.

Henry A. Lieberman and Albert R. Meyer. 2014. Visualizations for mental health topic models. In *Massachusetts Institute of Technology Master's Thesis*.

Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. 2011. Detection of clinical depression in adolescents speech during family interactions. *IEEE Transactions on Biomedical Engineering* 58(3):574–586.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. *order* 2:8.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. pages 118–127.

Stuart A Montgomery and MARIE Asberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry* 134(4):382–389.

Elliot Moore, Mark Clements, John W Peifer, Lydia Weisser, et al. 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *Biomedical Engineering, IEEE Transactions on* 55(1):96–107.

Michelle Renee Morales and Rivka Levitan. 2016a. Mitigating confounding factors in depression detection using an unsupervised clustering approach. In *Computing and Mental Health Workshop*.

Michelle Renee Morales and Rivka Levitan. 2016b. Speech vs. text: A comparative analysis of features for depression detection systems. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, pages 136–143.

Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. 2011. Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depression and anxiety* 28 6:447–55.

James C. Mundt, Adam P. Vogel, Douglas E. Feltner, and William R. Lenderking. 2012. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry* 72(7):580–587. https://doi.org/10.1016/j.biopsych.2012.03.015.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing* 5(3):217–226.

James W Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science* 8(3):162–166.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net* .

Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one* 9(1):e85733.

Sameer Pradhan, Noemie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. volume 199, pages 54–62.

Tom Pyszczynski and Jeff Greenberg. 1987. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin* 102(1):122.

Andrew G. Reece and Christopher M. Danforth. 2016. Instagram photos reveal predictive markers of depression. *CoRR* abs/1608.03282.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. *NAACL HLT 2015* page 99.

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

A John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, et al. 2003. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry* 54(5):573–583.

Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013a. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*. pages 847–851.

Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 32(10):648–658.

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. 2013b. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, pages 1–8.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. 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.

H. H Stassen, S Kuny, and D Hell. 1998. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology* 8(4):303–310. https://doi.org/10.1016/S0924-977X(97)00090-4.

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

Victoria Stodden and Sheila Miguez. 2013. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *SSRN* .

Douglas E Sturim, Pedro A Torres-Carrasquillo, Thomas F Quatieri, Nicolas Malyska, and Alan McCree. 2011. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Interspeech*. pages 2981–2984.

Andrea Carolina Trevino, Thomas Francis Quatieri, and Nicolas Malyska. 2011. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing* 2011(1):1–18.

Michel Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *4th Audio/Visual Emotion Challenge Proc.*. ACM, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*. ACM, pages 3–10.

V. Venek, S. Scherer, L.-P. Morency, A. Rizzo, and J. P. Pestian. 2016. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing* https://doi.org/10.1109/TAFFC.2016.2518665.

Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining*, Springer, pages 201–213.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 11–18.

James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. 2014a. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *4th Audio/Visual Emotion Challenge Proc.*. ACM, pages 65–72.

Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 89–96.

Jörg Zinken, Katarzyna Zinken, J Clare Wilson, Lisa Butler, and Timothy Skinner. 2010. Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry research* 179(2):181–186.