Natural Language Engineering **1** (1): 1–??. Printed in the United Kingdom

Natural Language Processing in Mental Health applications using non-clinical texts

RAFAEL A. CALVO[1] DAVID N. MILNE[1]
SAZZAD M. HUSSAIN[1,2] and HELEN CHRISTENSEN[3]

[1] *School of Electrical and Information Engineering, The University of Sydney, Australia*
email: Rafael.calvo@sydney.edu.au

[2] *Commonwealth Scientific and Industrial Research Organisation, CSIRO, Australia*

[3] *Black Dog Institute, University of New South Wales, Australia*

(*Received 20 March 1995; revised 30 September 1998*)

Abstract

Natural Language Processing (NLP) techniques can be used to make inferences about peoples' mental states from what they write on Facebook, Twitter and other social media. These inferences can then be used to create online pathways to direct people to health information and assistance and also to generate personalized interventions. Regrettably, the computational methods used to collect, process and utilise online writing data, as well as the evaluations of these techniques, are still dispersed in the literature. This paper provides a taxonomy of data sources and techniques that have been used for mental health support and intervention. Specifically, we review how social media and other data sources have been used to detect emotions and identify people who may be in need of psychological assistance; the computational techniques used in labeling and diagnosis; and finally, we discuss ways to generate and personalize mental health interventions. The overarching aim of this scoping review is to highlight areas of research where NLP has been applied in the mental health literature and to help develop a common language that draws together the fields of mental health, human-computer interaction and natural language processing.

1 Introduction

People write to communicate with others. In addition to describing simple factual information, people also use writing to express their activities, and convey their feelings, mental states, hopes and desires. Recipients then use this written information from emails and other forms of social media texts to make inferences, such as what someone else is feeling, which in turn influences interpersonal communication. Even when writing is shaped by the way someone *wants* to be perceived, this text still provides important cues to help friends and family recognise important life events for them to respond to with support and encouragement.

When people write digitally (e.g. on email or social media), their texts are processed automatically. Natural language processing (NLP) techniques make inferences about what people say and feel, and these inferences can trigger messages

or other actions. One of the most common uses of NLP is in the marketing sector where companies analyse emails and social media to generate targeted advertising and other forms of '*interventions*' (generally aimed at changing our behaviour towards buying something or following a link).

However, potential applications of NLP techniques extend beyond marketing. For example, NLP techniques have been identified as an important area of growth within the artificial intelligence in medicine community (Peek, Combi, Marin, and Bellazzi, 2015). In this paper we discuss the potential for NLP techniques to be utilised in the mental health sector. Mental health applications are designed to support mental health and wellbeing in an online environment. These applications are reliant on interdisciplinary collaboration between researchers and practitioners from areas such as computational linguistics, human-computer interaction and mental health (and mental health service delivery). Interdisciplinary collaborations benefit from the development of a common language and body of research evidence (Calvo, Dinakar, Picard, and Maes, 2016). Regrettably, there is a paucity of organised literature at this intersection between NLP, Human-Computer Interaction and mental health research. This paper aims to help researchers in these areas envision and work towards new mental health applications. The paper is structured as follows:

- Section 2 provides a brief explanation of the methodology used to identify relevant literature, and the inclusion and exclusion procedures applied to this scoping study.
- Section 3 describes the types of textual data that has been the focus of research to date. In this section the different data sources (ranging from lengthy diaries to brief tweets), and existing datasets that have been gathered and annotated (with moods, suicidal intent, etc.) using NLP are discussed.
- Section 4 describes the techniques that researchers have applied to make inferences or generate diagnoses about the emotional state or mental health of the authors of these texts. This work draws widely upon text classification and NLP.
- Section 5 identifies the different types of automated interventions for supporting mental health. These range from simple canned (i.e. the same for all users) interventions to personalised interventions. Where possible, we describe the specific studies in order to evaluate the reported efficacy of these interventions.

We conclude this section with a discussion of the gaps in the literature and future opportunities in this research domain.

2 Methods and Overview

Systematic reviews such as those published by the Cochrane Collaboration require an understanding of the existing literature and its gaps and involves a process that investigates a research question and develops and applies a framework for exploring a question utilising existing literature (Armstrong, Hall, Doyle, and Waters, 2011).

When an area is complex, very broad, or has not been reviewed before, a scoping review that maps the key concepts maybe more appropriate. Here, we aim to answer the following question: *What Natural Language Technologies have been used with user generated data in the area of mental health?* The multidisciplinary nature of this question, and the differences in terminology used across disciplines, means a systematic review methodology is not practical.

In the period between August 2014 and May 2015, the authors performed multiple searches using Google Scholar. Relevant research published after this period was added during the reviewing process. Google Scholar was chosen because it indexes journals, conferences and patent documents. Conference papers and patents are particularly important sources because a significant amount of the research in NLP and computer science is not published in journals, which is the standard in the mental health domain.

As is common in scoping reviews and when the review covers several research fields a definite set of search terms was not used. We performed multiple iterations of widening search terms and then identified and selected important and recurrent themes. In the first iteration approximately 50 papers were identified. From this literature, three stages of research reporting emerged: *Data*, *Labeling* and *Intervention*. Inclusion criteria for each of these stages involved:

- In *Data*, we focus exclusively on textual data (as opposed to physiological signals, activity, etc.) that has been analyzed for mental health applications. We also focus on texts that have been written by users (e.g. mostly consumers, occasionally patients) rather than doctors or researchers. There is a significant amount of research focused on using NLP to process clinical notes, medical records, and academic research papers, but they do not contribute to the focus of this review (user generated texts) - for a review see (Abbe, Grouin, Zweigenbaum, and Falissard, 2015). Initial search terms included the name of data sources often used in NLP projects (Twitter, Facebook, etc) AND Mental Health AND Natural Language Processing (written as facebook + "mental health" + "natural language processing"). Google Scholar generally sorts papers by the number of citations so papers with few or no citations might have been missed. Links were followed and if found appropriate added to a database. We run searches for (Facebook OR Twitter) AND (Mental Health) AND (Natural Language Processing) in PubMed, but did not get any results. The search for Mental Health AND Natural Language Processing gave 35 results. The references in highly relevant papers were scanned for title and authors that seemed relevant. This process was followed for each of the three areas (i.e. stages) by one of the authors. In later iterations all authors were involved.
- In *Automated Labeling*, we focus on applications of NLP to the textual data collected in the previous section. We searched for papers that used the different components of a classification system: feature extraction, feature selection and classification and mental health and NLP. There is an extensive machine learning literature on document classification techniques that has been re-

viewed previously (Kotsiantis, 2007; Sebastiani, 2002). For the purposes of this review, only those studies that included a mental health application were considered within the scope of the present discussion.

- In *Intervention*, we considered studies that showed ways in which NLP could be used in computer interventions to support mental health. Hypothetical uses of NLP in psycho-education are discussed but we excluded those psycho-education interventions where texts or multimedia are presented to clients without personalization or where they did not utilise NLP. Further, while we have focused on mental health interventions, because of the scarcity of literature in this area, we extended the scope of the literature search to include publications from other health areas that highlight novel uses of NLP that could also be applied as a mental health intervention.

Some resources, such as conference proceedings and literature reviews were useful to bootstrap our literature search. These include the First Workshop on Computational Linguistics and Clinical Psychology (Resnik, Resnik, and Mitchell, 2014) held within the annual meeting of the American Association on Computational Linguistics, the leading NLP conference. The papers published in the proceedings from this conference highlight the breath of current topics being considered: depression, Alzheimer's, autism, violence and aphasia. For the purposes of this scoping paper, we have taken a narrower, and standard definition of mental illness: disorders that affect cognition, mood and behaviours, including depression, anxiety disorders, eating disorders and addictive behaviours.

The area of affective computing and its literature on emotion detection from texts was also useful, particularly research (e.g. Calvo and D'Mello 2010; Strapparava and Mihalcea 2014) discussed in Section 4. These literature reviews focus on the computational aspects and do not generally consider the differences between data sources or the interventions. Further, although we found several reviews of eHealth interventions, we have limited this discussion to include only those with NLP features, such as the one by Barak and Grohol (2011) in Section 5.

2.1 Overview

Table 1 provides an overview of papers that have mined textual data for insights into the emotional state and mental health of the writer. The first column describes the objective of each investigation, and is split by whether they relate to individual *texts* (e.g. a blog post or tweet), to individual *authors* (by aggregating the their posts and any other profile information), or overall *populations* (i.e. to measure broad trends over large communities of authors). The table also describes the source of data used (typically social media, but there are exceptions). Most of the features considered by researchers and algorithms are based on linguistic analysis of text, but this is often supplemented by behavioural data (e.g. login times), social data (e.g. friends and followers) and demographics (e.g. age, gender, location). There are many different approaches for obtaining ground truth data to supervise and evaluate; it can be obtained directly from text authors by explicitly asking them

objective	references	data source	features	gold standard	section
Text level					
to detect specific emotions	Strapparava and Mihalcea (2008)	LiveJournal	linguistic	directly self-reported	3.3
	Pestian et al. (2012) and task participants	Suicide notes	linguistic	manual annotation	3.5
to detect mental health topics	Nguyen et al. (2014)	LiveJournal	linguistic	indirectly self-reported	3.3
to detect distress/suicide ideation	Homan et al. (2014)	Twitter	linguistic	manual annotation	3.1
	O'Dea et al. (2015)	Twitter	linguistic	manual annotation	3.1
to measure stigma	Li et al. (2015)	Weibo	none (observational)	manual annotation	3.4
to triage concerning content	Milne et al. (2016) and task participants	ReachOut	linguistic, behavioural	manual annotation	3.6
Author level					
to measure mood valence	Sadilek et al. (2013)	Twitter	linguistic, behavioural, social	none	3.1
to measure emotion contagion	Coviello et al. (2014), Kramer et al. (2014)	Facebook	linguistic	none	3.2
to detect depression	De Choudhury et al. (2013)	Twitter	linguistic, behavioural, social	directly self-reported	3.1
	Coppersmith et al. (2015) and participants	Twitter	linguistic	directly self-reported	3.1
to predict post-partum depression	De Choudhury et al. (2013a)	Twitter	linguistic, behavioural, social	indirectly self-reported	3.1
	De Choudhury and Counts (2014)	Facebook	linguistic, behavioural, social	directly self-reported	3.2
to detect distress/suicide ideation	Homan et al. (2014a)	TrevorSpace	social	directly self-reported	3.4
	Masuda et al. (2013)	Mixi	demographic, behavioural, social	indirectly self-reported	3.4
Population level					
to measure mood valence	Golder and Macy (2011)	Twitter	linguistic	none	3.1
	Kramer (2010)	Facebook	linguistic	external statistics	3.2
	Dodds et al. (2011)	Twitter	linguistic	none	3.1
	Mitchell et al. (2013)	Twitter	linguistic	external statistics	3.1
	Schwartz et al. (2013)	Twitter	linguistic	external statistics	3.1
to detect specific emotions	Larsen et al. (2015)	Twitter	linguistic	external statistics	3.1
to measure depression	De Choudhury et al. (2013b)	Twitter	linguistic, behavioural, social	external statistics	3.1

Table 1. Overview of papers that mine text for insight into author's moods and mental health.

to self-diagnose (e.g. by completing a survey), or indirectly from their behaviour (e.g. by joining a depression support group). Some researchers rely on arduous manual annotation, while others don't use ground truth at all, and instead perform observational studies.

At the *text* level, Pestian, Matykiewicz, Linn-Gust, South, Uzuner, Wiebe, Cohen, Hurdle, and Brew (2012) host a shared task for mining emotions from suicide notes. Nguyen, Phung, Dao, Venkatesh, and Berk (2014) separate out LiveJournal posts that discuss depression and related topics. Homan, Johar, Liu, Lytle, Silenzio, and Alm (2014) and O'Dea, Wan, Batterham, Calear, Paris, and Christensen (2015) detect posts containing suicide ideation and distress, and Li, Huang, Hao, ODea, Christensen, and Zhu (2015) investigate unhelpful, stigmatizing reactions to suicide on the Chinese social media platform Weibo. Milne, Pink, Hachey, and Calvo (2016) host a shared task for identifying and prioritizing concerning content on ReachOut.com's peer support forum.

At the *author* level, Sadilek, Homan, Lasecki, Silenzio, and Kautz (2013) measure temporal changes in mood valence of Twitter users, while Coviello, Sohn, Kramer, Marlow, Franceschetti, Christakis, and Fowler (2014) and Kramer, Guillory, and Hancock (2014) investigate how moods spread across social connections in Facebook. The various works of De Choudhury et al., and the participants of the shared task hosted by Coppersmith, Dredze, Harman, Hollingshead, and Mitchell (2015) all attempt to make clinical diagnoses (for depression, post-traumatic stress and postpartum depression) from social media data. Homan, Lu, Tuurmcrochesteredu, Lytle, Lytle, Rochester, Silenzio, and Silenzio (2014b) and Masuda, Kurahashi, and Onari (2013) aim to identify people who are in current distress or are contemplating suicide.

Most of the *population*-level studies use rough sentiment analysis to measure the mood valence (i.e. positive or negative affect) of Twitter and Facebook, and analyse text features almost exclusively. Golder and Macy (2011) and Dodds, Harris, Kloumann, Bliss, and Danforth (2011) are purely observational, but the others check that geographic and temporal variations correlate well against external statistics. For example, Schwartz, Eichstaedt, Margaret L. Kern, Dziurzynski, Park, Lakshmikanth, Jha, Seligman, and Ungar (2013) and Larsen, Boonstra, Batterham, ODea, Paris, and Christensen (2015) compare against surveys of life satisfaction, De Choudhury, Counts, and Horvitz (2013b) use depression incidence and prescription rates for antidepressants, and Mitchell, Frank, Harris, Dodds, and Danforth (2013) compare against gun violence, wealth, obesity, and several other indexes.

Table 1 also lists the gold standard for each study. The gold standard is the data set used to compare against. In some cases it is *directly self-reported* by the text author (e.g. via a mood diary) or *indirectly self-reported* via their behavior (e.g. joining support group for anxiety). In other cases the researchers must perform *manual annotation*, or cross-reference the texts with *external statistics* such as prescription rates of anti-depressants.

3 Data

This section describes the data sources that researchers have used for NLP, focusing on texts from which one might infer the author's mood, or diagnose a mental health issue. Considering that Tweets, blog and social media posts differ in many quantitative facts: length of texts, vocabulary used and other features that must be considered for processing, this section aims to distinguish the types of media used. Further, different data sources also differ in their reason people write them.

We start with research on datasets not explicitly focused on mental health; studies that look for emotions and for subtle cues of mental health in general day-to-day sources such as Twitter (Section 3.1), Facebook (Section 3.2), blogs (Section 3.3), and other social media (Section 3.4).

In Section 3.5 we discuss a series of studies that used suicide notes as the data source and describe how they were collected and used to train algorithms that, for example, detect real from fake ones. Other studies, such as those described in Section 3.6 have used texts written by people with mental illness seeking help and recovery, for example, personal diaries and social sharing of their problems. The benefits of writing for mental health has been recognised in the literature Chung and Pennebaker (2007); Pennebaker and Chung (2007); Pennebaker et al. (1988), and many people struggling with depression or anxiety write about their thoughts and experiences. Some choose to do so in online support groups and forums (described Section 3.6), making them available for researchers to analyse.

3.1 Twitter

Twitter—one of the most popular social networking or 'micro-blogging' sites—distinguishes itself as a source of textual data by accessibility and sheer volume. Almost all activity on Twitter is public by default, and its users simply broadcast their messages to whoever wants to listen. This openness yields a huge source of data—some 90 million tweets per day—that researchers have been quick to capitalize on.

Each tweet is a short 140 character message posted by an individual. Twitter's APIs allow for real-time monitoring, so it is possible to track patterns over time with very low latency. Only a fraction of tweets (<1%) contain geo-location data, but it is possible to (roughly) locate tweets using information from the poster's profile. The patterns of following, replying to and otherwise engaging with accounts forms a social network of sorts that can be mined in addition to Twitter's textual, temporal and geographical features.

Most studies were limited to Twitter's free APIs, which deliver either a 1% random sample (known as the *garden hose*) or require specific vocabulary of terms and accounts to be monitored. Twitter's terms and conditions prohibit redistribution, so there are no shared datasets and very little replication or direct comparison between the studies described below. While it is possible for researchers to share Twitter data without violating these conditions, there remains the question of whether it

would be ethical to do so because of difficulties in anonymizing data Horvitz and Mulligan (2015).

Many studies of Twitter, such as Signorini et al. (2011) and Paul and Dredze (2011) focus on physical health and disease, and are thus beyond the scope of the current discussion. Also out of scope are the many other studies, which investigate feelings and emotions as they pertain to a specific subject (e.g. a product or a politician). This section focuses on investigations that address the mental health and emotional wellbeing of the people who tweet.

One might expect that tweets would provide limited insight given their brevity and public character. It is perhaps surprising then, that many researchers have successfully utilised Twitter data as a source of insights into the epidemiology of emotions (e.g how they propagate) and mental illness. For example, Golder and Macy (2011) used 509 million publicly available posts written by 2.4M users from across the globe over a period of about 2 years. They showed that moods are more positive in the morning and decay during the day. They also found that people are happier on weekends and that seasonal moods change with day-length. These results hold few surpriseswe all know how much we love our weekends and crave sunlightbut they do start to validate Twitter as a reliable signal of affective functioning.

Schwartz, Eichstaedt, Margaret L. Kern, Dziurzynski, Park, Lakshmikanth, Jha, Seligman, and Ungar (2013) also validated the use of Twitter data in characterising geographic variations in wellbeing, compared to traditional phone surveys about life satisfaction (Lawless and Lucas, 2011). In this experiment, about a billion tweets were gathered and, where possible, mapped onto counties in the United States. The paper does not describe exactly how many tweets were successfully geo-located, but the resulting dataset included 1,300 counties for which they found at least 30 twitter users who had each posted at least 1,000 words. The authors then compared the Twitter data against phone interviews and demographics data for the area (socioeconomic surveys and Census). Topic models created with the Twitter data improved the accuracy of life satisfaction predictions based on the demographic controls (county-by-county scores for age, sex, ethnicity, income, education) which were in turn more predictive than the prevalence of words from emotion lexicons. In combination, the three approaches achieved a Pearson correlation of 0.54 with the scores obtained through random direct surveys of people in a county, while the controls on their own achieve only 0.44 correlation. *We Feel* (Larsen, Boonstra, Batterham, ODea, Paris, and Christensen, 2015) is a system that analyses global and regional changes in emotional expression. The evaluation consisted of 2.73 109 emotional tweets collected over 12-weeks, and automatically annotated for emotion (using LIWC), geographic location, and gender. The analysis confirmed regularities found in emotional expressions in diurnal and weekly cycles and these were reflected in the results for a PCA: the first component explained 87% of the variance and provided clear opposite loadings between positive and negative emotions. This study adds to the evidence that Twitter and possibly other publicly available social media data can provide insights into emotional wellbeing and illness across different populations.

De Choudhury, Counts, and Horvitz (2013b) built a classifier for determining

whether a Twitter post indicates depression. For the purposes of the current discussion, it is sufficient to say that the classifier produced a score that is higher for tweets that indicate depression than those that did not. The paper also describes a Social Media Depression Index (SMDI), which scores a group of tweets (e.g. those by a single author, or those from a city) by the prevalence of high scoring tweets. One experiment in this paper is similar to the Schwartz et al. (2013) study described above, but in this case the tweets were geo-located down to state level rather than county. The SMDI score was calculated for each state, and these scores achieved a 0.51 correlation against ground truth (defined as the data used as the training set) based on the Centre for Disease Control (CDC) calculations. It is important to note that this score is obtained entirely through analysis of Twitter; unlike (Schwartz et al., 2013) it does not make any use of background demographic data. In the same paper the authors also calculated the SMDI scores for the 20 unhappiest US cities and achieved a strong positive correlation (0.64) with the prescription rates for common antidepressants.

Many other studies follow a similar pattern of quantifying topics and emotions over entire populations using Twitter. For example, Dodds, Harris, Kloumann, Bliss, and Danforth (2011) introduced the Hedonometer, which cross-referenced a truly massive dataset of 4.6 billion tweets against their Language assessment by Mechanical Turk (LabMT) vocabulary of happy and sad terms. Like Golder and Macy (2011), their analysis of the resulting signal reveals unsurprising but reassuring patterns, such as increased happiness during the weekend. Mitchell, Frank, Harris, Dodds, and Danforth (2013) applied the Hedonometer using only tweets with exact GPS coordinates. At a state level, they found moderate correlations with incidences of gun violence (r=-0.66), the American health index (r=0.58), and the US Peace Index (r=0.52). At a city level, they also found associations amongst Hedonometer levels and indexes of wealth and obesity.

However, the studies described above consider only the broad picture. Analysis at this level may be more forgiving, because mistakes can be averaged out as tweets are aggregated across entire states, counties and cities. For researchers, this poses the question: are Twitter-derived measurements accurate enough to drill down to an individual?

One group from Microsoft Research has demonstrated two distinct approaches for conducting studies with individual Twitter users. In the first experiment, De Choudhury, Gamon, Counts, and Horvitz (2013) developed a classifier that estimates the risk of a Major Depressive Disorder (MDD) before it happens. To gather a gold-standard dataset they conducted a crowdsourcing task that required 1,583 volunteers to complete diagnostic surveys—specifically the Center for Epidemiologic Studies Depression Scale (CES-D) and the Beck Depression Inventory (BDI)—and answer questions about their basic demographics and history with depression. The same task invited (but did not require) participants to provide details of their twitter accounts. The task yielded two datasets; 305 twitter accounts with no indication of depression, and 171 accounts that scored highly on the CES-D survey and had also been diagnosed with depression at least twice in the last year (a requirement to qualify for MDD). A total of 2.1M tweets (an average of 4,533 tweets

for each account) were captured within the year prior to the crowdsourcing task. From this data they identified behavioural features about engagement, emotion, language styles and medication used. These features were used to train the classifier, and distinguish between the depressed and non-depressed accounts. To summarize their findings, the depressed user accounts were:

- less likely to tweet or respond to others' tweets
- more likely to tweet late at night
- more likely to use first-person pronouns (i.e. tweet about themselves)
- less likely to use third-person pronouns (i.e. tweet about others)
- less likely to follow others or gather followers

In addition to these social and behavioural differences, the study also identified a lexicon of terms that were more common among the depressed accounts, such as mentions of medications, depressive symptoms and words related to disclosure. All of these signals were combined with basic demographic features (age, gender, income and education level) to develop a classifier that could distinguish between depressed and non-depressed accounts with a recall of 63% (the fraction of accounts that are automatically detected) and precision of 74% (the fraction that is labelled correctly).

In another study the same authors focused on measuring behavioural and emotional changes that could be indicative of postpartum depression (De Choudhury, Counts, and Horvitz, 2013a). Rather than surveying new mothers directly, they instead monitored Twitter automatically for birth announcements, using newspapers as a model. This identified 376 mothers after manual verification via crowdsourcing. The tweets of each mother were gathered for the three months prior to the birth notice and another three months after. In all, the dataset contained 37k pre-partum and 40k postpartum tweets. These were analysed as described above (De Choudhury et al., 2013). Each mother was characterized with 33 features, and the ground truth (the training set) was based on a heuristic threshold of changes, rather than a direct questionnaire of depression. Instead the analysis identified mothers for whom twitter-derived measures of social engagement, linguistic style, etc.—the same signals that indicated depression in the previous study—changed dramatically between the pre- and postpartum periods. It was assumed that these were mothers who were not adjusting well to parenthood, and were displaying signs of post-partum depression. A classifier was built that couldfrom the pre-partum data alonepredict the mothers who would later exhibit these dramatic changes with an accuracy of 71%.

Coppersmith, Dredze, Harman, Hollingshead, and Mitchell (2015) host a shared task that also attempts to make clinical diagnoses from Twitter data; in this case for depression and post-traumatic stress disorder. Their work, and that of the task participants, is described in Section 4.4.

Twitter posts can be combined with geolocation information to produce approximate models of people's emotion and context. For example, Sadilek, Homan, Lasecki, Silenzio, and Kautz (2013) used information from 6,237 users who shared their GPS location. They used linguistic features (i.e. based on the Linguistic Inquiry and Word Count LIWC tool), behavioural (such as the number and time of

tweets), social network (i.e. friends) and a summary quantity calculated from LIWC that categorised users into three states: positive, neutral and negative. Using this data they found evidence of temporal mood patterns, demonstrated emotional contagion in groups and were able to predict when users were going to be in one of those three states (positive, neutral, negative) over the next 10 days. Another study using the same dataset (Homan, Johar, Liu, Lytle, Silenzio, and Alm, 2014) selected 2,000 tweets around suicide risk factors (1,370 from the LIWC sad dimension and 630 with suicide specific terms) and annotated them. The annotations were performed by a novice and counselling psychologists and then judged by another novice. Each of the tweets was annotated as *happy*, *no distress*, *low distress* and *high distress*. The authors then trained a Support Vector Machine (SVM) to classify the tweets into the 4 categories and found F1-measures of 0.4-0.6 (the harmonic mean of precision and recall).

O'Dea, Wan, Batterham, Calear, Paris, and Christensen (2015) aim to separate out genuinely concerning tweets mentioning suicidal ideation from the large amount of flippancy and hyperbole that exists on the platform. They gathered a corpus of 2000 tweets containing phrases like *kill myself* and *tired of living* and manually coded them as *strongly concerning* (14%), *possibly concerning* (56%) and *safe to ignore* (29%). Agreement between annotators ranged between 0.47 and 0.64 kappa, indicating that the task was strongly subjective. An automated SVM classifier was able to automatically separate strongly concerning tweets with a precision of 80% but a recall of only 53%.

3.2 Facebook

Facebook is another huge potential source of data. It is used by an estimated 1.13 billion people each day,[1] and has become a major platform for promoting mental health campaigns and recruiting research participants to mental health studies. This section focuses on attempts to measure the emotional state of Facebook's users.

While Twitter users simply broadcast their messages to anyone who will listen, Facebook users exert more control to form closed audiences of friends and family members. This increased privacy may increase the user's openness and honesty. Therefore, while Facebook data may be better suited to monitoring emotional state, it is more difficult to obtain by researchers outside Facebook.

Kramer (2010) analysed 400 million status updates from English speaking users within the United States, and developed a measure of gross national happiness by scoring each status update against the positive and negative terms within it (as given by the Affective Norm for English Words (ANEW) vocabulary described in Section 4.2). The measure was roughly validated in a similar fashion as Golder and Macy (2011), by visually inspecting the data for expected patterns like peaks during holiday periods and weekends and troughs during national disasters. A more rigorous validation was conducted by recruiting 1300 Facebook users to complete

[1] Updated statistics are available at `http://newsroom.fb.com/company-info`

a survey about life satisfaction. However, the happiness scores derived from their status updates achieved only a weak correlation with life satisfaction survey scores (r=0.17). It is not clear whether this weak association was due to the simplicity of the algorithm (understandable, given the scale at which it was run), or to insufficient data (some of the Facebook users involved had as few as three status updates) or to factors beyond the scope of the study design.

Also relevant to this discussion is the study completed by Coviello, Sohn, Kramer, Marlow, Franceschetti, Christakis, and Fowler (2014) which explored how negative emotion spreads as a contagion between Facebook accounts. The experiment investigated rain as a trigger for changes in valence, which was measured by the prevalence of known positive and negative terms within status updates (similar to Kramer 2010). The experiment demonstrated that not only do people post fewer positive and more negative posts during rainy days, but also that these posts have a statistically significant effect on the valence of their friend's posts. Positive posts promoted positive posts and inhibited negative ones, and *vice-versa*; even when these friends were not experiencing the same weather patterns.

Another study of emotional contagion in Facebook showed how small changes in the filtering algorithm used in the Facebook Newsfeed can have a significant impact on the moods of users (Kramer, Guillory, and Hancock, 2014). The experiment compared two filtering algorithms; one reducing the number of positive status updates a user was exposed to, and another suppressing negative updates. The valence of a post was again identified using a similar approach as in Kramer (2010). In total, 690k users were randomly selected and exposed to these algorithms. Approximately 155k participants within each condition posted at least one status update within the week. From these status updates, the authors measured a small (<0.1%) but statistically significant shift towards positive terms (and away from negative ones) among those exposed to fewer negative posts. The opposite held true for the positivity-reduced condition. The study contributed interesting insights into how emotions propagate on social networks and the impact of design on people's emotions, two under-served areas of research. But the study triggered an understandable flurry of outrage and concern for failing to inform or obtain consent from participants before attempting to manipulate their moods, as reflected in their Facebook updates (Calvo, Peters, and D'Mello, 2015).

Moreno and Jelenchick (2011) manually evaluated a year's worth of status updates of 200 college students. Each update was coded against the DSM IV criteria for depression. The authors reported that 25% of profiles contained at least one depressive post, and 5 profiles had periods of multiple depressive posts each spanning several months; the authors considered these periods to match DSM criteria for a major depressive episode. This figure was expected, given 30% of college students report feeling depressed and unable to function each year (American College Health Association, 2009).

The underrepresentation of depression terminology in the Facebook posts would seem to indicate that few people are comfortable sharing their depressive symptoms explicitly on Facebook. This does not invalidate Facebook data as a means of recruiting information that could be developed into a diagnostic tool, but it does point

out the need to look for subtler cues. To our knowledge, the only published attempt to do so automatically is by De Choudhury and Counts (2014), who recruited new mothers willing to share their Facebook data. In total they gathered 28 mothers who had been diagnosed with post-natal/post-partum depression (PPD) and 137 mothers with no experience of depression (a PHQ-9 Patient Health Questionnaire, was used to exclude mothers who displayed symptoms of depression but had not been clinically diagnosed with PPD). Their accounts were mined for activity for 50 weeks prior to and 10 weeks after birth.

From this data, the authors observed that mothers suffering from PPD were:

- less likely to make posts or share media
- less likely to engage with or tag media left by others
- less likely to receive comments or likes from friends
- more irregular in the times that they interacted
- more likely to use first person pronouns (i.e. talked more about themselves)
- more likely to ask questions

Interestingly the researchers did not find a statistically significant difference in the use of positive or negative emotion terms, as measured by LIWC. This is inconsistent with their previous work with Twitter (De Choudhury et al., 2013a), one explanation would suggest that Facebook posts are less emotionally expressive than tweets. Unfortunately we cannot make any direct comparisons between Twitter and Facebook, because the authors did not develop or evaluate a binary classifier in the same way.

3.3 Blogs and Journals

Blogs are used by many as personal diaries, and as a way to reflect and to share daily experiences. Increasingly, people can annotate their own posts with metadata (i.e. labels) about their emotions. Livejournal.com was one of the first to provide this self-annotation feature to users and it has been used in multiple studies. For example, Strapparava and Mihalcea (2007) extracted a corpus of 8,671 LiveJournal posts labelled (by the post authors) with 6 emotions. The study sits slightly outside of the scope of this review however, because they did not attempt to classify the blog posts themselves or the emotional state of their authors. Instead the posts were used as background training data to classify news headlines (from the SemEval 2007 dataset) and the emotion they were intended to provoke in the reader. Nevertheless, this work is described in more detail in Section 4.4.

Nguyen, Phung, Dao, Venkatesh, and Berk (2014) aim to automatically separate out LiveJournal posts that talk about tough times. For ground truth, they extracted a depression dataset of 38k posts from sub-communities for depression, self-harm, bereavement etc. and a control set of 230k posts from other, less dire sub-communities. They experimented with a wide range of features, including LIWC (linguistic, social, affective, cognitive, perceptual, biological, relativity, personal concerns and spoken), affect (also based on LIWC), and topic models generated by LDA. They were able to automatically determine whether each individual

blog post belonged to the depression or control set with an accuracy of 93%, and the best features were the LDA topics.

3.4 Other social media data

As new social media platforms are created they are used in mental health studies. For example, the nature of personal disclosure has been studied on Reddit (De Choudhury and De, 2014), an online media platform similar to discussion forums but with an optional throwaway account type that improves anonymity. The results suggest that anonymity promotes disclosure, despite the often caustic nature of the discussions generated around a post. Although the number of throwaway accounts is proportionally small, the fact that 61% were used only once suggest these users do not want to leave any trail behind, but as the authors point out, this also means they cannot receive much social support.

Different countries often have different languages. They also vary on their socioeconomic variables, their conceptions of mental health, stigma and therefore support of those who are ill. Therefore, studying the social networks popular in different countries is important. Mixi, the most popular social network platform in Japan, which is the Organization for Economic Co-operation and Development (OECD) country with the highest suicide rate, was used to study the relationship between a person's social network (i.e. relationships) and suicide ideation (Masuda, Kurahashi, and Onari, 2013). According to the authors the effect of the age, gender, and number of friends on suicide ideation was small.

Another interesting perspective is to look into peoples' attitudes towards suicide (e.g stigma) as they can inform the type of suicide prevention interventions. A study using data from Weibo, a Chinese social network platform (Li, Huang, Hao, ODea, Christensen, and Zhu, 2015) analysed the social attitudes of people publicly commenting on others who made public their intention to commit suicide and found stigma was widespread and terms such as *deceitful*, *pathetic*, and *stupid* were often used.

Those who suffer stigma maybe at risk of mental health issues and their communities might require special attention, for example researchers have studied TrevorSpace, a social network popular amongst lesbian, gay, transgender and bisexual communities (Homan, Lu, Tuurmcrochesteredu, Lytle, Lytle, Rochester, Silenzio, and Silenzio, 2014a). The study did not use the text but only the connections amongst individuals and showed that some features are predictive of distress or mental-ill health.

3.5 Suicide Notes

Suicide notes were first studied in the 19th century by Durkheim, a pioneer of suicide risk research (Durkheim, 1897). Shneidman created the field of suicidology around 1949 and collected and systematically studied suicide notes (Shneidman and Farberow, 1957). In 1957, Shneidman and Farberow published a book that explored the reasons why people kill themselves, and collected genuine suicide notes

paired with fake suicide notes written by a healthy control group that were demographically matched to the suicidal authors. The aim was to explore the difference between why people think others suicide, and why individuals actually do.

Pestian, Nasrallah, Matykiewicz, Bennett, and Leenaars (2010) extracted from this book a corpus of 33 genuine (written by people who completed suicide) and 33 fake notes (written by people simulating a suicide note). They showed how difficult they are to distinguish manually: in their experiments mental health professionals achieved an accuracy of 63%, while psychiatry trainees were only 49% accurate (i.e. worse than random chance). The authors were able to train a classifier that achieved 74% accuracy. However, the classifier relied somewhat on emotion annotations (e.g. guilt, hopeless, regret) that were made by a panel of three mental health professionals. The authors did not report the accuracy of the algorithm without these handcrafted annotations.

Removing this reliance on expert annotation appears to be the focus of Pestian, Matykiewicz, Linn-Gust, South, Uzuner, Wiebe, Cohen, Hurdle, and Brew (2012), which presents a shared task for annotating suicide notes. Shared tasks are common computer science activities where research groups compete to solve the same problem. The task used a corpus of genuine notes written by people who took their own life, collected between 1950 and 2011. The notes were manually transcribed, anonymized and carefully reviewed, before being shared with a panel of 1500 vested volunteers who were recruited from online communities and Facebook pages created to support grieving friends and family members. These volunteers were asked to annotate the notes for occurrences of negative sentiments (abuse, anger, blame, fear, guilt, hopelessness, sorrow), positive sentiments (forgiveness, happiness, peacefulness, hopefulness, love, pride, thankfulness) and neutral sentiments (instructions, information). In total, 900 suicide notes were each annotated separately by three volunteers.

In this shared task, 106 researchers in 24 teams tried to automatically reconstruct these manual annotations. They were given a training set of 600 annotated letters to develop their algorithm, which was evaluated on the remaining 300 letters (these were not released until algorithms were finalized).

There are two ways of calculating Precision, Recall and their harmonic mean F1 (a commonly used measure of classification accuracy). Macroaverages of these are the simple average over classes. Microaverages are obtained pooling per-document (i.e. post) decisions across classes, and computing P, R or F1 on the pooled contingency table. All of the top 10 competing teams achieved micro-averaged F1-measures between 53% and 62%. An ensemble combining several of the systems could theoretically achieve an F1-measure of 76%. Section 4.4 describes the approaches of the top few teams. The F1-measure has its best value at 1 and worst score at 0. Micro-averaged F1 was used to rank the teams.

This shared task used suicide notes written offline, mostly before the Internet era. But research on suicide notes written for the Internet have shown consistent findings with those written offline (Barak and Miron, 2005) and has been used as evidence for the development of support services (Barak, 2007).

Writing about suicide in the Internet age brings new challenges. Christensen,

Batterham, Mackinnon, Griffiths, Hehir, Kenardy, Gosling, and Bennett (2014) conducted a review of research studies that focused on three particular challenges: the use of online screening for suicide, the effectiveness of eHealth interventions aimed to manage suicidal thoughts, and newer studies aimed to proactively intervene when individuals at risk of suicide are identified by their social media postings. The review highlighted the need for more evidence on the effectiveness of eHealth interventions for suicide prevention.

3.6 Online forums and support groups

Online peer-to-peer communities and support groups are one of the most promising ways of reaching out to more people, and allow them to discuss their health issues (Eysenbach, Powell, Englesakis, Rizo, and Stern, 2004). These forums may be particularly good for reaching young people (O'Dea and Campbell, 2010). Although the evidence on their efficacy is not yet strong, this is due in part to the difficulty of running systematic trials (Eysenbach, Powell, Englesakis, Rizo, and Stern, 2004; Pistrang, Barker, and Humphreys, 2008). One of the limitations of peer support style groups is that they need trained moderators who can help manage the community and can provide informed feedback when needed. When the communities grow this becomes a challenge.

The CLPsych 2016 shared task (Milne, Pink, Hachey, and Calvo, 2016) aimed to address such challenges of scale by allowing moderators to focus their efforts where it is most needed. It collected forum posts from ReachOut.com–a site for young Australians facing tough times–and asked participants to automatically triage them as *green* (no intervention required), *amber* (a moderator should ideally respond, but not urgently), *red* (a moderator should respond as soon as they can), or *crisis* (the post indicates someone is at risk of harm).

The task attracted 60 submissions from 15 teams of researchers. Each team was initially given 947 annotated posts to develop and train their algorithms, and later given 280 posts–with annotations only seen by the task coordinators–for evaluation. The three best performing teams obtained a macroaveraged f-measure of 0.42, but used very different approaches to do so. Kim, Wang, Wan, and Paris (2016) combined relatively few features (unigrams and post embeddings) with an ensemble of SGD classifiers. Brew (2016) used traditional n-gram features with a well-tuned SVM classifier, and achieved the best separation of urgent posts (*crisis* and *red* vs. *amber* and *green*) with 0.69 f-measure. Malmasi, Zampieri, and Dras (2016) gathered a larger number of features from not only the post itself, but also the preceding and following ones, and achieved the best separation of *flagged* posts (*crisis*, *red* and *amber* vs. *green*) with 0.87 f-measure. The dataset is available[2] for researchers to continue developing their algorithms.

This triage task was a continuation of Moderator Assistant (Liu, Calvo, Daven-

[2] Researchers can apply for access to the ReachOut triage dataset at `http://bit.ly/triage-dataset`

port, and Hickie, 2013) a system that uses NLP to detect people in distress and helps moderators prioritise their workload (described in more detail in Section 5.5).

Other papers have analysed patients' self-narratives and provide further evidence of the potential of NLP techniques in PTSD diagnosis. For example, He, Veldkamp, Glas, and de Vries (2015) collected in an online survey in a forum for those seeking mental health aid and evaluated methods using n-grams and unigrams with Decision Trees, Nave Bayes, SVM and Product Score Model (PSM). Unigrams with PSM had the highest accuracy (0.82).

3.7 Summary

- Most popular forms of social media have been used as data sources for mental health applications.
- Number of users, language (i.e. English) and availability of APIs increase the chances of a platform being used. Twitter is the most widely used source of data mainly because the collection of public data is easy. Facebook is also common, often used by authors who also work for (or in partnership with) the company.
- We only searched papers in English, and these mostly talked about content written in English. A few exceptions (e.g. Japanese) are mentioned. It would seem that NLP in non-English languages is an unexplored area. This may be related to the lower quality, or absence of Natural Language Processing tools in languages other than English.
- One of the most studied corpora of user-generated texts is a suicide notes collection because it was used in a shared task competition. Share tasks allow researchers from multiple disciplines to collaborate on a particular problem.

4 Automated labeling

We use the term *labeling* in this section to mean the identification of emotions, moods and risk profiles that might indicate mental health problems or profiles that could be used to target interventions. It goes beyond diagnosis, as the term to refer to DSM V classifications of mental illness, although we do note when studies attempt to achieve specific psychiatric diagnosis. Labeling is generally done using document classification techniques that take certain features as input and map them to a set of labels. The techniques used include tokenization, feature extraction and selection, parsing and machine learning classification. We provide a summary of key publications organized according to the tasks required to build an automated labeling system.

4.1 Gathering training data

Often the first step to build a classifier that can automatically detect emotions and mental health issues is to gather labeled data. This is a requirement for supervised machine learning algorithms that use this data for training and evaluation. The

challenges of collecting and labeling data, particularly in realistic scenarios (i.e. in the wild) include time and cost (proportional to the amount of data being labeled), validity (particularly dependent on how the data was collected) and reliability (often requiring multiple trained annotators per document). These challenges have been discussed elsewhere, for example in the literature on Affective Computing (Riva, Calvo, and Lisetti, 2014). Semi-labeled and unlabeled data can also be used to train classifiers. For example, using Expectation-Maximization and Nave-Bayes classifiers (Nigam, McCallum, Thrun, and Mitchell, 2000) contributed a technique that reduced classification errors by up to 30%.

4.2 Feature extraction

The literature would suggest that a combination of language and other forms of features is the most promising. This will likely be dependent on the application. This section describes these in more detail

Demographic features. Textual data can be complemented with socio-demographic data of individuals. Variables like county-by-county scores for age, sex, ethnicity, income, education and geolocation have been used in studies showing that these features can improve the accuracy of the classifier (Schwartz et al., 2013). These results are in line with evidence suggesting that social expressions of emotion are age and gender dependent, while location influences time and weather, which in turn, can also influence emotions.

Lexical features. Emotions are often inferred by the percentage of emotional terms in the posts. These are often computed using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015; Tausczik and Pennebaker., 2010). LIWC is a software tool that provides psychologically-grounded lists of positive and negative emotional terms (amongst other categories reflecting writing style). The assumption, which is supported by research evidence (Strapparava and Mihalcea, 2014; Pennebaker, 2011), is that emotion terms reflect emotion feelings (i.e., if you use more positive words, it is because you feel more positive). Language and feelings are indeed related, however, it appears that the relationship is weak. LIWC counts the number of words in a category (e.g. affective or social). The words are added to a category by the developers who often provide 'validity judgments' the correlations of judges ratings of the category with the LIWC variable. In particular, the correlation between the specific emotional words in LIWC and the emotional ratings by human 'validity judges' is a modest 0.41 for positive and 0.31 for negative terms; and these are amongst the lowest in the LIWC dictionaries (Pennebaker, Boyd, Jordan, and Blackburn, 2015).

Emotion features from LIWC are often taken as good approximations of what people are feeling, thus, LIWC has been used in several mental health projects. But since the labeling is not done for clinical purposes the thesaurus does not directly indicate if, for example, a user wrote the word *sad*, has a higher probability of illness as would be measured by PHQ9 score or by the diagnosis of a clinician. Labelling of this type is generally done for a subset of each corpora or other information for the ground truth. For example, Kramer et al. (2014)'s study that manipulated Facebook

News Feed filters (described in Section 3.2) used emotion terms from LIWC. Other emotional thesauri include WordNet (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990), particularly its extension WordNet-affect (Strapparava and Valitutti, 2004) and normative databases such as the Affective Norm for English Words (ANEW) (Bradley and Lang, 1999). These thesauri can be used to group words, emotion words, for example, according to their valence or activation (i.e. arousal) (Calvo and Kim, 2013).

Behavioral features. When people write they leave a trace of their behaviors. This metalinguistic information includes the time they wrote, for how long, if the writing was reply to another post, etc. The study by Sadilek et al. (2013) using geolocation and twitter data (discussed in Section 3.1) is an example of how this type of data can be used in the modeling of moods.

Social features. These include the number of relationships (e.g. Facebook friends) and the topological map (i.e. who is related with who) that have been found to correlate to emotional and mental health constructs. The study by Sadilek et al. (2013) used social features to model emotional contagion within groups. Others have shown how they can be used to improve the prediction of mental health states (Masuda et al., 2013; Homan et al., 2014).

4.3 Feature selection techniques

Feature selection is required to reduce the number of features (ie. the input space) used in the classification task. Without feature selection classification algorithms suffer from overgeneralization and are often less efficient (Witten and Frank, 2005). There are multiple ways of reducing the dimensionality of the problem including Principal Component Analysis, regression models, and other statistical approaches. In text classification Bi-normal separation (BNS) methods outperformed all other methods in an evaluation of twelve feature selection methods (e.g. Information Gain) over a benchmark of 229 text classification problems (Forman, 2003). The same study compared different techniques when only one data set was available (the most frequent scenario) and showed that generally BNS outperformed other approaches. The exception was when precision was the goal in which case Information Gain yielded the best results.

4.4 Evaluating labeling systems

There are too many machine learning classification techniques to be discussed here, so we limit our discussion to those that have been used in mental health research. Complete descriptions of the algorithms can be found in machine learning textbooks or reviews (Strapparava and Mihalcea, 2014; Sebastiani, 2002).

One of the most comprehensive ways to evaluate and compare classification systems through shared task competitions between independent research groups. This approach is used in the computer science and NLP communities to help identify the most accurate system. To make the process fair and reliable, the training data for the system is shared with the competing teams, but the test data is only made

available to the shared task teams after the evaluation of the algorithms has been completed.

The suicide notes task described earlier is possibly the best example of shared task competitive challenges applied to mental health data. However, this approach has also been applied to related emotion detection tasks. For example, the SemEval-2007 task Strapparava and Mihalcea (2007), where emotions were detected in a collection of news stories. Albeit not on mental health and not covered here, the SemEval-2007 and other similar shared tasks provide important insights for those working on mental health applications.

Micro and macro averages of precision, recall and F1-measures were calculated for each team. Micro-averages weigh equally all data points so more common categories are weighted more heavily. In the suicide notes shared task, a team from the Open University in the UK had the best results with a micro-averaged F1=0.61, the least performing team had F1=0.30, showing a large variation in the results when different techniques are used. The teams used different techniques including Part Of Speech (POS) taggers, thesaurus like WordNet (Strapparava and Valitutti, 2004), emotional lexicons and LIWC. The winning team's submission (Yang, Willis, De Roeck, and Nuseibeh, 2012) is an example of the complexity of NLP systems. A text processing tool fed the notes to subsystems that identified instances of emotions, both at the token and the sentence levels. There were components to detect negations, emotion cues and terms. It also used numerous machine learning techniques that have shown to be successful in other applications, including Conditional random Fields (CRF), Support vector Machines (SVM), Nave Bayes (NB) and Maximum Entropy (ME). The final output was the emotion labels at the sentence level.

The runner-up (Xu, Wang, Liu, Tu, Sun, Tsujii, and Chang, 2012) also used a combination of techniques. A key distinguishing factor in this submission was perhaps the augmentation of the dataset with semantically similar blog posts from LiveJournal and the SuicideProject (where bloggers can annotate data with emotions). They also used CRF and SVM algorithms in a binary fashion one classifier for each label.

Since the other teams (Cherry, Mohammad, and De Bruijn, 2012; Luyckx, Vaassen, Peersman, and Daelemans, 2012; McCart, Finch, Jarman, Hickling, Lind, Richardson, Berndt, and Luther, 2012; Spasić, Burnap, Greenwood, and Arribas-Ayllon, 2012) reported details of their algorithms and outcomes, the shared task also provides insights into the techniques that did not produce satisfactory results. For example, Yu, Kübler, Herring, Hsu, Israel, and Smiley (2012) reported low accuracy using Wordnet, character n-grams and word n-grams (a contiguous sequence of n terms). They concluded that character n-grams and dependency pairs provide good features for emotion detection in suicide notes yet word n-grams and POS n-grams were less robust.

Another shared task was a 2015 Computational Linguistics and Clinical Psychology (CLPsych) shared task (Coppersmith, Dredze, Harman, Hollingshead, and Mitchell, 2015). The data used for the task, and a hackathon held at John Hopkins University, consisted of anonymised Tweets written by 1,746 users who

stated a diagnosis of depression or post traumatic stress disorder (PTSD), with demographically-matched community controls. The shared task consisted of three binary classification experiments: (1) depression versus control, (2) PTSD versus control, and (3) depression versus PTSD. The shared task had submissions from 4 research groups: University of Maryland who focused on topic models (Resnik, Armstrong, Claudino, and Nguyen, 2015), University of Pennsylvania which used a variety of methods (Preoţiuc-Pietro, Sap, Schwartz, and Ungar, 2015), University of Minnesota who used a rule based approach (Pedersen, 2015) and a submission by the workshop organisers and Microsoft that used character language models. Due to length limitations we recommend reading those papers for further details.

A rarer evaluation approach involves studying a classification system while it is in use. Section 5.5.1 describes a triage system that uses the classifiers in real life. Their evaluation should in general go beyond the classifier accuracy and explore aspects such as how users perceived the triage system, and how they evaluate the possible usage scenarios, etc.

4.5 Summary

- All feature selection and classification algorithms common in the NLP literature have been tried. Overall, we could not ascertain whether a particular technique was superior to other techniques.
- LIWC is the most widely used tool for extracting features from text.
- An understudied area is that of triage systems used in real time, with real users. Most studies are performed offline, where the diagnosis was not used to respond to actual users.

5 Interventions

The Internet has provided a medium for low cost psychotherapeutic interventions that psychologists have been incorporating into their practice for over a decade. In this section we describe examples of how Natural Language Processing and Generation are used to automatically create interventions.

Barak and Grohol (2011) provided the taxonomy of Internet-based interventions used in this section: psycho-educational websites, Interactive self-guided interventions, online counselling and psychotherapy, online support groups and blogs and other types, predicting the increase in Virtual reality, gaming and SMS and texting therapies.

In a comprehensive review (Barak, Hen, Boniel-Nissim, and Shapira, 2008) the authors considered 92 internet-based intervention studies (and included 64 in the meta-analysis) and reported a mean weighted treatment effect size of 0.53. We reviewed the 92 papers, however only one study (Owen, Klapow, Roth, Shuster, Bellis, Meredith, and Tucker, 2005) was found to use NLP techniques. Yet, the effect size was similar to effect sizes reported for traditional face-to-face therapies. The potential of further improving these results by using state-of-the-art NLP techniques is momentous.

The lack of studies describing text-based analysis techniques represents an area full of opportunity for the mental health intervention sector. NLP techniques could augment therapist-based mental health interventions. For example, NLP techniques could potentially be used to detect links between behaviors and emotions described by the client, or to help people recognize unhealthy thinking patterns. Since the number of studies is so small we expanded our scope to include studies from the physical health literature to demonstrate the potential of this area of research.

5.1 Web-based Psycho-educational Interventions

Psycho-education refers to the educational interventions where patients and their families can learn about the illness and the techniques and resources available to help them. Most examples of psycho-education interventions are fixed (i.e. not personalized) descriptions of symptoms and treatment written for a general audience. Although these have been excluded from this review we explore here how they could be made adaptive and personalized using NLP. Exploring new approaches is important since research shows that web-based psycho-education interventions have therapeutic value and show positive outcomes (Ritterband, Gonder-Frederick, Cox, Clifton, West, and Borowitz, 2003). Internet-based interventions (CBT, psychoeducation and other) have shown to be effective in treating depression and anxiety (Donker, Griffiths, Cuijpers, and Christensen, 2009; Spek, Cuijpers, Nyklícek, Riper, Keyzer, and Pop, 2007), eating disorders (Neve, Morgan, Jones, and Collins, 2010), smoking and alcohol consumption (Bewick, Trusler, Barkham, Hill, Cahill, and Mulhern, 2008; Myung, McDonnell, Kazinets, Seo, and Moskowitz, 2009) among other conditions.

Other application domains can provide ideas for innovative approaches in mental health. For example, in the same way that NLP has been used in educational technologies one could expect them to be used in psycho-education. Regrettably this has not been the case. A search on Google Scholar for Natural Language Processing psycho-education, computational linguistics psycho-education did not return any positive results. NLP is used in education, for example, in automated assessment tools where users can answer questions and get formative feedback to improve their understanding. In health this could be done to improve their understanding of an illness. The content of websites could also be improved by measuring the complexity of the language used, allowing content editors to adapt the language to different age and socioeconomic groups.

A common marketing strategy is to personalize information with automatically generated text, for example using simple templates to generate mailouts. Evaluations have shown that this type of semi-automatic and fully-automatic texts (i.e. letter, interventions) compare well with those generated by humans on a number of measures such as tone, rhythm and flow, repetition and terminology (Coch, 1996). Within health, personalization has been considered critical to patient-centered care and a number of studies have evaluated Natural language Generation (NLG) techniques in the authoring and personalization of webpages containing patient education materials. Unfortunately the personalization of materials has been tried only in

the context of physical health (e.g. surgical procedures, cancer, etc). In the context of mental health, personalizing information (i.e. interventions) could also be helpful for family members, moderators of peer-support groups, the general public in social media, etc. Automated text generation methods could provide the moderators with information that they can use for quickly customizing and replying. This may even be useful for mental health clinicians, where information about a specific patient can be presented in the form of a report.

There are several possible tools used to build NLG systems including SimpleNLG (Gatt, Portet, Reiter, Hunter, Mahamood, Moncur, and Sripada, 2009), a simple Java-based generation framework. The design of these tools is generally informed by the work of Reiter and Dale (2000) who defined a general architecture used in most of the applications we reviewed. The architecture has three components connected together into a pipeline: 1) A Document Planer determines the content and structure of a document. 2) A Microplanner decides how to communicate the content and structure chosen by the Document Planer. This involves choosing words and syntactic structures. 3) Surface Realiser maps the abstract representations used by the Microplanner into an actual text.

One of the few examples of mental health applications is PyschoGen (Dockrey, 2007), an NLG based system that changes its output based on the emotional state (set by the user). The experimental system followed the standard document planning → microplanning → realization pipeline. The goal of the project was not to send the text generated to real users or even generate highly natural output, but rather explore how emotional data could be generated. This type of approach could be suitable for mental health interventions that express empathy and compassion in line with the concept of client-centric health information and resources. Since there is little research on NLG applications to mental health we describe here some related to physical health.

Information Therapy (DiMarco, Covvey, Cowan, DiCiccio, Hovy, Lipa, and Mulholland, 2007) is a system providing preoperative information, that largely consists of personalizing resources that are normally distributed via brochures - each discussing a surgical procedure. The system was designed using a collection of reusable texts, each annotated with linguistic and formatting information, then the NLG tools automatically selected, assembled and revised the reader-appropriate pieces of text. This approach could easily be adapted to psycho-education interventions.

Similarly tailored information systems used in diabetes, migraine or cancer could be used to help those trying to learn about mental illness (though perhaps not for the patients themselves). Bental and colleagues (Bental and Cawsey, 2002; Bental, Cawsey, and Jones, 1999) have described a variety of tailored patient education systems that may act as examples. For diabetes patients PIGLIT provided personalized information about the disease, their hospital etc. (Binstead, Cawsey, and Jones, 1995). Similar work used hypertext pages for Migraine patients (Buchanan, Moore, Forsythe, Carenini, Ohlsson, and Banks, 1995), cancer patients (Jones, Pearson, McGregor, Cawsey, Barrett, Craig, Atkinson, Gilmour, and McEwen, 1999) and to accompany prescriptions (De Carolis, de Rosis, Grasso, Rossiello, Berry, and Gillie, 1996) have also been described. Tailored patient information systems have shown

to be more effective than non-tailored computer system (or leaflets). For example, a Randomized Control Trial (RCT) of cancer patients (Jones et al., 1999) (N=525) with three conditions (paper brochures, personalized and general information hypertext) showed that more patients offered the personalised pages felt that they had learned more (and used it more), that the information was relevant to them, and even shared the information with others. Not all the results of NLG applications have been positive. A clinical trial generating personalized smoking cessation letters (Reiter, Robertson, and Osman, 2003) after participants (N=2553) responded to a questionnaire indicated that those who received personalized information were no more likely to stop smoking than those who received a non-tailored letter.

5.2 Interactive, Self-Guided Interventions

While content in tailored information systems is directed one-way - from the system to the user - similar techniques can be used to produce interactive psychoeducation interventions. This includes interactive software (websites, mobile apps, SMS) that offer an individualized step-by-step structure of information and guidance a form of online self-help activity. The interactive system may provide a form of education as in the previous category, but may also only provide a structured set of activities for the user to complete. Back in the 1960s a system called Eliza (Weizenbaum, 1966) was one of the first Artificial Intelligence systems for this type of intervention. It emulated a Rogerian psychological perspective and used an early form of Artificial Intelligence to engage the user in a dialogue of question and answer and asked empathic questions, based on the previous dialogue move. There has been some research around this model, where a software agent plays the role of the therapist (Bohannon, 2015).

Structured interventions such as Computerized Cognitive Behavior Therapy (CCBT) have received the most attention over the last 20 years. Although we could not find any CCBT interventions utilizing NLG, positive results from their application in related areas (Barak et al., 2008) make them a prime candidate for future research. For example, the efficacy of Moodgym (Christensen, Griffiths, and Jorm, 2004), one of the pioneering tools for online CBT, showed that it was effective in reducing depression and dysfunctional thinking and in increasing understanding about CBT techniques. Since then many other CBT-based interventions have been developed and evaluated for the treatment of different mental health problems (Barak and Grohol, 2011). Enough evidence for the effectiveness of online CBT has now been accumulated that in the UK computerized CBT is included in the public healthcare coverage (Kaltenthaler, Brazier, De Nigris, Tumur, Ferriter, Beverley, Parry, Rooney, and Sutcliffe, 2006) and the market of commercial applications is booming (Aguilera and Muench, 2012). NLP techniques could possibly be used to personalize the activities or to automatically process personal diaries.

A limitation of current CBT interventions that could be addressed with NLG systems is that they can only help when the user proactively goes to the website or installs the app. They cannot use any of the information users generate during their everyday life, such as Facebook posts, tweets, emails etc. The NLP techniques

described in Section 4 can be used to gather some of this information from these sources, which could be used to make inferences and generate appropriate interventions to be delivered privately through the same social media.

5.2.1 Relational agents

Relational agents (Bickmore and Picard, 2005; Bohannon, 2015) are tools where patients have conversations with a computer. The term *relational* aims to highlight a higher aim than similar *conversational* agents designed for question-answering or short dialogue situations (e.g. Apple's Siri). The research used to design relational agents feeds from psychology, sociolinguistics, communication and other behavioural sciences and focuses on how to build long-term relationships. The language used by the agents can vary in sophistication, going from multiple-choice questions, pre-recorded texts or spoken language and most often include an NLG component.

Relational agents, and sometimes virtual reality (VR) environments, can interact and 'speak' to humans, ie. they are *Natural Language capable*. To do this they must understand spoken, or sometimes, written language, they must be able to manage a dialog (e.g. turn taking) and generate the output text/voice. Generally the speech signal is converted to text using a speech recognition system, and the speech is generated from text using a text-to-speech system. Although details of this work is beyond the scope of the present discussion, Kenny, Parsons, Gratch, Leuski, and Rizzo (2007) provide useful details on the architecture for one of these systems.

Briefly, dialogue systems work by understanding the users' dialogue move using the document classification techniques described in Section 4. These dialogue systems lookup for the closest match to the users' dialogue move (i.e. turn) and respond to it. If there is no statement close enough a default statement is used (e.g. "*sorry, I do not understand could you say that again*").

For relational systems to work, they have to engage patients and build a good therapist-patient relationship (Okun and Kantrowitz, 2014). It is believed that this relationship, and engagement in general, depend on the quality of the agents language, voice, metalinguistic and emotional expressions (e.g. face and body) (Bohannon, 2015). Again this is similar to what has been done and evaluated in education. For example, D'Mello, Lehman, and Graesser (2011) develop affect-aware Intelligent Tutoring Systems, somewhat similar to relational agents, and show that when the system detects and responds to emotions it is better at promoting learning and engagement.

Maintaining engagement with any behaviour change tool is challenging, even more so with agents that are expected to be used for a semester, a year or a whole life (Bickmore, Schulman, and Yin, 2010; Bickmore and Gruber, 2010). Although short-term engagement is somewhat easier it is not necessarily a good predictor of health outcomes (Bickmore et al., 2010). A way to study engagement with interventions involves using personal relationship research (Bickmore and Picard, 2005) and medical psychology in general.

Bickmore and Gruber (2010) evaluated a number of systems used in health coun-

selling and behaviour change interventions. They found that the agents can be used in CBT interventions, can be engaging and help in behaviour change, and they can also reduce cost or facilitate communication when human resources are scarce. The narrative is considered particularly important to the success of the intervention and can be designed to be engaging and possibly educational.

Some virtual agents use novel forms of data to personalize the dialogues. Help4Mood (Martínez-Miranda, Bresó, and García-Gómez, 2012b) is an interactive virtual agent aiming to help people recover from depression in their own homes. It uses subjective assessments, standard mood and depression questionnaires and diaries to personalize the conversations. An interesting feature of Help4Mood is that it also tracks aspects of behaviour such as sleep and activity levels that can be used to shape the conversations. It is designed for use with other forms of counselling and therapy and the information collected can be provided to therapists.

Factors that influence the quality of relationships with an agent include the User Interface (Bickmore and Mauer, 2006). For example, a text avatar (Rincón-Nigro and Deng, 2013), similar to a chatbox, would have different qualities than an embodied avatar. Some form of empathy in the avatar has been recognised as important and incorporated in several agents (Martínez-Miranda, Bresó, and García-Gómez, 2012a; Bickmore, Gruber, and Picard, 2005).

5.3 Text messaging

High tech options like relational agents and virtual reality environments are not always the most appropriate due to cost, reach or because they cannot be incorporated into everyday life. Other media, such as text messaging (SMS) can be used to send feedback using a mixture of preprogrammed parts and individually tailored information (Bauer, Percevic, Okon, Meermann, and Kordy, 2003). For example, automated text messaging was used by Aguilera and Muñoz (2011) to support CBT in an adult, low income population where smart phones and Internet access are not as common. It was aimed at increasing homework adherence, improving self-awareness, and helping track patient progress by engaging patients through questions and simple activities. The small trial showed evidence that automated text messaging 'conversations' could be a low cost approach to helping low income populations.

Fathom (Dinakar et al., 2014) is a system to help counsellors who provide help through SMS interventions (Crisis Text Line). Fathom uses topic models, a statistical modelling technique to track the evolution of themes in a conversation, in real time. The models were being built with 8106 conversations held by 214 counsellors. The evaluation did not yet include the topic maps with the full data set but consisted of understanding the experiences of seven counsellors using the system: they all found it most useful and had few feature requests.

A study to help patients with Bulimia Nervosa (Bauer, Percevic, Okon, Meermann, and Kordy, 2003) used canned and personalised messages. In this study patients would send a weekly message to the system and receive automated feedback that contained a template and personalised segments. The pilot program provided

evidence that it was well received and could be useful in aftercare treatment of patients. Regrettably effectiveness was not reported.

Academic evaluations of commercial tools like Buddyapp were not found. Buddyapp involves writing a diary and engaging with a self-help CBT type intervention by sending text messages, so these are not included here.

5.4 Online Counseling

Technology in this category—often referred as *e-therapy*—is used to mediate the interactions between patients and clinicians (e.g. Email, Skype) (Grohol, 2004) and can replace a face-to-face session. The forms of communication or interaction can be more flexible for counseling and can be asynchronous (e.g. email, forums, etc.) or synchronous (e.g. chat, video conference, etc.).

Chatboxes are an interesting form of therapy. As shown by Dowling and Rickwood (2013) in a systematic review of online counseling via chat they can be effective in helping patients. For example, clients using the online chat service have shown positive attitude towards counseling (Finn and Bruce, 2008). We could not find any recent study using NLP enhanced chatboxes, and regard this as an important area that requires further evaluation.

5.5 Online Support Groups and Blogs

Online mental health support groups have become increasingly popular since the late 1990s. They enable people in distress to find others with similar needs and problems, to share feelings and information, receive support, provide advice, and develop a sense of community. For example, the feedback received by peer-supporters on social media can help individuals reflect on their thoughts, which offers a type of mental health intervention through the feedback loop (Grohol, Anthony, Nagel, and Goss, 2010; Hoyt and Pasupathi, 2008). Online peer-support groups are effective in mental health and the degree to which a user engages with the group through posting new messages or replying to others, has shown to relate to the quality of the health outcome (Barak, Boneh, and Dolev-Cohen, 2010). The review by Griffiths, Calear, and Banfield (2009) focused on measuring the effectiveness of online support groups. They reviewed 31 papers (involving 28 trials) and found that 62.5% reported a positive effect on depressive symptoms, although only 20% used control groups. Regrettably the coding did not include any technical features of the studies, such as use of NLP.

Moderator Assistant (Liu et al., 2013), described earlier, provides trained moderators the ability to identify people at risk and to respond with resources (e.g. link to mental health related information). These moderators help keep the community together and raise the value of conversations. Besides the triage system described in Section 5.5.1, Moderator Assistant can also automatically create draft interventions that moderators are expected to edit using Natural Language Generation techniques (Hussain, Calvo, Ellis, Li, Ospina-Pinillos, Davenport, and Hickie, 2015). These drafts can improve the efficiency and quality of the feedback provided.

For example, they can use therapeutic language and focus on issues that reflect the ethos of the organisation. If the organisation wants to focus on certain illnesses or treatments it can develop specialized content.

5.5.1 Triage systems

Most research has used real data but largely neglected the translation of insights to real-life application of the diagnostic outcome. An understudied area is that of real use triage systems, used 'in the wild', where individuals considered 'at risk' are identified in order to receive some form of assistance. We have already discussed studies by Choudhury and others who used Twitter (De Choudhury et al., 2013a) and Facebook (De Choudhury and Counts, 2014) data and where recent mothers at risk of depression were identified, yet these studies did not evaluate how those people could be offered assistance, or whether this would even be acceptable.

For instance, a triage system that used social media data to alert people when someone in their network was depressed or suicidal was Radar, created by the Good Samaritans in the UK (Horvitz and Mulligan, 2015). The system quickly became an example of how difficult it is to use this data, even when it is for a noble cause, when the appropriate research has not been carried out as to how to provide feedback. The system used data from Facebook friends in ways they may not know about and this raised concerns about privacy. It also raised important issues around disclosure, such as who should find out that a person is not doing well or depressed. This information can be used by friends or foes, and even when it is by someone who wants to help, the person might not be the best qualified or might even get harmed herself - e.g. suicides occasionally happen in clusters of friends (Haw, Hawton, Niedzwiedz, and Platt, 2013).

Another approach is for this type of technology to be used only by people who have been trained to help with mental health issues. For example, Moderator Assistant (Liu et al., 2013) is a triage system for moderators at ReachOut.com a mental health organisation based in Australia. Some ReachOut.com users seek help by posting in discussion forums that have moderators who read the posts and where appropriate, respond. Their triage system aimed to help prioritise responses by identifying posts that needed urgent response together with information on the topic/issue. For the moderator, this might involve replying to the post, contacting the post author directly, and sometimes removing posts that do not comply with ReachOut.com policies. Given that this is often a complex decision, the triage tool also allows multiple moderators to communicate around a particular post.

5.6 Summary

- Text driven psychological interventions are the possibly the most common form of Internet intervention.
- Natural Language Processing techniques are rarely used as part of interventions.
- Psycho-educational and self-guided interventions have many similarities to

learning technologies where NLP have been widely used. There is an opportunity for one area to learn from experiences in the other.
- Relational agents have been increasingly common, often in the form of embodied conversational agents or companions. The increased popularity of chatbots and automated messaging systems will possibly support growth in this line of research.
- Text messaging, often considered as a low tech option is very useful, particularly in certain populations.
- Online counselling and peer support groups provide human-to-human communication where the computer mediation can add significant value through automatic triaging, summarisation and personalization of the interventions.

6 Conclusions

The review was aimed to provide a taxonomy of how NLP has been used in mental health applications and potential future opportunities for its integration into online mental health tools. This application domain is highly interdisciplinary, with the technical aspects of building systems and the mental health challenges of helping those who most need it. A common problem is that the research literature in one area is often not known to researchers in the other which makes collaboration difficult. In fact, often these collaborations can be hindered by differences in the language, terminology and methodology. We covered three areas in which multidisciplinary teams could collaborate:

1. data collection
2. processing or diagnosing, and
3. generation of automated mental health interventions

Multidisciplinary teams have used data from Twitter, Facebook, blogs and other social media services to learn about people's behavior, emotions, social communications and more (Section 3). Most of the research has focused on English language, and more research is needed on other languages, particularly taking into account how cultural factors such as mental health views and stigma may influence the outcomes.

What we called '*labeling*' (Section 4) encompasses triaging people at risk, the diagnosis of specific mental health conditions and even automatic labeling of suicide letters. There is probably other ways in which information extraction and text classification techniques can be used, but we have limited our discussion to those described in the mental health literature with a focus on depression and suicide. Although the techniques are similar to those used in other application domains one has to be aware of the differences. For example, while classification accuracy in marketing is important, a false negative (i.e. erroneously not identifying an individual as belonging to a group) in marketing only means an opportunity is lost, while in mental health it could be more serious such as not identifying someone who is in need of help. The trade-offs between precision and recall will likely be different in these applications.

Although there has been great progress in the data collection and processing, there has been insufficient research on the uses of NLG in mental health interventions, with the notable exception being the use of relational agents.

6.1 Research Limitations

This review is not an exhaustive compilation of all the work in NLP and mental health. The collaboration between psychologists and computing researchers is rapidly evolving, and new studies at the intersection of these fields are published regularly. For example, a recent special issue of Science Magazine dedicated to progress in artificial intelligence (AI) had several mentions to this area (Bohannon, 2015; Hirschberg and Manning, 2015; Horvitz and Mulligan, 2015).

Because of the character of the data, and the difficulty in maintaining the anonymity of those who write the texts (Horvitz and Mulligan, 2015) ethical considerations are crucial. In fact, researchers should be aware of difficulty of applying methods common in computer science such as shared task competitions or data sharing. We have not covered the ethical implications of being able to identify people in need. This is an important area that requires a dedicated article. Once methods are developed to accurately process information data in mental health, a clear area that requires further research is how this information can be relayed sensitively and ethically back to participants.

RAC and SH are supported by the Young and Well Cooperative Research Centre, which is established under the Australian Government's Cooperative Research Centres Program. RAC is supported by an Australian Research Council Future Fellowship FT140100824. RAC and DM is supported by an Australian Research Council Linkage Project. HC is supported by an NHMRC Fellowship 1056964.

References

Abbe, A., Grouin, C., Zweigenbaum, P. and Falissard, B. 2015. Text mining applications in psychiatry: a systematic literature review. *International journal of methods in psychiatric research*, 86–100.

Aguilera, A. and Muench, F. 2012. There's an App for that: Information technology applications for cognitive behavioral practitioners. *The Behavior therapist/AABT* **35**(4), 65–73.

Aguilera, A. and Muñoz, R. F. 2011. Text messaging as an adjunct to CBT in low-income populations: A usability and feasibility pilot study. *Professional Psychology: Research and Practice* **42**(6), 472–8.

American College Health Association 2009. National College Health Assessment Spring 2008 Reference Group Data Report. *Journal of American College Health* **57**, 477–88.

Armstrong, R., Hall, B. J., Doyle, J. and Waters, E. 2011. "Scoping the scope" of a cochrane review. *Journal of Public Health* **33**(1), 147–50.

Barak, A. 2007. Emotional support and suicide prevention through the Internet: A field project report. *Computers in Human Behavior* **23**(2), 971–84.

Barak, A., Boneh, O. and Dolev-Cohen, M. 2010. Factors underlying participants gains in online support groups. *Internet in psychological research*, 13–47.

Barak, A. and Grohol, J. M. 2011. Current and Future Trends in Internet-Supported Mental Health Interventions. *Journal of Technology in Human Services* **29**(3), 155–96.

Barak, A., Hen, L., Boniel-Nissim, M. and Shapira, N. 2008. A Comprehensive Review and a Meta-Analysis of the Effectiveness of Internet-Based Psychotherapeutic Interventions. *Journal of Technology in Human Services* **26**(2-4), 109–60.

Barak, A. and Miron, O. 2005. Writing characteristics of suicidal people on the Internet: A psychological investigation of emerging social environments. *Suicide and Life-Threatening Behavior* **35**(5), 507–24.

Bauer, S., Percevic, R., Okon, E., Meermann, R. U. and Kordy, H. 2003. Use of text messaging in the aftercare of patients with bulimia nervosa. *European Eating Disorders Review* **11**(3), 279–90.

Bental, D. and Cawsey, A. 2002. Personalized and adaptive systems for medical consumer applications. *Communications of the ACM* **45**(5), 62–3.

Bental, D. S., Cawsey, A. and Jones, R. 1999. Patient information systems that tailor to the individual. *Patient Education and Counseling* **36**, 171–80.

Bewick, B. M., Trusler, K., Barkham, M., Hill, A. J., Cahill, J. and Mulhern, B. 2008. The effectiveness of web-based interventions designed to decrease alcohol consumptiona systematic review. *Preventive medicine* **47**(1), 17–26.

Bickmore, T. and Gruber, A. 2010. Relational agents in clinical psychiatry. *Harvard review of psychiatry* **18**(2), 119–30.

Bickmore, T., Gruber, A. and Picard, R. 2005. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient education and counseling* **59**(1), 21–30.

Bickmore, T. and Mauer, D. 2006. Modalities for building relationships with handheld computer agents. In *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06*, New York, pp. 544–549.

Bickmore, T., Schulman, D. and Yin, L. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* **24**(6), 648–66.

Bickmore, T. W. and Picard, R. W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* **12**(2), 293–327.

Binstead, K., Cawsey, A. and Jones, R. 1995. Generating personalised information using the medical record. In *Proceedings of Artificial Intelligence In Medicine*, Berlin, New Yor, pp. 29–41.

Bohannon, J. 2015. The synthetic therapist. *Science* **349**(6245), 250–1.

Bradley, M. M. and Lang, P. J. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida.

Brew, C. 2016. Classifying reachout posts with a radial basis function svm. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 138–142.

Buchanan, B. G., Moore, J. D., Forsythe, D. E., Carenini, G., Ohlsson, S. and Banks, G. 1995. An intelligent interactive system for delivering individualized information to patients. *Artificial intelligence in medicine* **7**(2), 117–54.

Calvo, R. and D'Mello, S. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* **1**(1), 18–37.

Calvo, R. A., Dinakar, K., Picard, R. and Maes, P. 2016. Computing in mental health. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 3438–3445.

Calvo, R. A. and Kim, S. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence* **29**(3), 527–43.

Calvo, R. A., Peters, D. and D'Mello, S. 2015. When technologies manipulate our emotions. *Communications of the ACM* **58**(11), 41–2.

Cherry, C., Mohammad, S. M. and De Bruijn, B. 2012. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical informatics insights* **5**(Suppl 1), 147–54.

Christensen, H., Batterham, P., Mackinnon, A., Griffiths, K. M., Hehir, K. K., Kenardy, J., Gosling, J. and Bennett, K. 2014. Prevention of generalized anxiety disorder using a web intervention, iChill: randomized controlled trial. *Journal of medical Internet research* **16**(9), e199.

Christensen, H., Griffiths, K. M. and Jorm, A. F. 2004. Delivering interventions for depression by using the internet: randomised controlled trial. *BMJ* **328**(7434), 265–8.

Chung, C. and Pennebaker, J. 2007. The psychological functions of function words. *Social communication*, 343–59.

Coch, J. 1996. Evaluating and comparing three text-production techniques. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pp. 249–254.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. and Mitchell, M. 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*, Denver, Colorado, pp. 31–39.

Coviello, L., Sohn, Y., Kramer, A. D., Marlow, C., Franceschetti, M., Christakis, N. A. and Fowler, J. H. 2014. Detecting emotional contagion in massive social networks. *PloS one* **9**(3), e90315.

De Carolis, B., de Rosis, F., Grasso, F., Rossiello, A., Berry, D. C. and Gillie, T. 1996. Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine* **8**(2), 123–45.

De Choudhury, M. and Counts, S. 2014. Characterizing and predicting postpartum depression from shared facebook data. *Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing*, 626–38.

De Choudhury, M., Counts, S. and Horvitz, E. 2013a. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3267–3276.

De Choudhury, M., Counts, S. and Horvitz, E. 2013b. Social media as a measurement tool of depression in populations. pp. 47–56.

De Choudhury, M. and De, S. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 71–80.

De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. 2013. Predicting depression via social media. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 128–137.

DiMarco, C., Covvey, H., Cowan, D., DiCiccio, V., Hovy, E., Lipa, J. and Mulholland, D. 2007. The development of a natural language generation system for personalized e-health information. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, pp. 2339–2344.

Dinakar, K., Chaney, A. J. B., Lieberman, H. and Blei, D. M. 2014. Real-time topic models for crisis counseling. In *Twentieth ACM Conference on Knowledge Discovery and Data Mining, Data Science for the Social Good Workshop*.

D'Mello, S. K., Lehman, B. and Graesser, A. 2011. A motivationally supportive affect-sensitive autotutor. In *New perspectives on affect and learning technologies*, pp. 113–126.

Dockrey, M. 2007. Emulating Mental State in Natural Language Generation Systems. Technical report, University of British Columbia.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. and Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* **6**(12), e26752.

Donker, T., Griffiths, K. M., Cuijpers, P. and Christensen, H. 2009. Psychoeducation for depression, anxiety and psychological distress: a meta-analysis. *BMC Medicine* **7**(1), e79.

Dowling, M. and Rickwood, D. 2013. Online counseling and therapy for mental health problems: A systematic review of individual synchronous interventions using chat. *Journal of Technology in Human Services* **31**(1), 1–21.

Durkheim, E. 1897. *Suicide : a study in sociology.* [1951] Free Press.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C. and Stern, A. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj* **328**(7449), 1166–72.

Finn, J. and Bruce, S. 2008. The LivePerson model for delivery of etherapy services: A case study. *Journal of Technology in Human Services* **26**(2-4), 282–309.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research* **3**, 1289–305.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W. and Sripada, S. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *A.I. Communications* **22**(3), 153–86.

Golder, S. and Macy, M. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* **333**(6051), 1878–81.

Griffiths, M. K., Calear, L. A. and Banfield, M. 2009. Systematic review on internet support groups (ISGs) and depression (1): Do ISGs reduce depressive symptoms? *Journal of Medical Internet Research* **11**(3), e40.

Grohol, J. M. 2004. Online counseling: A historical perspective. In *Online counseling: A handbook for mental health professionals*, pp. 51–68. Elsevier Academic Press San Diego, CA.

Grohol, J. M., Anthony, K., Nagel, D. A. N. and Goss, S. 2010. Using Websites, blogs and wikis in mental health. *The use of technology in mental health applications ethics and practice*, 68–75.

Haw, C., Hawton, K., Niedzwiedz, C. and Platt, S. 2013. Suicide clusters: a review of risk factors and mechanisms. *Suicide and life-threatening behavior* **43**(1), 97–108.

He, Q., Veldkamp, B. P., Glas, C. A. W. and de Vries, T. 2015. Automated assessment of patients self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*.

Hirschberg, J. and Manning, C. D. 2015. Advances in natural language processing. *Science* **349**(6245), 261–6.

Homan, C. M., Johar, R., Liu, T., Lytle, M., Silenzio, V. and Alm, C. O. 2014. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. *ACL 2014*, 107–1.

Homan, C. M., Lu, N., Tuurmcrochesteredu, N. L. X., Lytle, M. C., Lytle, M., Rochester, U., Silenzio, V. M. B. and Silenzio, V. 2014a. Social Structure and Depression in TrevorSpace. *Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing - CSCW '14*, 615–24.

Homan, C. M., Lu, N., Tuurmcrochesteredu, N. L. X., Lytle, M. C., Lytle, M., Rochester, U., Silenzio, V. M. B. and Silenzio, V. 2014b. Social Structure and Depression in TrevorSpace. *Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing - CSCW '14*, 615–24.

Horvitz, E. and Mulligan, D. 2015. Data, privacy, and the greater good. *Science* **349**(6245), 253–5.

Hoyt, T. and Pasupathi, M. 2008. Blogging about trauma: Linguistic measures of apparent recovery. *E-Journal of Applied Psychology* **4**(2), 56–62.

Hussain, M. S., Calvo, R. A., Ellis, L., Li, J., Ospina-Pinillos, L., Davenport, T. and Hickie, I. 2015. Nlg-based moderator response generator to support mental health. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1385–1390.

Jones, R., Pearson, J., McGregor, S., Cawsey, A. J., Barrett, A., Craig, N., Atkinson, J. M., Gilmour, W. H. and McEwen, J. 1999. Randomised trial of personalised computer based information for cancer patients. *Bmj* **319**(7219), 1241–7.

Kaltenthaler, E., Brazier, J., De Nigris, E., Tumur, I., Ferriter, M., Beverley, C., Parry, G., Rooney, G. and Sutcliffe, P. 2006. Computerised cognitive behaviour therapy for depression and anxiety update: a systematic review and economic evaluation. *Health technology assessment* **10**(33), 1–186.

Kenny, P., Parsons, T. D., Gratch, J., Leuski, A. and Rizzo, A. A. 2007. Virtual patients for clinical therapist skills training. In *Intelligent Virtual Agents*, pp. 197–210.

Kim, S. M., Wang, Y., Wan, S. and Paris, C. 2016. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 128–132.

Kotsiantis, S. 2007. Supervised machine learning: A review of classification techniques. *Informatica* **31**, 249–68.

Kramer, A. D. 2010. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 287–290.

Kramer, A. D., Guillory, J. E. and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* **111**(24), 8788–90.

Larsen, M. E., Boonstra, T. W., Batterham, P. J., ODea, B., Paris, C. and Christensen, H. 2015. We feel: mapping emotion on twitter. *IEEE journal of biomedical and health informatics* **19**(4), 1246–52.

Lawless, N. and Lucas, R. 2011. Predictors of regional well-being: a county level analysis. *Social Indicators Research* **101**(3), 341–57.

Li, A., Huang, X., Hao, B., ODea, B., Christensen, H. and Zhu, T. 2015. Attitudes towards suicide attempts broadcast on social media: an exploratory study of chinese microblogs. *PeerJ* **3**, e1209.

Liu, M., Calvo, R. A., Davenport, T. and Hickie, I. 2013. Moderator assistant: helping those who help via online mental health support groups. In *Joint Workshop on Smart Health and Social Therapies, OzChi*.

Luyckx, K., Vaassen, F., Peersman, C. and Daelemans, W. 2012. Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomedical informatics insights* **5**(Suppl 1), 61–9.

Malmasi, S., Zampieri, M. and Dras, M. 2016. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 133–137.

Martínez-Miranda, J., Bresó, A. and García-Gómez, J. M. 2012a. Modelling Therapeutic Empathy in a Virtual Agent to Support the Remote Treatment of Major Depression. In *ICAART (2)*, pp. 264–269.

Martínez-Miranda, J., Bresó, A. and García-Gómez, J. M. 2012b. The Construction of a Cognitive-Emotional Module for the Help4Moods Virtual Agent. *Information and Communication Technologies applied to Mental Health*, 34–9.

Masuda, N., Kurahashi, I. and Onari, H. 2013. Suicide ideation of individuals in online social networks. *PloS one* **8**(4), e62262.

McCart, J. A., Finch, D. K., Jarman, J., Hickling, E., Lind, J. D., Richardson, M. R., Berndt, D. J. and Luther, S. L. 2012. Using ensemble models to classify the sentiment expressed in suicide notes. *Biomedical informatics insights* **5**(Suppl 1), 77–85.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. 1990. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography* **3**, 235–44.

Milne, D. N., Pink, G., Hachey, B. and Calvo, R. A. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, pp. 118–127.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. and Danforth, C. M. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* **8**(5), e64417.

Moreno, M. and Jelenchick, L. 2011. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety* **28**, 447–55.

Myung, S.-K., McDonnell, D. D., Kazinets, G., Seo, H. G. and Moskowitz, J. M. 2009. Effects of Web-and computer-based smoking cessation programs: meta-analysis of randomized controlled trials. *Archives of Internal Medicine* **169**(10), 929–37.

Neve, M., Morgan, P. J., Jones, P. R. and Collins, C. E. 2010. Effectiveness of webbased interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with metaanalysis. *Obesity Reviews* **11**(4), 306–21.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S. and Berk, M. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing* **5**(3), 217–26.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine learning* **39**(2-3), 103–34.

O'Dea, B. and Campbell, A. 2010. Healthy connections: online social networks and their potential for peer support. *Studies in health technology and informatics* **168**, 133–40.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C. and Christensen, H. 2015. Detecting suicidality on Twitter. *Internet Interventions* **2**(2), 183–8.

Okun, B. and Kantrowitz, R. 2014. *Effective helping: Interviewing and counseling techniques*. Cengage Learning.

Owen, J. E., Klapow, J. C., Roth, D. L., Shuster, J. L., Bellis, J., Meredith, R. and Tucker, D. C. 2005. Randomized pilot of a self-guided internet coping group for women with early-stage breast cancer. *Annals of Behavioral Medicine* **30**(1), 54–6.

Paul, M. and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 265–272.

Pedersen, T. 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*, Denver, Colorado, pp. 46–53.

Peek, N., Combi, C., Marin, R. and Bellazzi, R. 2015. Artificial Intelligence in Medicine Thirty years of artificial intelligence in medicine (AIME) conferences : A review of research themes. *Artificial Intelligence In Medicine* **65**(1), 61–73.

Pennebaker, J., Kiecolt-Glaser, J. and Glaser, R. 1988. Disclosure of traumas and immune function: health implications for psychotherapy. *Journal of Consulting and Clinical Psychology* **56**(2), 239–45.

Pennebaker, J. W. 2011. *The secret life of pronouns: How our words reflect who we are*. New York, NY: Bloomsbury Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K. and Blackburn, K. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Pennebaker, J. W. and Chung, C. K. 2007. Expressive writing, emotional upheavals, and health. *Handbook of health psychology*, 263–84.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A. and Leenaars, A. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights* **2010**(3), 19–28.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J. and Brew, C. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* **5**(Suppl 1), 3–16.

Pistrang, N., Barker, C. and Humphreys, K. 2008. Mutual help groups for mental health problems: A review of effectiveness studies. *American journal of community psychology* **42**(1-2), 110–21.

Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A. and Ungar, L. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, pp. 40–45.

Reiter, E. and Dale, R. 2000. *Building natural language generation systems*. Boston: MIT Press.

Reiter, E., Robertson, R. and Osman, L. M. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* **144**(1), 41–58.

Resnik, P., Armstrong, W., Claudino, L. and Nguyen, T. 2015. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, pp. 54–60.

Resnik, P., Resnik, R. and Mitchell, M. (Eds.) 2014. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics.

Rincón-Nigro, M. and Deng, Z. 2013. A text-driven conversational avatar interface for instant messaging on mobile devices. *IEEE Transactions on Human-Machine Systems* **43**(3), 328–32.

Ritterband, L. M., Gonder-Frederick, L. A., Cox, D. J., Clifton, A. D., West, R. W. and Borowitz, S. M. 2003. Internet interventions: In review, in use, and into the future. *Professional Psychology: Research and Practice* **34**(5), 527–34.

Riva, G., Calvo, R. A. and Lisetti, C. 2014. Cyberpsychology and Affective Computing. In R. Calvo, S. D'Mello, J. Gratch, and A. Kappas (Eds.), *Handbook of Affective Computing*, pp. 547–558. New York: Oxford University Press.

Sadilek, A., Homan, C., Lasecki, W., Silenzio, V. and Kautz, H. 2013. Modeling Fine-Grained Dynamics of Mood at Scale. In *WSDM*, pp. 3–6.

Schwartz, H. A., Eichstaedt, J. C., Margaret L. Kern, L., Dziurzynski, M. A., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P. and Ungar, L. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp. 583–591.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.* **34**(1), 1–47.

Shneidman, E. and Farberow, N. (Eds.) 1957. *Clues to suicide*. New York: Harper and Row.

Signorini, A., Segre, A. M. and Polgreen, P. M. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one* **6**(5), e19467.

Spasić, I., Burnap, P., Greenwood, M. and Arribas-Ayllon, M. 2012. A naïve Bayes Approach to classifying Topics in suicide notes. *Biomedical informatics insights* **5**(Suppl 1), 87–9.

Spek, V., Cuijpers, P. I. M., Nyklícek, I., Riper, H., Keyzer, J. and Pop, V. 2007. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological medicine* **37**(03), 319–28.

Strapparava, C. and Mihalcea, R. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 70–74.

Strapparava, C. and Mihalcea, R. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1556–1560.

Strapparava, C. and Mihalcea, R. 2014. Affect Detection in Texts. In R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas (Eds.), *The Oxford Handbook of Affective Computing*, Chapter 13, pp. 184–203. New York: Oxford University Press.

Strapparava, C. and Valitutti, A. 2004. WordNet-Affect: an affective extension of WordNet. In *LREC 2004 - Fourth IInternational Conference on Language Resources and Evaluation*, Volume 4, Lisbon, pp. 1083–1086.

Tausczik, Y. R. and Pennebaker., J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* **29**(1), 24–5.

Weizenbaum, J. 1966. ELIZAa computer program for the study of natural language communication between man and machine. *Communications of the ACM* **9**(1), 36–45.

Witten, I. H. and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xu, Y., Wang, Y., Liu, J., Tu, Z., Sun, J.-T., Tsujii, J. and Chang, E. 2012. Suicide note sentiment classification: a supervised approach augmented by web data. *Biomedical informatics insights* **5**(Suppl 1), 31–4.

Yang, H., Willis, A., De Roeck, A. and Nuseibeh, B. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical informatics insights* **5**(Suppl 1), 17–30.

Yu, N., Kübler, S., Herring, J., Hsu, Y.-Y., Israel, R. and Smiley, C. 2012. LASSA: Emotion detection via information fusion. *Biomedical informatics insights* **5**(Suppl 1), 71–6.