Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression*

Christine Howes, Matthew Purver
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK
{c.howes,m.purver}@qmul.ac.uk

Rose McCabe
University of Exeter Medical School
Exeter, UK
r.mccabe@exeter.ac.uk

Abstract

Mental illnesses such as depression and anxiety are highly prevalent, and therapy is increasingly being offered online. This new setting is a departure from face-to-face therapy, and offers both a challenge and an opportunity – it is not yet known what features or approaches are likely to lead to successful outcomes in such a different medium, but online text-based therapy provides large amounts of data for linguistic analysis. We present an initial investigation into the application of computational linguistic techniques, such as topic and sentiment modelling, to online therapy for depression and anxiety. We find that important measures such as symptom severity can be predicted with comparable accuracy to face-to-face data, using general features such as discussion topic and sentiment; however, measures of patient progress are captured only by finer-grained lexical features, suggesting that aspects of style or dialogue structure may also be important.

1 Introduction

Mental illnesses such as depression and anxiety have been called "the biggest causes of misery in Britain today" (Layard, 2012). The main avenue of treatment for such conditions is talking therapies, such as Cognitive Behavioural Therapy (CBT); however, there is far greater demand than can currently be met, and currently only 25% of sufferers in the UK receive treatment. Therapy is therefore increasingly being delivered online: this helps to improve access and reduce waiting times, and is just as effective as standard therapy (Kessler et al., 2009). However, this new online setting provides a challenge of evaluation and optimisation (Hanley and Reynolds, 2009; Beattie et al., 2009). Online therapy is a significant departure from face-to-face therapy, and it is not yet known exactly what features or approaches are likely to lead to successful outcomes, or help identify negative outcomes such as risk to the patient or others. Current methods (e.g. controlled studies) are expensive and time-consuming; we need fast, accurate methods to ensure treatment can be made effective and efficient in this new context.

Professional communication varies widely (McCabe et al., 2013b) and aspects of doctor-patient interaction and language are known to influence outcomes such as patient satisfaction, treatment adherence and health status (Ong et al., 1995; McCabe et al., 2013a). For therapists, automated methods to analyse therapist-client communication are of interest as there is little known about *how* the quality of communication influences patient outcome. Identifying patterns of effective communication – both in terms of what is spoken about and how it is spoken about – would help guide training of therapists. Moreover, it may assist in identifying successful therapy and perhaps, more importantly, where communication is not therapeutic and patients are failing to improve. This may warrant a different or more intensive therapeutic intervention. Applying computational linguistic techniques to therapy data could therefore offer potential to produce tools which can aid clinicians in predicting outcomes, diagnosing severity of symptoms and/or evaluating progress. Recent work on spoken therapy dialogue has shown promising results in a range of mental health tasks, including diagnosis of post-traumatic stress disorder (PTSD) and depression (DeVault et al., 2013; Yu et al., 2013),

*This work was partly supported by the ConCreTe project. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 7–16,
Baltimore, Maryland USA, June 27, 2014.

and prediction of outcomes in schizophrenia treatment (Howes et al., 2013).

Online therapy data provides a new challenge – language and interaction styles differ to face-to-face – but also an opportunity in the availability of large amounts of text data without the need for automatic speech recognition or manual transcription. Here, we present an initial investigation into the application of computational linguistic techniques to online therapy for depression and anxiety. We find that important measures such as symptom severity can be predicted with comparable accuracy to face-to-face data, and that general aspects such as discussion topic and sentiment are useful predictors; and suggest some ways in which techniques can be adapted for improved performance in future.

2 Background

2.1 Computational analysis & mental health

Research into computer-based diagnosis in mental health goes back at least to the 1960s – see (Overall and Hollister, 1964; Hirschfeld et al., 1974) amongst others – but most systems rely on doctor- or patient-reported data rather than naturally occurring language. Much recent work similarly uses self-reported clinical and socio-demographic data, e.g. to predict treatment resistance in depression (Perlis, 2013). Some recent natural language processing (NLP) research examines features of the language used by patients when discussing conditions or treatment, e.g. discovering topics and opinions from online doctor ratings (Paul et al., 2013) or social media (Paul and Drezde, 2011).

However, aspects of the communication *during treatment itself* are also associated with patient outcomes (Ong et al., 1995). In the mental health domain, recent work suggests that, for patients with schizophrenia both conversation structure (*how* communication proceeds in therapy), and content (*what* is talked about), can affect outcomes (McCabe et al., 2013a; John et al., under review). NLP research has now begun to examine both. Wallace et al. (2013) model speech acts to characterise doctor-patient consultations on medication adherence; Angus et al. (2012) use unsupervised topic models to visualise shared content in clinical dialogue; Cretchley et al. (2010) use a similar approach for a qualititative analysis of topic and communication style between patients with schizophrenia and carers. DeVault et al. (2013) use features of speech, and Yu et al. (2013) multimodal features, from video-mediated dialogue to detect depression and PTSD with promising accuracies (0.66 to 0.74 depending on condition and task). In face-to-face therapy for schizophrenia, Howes et al. (2012; 2013) use a combination of supervised and unsupervised approaches to predict a range of diagnostic and outcome measures, including future adherence to treatment (accuracy 0.70); fine-grained lexical features gave reasonable accuracy, with more general topic features giving weaker prediction of some outcomes.

2.2 Topic modelling

One focus of research for mental health is therefore on methods for analysing content (*what* is talked about). Traditional methods, while effective, involve time-consuming hand-coding of data (Beattie et al., 2009; John et al., under review); NLP techniques can reduce this requirement. Unsupervised probabilistic models (e.g. Latent Dirichlet Allocation (LDA) Blei et al. (2003) and variants) have been widely applied to learn topics (word distributions) from the data itself, connecting words with similar meanings and even distinguishing between uses of words with multiple meanings (Steyvers and Griffiths, 2007). Such techniques have been applied successfully to structured dialogue e.g. meetings and tutorials (Purver et al., 2006; Eisenstein and Barzilay, 2008), and more recently to dialogues in the clinical domain (Cretchley et al., 2010; Howes et al., 2013), with topics found to identify important themes within therapy conversation such as medication, symptoms, family and social issues, and to correlate with outcomes.

2.3 Sentiment and emotion analysis

One aspect of conversation process and style is the affect or emotion present. NLP research has generally approached this via the task of *sentiment* detection, distinguishing positive from negative (and sometimes neutral) stance (Pang and Lee, 2008). Methods generally take either a knowledge-rich approach (relying on e.g. dictionaries of sentiment-carrying words (Pennebaker et al., 2007)), or a data-rich approach via (usually supervised) machine learning over datasets of sentiment-carrying text (e.g. Socher et al. (2013)). The former can provide deeper insights, but are less robust in the face of unexpected vocabulary, unusual or errorful spelling; the latter are more ro-

bust but require training from large datasets. Recent research has attempted finer-grained distinctions, e.g. detecting specific emotions such as anger, surprise, fear etc; again, approaches can be characterised as dictionary-based or machine-learning-based (Chuang and Wu, 2004; Seol et al., 2008; Purver and Battersby, 2012; De Choudhury et al., 2012). The resulting sentiment or emotion ratings have been widely used to determine aspects of personality and mental state in various domains. In social media text, Quercia et al. (2011; 2012) found correlations between sentiment and levels of popularity, influence and general well-being; O'Connor et al. (2010) with measures of public opinion. Closer to our application, Liakata et al. (2012) show that these methods can be applied to analyse emotion in suicide notes.

2.4 Research questions

Here, similar to (DeVault et al., 2013; Howes et al., 2013), our primary question is whether these approaches can be usefully applied to diagnose conditions and predict outcomes, but in a new modality – online text-based therapy – which may require different and/or more robust methods. In addition, we would like to gain some insight into which features of language and interaction might be predictive, in order to help clinicians improve therapeutic methods, and to assess how general and transferable any model might be. Our main questions here are therefore:

- What features of text-based online therapy dialogue might help predict symptoms and/or outcomes? Specifically, how predictive are conversation topic and emotional content?
- Can we detect them accurately and reliably, using approaches generalisable to large datasets, across different subjects and conditions?
- Can the features provide any insights into the treatment process and/or the online modality?

3 Method

3.1 Data

The data used in this study consisted of the transcripts from 882 Cognitive Behavioural Therapy (CBT) treatment dialogues between patients with depression and/or anxiety and their therapists using an online text-based chat system. The transcripts are from online CBT provided by Psychology Online, who deliver 'live' therapy from a qualified psychologist accessed via the internet (`http://www.psychologyonline.co.uk`). Of the 882 transcripts, 837 are between therapists and patients who were in an ongoing treatment program or had completed their treatment by the time our sample was collected. There are 167 patients in this sample (125 females and 42 males), with 35 different therapists (for 2 patients the identity of the therapist is unknown). The number of transcripts per patient ranges from 1 to 14, with a mean of 5.011 (s.d. 2.73). For all of the measures based on the transcripts, as outlined below, we included all text typed by both the therapist and the patient. In addition to the transcripts themselves, each patient normally filled out two questionnaires prior to each session with their therapist. These are described below.

3.2 Outcomes

Patient Health Questionnaire (PHQ-9) This is a self-administered diagnostic instrument for common mental disorders (Kroenke and Spitzer, 2002). The PHQ-9 is the depression module, which scores each of the 9 DSM-IV criteria as '0' (not at all) to '3' (nearly every day). A higher score indicates higher levels of depression, with scores ranging from 0-27. It has been validated for use (Martin et al., 2006).

Generalised Anxiety Disorder scale (GAD-7) Similarly, the GAD-7 (Spitzer et al., 2006) is a brief self-report scale of generalised anxiety disorder. This is a 7-item scale which scores each of the items as '0' (not at all) to '3' (nearly every day). A higher score indicates higher levels of anxiety.

Outcome measures For the data in our sample, PHQ-9 and GAD-7 were highly correlated ($r = 0.811$, $p < 0.001$) so for the results reported below we focus on PHQ-9. As each patient filled in the PHQ-9 before each consultation, we used two different outcome measures: *PHQ now* – the PHQ-9 score of the patient for the questionnaire completed immediately prior to the consultation; and *PHQ start-now* – the difference between the PHQ-9 score prior to any treatment and *PHQ now*, i.e. a measure of progress (how much better or worse the patient is since the start of their treatment). Although these two measures are numerical, one of the general aims of our research is to identify patients at risk. We therefore binarised the outcome measures and treated our task

as a categorisation problem to identify the group of interest. For *PHQ now*, these were patients with moderate to severe symptoms; for *PHQ start-now*, patients whose PHQ score had not improved.

3.3 Topics

The transcripts from the 882 treatment consultations were analysed using an unsupervised probabilistic topic model, using MALLET (McCallum, 2002) to apply standard Latent Dirichlet Allocation (Blei et al., 2003), with the notion of *document* corresponding to a single consultation session, represented as the sequence of words typed by any speaker. Stop words (common words which do not contribute to the content, e.g. 'the', 'to') were removed as usual (Salton and McGill, 1986), but the word list had to be augmented for text chat conventions and spellings (e.g. unpunctuated "ive"). Additionally, common mispellings were mapped to their correctly spelled equivalents using Microsoft Excel's in-built spellchecker. This was due to the nature of text chat, in contrast to transcribed speech or formal text – the word 'questionnaire', for example, was found to have been typed in 21 different ways. Following (Howes et al., 2013) we set the number of topics to 20,[1] used the default setting of 1000 Gibbs sampling iterations, and enabled automatic hyperparameter optimisation to allow an uneven distribution of topics via an asymmetric prior over the document-topic distributions (Wallach et al., 2009).

As Howes et al. (2013) did in face-to-face therapy, we found most topics were composed of coherent word lists, with many corresponding to common themes in therapy e.g. family (Topic 12), symptoms (16), treatment process (2, 14), and issues in work and social life (19, 5) – see Table 5.

3.4 Sentiment and emotion analysis

Each turn in the transcripts was then annotated for strength of positive and negative sentiment, and level of anger. We compared three approaches: the dictionary-based LIWC (Pennebaker et al., 2007) and two machine learning approaches, the Stanford classifier based on deep neural nets and parse structure trained on standard text (Socher et al., 2013), and one based on distant supervision over social media text, Sentimental (Purver and Battersby, 2012).[2] None are specifically designed for therapy dialogue data; however, given the unorthodox spelling and vocabulary used in text chat, we expect machine-learning based approaches, and training on "noisy" social media text, to provide more robustness.

We used each to provide a positive/negative/neutral sentiment value; for LIWC, we took this from the relative magnitudes of the `posemo` and `negemo` categories. Two human judges then rated the 85 utterances in one transcript independently. Inter-annotator agreement was good, with Cohen's kappa = 0.66. Agreement with LIWC was poor (0.43-0.45); with Stanford better (0.51-0.54); but best with Sentimental (0.63-0.80). For anger, LIWC gave only one utterance a non-zero rating, while Sentimental provided a range of values. We therefore used Sentimental in our experiments. Raw values per turn were scaled to [-1,+1] for sentiment (-1 representing strong negative sentiment, +1 strong positive), and [0,1] for anger; we then derived minimum, maximum, mean and standard deviation values per transcript.

3.5 Classification experiments

We performed a series of experiments, to investigate whether various features of the transcripts could enable automatic detection of patient responses to the PHQ-9. The full range of possible features were calculated for each transcript – see Table 1. As well as topic, sentiment and emotion features as detailed above, we include raw lexical features to characterise details of content, and some high-level features (amount of talk; patient demographics; and therapist identity, known to affect outcomes).

In each case, we used the Weka machine learning toolkit (Hall et al., 2009) to pre-process data, and a decision tree classifier (J48), a logistic regression model and the support vector machine implementation LibLINEAR (Chang and Lin, 2001) as classifiers. *PHQ now* was binarised based on the classification in Kroenke and Spitzer (2002), whereby scores of 10 or over are moderate to severe and scores of less than 10 are mild. *PHQ start-now* was binarised according to whether there was an improvement (reduction) in the PHQ score or not. Positive scores indicate

[1] An arbitrary decision, but Howes et al. (2013) chose it to match the number defined by manual coders in a therapy domain.

[2] Available from `liwc.net`, `nlp.stanford.edu` and `sentimental.co` respectively.

Feature set	Description
AgentID	Identity of the therapist
High level	Client gender; client age group; session number; client/agent number of words and turns used; proportion of all words per participant
Topic	Probability distribution of topics per transcript (one value per topic per transcript)
Sentiment	Overall sentiment mean, standard deviation, minimum and maximum; overall anger mean, standard deviation, minimum and maximum
Word	Unigrams, for all words that appeared in at least 20 of the transcripts, regardless of speaker; the features were the normalised counts of each word
N-gram	As word, but including unigrams, bigrams and trigrams

Table 1: Feature sets for classification experiments

an improvement; scores of 0 or lower indicate no change or a worsening of PHQ score. Each outcome indicator was tested with different feature sets using 10-fold cross-validation.[3]

4 Results

4.1 Correlations

First, we examined statistical associations between our four outcome measures and our available features (see Section 3). R-values are shown for all significant correlations (at the $p < 0.05$ level) in Tables 2-4. For the *PHQ now* measure, a positive correlation means a greater value of the feature is associated with a greater value of the PHQ score (i.e. a higher level of symptoms). For the *PHQ start-now* measures, a positive correlation means that a greater value of the feature is associated with a greater improvement in the PHQ score since the start of treatment. Correlations greater than ± 0.2 are shown in bold.

High-level With patients with a worse (higher) PHQ score (*PHQ now*), more words and turns are typed by both participants. Better overall progress scores are also weakly associated with the amount of talk, with fewer turns typed by both participants if patients' PHQ score has improved by a greater amount since the start of their treatment program (see Table 2).

Sentiment As shown in Table 3, more negative sentiment expressed in the transcripts (mean and minimum), a higher variability of sentiment between negative and positive (s.d.), and greater levels of anger (mean and maximum) are associated with worse PHQ scores. More positive sentiments (mean and maximum) are also associated with better progress.

Topic Topics 2, 6, 9, 10, 16 and 17 are negatively correlated with PHQ scores, i.e. higher levels of these topics are associated with better PHQ (see Table 4). Some of these topics involve words related to assessing the patient's progress and feedback, e.g. topic 2 includes *session*, *goals* and *questionnaires*, and topic 17 includes *good*, *work* and *positive*. Others relate to specific concerns of the patient, e.g. topic 6 (*worry*, *worrying* and *problem*) and topic 16 (*anxiety*, *fear* and *sick*). The top twenty words assigned to each topic by LDA, and the direction of significant correlations are shown in Table 5.

Conversely, topics 4, 5, 7, 8, 11 and 18 are positively correlated with PHQ scores, meaning more talk assigned to these topics is associated with worse PHQ. Several of these topics relate to specific issues, such as topic 5 (*sleep*, *bed*, *night*) and topic 18 (*eating*, *food*, *weight*). Some of these topics display overlap with the previous group (e.g. topics 2 and 4 both contain words reviewing progress such as *session*, *week*, *next* and *last*); this suggests that some topics (e.g. progress or particular issues) are discussed in importantly (and recognisably) different ways or contexts (possibly different emotional valences – see below), and these differences are being identified by the automatic topic modelling.

Similarly, greater amounts of talk in topics 2, 15 and 17 are weakly associated with better progress. These are the topics identified above as involving words related to assessing progress, and feedback. Greater amounts of talk in topic 8 (*checking*, *OCD*, *anxiety*, *rituals*) is associated with worse progress.

Cross-correlations between topic and sentiment features Previous work has hypothesised that automatically derived topics may differ from hand-coded topics in picking up additional factors of the communication such as valence (Howes et al., 2013). To explore this on a global level (i.e.

[3] We partition the data into 10 equal subsamples, and use each subsample as the test data for a model trained on the remaining 90%. This is repeated for each subsample (the 10 folds), and the test predictions collated to give the overall results. This partitioning is done by transcript: different transcripts from the same patient may therefore appear in training and test data within the same fold; our use of low-dimensional topic/sentiment features should minimise over-fitting, but future work will investigate the extent of this effect.

Measure	PHQ now	PHQ start-now
Agent number of words	**0.231**	
Client number of words	0.195	
Agent number of turns	0.149	-0.080
Client number of turns	0.193	-0.071

Table 2: Significant correlations of high-level features and outcomes

Measure	PHQ now	PHQ start-now
Sentiment mean	**-0.237**	0.119
Sentiment s.d.	0.161	
Sentiment minimum	-0.167	
Sentiment maximum		0.074
Anger mean	0.185	
Anger s.d.	0.074	
Anger minimum		
Anger maximum	0.192	

Table 3: Significant correlations of sentiment features and outcomes

at the level of the transcript, rather than at the finer-grained level of the turn) we examined cross-correlations between sentiment and topic. This initial exploration offers support for this hypothesis, as can be seen in Table 6. For example, topics 3 and 4 both contain words relating to feelings and thoughts, but topic 3 is positively correlated with sentiment, while topic 4 is negatively correlated. These correlations indicate a complex relationship between topic and sentiment which should be explored further in future research; a joint topic-sentiment model might be appropriate e.g. (Paul et al., 2013). Although some topics pattern consistently with sentiment (e.g. topic 12, with words about relatives and relationships, is associated with negative sentiments and higher levels of anger) some do not (e.g. topic 19 is associated with more positive sentiment, but greater anger). Examination suggests that this topic involves discussions *about* feelings of anger, but not necessarily *expressing* anger, and also may include talk on how to deal with such feelings (with words like *assertive*). These observations may indicate that in this domain, in which people explicitly talk about their feelings, fully accurate sentiment and emotion analysis may require a different approach than in domains such as social media analysis.

4.2 Classification experiments

Results of classification experiments on different feature sets are shown in Tables 7-9. For each experiment, the weighted average f-score is shown, with the f-score for the class of interest shown in brackets. For *PHQ now* the class of interest is patients with high (moderate to severe) PHQ-9 scores; for *PHQ start-now* we are concerned with patients who are *not* getting better. As a baseline, the proportion of the data in the class of interest in each case is shown in the first column in Table 7 – note that these are not exactly 50%, but reflect the actual proportions in the data (see Section 3.5).

High-level As can be seen in Table 7, if we use a feature set consisting of high-level features and AgentID, we are able to predict *PHQ now* and *PHQ start-now* reasonably well (> 0.7). However, given the nature of the data, it is uncommon for a therapist to have many clients of the same age group and gender; these features can therefore act as a reasonable proxy for identifying individual patients, meaning that this result is somewhat spurious. Also, although identity of therapist is an important factor in therapeutic outcomes (McCabe et al., 2013a; McCabe et al., 2013b), we would like to identify aspects of the communication that explain *why* particular therapists are more successful than others, and generalise our findings to new therapists. AgentID was therefore removed in all subsequent experiments.

Sentiment and topic As shown in Table 8, using the proportions of derived topics by transcript as features does allow us to predict whether a patient has a high *PHQ now* score reasonably well; but sentiment alone performs poorly. Combining sentiment and topic features, however, allows us to predict *PHQ now* with scores of around 0.7 (i.e. approaching the accuracy achieved using high-level and AgentID features above). Prediction of the progress measure is less effective.

Words and n-grams For the symptom measure, using words and n-grams gives f-scores in

Measure	PHQ now	PHQ start-now
Topic 2	-0.157	0.112
Topic 4	0.124	
Topic 5	0.176	
Topic 6	-0.117	
Topic 7	**0.217**	
Topic 8	0.093	-0.126
Topic 9	-0.077	
Topic 10	-0.149	
Topic 11	0.140	
Topic 12	0.080	
Topic 15		0.072
Topic 16	-0.112	
Topic 17	**-0.211**	0.079
Topic 18	0.121	

Table 4: Significant correlations of topic features and outcomes

Topic	PHQ +/-	Sentiment +/-	Anger +/-	keywords
Topic 0		-	+	good thought re well also mindfulness hw thoughts now vc maybe prob message neg just wk one self bit
Topic 1				people good others self evidence thought enough wrong negative esteem thinking say confidence beliefs person true someone belief situation
Topic 2	-	+	-	session send goals next week last sent read great think questionnaires also homework goal appointment set time cbt able
Topic 3		+		thoughts thinking unhelpful helpful look thought behaviours go feelings may think anxiety negative try aware behaviour agenda start self
Topic 4	+	-		feel think like just good really week now know last session next say felt people thoughts going feeling bit
Topic 5	+	-	+	sleep bed day week work get night mood time diary see better much sleeping activity house routine done activities
Topic 6	-			worry worrying worries bit stop train worried problem go example idea control hierarchy driving exposure home happen worst car
Topic 7	+	-		help feel gp depression thank understand therapy now feeling life today think problems able little message medication sorry make
Topic 8	+			check checking ocd thoughts anxiety try something difficult danger brain week sense threat helpful away rituals anxious elephant images
Topic 9	-		-	think time like much way sure see though know look lot sounds well also right thing sorry sense different
Topic 10	-	+		thought thoughts anxiety really situation situations one week next example social experience record great emotions thanks notice see make
Topic 11	+		+	things get time go need like want now just something feel know one work good day going give next
Topic 12	+	-	+	mum relationship husband life family dad parents never love feelings children said years mother much hard way told sister
Topic 13				really week think appointment homework however lets teeth questions great just ready start may dentist set end sure therapy
Topic 14		+	-	great right sure appointment just thank well tonight loo lol good say really cool get going sorry transcript absolutely
Topic 15		+	-	things like get bit good sounds feeling also something really great today think idea send week useful anything make
Topic 16	-		-	anxiety panic breathing get anxious feeling going go attack fear physical control try happen sick symptoms times cope distraction
Topic 17	-	+	-	good work well positive back help really time still last much weeks use thanks session better keep done things
Topic 18	+			eating eat food weight day week meal lunch dinner pie energy good mum put table public walk believe ate
Topic 19		+	+	work job anger angry school stress thanks wife team stuff issues also boss year assertiveness assertive meeting kids times

Table 5: Top 20 words per topic

line with those using only the reduced dimensionality of sentiment and topic. This is surprising; one might expect finer-grained lexical features (which provide more information via a much higher-dimensional feature space) to increase predictivity, as per Howes et al. (2013); on the other hand, it is also promising as it suggests that meaningful generalisations can be drawn out of this data using NLP techniques.

For the progress measure, on the other hand, n-gram features perform better than topic/sentiment (though not as well as on the symptom measures); this suggests that there are aspects of the communication that can assist in predicting patient progress, but that they are not captured by the topic and sentiment information as currently defined. This suggests that dialogue structure or style may play a role; one possibility for exploration is to look at topic and/or sentiment at a finer-grained level and examine their dynamics (e.g. are positive sentiments expressed near the start or end of a consultation linked to better progress)?

5 Discussion

Standard topic, sentiment and emotion modelling can be usefully applied to online text therapy dialogue, although care is needed choosing and applying a technique suitable for the idiosyncratic language and spelling. The resulting information allows us to predict aspects of symptom severity and patient progress with reasonable degrees of accuracy (similar to those achieved with face-to-face data (DeVault et al., 2013; Howes et al., 2012)), without requiring knowledge of therapist identity. However, some measures of patient progress are predicted better with fine-grained, high-dimensional lexical features, suggesting that insight into style and/or dialogue structure is required, beyond simple topic or sentiment analysis.

	Sentiment				Anger			
Measure	mean	s.d.	min	max	mean	s.d.	min	max
Topic 0	-0.083	0.189	**-0.234**	**0.206**	**0.329**	**0.343**	-0.144	**0.267**
Topic 1				0.087		0.083		
Topic 2	**0.245**	-0.180	**0.202**	-0.135	-0.175	-0.109	0.076	-0.176
Topic 3	0.113	**-0.213**	0.159	-0.135		-0.123	0.110	0.095
Topic 4	**-0.350**	**0.324**	**-0.201**	0.099		0.074		
Topic 5	-0.079				0.119			
Topic 6				0.068				
Topic 7	-0.083			-0.167		-0.109	0.110	
Topic 8		0.078		0.123			-0.104	
Topic 9		-0.072			-0.071		-0.075	
Topic 10	0.100	-0.167	0.133	-0.073				
Topic 11		0.086			0.161	0.132		0.121
Topic 12	**-0.338**	0.182	-0.156		**0.233**	0.092	-0.087	0.146
Topic 13		-0.111		-0.112		**-0.243**	0.077	-0.089
Topic 14	0.112	0.156	-0.183	0.186	-0.087	0.225	-0.116	0.204
Topic 15	0.140	-0.179	0.072	-0.064	-0.161	-0.156		-0.070
Topic 16					-0.090	-0.089	0.073	-0.115
Topic 17	**0.385**	-0.156	**0.267**	-0.116	**-0.408**	-0.139	0.078	**-0.288**
Topic 18						-0.071		
Topic 19	0.177				**0.209**			

Table 6: Significant correlations between topic and sentiment features

	Baseline	Agent		High-level (H/L)			
Measure	Proportion	OneR	(Worse)	inc Agent J48		exc Agent J48	
PHQ Now	*40.5%*	0.584	(0.360)	**0.738**	(0.637)	0.640	(0.561)
PHQ Start-now	*38.1%*	0.639	(0.446)	**0.707**	(0.611)	0.545	(0.299)

Table 7: Weighted average f-scores of outcomes using different high-level feature groups (figures in brackets are the f-scores for the class of interest; i.e. *PHQ Now* – patients with higher/more symptomatic PHQ; *PHQ Start-now* – patients showing no change or a worsening in PHQ)

		Sentiment				Topic				Sentiment + Topic			
		inc H/L		exc H/L		inc H/L		exc H/L		inc H/L		exc H/L	
J48	PHQ Now	0.625	(0.528)	0.610	(0.437)	0.642	(0.548)	0.650	(0.512)	0.641	(0.544)	0.638	(0.522)
	PHQ Start-now	0.630	(0.412)	0.508	(0.094)	0.628	(0.479)	0.477	(0.024)	0.619	(0.474)	0.526	(0.147)
Logistic Regr.	PHQ Now	0.626	(0.497)	0.610	(0.432)	**0.689**	(0.585)	0.658	(0.537)	**0.707**	(0.613)	**0.674**	(0.559)
	PHQ Start-now	0.532	(0.218)	0.605	(0.025)	0.593	(0.369)	0.569	(0.283)	0.591	(0.377)	0.557	(0.295)

Table 8: Weighted average f-scores using sentiment/topic features (figures in brackets are the f-scores for the class of interest)

	Words				N-grams			
Measure	inc H/L		exc H/L		inc H/L		exc H/L	
PHQ NOW	0.655	(0.575)	0.676	(0.614)	**0.696**	(0.615)	**0.686**	(0.616)
PHQ Start-now	0.616	(0.528)	0.623	(0.506)	0.626	(0.459)	**0.645**	(0.532)

Table 9: Weighted average f-scores using raw lexical features (words/ngrams) using LibLINEAR (figures in brackets are the f-scores for the class of interest)

References

D. Angus, B. Watson, A. Smith, C. Gallois, and J. Wiles. 2012. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLoS ONE*, 7(6):1–12.

A. Beattie, A. Shaw, S. Kaur, and D. Kessler. 2009. Primary-care patients' expectations and experiences of online cognitive behavioural therapy for depression: a qualitative study. *Health Expectations*, 12(1):45–59.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

C.-C. Chang and C.-J. Lin, 2001. *LIBSVM: a library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Z.-J. Chuang and C.-H. Wu. 2004. Multi-modal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, 9(2):45–62, August.

J. Cretchley, C. Gallois, H. Chenery, and A. Smith. 2010. Conversations between carers and people with schizophrenia: a qualitative analysis using Leximancer. *Qualitative Health Research*, 20(12):1611–1628.

M. De Choudhury, M. Gamon, and S. Counts. 2012. Happy, nervous or surprised? Classification of human affective states in social media. In *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM)*.

D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. S. Rizzo, and L.-P. Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202.

J. Eisenstein and R. Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGDKDD Explorations*, 11(1):10–18.

T. Hanley and D. Reynolds. 2009. Counselling psychology and the internet: A review of the quantitative research into online outcomes and alliances within text-based therapy. *Counselling Psychology Review*, 24(2):4–13.

R. Hirschfeld, R. L. Spitzer, and M. R.G. 1974. Computer diagnosis in psychiatry: A Bayes approach. *Journal of Nervous and Mental Disease*, 158:399–407.

C. Howes, M. Purver, R. McCabe, P. G. T. Healey, and M. Lavelle. 2012. Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*.

C. Howes, M. Purver, and R. McCabe. 2013. Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical Informatics Insights*, 6(Suppl. 1):39–50, July.

P. John, M. Lavelle, S. Mehnaz, and R. McCabe. under review. What do psychiatrists and patients with schizophrenia talk about and does it matter? *Psychiatric Bulletin*.

D. Kessler, G. Lewis, S. Kaur, N. Wiles, M. King, S. Weich, D. Sharp, R. Araya, S. Hollinghurst, and T. Peters. 2009. Therapist-delivered internet psychotherapy for depression: a randomised controlled trial in primary care. *Lancet*, 374:628–634.

K. Kroenke and R. L. Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann*, 32(9):1–7.

R. e. a. Layard. 2012. How mental illness loses out in the NHS. Technical report, Mental Health Policy Group, Centre for Economic Performance, London School of Economics, June.

M. Liakata, J.-H. Kim, S. Saha, J. Hastings, and D. Rebholz-Schuhmann. 2012. Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical Informatics Insights*, 5(1):175–184.

A. Martin, W. Rief, A. Klaiberg, and E. Braehler. 2006. Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. *General hospital psychiatry*, 28(1):71–77.

R. McCabe, P. G. T. Healey, S. Priebe, M. Lavelle, D. Dodwell, R. Laugharne, A. Snell, and S. Bremner. 2013a. Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia. *Patient Education and Counselling*.

R. McCabe, H. Khanom, P. Bailey, and S. Priebe. 2013b. Shared decision-making in ongoing outpatient psychiatric treatment. *Patient education and counseling*, 91(3):326–328.

A. K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th AAAI International Conference on Weblogs and Social Media*, pages 122–129.

L. Ong, J. De Haes, A. Hoos, and F. Lammes. 1995. Doctor-patient communication: a review of the literature. *Social science & medicine*, 40(7):903–918.

J. Overall and L. Hollister. 1964. Computer procedures for psychiatric classification. *Journal of the American Medical Association*, 187:583–585.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

M. Paul and M. Drezde. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

M. Paul, B. Wallace, and M. Dredze. 2013. What affects patient (dis)satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. In *Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

J. W. Pennebaker, R. J. Booth, and M. E. Francis. 2007. Linguistic inquiry and word count (LIWC): A computerized text analysis program. Austin, TX: LIWC.net.

R. H. Perlis. 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, 74(1):7–14. Sources of Treatment Resistance in Depression: Inflammation and Functional Connectivity.

M. Purver and S. Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491.

M. Purver, K. Körding, T. Griffiths, and J. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 17–24.

D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. 2011. In the Mood for Being Influential on Twitter. In *Proceedings of the 3^{rd} IEEE Conference on Social Computing (SocialCom)*.

D. Quercia, J. Crowcroft, J. Ellis, and L. Capra. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 965–968.

G. Salton and M. McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Y.-S. Seol, D.-J. Kim, and H.-W. Kim. 2008. Emotion recognition from text using knowledge based ANN. In *Proceedings of ITC-CSCC*.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, 166(10):1092–1097.

M. Steyvers and T. Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

B. C. Wallace, T. A. Trikalinos, M. B. Laws, I. B. Wilson, and E. Charniak. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1775.

H. M. Wallach, D. M. Mimno, and A. McCallum. 2009. Rethinking LDA: Why priors matter. In *NIPS*, volume 22, pages 1973–1981.

Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell. 2013. Multimodal prediction of psychological disorder: Learning nonverbal commonality in adjacency pairs. In *Proceedings of the SemDial 2013 Workshop*, pages 193–202.