HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2018 September 01.

Published in final edited form as:

IEEE J Biomed Health Inform. 2017 September ; 21(5): 1449–1459. doi:10.1109/JBHI.2016.2601123.

Predicting Social Anxiety Treatment Outcome based on Therapeutic Email Conversations

Mark Hoogendoorn[1,2], **Thomas Berger**[3], **Ava Schulz**[3], **Timo Stolz**[3], and **Peter Szolovits**[2]

[1]VU University Amsterdam, Department of Computer Science, De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands [2]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, 32-254, Cambridge, MA 02139, USA [3]Universität Bern, Klinische Psychologie und Psychotherapie, Fabrikstrasse 8, 3012 Bern, Switzerland

Abstract

Predicting therapeutic outcome in the mental health domain is of utmost importance to enable therapists to provide the most effective treatment to a patient. Using information from the writings of a patient can potentially be a valuable source of information, especially now that more and more treatments involve computer-based exercises or electronic conversations between patient and therapist. In this paper, we study predictive modeling using writings of patients under treatment for a social anxiety disorder. We extract a wealth of information from the text written by patients including their usage of words, the topics they talk about, the sentiment of the messages, and the style of writing. In addition, we study trends over time with respect to those measures. We then apply machine learning algorithms to generate the predictive models. Based on a dataset of 69 patients we are able to show that we can predict therapy outcome with an Area Under the Curve (AUC) of 0.83 halfway through the therapy and with a precision of 0.78 when using the full data (i.e., the entire treatment period). Due to the limited number of participants it is hard to generalize the results, but they do show great potential in this type of information.

I. Introduction

Mental disorders are highly prevalent (see e.g., [1]), but only a small proportion of individuals in need receive treatment (see e.g., [2]). One solution to the many challenges faced by mental health care is the use of new technology such as the Internet. Internet-based interventions can be delivered at low cost to large populations and they can be used flexibly and accessed easily. Research on Internet interventions has grown rapidly during the past 10 years, and more than 100 controlled trials evaluating Internet-based treatments in the field of anxiety disorders, mood disorders and other conditions have found promising outcomes (see e.g., [3], [4]).

Although the efficacy of Internet interventions has been shown in a variety of research papers, the domain still suffers from a lack of predictors to determine the expected

effectiveness of therapy for *specific* patients. Given the newly developed treatments, which frequently record a lot of data about the behavior of the patient, the road to more accurate predictive models for individual patients has been opened. Many of these treatments however contain interactions with the system or the therapist that involve free text (e.g., written assignments, email or chat conversations with the therapist). This data is less straightforward to use for predictive modeling although it is potentially a valuable source of information.

In the area of Natural Language Processing (NLP) ample approaches have been proposed to extract meaningful information from text. Examples from the medical domain include [5], [6], [7], [8] and [9]. Typical approaches include counting words, identifying topics, and coupling the terms to a domain specific ontology. Using this structured knowledge extracted from the text, machine learning algorithms can be applied to generate predictive models. However, very few approaches have been applied in the mental health domain. The findings of the published few studies in the mental health domain (see e.g., [10], [11], and [12]) do show that this is a promising avenue.

In this paper, we study predictive modeling of social anxiety symptoms. The dataset originates from a clinical trial in which patients engaged in a self-help treatment program and were supported by a secured email facility to interact with a therapist [13]. The therapist pro-actively approached the patients on a weekly basis. The treatment covered a period of twelve weeks in total and the dataset covers 69 patients, all diagnosed with a social anxiety disorder. We try to predict a successful therapeutic outcome (according to a significant improvement in the Social Phobia Measure [14] from the start to the end of therapy) at three points in time: (1) at the start of the therapy by means of socio-demographic data; (2) halfway through the therapy (6 weeks) by using the socio-demographic data and the emails sent by the patient up to that time point, and (3) at the end of the therapy by using the socio-demographic data and all email data originating from the patient. To extract useful predictors from the emails, we deploy a range of techniques, including basic emailing behavior (e.g., response time, length of emails), word usage, writing style, sentiment, and topic modeling. We not only look at the average score for these predictors, but also study their trends over time.

This paper is organized as follows. We start by exploring related work in more detail in Section II. The methodology is described in three sections: a more extensive description of the dataset is given in Section III, an explanation of the approach used to extract predictors from the email conversations is in Section IV, and the experimental setup is described in Section V. The results are presented in Section VI. Finally, Section VII is a discussion.

II. Related Work

Within the medical domain, a substantial amount of work has been done aimed at the processing of medical notes. A variety of approaches have been proposed including systems that extract medical terms from text (e.g., MetaMap [15], the Health Information Text Extraction system (HITEx, cf. [16]), and cTAKES [17]). More tailored approaches have been used for specific diseases, such as [5] using the HITEx system to enrich coded data

with terms extracted from physician notes for Rheumatoid Arthritis, [6] using topic modeling on medical notes for ICU mortality prediction, and [7] using subgraph identification in medical texts to identify useful predictors. All these approaches show that notes can truly complement other (more structured) data that is available within the medical domain. Note that all systems have been developed for the English language.

These projects have focused on notes written by medical professionals, unlike the subject of study in this paper, where we try to distill information from email written by patients. There are however some exceptions to this focus on professional writing. In [10] suicide notes (written in English) of patients are analyzed using NLP techniques. Their main aim was to distinguish genuine suicide notes from notes written by healthy subjects that were asked to write a note as if they were going to commit suicide. The authors see this as a first step towards modeling the risk of repeated suicide attempts. The results showed a correct ability to classify notes in 73% of the cases using an NLP machine learning setup versus 63% obtained by medical experts. Features they used include the part-of-speech tags, readability of the notes, and emotional values associated with the words used. In [11] essays written by college students at the University of Texas at Austin were analyzed using NLP techniques. They aimed to differentiate between essays written by students who were depressed at the time of writing, those who were not depressed and had never been depressed before, and those who had a history of depression but did not fulfill the criteria at the time of writing. They showed, for a group 124 participants, that the depressed students used the word “I” and words with a negative valence more frequently than the other groups. To extract these measurements from the text they used the Linguistic Inquiry and Word Count text analysis program developed by one of the authors. A more extensive description of the program, including a more extensive evaluation, can be found in [18].

A number of more manual approaches to extracting meaningful features from text have also been described. For anxiety, in[19] (in 1960!) showed the importance of word usage (in English), as he was able to distill an anxiety score by using an adjective check list. In [12] it is shown that an increased usage of what they refer to as “discrepancy words” (wordings referring to discrepancies between actual and ideal selves, e.g., should, wish, hope) was predictive for depression improvement. The language of the writings was Dutch. Adherence to the treatment was found to be related to more words used on the application form, more social words, and fewer discrepancy words used in general.

III. Dataset Description & Preparation

The dataset we study in this paper derives from 69 patients, all diagnosed with a social anxiety disorder, who participated in a randomized controlled trial on an internet-based guided self-help intervention for social anxiety disorder [13]. The self-help program was based on the established cognitive-behavioral model by Clark and Wells [20] and has been shown efficacious in previous studies ([21], [22], [23]). While participants worked their way through the self-help program, therapists assisted and supported them via email. Once a week, therapists wrote an email with feedback to the participants on their behavior and progress in the self-help program. The most important aspects of this feedback were recognition and reinforcement of the participants’ independent work with the program.

Participants were informed that they could contact their therapist whenever they wanted to. In case the participants wrote an email, therapists were instructed to answer within at most 3 days. The treatment covered a period of 12 weeks in total, supported by four female therapists. Participants were recruited in Switzerland, Austria and Germany through a study website and postings in internet forums. Thus, the sample was a self-selected group of individuals who had expressed interest in the treatment. For more details on the participants, see [13]. As an outcome measure the Social Phobia Measure is taken [14]. Since we are interested in differentiating between patients who recover and those who do not (as this is most valuable for therapists) we have created two categories of outcomes, namely patients who show a reliable improvement in the aforementioned measure, and patients who do not. Reliable change was determined according to the Reliable Change Index [24] by using the retest reliability reported for the German version of the Social Phobia Measure. Out of the 69 patients, 48 showed a reliable improvement, whereas 21 did not. See [13] for more details on the trial setup. Of course, an alternative would have been to predict the actual Social Phobia Score, but given the limited size of the dataset we felt this would be too fine grained to allow creation of any meaningful models.

The data available to us includes socio-demographic data of the patients, the emails sent by the patients and therapists including their time stamp and by whom each was sent (i.e., patient or therapist), and the outcome indicating whether reliable improvement was achieved. Note that we only focus on the content of the emails written by the patient in this research. Identifying information has been removed from the socio-demographic data, and the emails have been anonymized by hand. Table I shows more detail on the socio-demographic data that is available.

IV. Methodology

Our main goal is to generate models that allow prediction of therapeutic outcomes based on the socio-demographic data combined with the text of the emails obtained from the patient during treatment. Using the socio-demographic information in our predictive models is easy as they can just be considered attributes for our learning algorithm, although we do transform the nominal attributes (e.g., marital status) into binary attributes. In order to cope with the text of emails in an effective way however, dedicated algorithms are needed to transform raw text to usable attributes. This process and the techniques used are described in this section. We use a variety of different algorithms, each covering a specific aspect of an email. Essentially, we focus on the following aspects:

- **basic mailing behavior** focusing on the more straightforward metrics such as the length of emails written by the patient, whether the patient regularly responds to emails sent by the therapist, and how long such a reply takes.
- **word usage** covering the words being used in the email. Some words might be distinctive for whether the patient is on the right track or not.
- **writing style**, in which we focus on the grammatical style used by the patient. Particular ways of phrasing may be indicative of increased chances of recovering.

- **sentiment** addresses the positivity or negativity of the message conveyed in the email. We anticipate that more positive phrasing would point at recovery or therapeutic effectiveness.
- **topic**, specifying at a higher level of abstraction what topic an email concerns.

Once we have identified these attributes per email, we can aggregate them using simple averages over the emails (e.g., on average patient *n* uses the word *future* five times in an email). In addition, we look at changes of these values over time by means of trends (e.g., an increasing trend can be seen for patient *n* in the usage of the word *future*). Below, we expand on each of these aspects in more detail and explain precisely how we have operationalized them.

A. Basic Mailing Behavior

The first set of attributes concerns the basic mailing behavior of the patient, focusing on the length of emails, response rates, and response times. For the length of the email, we tokenize the text using the Python NLTK toolkit [25] and count the number of words identified. To identify whether an email from a therapist received a response, we examine all the emails between a patient and therapist in time order, starting with the first email sent by the therapist. When we encounter an email sent by the therapist, we check whether the next email in the ordered sequence is an email from the patient. If so, we consider it to be a reply to the therapist's email, and if not, we assume the reply to be missing. Note that this is the only way in which we use the email sent by the therapist. In all other approaches described below we only consider the emails sent by the patient.

B. Word Usage

The next set of attributes concerns the content of the email, and more specifically the usage of particular words by the patient. For this, we tokenize the text per email, stem the words, remove the stop words using a standard stop word list (from the NLTK toolkit), and perform counts of the words that remain. We then consolidate the words identified in the set of all emails and use this as the set of attributes. The number of occurrences of a word in an email is the value of the attribute associated with that word for the specific email. We also tried normalizing the value based on the length of the email, but experiments showed performance was substantially lower when normalization was applied. Due to the limited size of our dataset we want to avoid words that are very specific to a single or just a few patients. To handle this, we apply a criterion that a word is only considered when it occurs in emails of a sufficiently high number of distinct patients (see Section V for the precise setting). An alternative would have been to use the TF-IDF score.

C. Writing Style

We analyze the writing style of an email by means of the relative occurrence of specific part-of-speech tags. Since we are facing emails written in German, we use the pattern.de Python toolkit*. We iterate though the email and parse each sentence using the toolkit (after having identified the sentences using the NLTK sentence tokenizer). It assigns highly detailed tags

*http://www.clips.ua.ac.be/pages/pattern-de

to each word in the sentence based on the guidelines proposed in [26], e.g., a variety of different types of adjectives, nouns, adverbs, verbs, and pronouns. This provides a total of 36 part-of-speech tags. We then count the occurrence of each type in the email and normalize it based on the length of the email.

D. Sentiment Analysis

While the identification of the overall sentiment of an email deserves a paper on its own, we apply a relatively simple approach to identify the sentiment of the email. We apply three different heuristic schemes taken from [27]. We start with identifying the sentiment scores of the sentences present in the email. This value is determined based on the sentiment scores of each of the stemmed words in the sentence. We use a German sentiment annotated corpus of words which combines the work reported in [28], [29], and [30]†. We take the word sentiment score from [29]. If the word is not present in the corpus we take the value from [30] and if it is missing again we take the value from [28]. If the word is not present in any of the corpora we do not consider the word in our calculations. Once we have obtained the value for the words in the sentence we follow three approaches (in line with [27]):

- **Voting** We assign a positive sentiment to the sentence if the majority of the words is positive (and vice versa). In case there is a tie we ignore the sentence. In addition, we ignore the sentiment of negation words in this voting scheme (in German the negation words considered are *nicht* and *keine*).
- **Neg(1)** In this approach we *do* take the presence of negations into account. In the Neg(1) approach if at least one negation is present we flip the sentiment of the sentence following from the voting scheme.
- **Neg(n)** Instead of considering at least one occurrence of negation, in the Neg(n) approach we count all negations and flip the sentiment of the majority vote in case of an odd number of negations. Note that this is a huge simplification.

Although the approach described above is rather simplistic, in [27] this approach has been shown to be competitive with more sophisticated heuristics that take more information about the structure of the sentence into account.

Once we have obtained the score per sentence, we assign the number of sentences with a positive sentiment divided by the total number of sentences for which we were able to derive a sentiment score as the sentiment score of the email.

E. Topic Modeling

Finally, we apply topic modeling, and more specifically the parameterized Latent Dirichlet Allocation (LDA) approach (cf. [31]). These topics are expressed by the words that are covered by the topic and a weight is given to each of the words. Given the words in the email, a score for each of the topics can be assigned based on the weights assigned to the words for each specific topic. This hopefully provides a more high-level view of the email. Each topic forms an attribute and the value assigned for an email is the score on that

†downloaded from: https://sites.google.com/site/iggsahome/downloads

particular topic. In order to get useful topics, we filter the opening and closing sentences of the emails, remove stop words, and removed three words that were not part of the stopwords list but were considered to be stopwords (imm (the stemmed variant of immer: always), fur (for), and dass (that)).

F. Aggregation

Now that we have obtained scores for individual emails, we can start aggregating the values to be able to apply machine learning algorithms. We assume that we investigate a specific time interval $[t_{start}, t_{end}]$, for instance the first 6 weeks of therapy, and want to determine a score *per patient* for each attribute during this interval. Hence, to determine the score for the patient, we consider each email sent by the patient at a time point t_i that satisfies the condition $t_{start} \leq t_i \leq t_{end}$. For each of the selected emails, we determine the value for each attribute a_j, where the value at t_i is referred to as $a_j(t_i)$. Assuming n ordered emails and their associated timestamps that satisfy this condition, we now end up with a set of time points $\{t_0, \ldots t_n\}$ and values $\{a_j(t_0), \ldots, a_j(t_n)\}$. We consider two options for aggregation of the attribute a_j for the patient:

- **Average**: we simply sum the values up and divide the by number of emails, i.e. $\frac{a_j(t_0)+\ldots+a_j(t_n)}{n}$
- **Trend**: we perform a linear regression using the time points and values identified and use the coefficient of the resulting equation as the value for the attribute, expressing an increase or decrease in the score over time.

A summary of the metrics used is shown in Table II. The precise setting for the time intervals used in our experiments is explained in the next section.

V. Experimental Setup

This Section explains the approach followed to generate and evaluate predictive models based on our previously explained methodology, including the selected parameter settings.

A. Approach

The ability to predict outcomes early in the course of therapy can be valuable because it can allow therapists to adjust therapy as soon as possible to improve the chances of a speedy recovery. We consider predicting the outcome of therapy (i.e., whether there was a reliable improvement during the course of therapy in the Social Phobia Measure) at three different time points from the features described in Section IV. These are:

1. at the *start* of the therapy (i.e. only having the socio-demographic data (*SD*) available).
2. after the *first 6 weeks* of therapy (i.e. using only *SD* and the emails exchanged during the first 6 weeks)
3. after the *full 12 weeks* of therapy (i.e. using all data).

At the start, each patient is represented by a vector of the socio-demographic data (*SD*). At each of the subsequent times, we represent each patient by that vector extended by additional components that encode either the average (*A*) or the trend (*T*) of each feature derived from mailing behavior, word usage, writing style, sentiment analysis and topic modeling, aggregated over the relevant time period. The outcome is 1 if there was a reliable improvement, 0 otherwise.

Our first step in model building is to identify which of the large set of aggregated attributes in the patient vectors correlates strongly enough with outcome to be included in the model. For each time point, we consider only the top k attributes with the highest Pearson correlation to outcome, excluding attributes that were highly correlated (Pearson correlation ≥ 0.70) with other attributes previously selected. Initial experiments had suggested that such pre-selection resulted in models with the best performance.

We then apply a variety of machine learning algorithms, including a decision tree learning algorithm (CART, cf. [32]), logistic regression, and random forests [33], selected based on their ability to deliver insightful models with (reasonably) good performance. We build models using the various machine learning methods based on solely *SD*, *SD* and *A*, and *SD* and *T*, with data aggregated to the start, midpoint and end of therapy. (Clearly, there are no *A* or *T* data at the start, and models based on only *SD* data are equivalent at every time point.)

To train and test the models, we use five fold cross validation. The average Area Under the Receiver Operating Characteristic (ROC) Curve over the five folds is used as the primary score of the models. In addition, we calculate the average precision, recall, and F1 score over the five folds. We produce these scores for a threshold value for each of the models in the five folds that yields a false positive rate of 0.4, as this is the point where the ROC curve generally starts to flatten.

B. Parameter Settings

There are a number of tunable parameters within our approach.

- For topic modeling, we choose 25 topics to use in characterizing the email topics, based on initial experimentation.
- We choose $k = 20$ as the number of the attributes most highly correlated to outcome that should be included in our models, determined by rigorous experimentation. Although this choice is not ideal for every subset of the data (e.g., for *SD* data alone, a much smaller number would be better), we prefer to choose a uniform value over all experiments, and 20 yielded best overall results.
- We select all words in the *word usage* category that occur in emails of at least four distinct patients (~5% of all patients).
- Least square error is used as the minimization criterion for logistic regression.
- The Gini impurity measure is used as a splitting criterion for CART. We also use a tree depth = 5, with a minimum of four samples per leaf node, given the limited number of patients in the current dataset.

- Random forest uses the same settings as CART, with a forest size of 10 (a relatively low number, but the main goal is to deliver an understandable model).
- Given the unbalance in the dataset, all algorithms generated models using a weighing strategy, giving more importance to less frequently occurring classes, inversely proportional to their occurrence in the data.

Note that the results presented in the next Section are strongly influenced by the choices made above due to the limited size of the dataset. To generate more solid conclusions, a substantially bigger dataset would be required, which is future work.

VI. Results

The experimental results are presented below. First, we focus on an analysis of the correlations between the various attributes and the target (i.e., the first step identified in Section V), followed by the results of the predictive modeling.

A. Analysis

The focus in this analysis is to gain insight into correlations of the metrics that were defined in Section IV and the effectiveness of the intervention. We will focus on the different types of attributes and explore the top correlations. Later, only these top predictors will be used to compose predictive models. We will report Pearson correlations together with their 2-tailed p-values. Note that the sample size used to calculate all correlations scores throughout this section equals the number of patients as we aggregate all data of a patient into a single record and use all patients to compute the scores. In this analysis, independence of the attributes will be assumed, which makes it easier to explore, although this assumption is most likely violated for many attributes. In the learning algorithm, attributes are only considered as predictors if they are not strongly correlated to attributes that have already been selected.

a) Socio-Demographic Data—We first explore the socio-demographic (*SD*) data. As noted above, these data and models based on them do not change over the course of therapy. While a wider range of data is available, Table III shows the five attributes with the most extreme (i.e., most positive or negative) correlation values. It can be seen that all are weak correlations (and this will be true for many of the attributes considered), yet some interesting aspects do arise. A current major depression has a weak positive correlation with recovery. A higher score on the Social Phobia Measure before the intervention is positively correlated with a successful outcome. Furthermore, being female is correlated with a successful therapy as well. Being retired or unemployed reduces the chances of recovery.

As mentioned before, due to the limited dataset combined with the relatively weak correlations, it is difficult to draw sound conclusions from these observations.

b) Basic Mailing Behavior—Table IV shows the correlations between the basic mailing characteristics of the patient and outcome. It presents correlation results for both the 6 and 12 week periods and for models using both averaging and trends. There seems to be a lack of clear and consistent correlations between these attributes and the target, although a higher

response rate also shows a weak correlation to therapy effectiveness. A trend of an increasing email length seems to be positively correlated with recovery when considering the full 12 weeks of data.

To understand why these correlations are relatively weak, consider the distributions of the values as averages taken over the full 12 weeks, visualized in Figure 1. Some of the patients who do not recover did not respond to the intervention at all, showing their possible lack of engagement in the therapy. The response time is more extreme for patients that did not recover: they tend to reply either faster or slower compared to the recovered patients. When it comes to email length, quite a large fraction of patients that do not recover write very brief emails, especially when compared to the recovered patients.

c) Word Usage—Usage of particular words by the patients is considered next. In total 3852 unique words were found before stemming, 3025 after stemming, 2932 after removing the stopwords and 515 after removing the insufficiently frequent words (i.e., retaining words occurring among 4 patients or more) for the 6 week period. For the 12 week period these numbers are 6102, 4651, 4557, and 876, respectively. Table V shows the correlations for both periods and both aggregation approaches. Only the top 3 predictors for the 6 and 12 week period are shown. Note that some predictors seen in the 12 week period might score higher for the 6 week period compared to the selected predictors, but they were excluded because they occur among an insufficient number of patients during the 6 week period. The results shown are highly dependent on the parameter setting of the minimum number of patients. However, to make the result at least somewhat more generalizable, this is considered an important criterion to apply.

The table shows that the words in general exhibit a bit higher correlations to outcome than those features seen in the previous categories (though they are still weak). Writing a lot about “diaries”, in which anxiety related thoughts and behaviors are recorded, and about “computer” proves to be a negative predictor for a successful outcome when considering the full period. A closer look at the usage of these words shows that the diary is often mentioned by participants when they have a problem sticking to this homework assignment. The word “computer” on the other hand is sometimes used in a context in which participants indicate that they are not satisfied with the treatment by writing critically about computers that can not replace humans. For the 6 week period, talking about feelings and the course of the therapy shows a negative correlation to outcome. More positive predictors include words like “to succeed” and “to take part”. When considering the trends, a positive trend in words such as “to take part” and “progressive muscle relaxation” shows positive correlations for the 12 week case while more negative correlations dominate the top for the 6 weeks case, including increased talking about relations to others and about rights and reasons.

d) Writing Style—The part-of-speech tags are an indication of the writing style, and are shown in Table VI. Usage of nouns and foreign words is observed to be weakly positively correlated with recovery. For personal pronouns an unexpected behavior is seen: while it shows a weak positive correlation for the first 6 weeks, it shows a negative correlation over the full 12 weeks. An increased use of adverbs is positively correlated with recovery. Using more adjectives as the therapy progresses is correlated with a negative outcome. For trends

related to determiners and pronouns the same unexpected behavior as described before (i.e., correlation sign changes) are again observed.

e) Sentiment Analysis—Although sentiment analysis was expected to be a promising avenue when considering mental health disorders, the results shown in Table VII demonstrate that the correlations are very weak. Particularly for the full period of 12 weeks, the correlations are negligible, although for the 6 week data a bit stronger correlations can be seen. When looking at the trends, it seems that becoming more positive as the therapy progresses is correlated with a successful outcome. This holds for the first 6 weeks as well as the entire period. There is not much difference between the different sentiment heuristics, although the voting based scheme seems to perform slightly better than the rest. These disappointing results might be caused by the relatively simple nature of the scheme, which is not suited for this specific type of text although it has been shown to work well on other datasets, or it might be specific to the combination of the German language and the approach.

f) Topic Modeling—The final category of attributes considered are those related to topic modeling. The results are shown in Table VIII. The top topics do show a bit stronger correlations compared to the other categories. For the full 12 weeks, topics covering the situation (topic 11, 23, and 24) seem to have a negative correlation with successful outcome, also in combination with fear. When considering the 6 week period, the combination of gratitude and fear is quite clearly negatively correlated with recovery whereas topic 18 (related to feeling good and being grateful) by itself is a negative predictor, but when it comes to an increasing trend it becomes a positive predictor.

To summarize this preliminary analysis, it can be observed that the top predictors of most categories of attributes show weak correlations with recovery whereas some do express a bit more substantial correlations (e.g., the topic modeling), though certainly not strong. Combinations of the attributes could of course still show a good predictive value.

B. Predictive Modeling

Our goal is to generate accurate predictive models for therapeutic success to improve decision support for therapists. The machine learning setup was tested on different time frames and different subsets of attributes. In this section, we present results of building predictive models for outcome from both the full 12 week data and from the first 6 weeks of data. The 12 week models represent an upper bound on how much information these methods can extract from all the data. They are not practically useful because by the end of that period, the actual outcome can be observed. The 6 week models can provide useful guidance for mid-course corrections in treatment. Fortunately, our results suggest that predictions from 6 weeks of data are equally accurate to those from the full 12 week data set, although the best models built at the two times differ in detail.

Table IX shows the results when running the benchmark (i.e., only socio-demographic data, *SD*) versus the results when using all email data for 12 weeks for different dataset compositions and algorithms. The table also shows the confidence intervals of the AUC's

reported, as well as precision, recall and F1 scores (selecting a threshold resulting in a false positive rate of 0.4).

The results show that predicting therapeutic outcome based on socio-demographic data performs essentially equally to random guessing. When considering the email behavior however, an AUC of up to 0.78 is obtained when exploiting all data available: the averages and trends of the email attributes combined with the socio-demographic data using the random forest approach. Overall, the performance of the logistic regression approach is reasonable as well, but the top performing model originates from a random forest with 10 trees. The single decision tree scores worst for most settings. As mentioned before, only 10 trees were selected to keep the model understandable for domain experts (a setting of 100 trees was also tried, but did not produce substantially better results). When considering the precision, recall, and F1 scores (given a threshold resembling a false positive rate of 0.4) a similar ranking can be seen. In the 20 attributes selected in the best model, 17 are related to particular words (8 of which are trends, and 9 averaged), 2 are related to topics and 1 is related to the writing style. Table X shows the predictors used and their correlations with the outcome. The best model performs significantly better than the socio-demographic data, despite the relatively small dataset.

Although these results are interesting, in a therapeutic setting it is important to know early in treatment whether it is best to continue, or whether it is best to change strategy to improve the chances for recovery. Table XI shows the results when using only the first 6 weeks of data. Interestingly, using only part of the data does not strongly influence the results: there is no significant difference between the AUC shown for these 6 weeks (0.83) compared to the full 12 weeks, which is a promising result, although it should be tested on a larger dataset to provide a definite conclusion. The same can be seen when considering the precision, recall, and F1 measures. In this case, the trend data seems most informative while the logistic regression approach performs best using that data. When considering the best model (socio-demographic with trends) out of 20 attributes, 1 was related to socio-demographic data (current major depression), 2 to writing style, 4 to topics, and the remainder of the attributes were trends in words. The predictors are shown in Table XII. It is an interesting observation that for the longer (12 week) period trends themselves do not seem to improve predictive performance while for the shorter (6 week) period the trends are valuable. Finding the cause of this is part of future work.

To conclude, it can be said that the texts help in predicting therapeutic outcome. Word usage seems to be the main contributing factor, and trends are important in both top scoring models. Though the results are promising, a more extensive dataset is needed to fully understand how solid the predictors are that we found in this endeavor.

VII. Discussion

In this paper, we have explored the usage and content of patient emails in the prediction of therapeutic outcome in an internet intervention for patients suffering from social anxiety disorder. Given the current era of computer-based therapies, such data is nowadays widely available yet hardly used. To extract useful predictors from the text-based data, various

approaches have been introduced, ranging from the extraction of basic mailing behavior, word usage, writing style, sentiment, and general topics of emails via topic modeling. As a means to aggregate the values of the different emails of patient, and to facilitate the application of standard machine learning algorithms, averaging of values over all emails belonging to a single patient as well as the identification of trends among the emails written by a patient were used.

Using this approach, reasonably accurate performance can be obtained when applied to the social anxiety dataset of 69 patients we used for this research. This shows the benefit of using the information contained within the emails. While it is difficult to draw sound conclusions on specific predictors (e.g., word usage, writing style) given the size of the dataset and dependence on the specific settings of the parameters of the approach, we did see that some interesting correlations could be observed in all different categories of attributes, which were explainable from a therapeutic perspective. Due to the limited amount of data we cannot make claims on the generalizability of these correlations. We would need to test on an independent dataset which was not available, hence this is a limitation of our current work. Furthermore, we observed that only using the initial 6 weeks of data showed a performance comparable to exploring the full history, and both significantly surpassed using only socio-demographic data. We would expect these results to generalize reasonably well to other (similar) languages such as English. The only part of our method which is very language dependent seems to be the writing style, and these predictors play a minor role. Furthermore, the language could also influence the predictive values of word usage. In order to judge the generalizability over different languages we would need to experiment with multiple datasets that cover different languages.

As illustrated by this research, Internet-based treatments open up new perspectives for the investigation and use of predictors of treatment outcome based on the conversation between patients and therapists. In contrast to face-to-face psychotherapies, in which the dialogue first needs to be recorded and transcribed in order to analyze it, conversations in Internet-based treatments are automatically recorded and immediately available for the analysis and the clinical application of predictive models. Future clinical applications of predictive models based on email conversations could thus continuously monitor emails from patients and provide therapists with immediate feedback and timely warnings about possible treatment failures in order to guide ongoing treatment. Research on face-to-face psychotherapies suggests that therapists are not alert to treatment failure and overconfident in their own clinical judgment [34]. “Real time” psychometric feedback over the course of treatment based on self-report measures has shown its effect on therapy outcome, especially for patients at risk for negative outcome ([35], [36]). In spite of these potential benefits of “real time” feedback, it appears that few clinicians engage in the continuous monitoring of treatment progress of their patient. Hatfield and Ogles [37] directly asked clinicians why they do not use this possibility and many therapists indicated practical reasons, such as “takes too much time” or “adds an extra burden to clients”. Predicting outcome from measures which do not place the burden of repeatedly filling out questionnaires on clients and which can easily be gathered on an ongoing basis such as email conversations may thus be especially promising for future clinical applications of “real time” feedback systems in Internet-based treatments. Moreover, and in contrast to self-report measures, conversations

between patients and therapists provide unobtrusive, non-reactive data. The early identification of negative developments in treatments may be of particular importance in Internet-based interventions because they can be implemented as a step in stepped care models, in which lower cost interventions such as Internet-delivered treatments are tried first, with more intensive and costly interventions such as face-to-face therapies reserved for those patients insufficiently helped by the initial interventions [38]. Since frustration and demoralization of therapeutic failure in a first step treatment is one of the major concerns regarding stepped care frameworks, the early identification of non-responders and an early stepping up to a more intensive treatment could prevent patients from having to undergo an ineffective, time-consuming treatment, and could increase the chance that the more intensive treatment will work [39].

For future work, we want to apply the approach we have used in this paper to a larger dataset in order to find more robust predictors. In addition, we would like to see whether we could expand our range of information we extract from the emails in different directions. We envision that the approach presented in [7] could be an interesting way forward to identify more of the semantics of the emails. In addition, trying to identify the overall complexity of the text (e.g., using the Flesch/Flesch-Kincaid readability test, cf [40]) could be another interesting metric worth exploring, although we do have information about the education level of the patients to start with. We also envision comparing the prediction by our models to those provided by therapists at different stages to see how human predictions compare to the computerized ones. Finally, we also want to study a prediction of the actual value of the Social Phobia Score after the therapy, which would require a larger dataset as well.

Acknowledgments

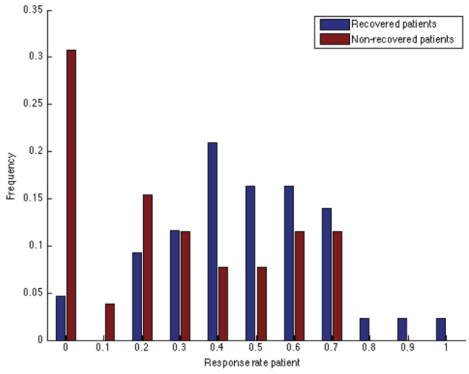
This research has been partly conducted as part of the E-COMPARED project, funded by the European Union under the FP 7 Framework with project number 603098. The study from which the dataset originated was party funded by the Swiss National Science Foundation (SNSF PP00P1 144824/1). Prof. Szolovits' work is supported by NIH grants R01-EB017205 and U54-HG007963.

References

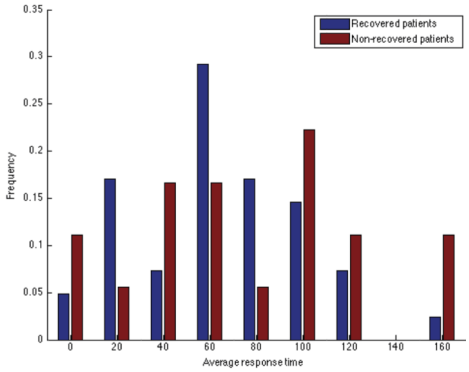
1. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, Charlson FJ, Norman RE, Flaxman AD, Johns N, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. The Lancet. 2013; 382(9904): 1575–1586.
2. Wang P, Aguillar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the who world mental health surveys. The Lancet. 2007; 370(9590):841–850.
3. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: A metaanalysis. PLoS ONE. 2010; 5(10):e13196. [PubMed: 20967242]
4. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Internet-based vs. face-to-face cognitive behaviour therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry. 2014; 13:288–295. [PubMed: 25273302]
5. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010; 62(8):1120–1127. [PubMed: 20235204]

6. Lehman, L-w, Saeed, M., Long, W., Lee, J., Mark, R. Risk stratification of icu patients using topic models inferred from unstructured progress notes. AMIA Annual Symposium Proceedings; American Medical Informatics Association; 2012. p. 505
7. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. Journal of the American Medical Informatics Association. 2014; 21(5):824–832. [PubMed: 24431333]
8. Caspar F, Berger T, Hautle I. The right view of your patient: A computer assisted, individualized module for psychotherapy training. Psychotherapy: Theory, Research, Practice, Training. 2004; 41(2):125–135.
9. Wolf M, Sedway J, Bulik CM, Kordy H. Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. International Journal of Eating Disorders. 2007; 40:711–717. [PubMed: 17683092]
10. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: A content analysis. Biomedical informatics insights. 2010; 2010(3): 19. [PubMed: 21643548]
11. Rude S, Gortner E-M, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognition & Emotion. 2004; 18(8):1121–1133.
12. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: Associations of clients? word use with adherence and outcome. Journal of affective disorders. 2014; 160:10–13. [PubMed: 24709016]
13. Schulz A, Stolz T, Berger T. Internet-based individually versus group guided self-help treatment for social anxiety disorder: protocol of a randomized controlled trial. BMC psychiatry. 2014; 14(1):115. [PubMed: 24735420]
14. Stangier U, Heidenreich T, Berardi A, Golbs U, Hoyer J. Die erfassung sozialer phobie durch die social interaction anxiety scale (sias) und die social phobia scale (sps). Z Klin Psychol Psychiatr Psychother. 1999; 28:28–36.
15. Aronson, AR. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proceedings of the AMIA Symposium; American Medical Informatics Association; 2001. p. 17
16. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making. 2006; 6(1):30. [PubMed: 16872495]
17. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010; 17(5):507–513. [PubMed: 20819853]
18. Tausczik YR, Pennebaker JW. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology. 2010; 29(1):24–54.
19. Zuckerman M. The development of an affect adjective check list for the measurement of anxiety. Journal of Consulting Psychology. 1960; 24(5):457. [PubMed: 13788885]
20. Clarke, D., Wells, A. A cognitive model of social phobia. New York: Guilford Press; 1995.
21. Berger T, Caspar F, Richardson R, Kneubhler B, Sutter D, Andersson G. Internet-based treatment of social phobia: A randomized controlled trial comparing unguided with two types of guided self-help. Behaviour Research and Therapy. 2011; 49:158–169. [PubMed: 21255767]
22. Berger T, Hohl E, Caspar F. Internet-based treatment for social phobia: A randomized controlled trial. Journal of Clinical Psychology. 2009; 65:1021–1035. [PubMed: 19437505]
23. Boettcher J, Berger T, Renneberg B. Does a pre-treatment diagnostic interview affect the outcome of internet-based self-help for social anxiety disorder? a randomized controlled trial. Behavioural and Cognitive Psychotherapy. 2012; 40:513–528. [PubMed: 22800984]
24. Jacobson N, Truax PP. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psycholog. 1991; 59:12–19.
25. Bird, S. Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics; 2006. Nltk: the natural language toolkit; p. 69-72.
26. Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). 1990

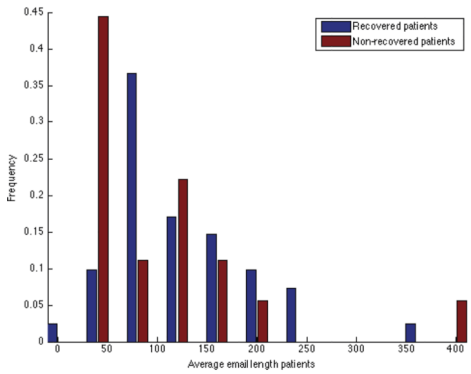
27. Choi, Y., Cardie, C. Learning with compositional semantics as structural inference for subsentential sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics; 2008. p. 793-801.
28. Clematide S, Gindl S, Klenner M, Petrakis S, Remus R, Ruppenhofer J, Waltinger U, Wiegand M. Mlsa-a multi-layered reference corpus for german sentiment analysis. LREC. 2012:3551–3556.
29. Scholz, T., Conrad, S., Hillekamps, L. Text, Speech and Dialogue. Springer; 2012. Opinion mining on a german corpus of a media response analysis; p. 39-46.
30. Remus R, Quasthoff U, Heyer G. Sentiws-a publicly available german-language resource for sentiment analysis. LREC. 2010
31. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research. 2003; 3:993–1022.
32. Breiman L, Friedman J, Olshen R, Stone C, Steinberg D, Colla P. Cart: Classification and regression trees. Wadsworth: Belmont, CA. 1983; 156
33. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.
34. Hannan C, Lambert M, Harmon C, Nielsen S, Smart D, Shimokawa K, Sutton S. A lab test and algorithms for identifying clients at risk for treatment failure. Journal of Clinical Psychology. 2005; 61:155–163. [PubMed: 15609357]
35. Lambert M. What have we learned from a decade of research aimed at improving psychotherapy outcome routine care? Psychotherapy Research. 2007; 17:1–14.
36. Shimokawa K, Lambert M, Smart D. Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of Consulting and Clinical Psychology. 2010; 78:298–311. [PubMed: 20515206]
37. Hatfield DR, Ogles BM. The use of outcome measures by psychologists in clinical practice. Professional Psychology: Research and Practice. 2004; 35:485.
38. Haaga D. Introduction to the special section on stepped care models in psychology. Journal of Consulting and Clinical Psychology. 2000; 68:547–548. [PubMed: 10965628]
39. Wilson G, Vitousek K, Loeb K. Stepped care treatment for eating disorders. Journal of Consulting and Clinical Psychology. 2000; 68:564–572. [PubMed: 10965631]
40. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. DTIC Document, Tech. Rep. 1975



(a) Response rate



(b) Response time (hours)



(c) Email length

Fig. 1.

Overview of email characteristics of recovered and non-recovered patients over the entire period of 12 weeks, the y-axis shows the fraction of patients, the x-axis the bins related to the specific measurement

TABLE I

Overview of data fields available in the dataset

Data field	Description
Patient data	
Patient ID	The ID of the patient
Age	The age of the patient
Gender	The gender of the patient (0=male, 1=female)
Marital status	Indicates whether the patient is living alone; living with partner (not married); living with partner (married); divorced, widow or widower
Highest education	Expresses the education level of the patient, ranging from compulsory school, apprenticeship, college or university
Employment	The patient can either be full time employed/working, part time employed/working, student, housewife, retired, or unemployed
Medication	Indicates if the patient is receiving medication to stabilize a depression or anxiety disorder
Past Psychotherapy	Whether psychotherapy was received in the past
Current Major Depression	Whether the patient was shown to have a Major Depression at the start of the treatment by the Structured Clinical Interview for DSM IV (SCID)
Prior Major Depression	Expresses whether the patient was shown to have had a Major Depression before the treatment by the SCID
Comorbidities	The total number of diagnosed disorders according to the SCID
Pre-intervention Social Phobia Measure	The score of the patient on the Social Phobia Measure before the start of the intervention
Outcome	1 in case of a reliable improvement on the Social Phobia Measure, 0 otherwise
Email data	
Patient ID	The ID of the patient related to this email.
Date and time sent	When the email was sent
Text	The text of the email
Sent by whom	Indicating whether the email was sent by the *therapist* or the *patient*

TABLE II

Measurements used for emails

Measurement	Description
Individual measurements	
Basic mailing behavior	Basic measurements about the mailing behavior of the patient, including the length of the emails written, whether emails from therapists receive responses and how long it takes to respond.
Word usage	A bag-of-words approach to determine the words that occur in the email
Writing style	The style of writing used in the email, identified through part-of-speech tags
Sentiment	Whether the email has a positive or negative sentiment
Topic	The score on the various topics in the email that follows from a topic modeling approach
Aggregate measurements	
Average	Perform averaging over all the emails of the attributes described before
Trend	Look at the development of the metrics identified over time

TABLE III

Top correlations among socio-demographic data

Predictor	**Pearson correlation and p-value**
current major depression	0.20 (p = 0.096)
pre-intervention social phobia measure	0.18 (p = 0.148)
employment retired	−0.17 (p = 0.168)
gender	0.16 (p = 0.192)
employment unemployed	−0.13 (p = 0.283)

TABLE IV

Top correlations among basic mailing behavior

Predictor	Pearson correlation and p-value (12 weeks)	Pearson correlation and p-value (6 weeks)
average		
email length patient	0.05 (p = 0.665)	0.03 (p = 0.782)
response rate	0.13 (p = 0.303)	0.16 (p = 0.187)
response time	−0.12 (p = 0.337)	0.07 (p = 0.568)
trend		
email length patient	0.18 (p = 0.146)	−0.01 (p = 0.936)
response rate	−0.14 (p = 0.264)	0.10 (p = 0.436)
response time	0.14 (p = 0.240)	−0.18 (p = 0.150)

TABLE V

Top correlations to outcome (top 3 for both periods) among words used by the patients. Note that a word must occur in the email of at least 4 different patients; as a result certainly highly correlated words might not have passed that threshold for the 6 week case, yet the correlations have been calculated and included in the table.

Predictor	**Pearson correlation and p-value (12 weeks)**	**Pearson correlation and p-value (6 weeks)**
average, 12 weeks		
protokolli (*to keep a diary about anxiety related thoughts and behaviour*)	−0.3 (p = 0.013)	−0.26 (p = 0.031)
mach (*make*)	0.27 (p = 0.027)	0.27 (p = 0.026)
comput (*computer*)	−0.26 (p = 0.030)	−0.12 (p = 0.321)
average, 6 weeks		
verlauf (*course*)	−0.22 (p = 0.075)	−0.24 (p = 0.051)
empfind (*feel*)	−0.20 (p = 0.102)	−0.23 (p = 0.062)
gelingt (*succeed*)	0.16 (p = 0.191)	0.22 (p = 0.065)
trend, 12 weeks		
mitmach (*to take part*)	0.26 (p = 0.030)	0.23 (p = 0.052)
uber (*above*)	0.26 (p = 0.033)	0.24 (p = 0.045)
pme (*pmr (progressive muscle relaxation)*)	0.26 (p = 0.034)	0.01 (p = 0.925)
trend, 6 weeks		
gegenub (*in relation to others/something*)	−0.18 (p = 0.137)	−0.30 (p = 0.012)
recht (*right*)	−0.03 (p = 0.821)	−0.25 (p = 0.042)
grund (*reason*)	0.08 (p = 0.501)	−0.22 (p = 0.067)

TABLE VI

Correlations for the part-of-speech tags observed in emails

Predictor	Pearson correlation and p-value (12 weeks)	Pearson correlation and p-value (6 weeks)
average		
noun, proper singular	0.19 (p = 0.122)	0.25 (p = 0.042)
wh-pronoun, personal	−0.18 (p = 0.141)	0.16 (p = 0.187)
foreign word	0.16 (p = 0.185)	0.20 (p = 0.105)
adverb	0.14 (p = 0.259)	0.23 (p = 0.052)
pronoun, personal	0.01 (p = 0.943)	0.21 (p = 0.088)
trend		
adverb	0.27 (p = 0.023)	0.29 (p = 0.014)
adverb, particle	0.23 (p = 0.062)	0.14 (p = 0.252)
wh-determiner	0.21 (p = 0.081)	−0.07 (p = 0.571)
pronoun, possessive	0.07 (p = 0.586)	−0.24 (p = 0.043)
adjective	−0.14 (p = 0.263)	−0.17 (p = 0.151)

TABLE VII

Correlations for the sentiment analysis

Predictor	Pearson correlation and p-value (12 weeks)	Pearson correlation and p-value (6 weeks)
average		
sentiment vote	0.04 (p = 0.773)	0.19 (p = 0.128)
sentiment neg(1)	0.02 (p = 0.875)	0.17 (p = 0.155)
sentiment neg(n)	0.02 (p = 0.847)	0.17 (p = 0.172)
trend		
sentiment vote	0.19 (p = 0.305)	0.16 (p = 0.309)
sentiment neg(1)	0.19 (p = 0.126)	0.15 (p = 0.232)
sentiment neg(n)	0.19 (p = 0.121)	0.12 (p = 0.202)

TABLE VIII

Correlations with outcome for the top 3 topics identified for the different settings. Note that different topics are identified for the two different settings due to the difference in texts used. Furthermore, only the four words with the highest weight are shown per topic.

Predictor	Pearson correlation and 2-tailed p-value
average, 12 weeks	
topic 11 (werd (*will, future tense*): 0.022, schon (*already*): 0.016, situation (*situation*): 0.013, sitzung (*session*): 0.012)	−0.31 (p = 0.009)
topic 2 (angst (*fear*): 0.009, werd (*will, future tense*): 0.009, gut (*good*): 0.009, schon (*already*): 0.009)	−0.24 (p = 0.049)
topic 1 (tag (*day*): 0.013, schon (*already*): 0.013, ubung (*exercise*): 0.012, mehr (*more*): 0.011)	0.19 (p = 0.120)
trend, 12 weeks	
topic 23 (situation (*situation*): 0.014, angst (*fear*): 0.013, sitzung (*session*): 0.012, schon (*good*): 0.009)	−0.22 (p = 0.070)
topic 7 (schon (*good*): 0.011, werd (*will, future tense*): 0.011, dank (*gratitude*): 0.010, imm (*always*): 0.009)	0.19 (p = 0.122)
topic 24 (situation (*situation*): 0.010, wied (*again*): 0.009, konn (*could*): 0.008, angst (*fear*): 0.007)	−0.18 (p = 0.136)
average, 6 weeks	
topic 6 (dank (*gratitude*): 0.013, schon (*already*): 0.012, angst (*fear*): 0.011, antwort (*answer*): 0.009)	−0.36 (p = 0.002)
topic 22 (imm (*in*): 0.013, funktioniert (*functions*): 0.013, wied (*again*): 0.013, angst (*fear*): 0.012)	0.18 (p = 0.140)
topic 18 (schon (*already*): 0.011, gut (*good*): 0.009, dank (*gratitude*): 0.008, all (*all*): 0.007)	−0.17 (p = 0.164)
trend, 6 weeks	
topic 18 (*see above*)	0.23 (p = 0.053)
topic 6 (*see above*)	−0.22 (p = 0.074)
topic 19 (ubung (*exericse*): 0.012, imm (*always*): 0.010, sitzung (*session*): 0.010, mal (*times*): 0.008)	0.21 (p = 0.080)

TABLE IX

AUC's (with 95% confidence interval), and Precision (P), Recall (R) and F1 observed with a threshold set to a value corresponding with a false positive rate of 0.4 using the full 12 weeks of data with different learning algorithms.

Data	Logistic Regression				Decision Tree				Random Forest			
	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1
SD	0.55 *(0.43–0.67)*	0.82	0.48	0.61	0.55 *(0.42–0.67)*	0.83	0.31	0.45	0.48 *(0.36–0.61)*	0.77	0.35	0.49
A	0.75 *(0.65–0.84)*	0.86	0.67	0.77	0.45 *(0.33–0.58)*	0.79	0.23	0.35	0.77 *(0.67–0.86)*	0.87	0.71	0.78
T	0.64 *(0.52–0.75)*	0.83	0.40	0.54	0.59 *(0.47–0.71)*	0.80	0.25	0.38	0.62 *(0.50–0.73)*	0.81	0.46	0.59
SD/A	0.75 *(0.65–0.84)*	0.86	0.67	0.75	0.47 *(0.35–0.60)*	0.83	0.21	0.33	0.65 *(0.53–0.76)*	0.87	0.56	0.68
SD/T	0.67 *(0.56–0.78)*	0.85	0.48	0.61	0.64 *(0.52–0.75)*	0.89	0.33	0.48	0.61 *(0.49–0.73)*	0.82	0.38	0.51
SD/A/T	0.72 *(0.62–0.83)*	0.85	0.60	0.71	0.63 *(0.52–0.75)*	0.83	0.40	0.54	0.78 *(0.69–0.87)*	0.87	0.69	0.77
A/T	0.72 *(0.62–0.83)*	0.85	0.60	0.71	0.64 *(0.52–0.75)*	0.83	0.40	0.54	0.76 *(0.67–0.86)*	0.88	0.73	0.80

For the data: SD=Socio-Demographic Data, A=Average value of attributes over individual mails, T=Trends, Random forest with 10 trees only.

TABLE X

Predictors in model with SD/A/T. Note that predictors have iteratively been selected based on their score on the Pearson correlations. If a new predictor was strongly correlated (≥ 0.7) with an already selected predictor, the predictor was not included.

Category	Predictors and their correlations.
word usage (average)	protokolli (*to keep a diary (about anxiety related thoughts and behavior)*, −0.3), mach (*make*, 0.27), comput (*computer*, −0.26), pme (*pmr; progressive muscle relaxation*, −0.26), darub (*about it*, 0.25), realist (*realist*, −0.24), geht (*go*, 0.23), teilhab (*to partake*, −0.23), voll (*full*, 0.22),
word usage (trend)	mitmach (*to take part*, 0.26), uber (*about*, 0.26), wirkt (*it works*, 0.26), zukunfst (*future*, 0.25), einstell (*attitude*, 0.24), gesprach (*conversation*, 0.24), bemuh (*effort*, −0.23), eben (*just*, −0.22)
writing style (trend)	adverb (0.27)
topics (average)	topic 11 (*werd (will, future tense): 0.022, schon (already): 0.016, situation (situation): 0.013, sitzung (session): 0.012*, −0.31), topic 2 (*angst (fear): 0.009, werd (will, future tense): 0.009, gut (good): 0.009, schon (already): 0.009*, −0.24)

TABLE XI

AUC's (with 95% confidence interval), and Precision (P), Recall (R) and F1 observed with a threshold set to a value corresponding with a false positive rate of 0.4 using the first 6 weeks of data with different learning algorithms.

Data	Logistic Regression				Decision Tree				Random Forest			
	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1
A	0.72 *(0.61–0.82)*	0.88	0.75	0.81	0.62 *(0.50–0.73)*	0.86	0.40	0.54	0.76 *(0.66–0.85)*	0.88	0.73	0.80
T	0.78 *(0.68–0.87)*	0.89	0.65	0.75	0.70 *(0.59–0.80)*	0.88	0.46	0.60	0.72 *(0.62–0.82)*	0.87	0.54	0.67
SD/A	0.74 *(0.64–0.84)*	0.88	0.79	0.84	0.65 *(0.53–0.76)*	0.87	0.42	0.56	0.81 *(0.72–0.89)*	0.88	0.75	0.81
SD/T	0.83 *(0.75–0.91)*	0.90	0.75	0.82	0.70 *(0.59–0.80)*	0.88	0.46	0.60	0.77 *(0.67–0.86)*	0.88	0.48	0.62
SD/A/T	0.71 *(0.60–0.81)*	0.88	0.75	0.81	0.64 *(0.53–0.76)*	0.90	0.40	0.55	0.72 *(0.62–.82)*	0.87	0.69	0.77
A/T	0.71 *(0.60–0.81)*	0.88	0.75	0.81	0.64 *(0.53–0.76)*	0.90	0.40	0.55	0.73 *(0.63–0.84)*	0.89	0.67	0.76

For the data: SD=Socio-Demographic Data, A=Average value of attributes over individual mails, T=Trends, Random forest with 10 trees only.

TABLE XII

Predictors in model SD/T. Note that same including criteria have been used as explained for the 12 week setting.

Category	Predictors
word usage (trend)	gegenub (*in relation to others/something*, −0.30), recht (*right*, −0.25), grund (*reason*, −0.22), wunsch (*wish*, −0.21), langsam (*slowly*, −0.21), moglich (*possibly*, 0.21), hilflos (*helpless*, 0.20), feststell (*declaratory*, 0.20), angespannt (*tense*, 0.20), fortschritt (*progress*, 0.19), gerat (*get*, 0.19), umstand (*circumstances*, 0.19), info (*information*, 0.19)
topics (trend)	topic 18 (*schon (already): 0.011, gut (good): 0.009, dank (gratitude): 0.008, all (all): 0.007*, 0.23), topic 6 (*dank (gratitude): 0.013, schon (already): 0.012, angst (fear): 0.011, antwort (answer): 0.009*, −0.22), topic 19 (*ubung (exericse): 0.012, imm (always): 0.010, sitzung (session): 0.010, mal (times): 0.008*, 0.21), topic 20 (*situation (situation): 0.012, tag (day): 0.011, ganz (entire): 0.010, and (and): 0.009*, 0.20)
syntax (trend)	adverb, particle (0.29), pronoun, possessive (−0.24)
socio-demographic	current major depression (0.20)