Predicting Treatment Outcome from Patient Texts: The Case of Internet-Based Cognitive Behavioural Therapy

Evangelia Gogoulou[1] Magnus Boman[2] Fehmi Ben Abdesslem[1]
Nils Isacsson[3] Viktor Kaldo[3,4] Magnus Sahlgren[1]

[1] Research Institutes of Sweden (RISE)
[2]School of Electrical Engineering and Computer Science,
The Royal Institute of Technology(KTH) &
Department of Learning, Informatics, Management and Ethics, Karolinska Institutet
[3]Department of Clinical Neuroscience, Karolinska Institutet, &
Stockholm Health Care Services, Region Stockholm
[4]Faculty of Health and Life Sciences, Linnaeus University

{evangelia.gogoulou, fehmi.ben.abdesslem, magnus.sahlgren}@ri.se
{mab@kth.se}
{nils.isacsson, viktor.kaldo}@ki.se

Abstract

We investigate the feasibility of applying standard text categorisation methods to patient text in order to predict treatment outcome in Internet-based cognitive behavioural therapy. The data set is unique in its detail and size for regular care for depression, social anxiety, and panic disorder. Our results indicate that there is a signal in the depression data, albeit a weak one. We also perform terminological and sentiment analysis, which confirm those results.

1 Introduction

The Internet Psychiatry Clinic in Stockholm offers Internet-based Cognitive Behavioral Therapy (ICBT) for depression, social anxiety, and panic disorder with documented significant treatment effects (Titov, 2018). The treatment is delivered on a secure online platform that enables patients to submit answers to weekly questionnaires of primary symptoms, depression, and suicidal ideation, as well as interactive worksheets. Because patients normally are not in contact with the psychologists away from keyboard during treatment, the infrastructure platform logs a large amount of data for each patient, yielding a unique data set. Because Sweden was the first country to introduce this kind of ICBT, the clinical experience is considerable (Hedman et al., 2012), and experimental research of the kind reported in this paper rests on well-established procedures for sensemaking of machine learning at the clinic (Boman and Sanches, 2015). The overarching problem is how to be able to offer accelerated care, via identification of those predicted not to succeed with treatment, or at least not with enough decrease in symptoms, and giving these people more attention, as early as possible.

We here investigate the possibility of predicting outcome based *only* on patient texts, in various points in treatment. We formulate this as a unimodal binary text categorisation problem, where the categories are simplified to either success or failure. In this initial feasibility study, we evaluate the performance of a number of text categorisation methods for predicting treatment outcome. Even if our methodology is mostly exploratory, we also relate terminological and sentiment analysis of patients' text to the treatment outcome.

2 Related work

NLP has previously been used for predicting the outcome of psychiatric treatment in a number of encouraging studies. Stylometric analysis on patients' texts from the data set under study indicates variation through time in the sentiment sign (positive or negative) in patient text messages (Boman et al., 2019). Althoff et al. (2016) suggest that more references to the future, references to other individuals, and positive conversation perspective are all positively related to the success of counseling conversations. Cohan et al. (2018) construct a

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 575–580
April 19 - 23, 2021.

mental health-related corpus with Reddit posts and analyse users' language to identify self-reported diagnoses of nine different mental health conditions, including depression. Pestian et al. (2010) propose a classification model based on linguistic features, including part-of-speech tags and emotional concepts among others, that under some circumstances outperforms mental health professionals on identifying fake suicide notes. Gkotsis et al. (2016) study the usefulness of negation and affirmation in mental health records with the word *suicide* itself, while Song et al. (2020) identify statements related to suicidal behaviour in electronic health records, using a deep neural network. Trotzek et al. (2018) address the problem of early detection of depression using an ensemble of a Convolutional Neural Network and a classifier based on user-level linguistic metadata.

3 Data Set

The total data covers $6,821$ patients enrolled in a 12-week treatment for three different psychological disorders: Major Depressive Disorder (hereafter referred to as MDD, or depression), social anxiety, or panic disorder. Each treatment consists of weekly exercises, referred to as *homework reports*. For predicting treatment outcome, we focus on the depression treatment programme, which consists of 10 homework reports distributed over the 12 weeks. The homework contains questions about the progress made by the patient during treatment, such as "What is the most important thing you bring from module 1?" The questions can be both closed questions, where the patient selects pre-defined answers from a list, and open-ended questions, where the patient is invited to type in free text. The depression data covers $3,179$ patients. We remove patients who completed less than five homework reports (out of ten), patients without treatment outcome (who did not fill out the self-assessment at the end of the treatment), and empty texts. The resulting filtered data set contains $16,379$ texts from homework reports completed by $1,986$ patients. Each text is a concatenation of all answers written by a patient in one single weekly homework.

We also use patients' texts from the other treatments in order to enhance the domain pre-training for the embeddings. This data covers all $6,821$ patients participating in any of the three treatments, including $180,017$ free texts answers from the homework reports and $146,398$ direct messages, totalling almost 29 million $(28,993,089)$ tokens.

As target scores for predicting treatment outcome for the depression patients, we use the self-assessment MADRS-SR method (Fantino and Moore, 2009), which consists of ten questions scored between zero and six, and amounting to each patient having done between five and ten modules, yielding a total score from zero to 54. Therapists commonly use the cut-off score of 10 to partially define successful treatment. All diagnosed patients started the treatment above this score, and the treatment is defined as successful if the score drops below 10 at the end of the treatment (*remission*) (Hawley et al., 2002). When patients show a symptom reduction of at least 50%, this is defined as a *response*, and the treatment is also considered successful, even if the final score is above the cut-off value. In other cases, the treatment is considered a failure. Regarding the choice of 50% cut-off for the definition of the responder patients, it is supported by previous work (Karin et al., 2018), while the same value has been previously used by clinical researchers (Forsell et al., 2019). All patients are labelled as being subject to a treatment success or a treatment failure, as shown in Table 1.

Table 1: Patient labels in the prediction data set, based on the treatment outcome.

Outcome	Cases	Label	Total
Response	164	Success	1,075
Remission	111		
Response and remission	800		
No response, no remission	911	Failure	911

4 Methodology

We formulate the task of predicting outcome as a simple binary text classification problem, where the classes are either *success* or *failure*. Our goal is to assist clinicians in defining the best possible adaptive treatment strategy. Therefore, the class definitions used here are inherited from the treatment model employed at the clinic. We compare four different text representations: TF-IDF (Jones, 1972),[1] Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and Doc2Vec (Le and Mikolov, 2014), all of which are fed into a simple linear classifier trained with binary cross entropy loss using PyTorch.[2] All word embeddings are pro-

[1] https://scikit-learn.org
[2] https://pytorch.org/

duced with the Gensim implementation [3] and are pre-trained on patients' texts from all three treatments, as described in Section 3.

We use a naïve classification model that generates stratified predictions based on the class distributions as the baseline. The training set includes $1,588$ patients, with 860 labelled as success and 728 as failure, from the depression programme who have completed at least five weeks of homework reporting. All models are trained with 3-fold cross validation and early stopping, to avoid overfitting. After training, we evaluate each model on the test set, which consists of 398 patients (215 labelled as success and 183 as failure). Model performance is measured using the F1 score metric.

5 Results

The results from the classification experiment are presented in Table 2. The classification model based on TF-IDF consistently improves on the baseline, for all three values of the parameter #Homework reports. On the other hand, the models based on Word2Vec, FastText and Doc2Vec marginally reach this goal. Between all combinations of embeddings and number of homework reports considered, the TF-IDF model using one or three homework reports is the best performing model. Word2Vec exhibits better performance than FastText and Doc2Vec when the number of homework reports used is set to 3. Particularly remarkable is the relative F1 score difference between TF-IDF (number of homework reports equal to 1 or 3) and the rest of the models. This difference indicates the appropriateness of Bag-of-Words models in this experimental setup, compared to text representation methods that leverage semantic information. Interestingly, model performance does not always improve when more homework reports are considered. One possible explanation is that the resulting patient representation has less predictive power when more patient's text is considered, given the relatively small size of the training set. However, further research is necessary to confirm this.

Table 3 shows the test F1 score of the best performing model, which is TF-IDF, for each subgroup of the *success* class and all three values of the parameter #Homework reports. As defined in Section 3, the *success* class consists of patients that are *responders* or *remitters* or both *responders* and *remitters* (*responders+remitters*). It is observed that the F1 score calculated for the group of *remitters* is consistently lower than the ones achieved in the other two groups. This suggests that the model might benefit from a different classification setup, where *remitters* are considered as a separate class. Such analysis is left for future work.

As indicated by the classification results, simple lexical representations perform best in our experimental setup. A concrete follow-up question then becomes: Which vocabulary terms have a strong positive or negative correlation with the success class? To answer the question, we perform a linear regression analysis on the complete set of $1,986$ TF-IDF patient embeddings from our prediction data set, using the first three homework sets, and then compute the dot product between each word in the patient vocabulary and the regression coefficient. Vocabulary words with positive value of dot product contribute positively to the prediction of the success class and *vice versa*. Examples of terms with high positive or negative correlation with the success class are presented in Table 4.

There seems to be a vague relation between terminology and treatment outcome, which corroborates previous work by Althoff et al. (2016) that reports positive correlation between positively signed written language and successful counseling conversations. We therefore perform sentiment analysis on the complete list of homework reports written by $2,611$ depression patients, with at least one homework report completed. Since we are interested in the overall progression of sentiment in patients' answers, we are not constrained to include only patients with a specified number of homework reports completed. We train a Multinomial Naïve Bayes classifier for predicting the sentiment class (positive, neutral or negative) on the SenSALDO Swedish sentiment lexicon.[4] All words are represented with TF-IDF embeddings. The sentiment classifier is then used to identify the sentiment class of all words in the vocabulary, excluding stopwords and words with document frequency lower than 0.2. For each patient and concatenated set of homework reports, the number of word occurrences per sentiment class is calculated. We group the patients by the label outcome and compute the average relative frequency of the number of positive words divided by the number of negative words for each homework report and group. The progression of

[3] https://radimrehurek.com/gensim/

[4] https://spraakbanken.gu.se/en/resources/sensaldo

#Homework reports	Baseline	TF-IDF	Word2Vec	FastText	Doc2Vec
1	0.58	**0.69**	0.60	0.56	0.59
3		**0.69**	**0.62**	0.60	0.55
5		0.59	0.59	0.61	0.59

Table 2: Test F1 score for the three proposed classification models considering various sets of weekly homework reports as model input. The performance of the baseline (stratified classifier) is reported for comparison.

	TF-IDF #hw=1	TF-IDF #hw=3	TF-IDF #hw=5
Responders	0.97	0.97	0.72
Remitters	0.89	0.95	0.6
Responders+Remitters	0.99	0.98	0.76

Table 3: Test F1 score calculated for each subgroup of the *success* class when using TF-IDF as the text representation method. #hw denotes the number of homework reports used by the model.

the average relative frequency of words with positive sentiment throughout the homework reports is presented in Figure 1.

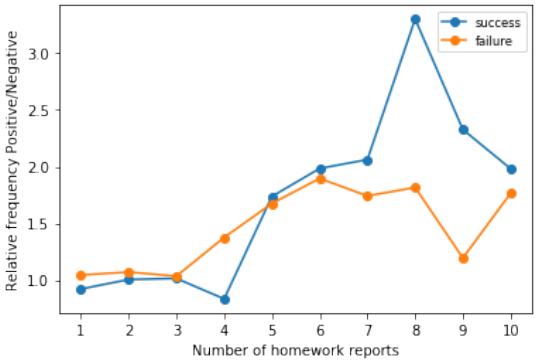

Figure 1: Average relative frequency of words with positive sentiment for all depression patients grouped by the outcome label.

Similar to (Althoff et al., 2016), we observe that more positive sentiment in the homework answers is related to successful treatment outcome. This reflects the treatment progress for patients that in the end are classified as successful. Specifically, homework report 8 had questions concerned with the overall opinion of the treatment, and the usefulness of the tools that the patients had been using. The large differences observed at homework report 8 confirm our intuition that patients showing substantial progress during the treatment, would evaluate its usefulness in more positive terms. Whereas the analysis presented in (Althoff et al., 2016) concerns open-ended counseling conversations, our data set contains patients' answers to targeted questions with positive or negative underlying sentiment. Two indicative questions are *"What's the most important thing you have learned from the past week?"* and *"Tell us about the activities you had planned for this week that you followed through. How was it?"*. The variation in sentiment in homework questions is reflected in the saturation of the relative frequency throughout the progression of homework reports.

6 Discussion

The results demonstrate that the task of predicting treatment outcome based on patient text is very difficult. It is interesting to note that embedding-based representations in some cases fail to perform better than random guessing, which we partly attribute to the comparably small amounts of training data for both domain adaptation and classification, and partly to the fact that such models tend to conflate paradigmatically similar terms, which in this application can signal very different treatment effects. This is demonstrated by terminological and sentiment analysis, which indicate a small but noticeable difference in vocabulary between successful and not successful patients. One hypothesis may be that antonyms are predictive of different outcomes, but have similar representations in the word embeddings, which makes it difficult for a simple linear classifier to separate them. The Bag of Words does not have this problem, which may be the reason it performs better in this specific application.

The selection of patients with a minimum predefined number of homework reports completed for the classification experiments serves the purpose of identifying the best prediction point in the home-

Positive words	forward, away, day, thoughts, initiative, act, wanted, feeling, together, energy, later
Negative words	sleep, rarely, unfortunately, sad, friend, rewarding, despite, life, walk, fill, different

Table 4: List of positively (up) and negatively (down) correlated words with the success class. Words have been translated to English from the original Swedish.

work completion task, but it also poses an important limitation: it reduces the variance in the outcome label. More specifically, success rate is expected to be higher among patients that have completed at least five homework reports. The more careful study of different patient subgroups, based on the number of completed homework reports, is left for future work.

Our current analysis ignores the set of homework assignment questions. Since much of the patient text in ICBT is generated on the prompt of a question, future work could study the correlation of question-answer pairs with the treatment outcome.

7 Conclusions

We have analysed the potential of using patient text from Internet-based Cognitive Behavioural Therapy as the only signal for predicting treatment outcome. When framing the problem as a binary classification task between treatment success and failure, we manage to beat stratified random guessing using a simple Bag of Words classifier. This demonstrates the feasibility of our approach. Interestingly, word embeddings fail to improve on this simple approach, and our best results are achieved using only data from the first couple of weeks of treatment. This has clinical significance because interventions during treatment are still meaningful, and could prove crucial, in the first four weeks of treatment. Additionally, we provide simple terminological and sentiment analysis, which also indicate that there is a signal in the patient text data, albeit a weak one.

The work reported on here should be seen as a feasibility study in that it shows the potential to predict treatment outcome based on patient text, even in a very small subset of the set of patients reporting and reflecting on what they do in treatment, and how treatment progresses. Our next step is to look into larger subsets than the one employed here, but that step will be taken in conjunction with moving from the well-established albeit simple methods used here to contextualised language models.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Magnus Boman, Fehmi Ben Abdesslem, Erik Forsell, Daniel Gillblad, Olof Görnerup, Nils Isacsson, Magnus Sahlgren, and Viktor Kaldo. 2019. Learning machines in internet-delivered psychological treatment. *Progress in Artificial Intelligence*, 8(4):475–485.

Magnus Boman and Pedro Sanches. 2015. Sensemaking in intelligent health data analytics. *KI-Künstliche Intelligenz*, 29(2):143–152.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bruno Fantino and Nicholas Moore. 2009. The self-reported montgomery-asberg depression rating scale is a useful evaluative tool in major depressive disorder. *BMC psychiatry*, 9:26.

Erik Forsell, Nils Isacsson, Kerstin Blom, Susanna Jernelöv, Fehmi Ben Abdesslem, Nils Lindefors, Magnus Boman, and Viktor Kaldo. 2019. Predicting treatment failure in regular care internet-delivered cognitive behavior therapy for depression and anxiety using only weekly symptom measures. *Journal of Consulting and Clinical Psychology*.

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

C Hawley, Tim Gale, and T Sivakumaran. 2002. Defining remission by cut off score on the madrs: Selecting the optimal value. *Journal of affective disorders*, 72:177–84.

Erik Hedman, Brjánn Ljótsson, and Nils Lindefors. 2012. Cognitive behavior therapy via the internet: a systematic review of applications, clinical efficacy and cost–effectiveness. *Expert review of pharmacoeconomics & outcomes research*, 12(6):745–764.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Eyal Karin, Blake F Dear, Gillian Z Heller, Milena Gandy, and Nickolai Titov. 2018. Measurement of symptom change following web-based psychotherapy: Statistical characteristics and analytical methods for measuring and interpreting change. *JMIR mental health*, 5(3):e10200.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

Xingyi Song, Johnny Downs, Sumithra Velupillai, Rachel Holden, Maxim Kikoler, Kalina Bontcheva, Rina Dutta, and Angus Roberts. 2020. Using deep neural networks with intra- and inter-sentence context to classify suicidal behaviour. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1303–1310, Marseille, France. European Language Resources Association.

Nickolai et al. Titov. 2018. ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet interventions*, 13:108 – 115.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.