A two-step approach for the prediction of mood levels based on diary data

Vincent Bremer[1][0000−0002−0304−5859], Dennis Becker[1][0000−0002−1437−6127], Tobias Genz[1][0000−0002−4198−4013], Burkhardt Funk[1][0000−0001−5855−2666], and Dirk Lehr[2][0000−0002−5560−3605]

[1] Institute of Information Systems, Leuphana University, Lüneburg, Germany
[2] Institute of Psychology, Leuphana University, Lüneburg, Germany

Abstract. The analysis of diary data can increase insights into patients suffering from mental disorders and can help to personalize online interventions. We propose a two-step approach for such an analysis. We first categorize free text diary data into activity categories by applying a bag-of-words approach and explore recurrent neuronal networks to support this task. In a second step, we develop partial ordered logit models with varying levels of heterogeneity among clients to predict their mood. We estimate the parameters of these models by employing MCMC techniques and compare the models regarding their predictive performance. This two-step approach leads to an increased interpretability about the relationships between various activity categories and the individual mood level.

Keywords: Text-Mining · Ordinal Logit · Diary Data.

1 Introduction

Mental issues are increasing around the world and access to healthcare programs are limited. Internet-based interventions provide additional access and can close the gap between treatment and demand [8]. In these interventions, participants often provide diary data in which they rank, for example, their mood levels and simultaneously report daily activities. Because various activities from walking a dog, to volunteering, cleaning the house, or having a drink out with friends affect mood in different and complex ways [9], we attempt to analyze the effects that different activities have on the mood level.

In this study, we propose a two-step approach for the analysis of free text diary data that is provided by participants of an online depression treatment [4]. The dataset consists of 440 patients who provided 9,192 diary entries. We utilize text-mining techniques in order to categorize the free text into defined activity categories (exercise, sickness, rumination, work related, recreational, necessary, social, and sleep related activities) and use individualized partial ordered logit models to predict the mood level. This two-step approach allows for interpretability of the effects between the activity categories and the mood level. Thus, besides studying these relationships, we contribute to the field of

machine learning by proposing a mixed method approach to analyze diary data. This short paper is based on a full paper already published in [3]. Here, more information about the methods, results, and discussion including a full list of references can be found.

2   Method

Figure 1 illustrates the two-step approach. In the first step, we utilize bag-of-words (BoW) categorization and extent the results by applying recurrent neuronal networks (RNN) [5] in order to categorize the free text into activity categories. We split all diary entries into sentences and identify the most frequent ($\geq$10 occurrences) 1- and 2-grams. Next, two of the authors manually associated the frequent 1- and 2-grams with an activity category. Only the 1- and 2-grams that are assigned identically by both authors are utilized for the BoW categorization. The sentences are then assigned to one or multiple activities based on the categorized n-grams. Since 8,032 sentences do not contain any of the n-grams, they cannot be categorized. We then train an Elman network (RNN) on the categorized sentences. The RNN classifies sentences that are not already assigned by the BoW categorization. Some sentences are not associated because these consist of words that do not appear in the training corpus. The results of the BoW categorization and the merged results of both approaches are then utilized as input for the second step.
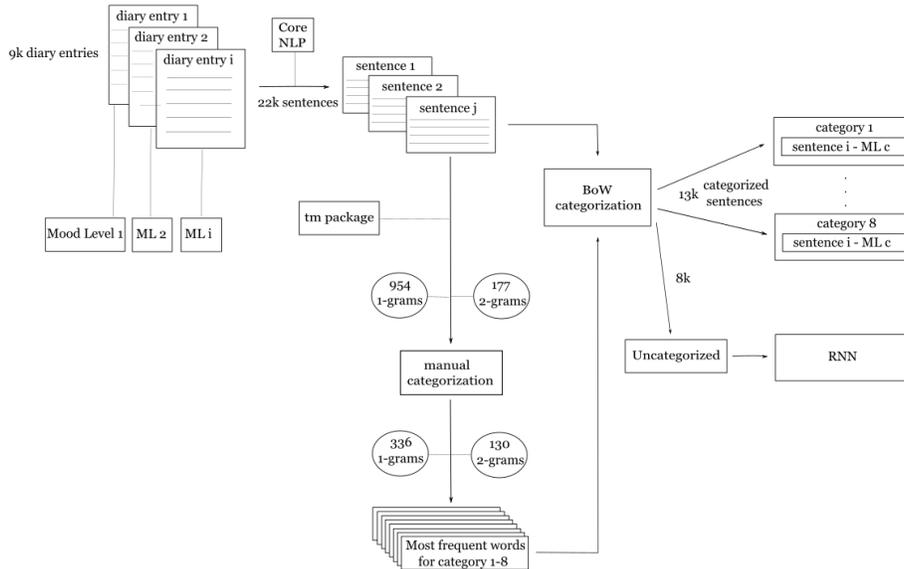

Fig. 1. The process of the two-step approach.

Because the mood level is ranked on a scale from one to ten, we use a partial ordered logit model for the prediction and the analysis of the effects between the assigned activity categories and the mood level. The ordered logit model is based on the proportional odds assumption (POA), which means that independent variables have the same effect on the outcome variable across all ranks of the mood level [7]. The partial ordered logit model, however, allows variables that violate this assumption to vary among the ranks. We test the assumption by a likelihood ratio test. The logit is then calculated as follows:

$$\ln(\theta_{ijt}) = \alpha_{ij} - \left(\underbrace{\sum_{a \in A_1} \beta_{aj} \, x_{ajt}}_{\text{POA holds}} + \underbrace{\sum_{a \in A_2} \beta_{aij} \, x_{ajt}}_{\text{POA violated}} \right),$$

where α_{ij} represents the threshold between the ranks of the mood level for $i = 1, \ldots, I = 9$ and $j = 1, \ldots, J = 440$. The activities of participant j at time t are represented by x_{ajt}, where $A_1 = \{$sleep related, recreational activities$\}$ and $A_2 = \{$exercise, sickness, rumination, social, work related, necessary activities$\}$. The parameters to be estimated are $\beta_{[\ldots]}$. The index j in α_{ij} addresses the problem of scale usage heterogeneity [6]. Additionally, we hypothesize that the effects of the activities vary among participants. Thus, we also include client specific β-parameters. For a robustness check, we also implement the partial ordered logit model without the consideration of heterogeneity among the participants (Model 1), only implement the individual α-parameters (Model 2), only client specific β-terms (Model 3), and the above specified model including both heterogeneity terms (Model 4). Therefore, we obtain four different models, which we compare regarding their predictive performance.

3 Results & Discussion

We compare the models by using the Deviance Information Criterion (DIC), which is especially suited for Bayesian models that are estimated by MCMC methods [2]. The results of the DIC indicates a superior performance for the model that includes both heterogeneity terms. According to [1], however, the DIC can be prone to select overfitted models. Thus, for applying an out-of-sample test, we randomly extract mood entries (680 sentences) and their corresponding activities from the data before training the model. We then predict the mood level of the individuals in the test data and utilize the Root Mean Square Error (RMSE) as well as the Mean Absolute Error (MAE) as performance indicators. We also report performance measures for a so called *Mean Model*; here, we use the average mood level of the training set as predictions for the test dataset (in this case the mood level 6).

As illustrated in Table 1, an increasing degree of heterogeneity reduces the prediction error. The additionally classified activities by the RNN do not contribute to an increased performance. This can potentially arise because the train-

Table 1. Model comparison with levels of heterogeneity for each text-mining approach.

Measure	Model 1 BoW	Model 1 RNN	Model 2 BoW	Model 2 RNN	Model 3 BoW	Model 3 RNN	Model 4 BoW	Model 4 RNN	Mean Model
RMSE	2.32	2.33	1.98	1.98	1.87	1.91	1.81	1.86	1.91
MAE	1.78	1.82	1.48	1.49	1.41	1.41	1.37	1.37	1.53

ing data used for the RNN, which is based on the BoW categorization, might not be accurate enough for the RNN to generate new knowledge. Model 4 for the BoW categorization shows the best predictive performance. Thus, we utilize this model for revealing the relationships between the activities and the mood level.

We find that the category sickness has a strong negative and significant effect on mood. Furthermore, our analysis suggests that the category rumination affects the mood level in a negative way and social activities have a positive effect on the mood level. The other activities are not significant. These results are consistent with literature in the field [9]. During the ECML, we will additionally present the results of a model that directly predicts the mood levels based on the free text data.

References

1. Ando, T.: Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika **94**(2), 443–458 (2007)
2. Berg, A., Meyer, R., Yu, J.: Deviance Information Criterion for Comparing Stochastic Volatility Models. Journal of Business & Economic Statistics **22**(1), 107–120 (2004)
3. Bremer, V., Becker, D., Funk, B., Lehr, D.: Predicting the individual mood level based on diary data. In: 25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, June 5-10, 2017. p. 75 (2017)
4. Buntrock, C., Ebert, D.D., Lehr, D., Cuijpers, P., Riper, H., Smit, F., Berking, M.: Evaluating the efficacy and cost-effectiveness of web-based indicated prevention of major depression: design of a randomised controlled trial. BMC psychiatry **14**, 25–34 (2014)
5. Elman, J.L.: Finding Structure in Time. Cognitive Science **14**(2), 179–211 (1990)
6. Johnson, T.R.: On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. Psychometrika **68**(4), 563–583 (2003)
7. McCullagh, P.: Regression Models for Ordinal Data. Journal of the Royal Statistical Society **42**(2), 109–142 (1980)
8. Saddichha, S., Al-Desouki, M., Lamia, A., Linden, I.A., Krausz, M.: Online interventions for depression and anxiety - a systematic review. Health psychology and behavioral medicine **2**(1), 841–881 (2014)
9. Weinstein, S.M., Mermelstein, R.: Relations between daily activities and adolescent mood: the role of autonomy. Journal of Clinical Child and Adolescent Psychology **36**(2), 182–194 (2007)