BioInfo@UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases

Alina Trifan[0000−0001−7613−1435], Pedro Salgado[0000−0001−6230−6484] and José Luís Oliveira[0000−0002−6672−6176]

DETI/IEETA, University of Aveiro, Portugal
{alina.trifan, psalgado, jlo}@ua.pt

Abstract. This paper describes the participation of the Bioinformatics group of the Institute of Electronics and Engineering Informatics of University of Aveiro in the shared tasks of CLEF eRisk 2020[1]. The eRisk initiative fosters Natural Language Processing research for the automatic detection of risk situations on the internet. Similar to the previous years, the challenge was organized in two tasks, which aimed the early detection of self-harm (T1) and severity of depression (T2) in online forums. We addressed these tasks both from a standard machine learning perspective and from a behavioural point of view. The results we obtained endorse the use of social monitoring as a possible complement to more traditional public health surveillance and intervention practices.

Keywords: social mining · early detection · depression · self-harm · psycholinguistic patterns.

1 Introduction

In the last decade the digitalization of social interactions has created opportunities for researchers and practitioners to use social media as a data source for learning from a different perspective about health and well-being. Social data, defined as data that is created by people with the goal of sharing it with others [24] is a quite recent term that, together with the advances in text mining and Natural Language Processing (NLP) fueled the development of a new research area known as social media mining. Research initiatives such as CLEF Early Risk [20] dynamize the scientific advances and societal impact that this research area can have. They foster collaborative work on the topic of mental health and social data, and push forward new discoveries and insights that can potentially benefit public health.

[1] http://early.irlab.org/

The relation between social media and well-being is well recognized, as users of social media networks often share very personal feelings and beliefs. There are numerous online communities that provide support and counseling for users in need. Most important, these social interactions lead to an impressive data lake that represents an opportunity for scientific advancement and social good [14]. Reliable predictive models allow early detection of heath conditions and pave the way for health interventions, by promoting relevant health services, or by delivering useful health information [30]. In a systematic review on social mining for mental health, Alonso et. al [2] conclude that the use of social mining applied to diseases such as dementia, schizophrenia and depression can be of great help to the clinical decision, diagnosis prediction and ultimately improve the patient's life quality. Because mental health issues are current societal issues they demand new prevention and intervention strategies. Early detection of mental illness is an essential step in the evolution of our current society.

This paper describes the participation of the BioInfo@UAVR team in the CLEF eRisk 2020 tasks. This is our second participation in these tasks and our approach built upon the methodology we used in 2019. As such, we combined standard machine learning algorithms with extended psycholinguistics and behavioral patterns derived from the literature. The methodology and associate results are presented in this paper, along with differences and improvements with respect to our previous participation, as well as a discussion on future work. The rest of this paper is organized as follows: Section 2 overviews the current background in social data mining. The next two sections are dedicated to the description of each of the tasks, and include both the methodologies used and the results obtained. We conclude the paper and discuss possible improvements and future work in Section 5.

2 Background

Mental and behavioral health, is an area of health with one of the largest gaps between the seriousness of the problem and the little information we have available. This makes it one of the most promising areas of research with social monitoring [24]. While the landscape of mental health has been changing over the last decades, the traditional clinical research still faces the lack of precise and timely diagnosis. A standard diagnosis of mental health issues relies mostly on patient interviews and clinical diaries. In order to overcome these gaps, researchers explore social data in an attempt to better understand a wide range of mental health disorders. As such, big data and artificial intelligence offer exciting opportunities for the screening and prediction of mental problems [17].

Mental disorders include many different illnesses, with depression being the most prominent. Moreover, self-harm and anxiety can lead to suicidal ideation. In some of the most varied medical research, data science approaches have allowed researchers to mine large healthcare datasets to detect patterns and to better understand a specific disease or its evolution [4,5,15,21,28,29,32,33]. Researchers have been using over the last decade publicly available social media messages and

interactions as a data source for studying a variety of mental health conditions [7–9, 12, 22].

Even if social media systems can deliver novel, reliable information, there is a challenge in determining how to act on this information. In areas without existing empirical data, where social monitoring systems deliver new information, careful validation and evaluation will be necessary to determine the extent to which the information can be relied on. A recent study by Ernala et al. [10] questions the validity of classification results when there is no medical confirmation of the diagnosis and raises a meaningful discussion on the methodologies used so far for identifying patients at risk in online forums. One of the first demonstration of suicide risk assesment through Reddit posts, matched with clinical knowledge was reported by Shing et al. [27] and paves the way into bridging computational social mining and clinical research in the area of mental health.

3 Task 1 - Early detection of signs of self-harm

Task 1 consisted in sequentially processing pieces of evidence and detect early traces of self-harm, as soon as possible. The collection contains writings of social media content from two categories of users: users that at some point in their history have harmed themselves and control users, that do not have any history of self-harming. A labelled training collection was released prior to the evaluation period. For the test stage a server that iteratively releases user writings was set up by the organization. After each round of writings that the server would release, a decision had to be emitted. Classifying a user as being prone to self-harm was considered an irreversible decision, while a decision of non-self-harming was open to updates in the following rounds of decisions. Self-harm ideation often relates to depression and poor mental health, therefore we were interested in exploring psycholinguistic features that are found in the written or oral expressions of depressed users.

3.1 Dataset description

The training and test collection for this task have the same format as the collection described in [18]. They represent collections of writings (posts or comments) from a set of social media users and, for each user, the collection contains a sequence of writings in chronological order. Unlike the same task that ran in 2019, this year edition provided a training dataset. The characteristics of the training set are presented in Table 1.

3.2 Metrics

The evaluation metrics that have been regularly used for the eRisk challenges is ERDE, the early risk detection measure proposed by Losada et al. [18]. As identified in last year's overview report [19], this measure has several drawbacks, which led to the inclusion of alternative evaluation metrics. As such, $F_{latency}$ a

Table 1. Task1 training dataset.

	Self-harm	Control
#subjects	41	299
#submissions	6 927	163 506
avg #posts/subject	169.0	546.8
avg #words/post	24.8	18.8

measure proposed by Sadeque et al. [26] was also used. This measure takes into consideration the effectiveness of the decision (estimated with the F measure) and the delay for emitting the decision. A perfect system would get an $F_{latency}$ of 1. These metrics are further complemented with a ranking evaluation of the systems after seeing k writings, with varying k.

3.3 Methods

For this task we submitted 3 different runs. Each team was allowed to submit up to five different runs. All runs had to complete one round's decisions prior to getting the next round writing. This means there could not be any transfer learning from one run to another. For all 3 runs, we followed a number of common steps in the preprocessing phase. The posts were lowercased and tokenized. Stopwords are filtered, based on the stopwords list of the Natural Language Toolkit[2].

For the first run, we also removed all non-alphabetic characters. For this approach, we followed a standard processing stream for text classification. We initially split the dataset into training and validation chunks, with a ratio of 2:1. We considered Bag of Words (BoW) and tf-idf based feature weighting with linear Support Vector Machine with Stochastic Gradient Descent and Passive Aggressive classifiers. We trained and validated both classifiers on the validation corpus. The SVM led to slightly better results in terms of F1 in the validation stage, so we retrained the model with the whole corpus (training + validation). We only started emitting decisions in the 10th round of server writings and we did all the classification online, without applying any offline knowledge. This means that in the first 9 rounds all decisions were emitted as 0. This threshold for the delay in emitting the decision was selected based on our previous participation, where we concluded that each user had a history of at least 10 writings.

Our second run was based on a mixture of machine learning and psycholinguistic features. The methodology is composed by five different feature extraction algorithms. The first two algorithms are intended to compute features within the data by measuring the frequencies of specific characters or words. The first one acts before any processing takes place and its purpose is to find emojis and punctuation symbols on the given text. The second one, receives as argument a list of self-harm related keywords. Synonyms as well as antonyms are extracted for every keyword using NLTK's wordnet [13]. Moreover, it also has a collection of sets of words. Some of the sets are absolutist words [1], first person words

[2] https://www.nltk.org/

and symptoms. We summarize in Table 2 the main linguistics features that we considered and the lexicon source. The list of absolutist words is presented in Table 3.

Table 2. Linguistic features and source lexica.

Feature	Source
Negative words	https://www.enchantedlearning.com/wordlist/negativewords.shtml
Positive words	https://www.enchantedlearning.com/wordlist/positivewords.shtml
Symptoms	https://www.valleybehavioral.com/disorders/self-harm/
Related diseases	https://www.valleybehavioral.com/disorders/self-harm/
harm lexicon	https://www.thesaurus.com/browse/harm
depression lexicon	https://www.thesaurus.com/browse/depression
anxiety	https://www.thesaurus.com/browse/anxiety

The third algorithm is a tf-idf vectorizer which turns the text into a tf-idf matrix. The forth and fifth algorithm use paragraph vectors based on gensim [3] Doc2vec [16], with two different models, Distributed Memory and Distributed Bag of Words. The output of these five algorithms is concatenated and the best features are extracted. These features are then fed to the Adaboost classifier, which led to better results in the validation stage among different classifiers that we trained.

Table 3. Absolutist words validated by Al-Mosaiwi et al. [1].

absolutely	all	always	complete	completely
constant	constantly	definitely	entire	ever
every	everyone	everything	full	must
never	nothing	totall	whole	

Our third and last run was a combination of the previous two in a sense that it made a decision based on the highest probability score output by the two first runs. Simply put, our third run would emit the decision whose score in the previous two runs was higher when the decisions emitted by the first two runs would not be identical.

3.4 Results

The results obtained are shown in Table 4, along with the best results in this task, for comparison. The results of all participating teams can be found in [20].

Our second run obtained the best scores among our three submission. This was somehow expected as it was the most complex one. It took into consideration

[3] https://radimrehurek.com/gensim/auto_examples/index.html

Table 4. Evaluation of BioInfo@UAVR's submission in Task 1. The best results were added for comparison. It is important to note that no single team reached the best results in all metrics.

	P	R	F1	ERDE5	ERDE50	latency	speed	latency-weighted F1
Run 1	.609	.375	.464	.260	.178	14	.949	.441
Run 2	.591	.654	.621	.273	.120	11	.961	.597
Run 3	.629	.375	.470	.259	.177	13	.953	.448
Best results	.913	1	.754	.134	.071	1	.1	.658

not only linguistic features, but also psycholinguistic and behavioral patterns. Unfortunately for us, during the test period we were only able to process roughly a quarter of the total writings. That was mainly due to our late submission of the runs and to some backup disk writings that slowed down our script. Following the competition's test phase, we processed off-line the whole corpus with the same algorithms that we submitted in our best-performing run (second run). We simulated the same round-based writing release and the results obtained in our off-line simulation setup are very close to the best results obtained during the on-line test stage. In this off-line test stage, we obtained a precision score of 0.80, a recall score of 0.58 and an F1 - score of 0.67. Furthermore, using a different classifier to build the pipeline, a deep learning model, the results were even better. Precision score of 0.75, recall score of 0.72 and F1 score of 0.73.

4 Task 2 - Estimating the level of depression

This task was aimed at exploring the viability of automatically estimating the severity of multiple symptoms associated with depression [20]. Given the user's history of writings, participants had to work out a solution for predicting the user's response to each individual question included in Beck's Depression Inventory Questionnaire (BDI) [3]. The questionnaire assesses the presence of feelings like sadness, pessimism, loss of energy, hunger/loss of appetite, etc. For each individual question, a numeric value between 0 and 3 is considered a valid answer, with the exception of two questions, whose possible answers were: 0, 1a, 1b, 2a, 2b, 3a or 3b.

4.1 Dataset description

The training dataset was the dataset used in the test stage of the task's first edition, CLEF eRisk 2019 [19]. It contained 20 files, with one file per user provided. For each user, a file containing the history of writings on a social network was also provided. An annotation file, or ground truth file was also provided, containing the answers of all users to each of the questions in the questionnaire. The number of writings per user varied from 30 to 1511. The average number of writings of the dataset was 548, with a median of 328.5.

4.2 Metrics

The organizers of this task collected questionnaires filled by social media users together with their history of writings. For each user, the history of writings was extracted right after the user provided us the filled in questionnaire. The questionnaires filled by the users were considered the ground truth and were used to assess the quality of the responses provided by each participating team.

The evaluation metrics reflected the differences between the answers of the questionnaire provided by the task participants and the ones provided by the users that were part of the dataset. Moreover, in the psychological domain it is customary to associate depression levels with categories. Depression levels are defined as the sum of all answers of the 21 questions of the questionnaire. The following depression categories were used for further extension of the evaluation metrics:

- minimal depression - [0–9]
- mild depression - [10–18]
- moderate depression - [19–29]
- severe depression - [30–63]

The following metrics were considered for the evaluation of the results [20]:

- *Hit Rate (HR)* - the ratio of cases where the automatic questionnaire has exactly the same answer as the real questionnaire.
- *Average Hit Rate (AHR)* - HR averaged over all users.
- *Closeness Rate (CR)* - the absolute difference between the real and the participant provided answer.
- *Average Closeness Rate (ACR)* - CR averaged over all users.
- *Difference between overall depression levels (DODL).*
- *Average DODL (ADODL)* - DODL averaged over all users.
- *Depression Category Hit Rate (DCHR)* - the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire.

4.3 Methods

Our approach for solving this task built upon algorithms that we used in CLEF eRisk 2019 edition for solving not only this task, but also Task 1. In the previous year we have used for the training stage of Task 1 a machine learning model trained on Yates et al. [31] Reddit depression dataset. This dataset consists of all Reddit users who made a post between January and October 2016, matching high-precision patterns of self-reported diagnosis (e.g. "I was diagnosed with depression"). The depressed users were matched by control users, who have never posted in a subreddit related to mental health and never used a term related to it. In order to avoid a straight-forward separation of the two groups, all posts of diagnosed users related to depression or mental health were removed.

The first step in this year's approach to addressing Task 2 was to predict whether a user was depressed using the classifier previously trained of the Yates et. al dataset. Next, we conjugated the score of this classification with several psycholinguistics and behavioral patterns, as presented next. For each category, a score was calculated for each user as a normalized value of the number of occurrences of the features considered for each category with respect to the total number of occurrences of the same features over the dataset. These scores were then normalized to the interval [0,3].

- Lexical category of a user's text - depressed users tend to have an overall more negative connotation of their texts [9,23]. To this purpose we employed Empath, an NLP framework for calculating the average polarity of a user's writings.

- Use of self-related words (e.g: I, myself, mine) - depressed users tend to use them more often in their writings [6, 25]

- Use of absolutist words - Al-Mosaiwi et al. [1] recently showed that anxiety, depression, and suicidal ideation forums contained more absolutist words than control forums. The list of absolutist words used is presented next in Table 3.

- Mentions of words related to mental disorders, (e.g.:depression, bipolar, schizophrenia, psychotic, ocd).

- Use of the words cry, guilt and their derivatives.

- Use of the words sleep, anxious and their derivatives.

- Use of the words irritated, fatigue, tired and their derivatives.

This list is based on the psycholinguistic patterns and semantic clusters that we used in our previous participation in this shared lab. Compared to the approach that we took in our first participation in this task, we decided to remove some of the features that we used last year and we explored the use of Empath [11]. A statistical analysis of the training corpus revealed that the non-depressed users had relatively low depression scores, as it would be expected. As such, the users that our trained model would consider non-depressive were scored with low scores in all categories. Regarding the psycholinguistic features, our follow-up analysis of our eRisk2019 submission revealed that some of the features we included last year did not significantly contribute to the overall scores.

4.4 Results

Task participants had to provide a result file with one line per user in the test dataset. Each line contained the username and 21 values that corresponded to

[3] https://github.com/Ejhfast/empath-client

the answers of the 21 questions included in Beck's Depression Inventory. The results obtained by our team are presented in Table 5, along with the best results obtained in this task, for each of the metrics. The results of all participating teams can be found in [20].While the general results obtained in this task have slightly improved since last year, they stand as proof of its difficulty. One important aspect to be mentioned is that our team obtained the best score for the AHR metric and second best score for ACR.

Table 5. Evaluation of BioInfo@UAVR's submission in Task 2. The best results for each metric were added for comparison. It is important to note that no single team achieved the best results for all metrics.

	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	38.30%	69.21%	76.01%	30.00%
Best scores	38.30%	69.41%	83.15%	35.71%

5 Conclusions and Future Work

We presented in this paper the results of our team's participation in the eRisk2020 shared tasks. Considering this is the second participation in this shared lab, our submissions were built upon the core approaches used in the previous edition. We extended the previous work by having considered more psycholinguistic and behavioral features, which led to more submissions for Task 1 and overall better results obtained in both tasks. While we recognize the potential that social mining has for signaling a user's mental health status and for the early detection of risk situation, we have come to understand that one possible limitation of our work is the lack of clinical knowledge. As researchers with computational backgrounds, who are often unfamiliar with existing practices in mental healthcare, we are in the frontline of developing new algorithms for social data. In order to better understand the tasks that we have in our hands and to improve the end solution we will focus on having the missing clinical perspective on our future participations.

Acknowledgments

This work was supported by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

References

1. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science p. 2167702617747074 (2018)

2. Alonso, S.G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D.C., Nozaleda, L.M., Franco, M.: Data mining algorithms and techniques in mental health: A systematic review. Journal of medical systems **42**(9), 161 (2018)
3. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Archives of general psychiatry **4**(6), 561–571 (1961)
4. Benton, A., Coppersmith, G., Dredze, M.: Ethical research protocols for social media health research. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp. 94–102 (2017)
5. Chen, L., Hossain, K.T., Butler, P., Ramakrishnan, N., Prakash, B.A.: Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. Data mining and knowledge discovery **30**(3), 681–710 (2016)
6. Chung, C., Pennebaker, J.W.: The psychological functions of function words. Social communication **1**, 343–359 (2007)
7. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 51–60 (2014)
8. Coppersmith, G., Leary, R., Whyne, E., Wood, T.: Quantifying suicidal ideation via language usage on social media. In: Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM (2015)
9. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. ICWSM **13**, 1–10 (2013)
10. Ernala, S.K., Birnbaum, M.L., Candan, K.A., Rizvi, A.F., Sterling, W.A., Kane, J.M., De Choudhury, M.: Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 134. ACM (2019)
11. Fast, E., Chen, B., Bernstein, M.S.: Empath: Understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 4647–4657. ACM (2016)
12. Fatima, I., Abbasi, B.U.D., Khan, S., Al-Saeed, M., Ahmad, H.F., Mumtaz, R.: Prediction of postpartum depression using machine learning techniques from social media text. Expert Systems p. e12409
13. Fellbaum, C.: Wordnet. The encyclopedia of applied linguistics (2012)
14. Giannotti, F., Trasarti, R., Bontcheva, K., Grossi, V.: Sobigdata: social mining & big data ecosystem. In: Companion Proceedings of the The Web Conference 2018. pp. 437–438 (2018)
15. Kim, Y., Huang, J., Emery, S.: Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. Journal of medical Internet research **18**(2) (2016)
16. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
17. Liang, Y., Zheng, X., Zeng, D.D.: A survey on big data-driven digital phenotyping of mental health. Information Fusion **52**, 290–307 (2019)
18. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 28–39. Springer (2016)
19. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)

20. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: A. Arampatzis, E. Kanoulas, T.T.S.V.H.J.C.L.C.E.A.N.L.C.N.F.e. (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Springer International Publishing (2020)
21. Loveys, K., Crutchley, P., Wyatt, E., Coppersmith, G.: Small but mighty: Affective micropatterns for quantifying mental health from social media language. In: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality. pp. 85–95 (2017)
22. MacAvaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., Goharian, N.: Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 168–173 (2018)
23. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD). vol. 2012, pp. 1–8. ACM New York, NY (2012)
24. Paul, M.J., Dredze, M.: Social monitoring for public health. Synthesis Lectures on Information Concepts, Retrieval, and Services **9**(5), 1–183 (2017)
25. Rude, S., Gortner, E.M., Pennebaker, J.: Language use of depressed and depression-vulnerable college students. Cognition & Emotion **18**(8), 1121–1133 (2004)
26. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 495–503. ACM (2018)
27. Shing, H.C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., Resnik, P.: Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 25–36 (2018)
28. Trifan, A., Antunes, R., Matos, S., Oliveira, J.L.: Understanding depression from psycholinguistic patterns in social media texts. In: European Conference on Information Retrieval. pp. 402–409. Springer (2020)
29. Vaterlaus, J.M., Patten, E.V., Roche, C., Young, J.A.: # gettinghealthy: The perceived influence of social media on young adult health behaviors. Computers in Human Behavior **45**, 151–157 (2015)
30. Wongkoblap, A., Vadillo, M.A., Curcin, V.: Researching mental health disorders in the era of social media: systematic review. Journal of medical Internet research **19**(6), e228 (2017)
31. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. p. 2968–2978. Association for Computational Linguistics (2017)
32. Yun, G.W., Morin, D., Park, S., Joa, C.Y., Labbe, B., Lim, J., Lee, S., Hyun, D.: Social media and flu: Media twitter accounts as agenda setters. International journal of medical informatics **91**, 67–73 (2016)
33. Zhang, J., Brackbill, D., Yang, S., Centola, D.: Identifying the effects of social media on health behavior: Data from a large-scale online experiment. Data in brief **5**, 453–457 (2015)