engineering proceedings

MDPI

Proceeding Paper

Multiple-Choice Question Answering Models for Automatic Depression Severity Estimation †

Jorge Gabín *, Anxo Pérez and Javier Parapar

Information Retrieval Lab, Centro de Investigación en Tecnoloxías da, Información e as Comunicacións (CITIC), Universidade da Coruña, 15071 A Coruña, Spain; anxo.pvila@udc.es (A.P.); javier.parapar@udc.es (J.P.)

* Correspondence: jorge.gabin@udc.es; Tel.: +34-881-01-12-76

† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Depression is one of the most prevalent mental health diseases. Although there are effective treatments, the main problem relies on providing early and effective risk detection. Medical experts use self-reporting questionnaires to elaborate their diagnosis, but these questionnaires have some limitations. Social stigmas and the lack of awareness often negatively affect the success of these self-report questionnaires. This article aims to describe techniques to automatically estimate the depression severity from users on social media. We explored the use of pre-trained language models over the subject's writings. We addressed the task "Measuring the Severity of the Signs of Depression" of eRisk 2020, an initiative in the CLEF Conference. In this task, participants have to fill the Beck Depression Questionnaire (BDI-II). Our proposal explores the application of pre-trained Multiple-Choice Question Answering (MCQA) models to predict user's answers to the BDI-II questionnaire using their posts on social media. These MCQA models are built over the BERT (Bidirectional Encoder Representations from Transformers) architecture. Our results showed that multiple-choice question answering models could be a suitable alternative for estimating the depression degree, even when small amounts of training data are available (20 users).

Keywords: depression prediction; social media; pre-trained language models; multiple-choice question answering

check for updates

Citation: Gabín, J.; Pérez A.; Parapar, J. Multiple-Choice Question Answering Models for Automatic Depression Severity Estimation. *Eng. Proc.* **2021**, *7*, 23. https://doi.org/10.3390/engproc2021007023

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 12 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The World Health Organization (WHO) [1] placed mental health as one of the most relevant components of health. Depression is one of the most common mental disorders. By itself, it affects more than 270 million people. Despite having many harmful effects, there are some effective known treatments. The main problem relies on providing early and effective risk detection.

One of the most reliable and frequent methods to measure depression severity is the Beck Depression Inventory-II (BDI-II) [2]. Although significant evidence exists regarding its performance, some aspects often affect the results of these questionnaires.

These days, health organizations are publishing these questionnaires so that users can fill them in by themselves. However, people with mental disorders usually do not dare to visit those web pages and fill in the questionnaires. In this new communication era, people use social networks to share their feelings and emotions. Hence, these platforms are a great way to collect data to identify disorders like depression [3].

In this context, we describe an approach to improve the automatic estimation of the degree of depression from users on social media. Our study presents the use of pre-trained language models [4] to predict the depression degree of subjects. We evaluated these models for the task "Measuring the Severity of the Signs of Depression" of the CLEF 2020 eRisk Track [5]. Our results achieved moderate performance among all the participants of the task.

2. Experiments

2.1. Datasets

In this study, we use the datasets provided by eRisk 2019 and 2020 for the task "Measuring the Severity of the Signs of Depression" [5,6]. Each dataset contains the history of 20 and 70 users, respectively, providing the users' actual responses to the questionnaire and its complete history of postings. We used the 2019 dataset as training data and the 2020 dataset for testing.

We also used RACE (Large-scale ReAding Comprehension Dataset From Examinations) [7] and SWAG (Large-Scale Adversarial Dataset for Grounded Commonsense Inference) [8], two general-purpose multiple-choice question answering datasets. After some preliminary comparisons, we selected the RACE dataset to perform the first fine-tune over BERT as the results obtained were slightly better.

2.2. Beck Depression Inventory-II (BDI-II)

Beck Depression Inventory-II (BDI-II) is a questionnaire formed by 21 items to measure the depression severity. For each item, the BDI-II provides four options (except items 15 and 17, which provide seven options) and sentences to explain their meanings. These options represent a scale from the absence of the symptom to a total identification.

2.3. Models

We used a modified BERT [9] model for Multiple-Choice Question Answering (MCQA). This model was built over the pre-trained `bert-base-uncased` model, modifying it to allow multiple-choice question answering. In [4], we can see the process followed to build the model and its comparison with other baseline models.

We also tried to pre-train the MCQA models provided by the Hugging Face library (such as RoBERTa for multiple-choice), but the results obtained were much worse than those obtained using the adaptation mentioned.

2.4. Our Approach

Pre-trained language models are usually trained on a large text corpus and then fine-tuned on a downstream task. Following this approach, in the training phase, we fine-tuned a pre-trained model using the RACE dataset. We additionally fine-tuned the model using training data from the 2019 eRisk task. For that, we built a custom dataset that contains every post from each user combined with each question from the BDI-II questionnaire with all its options (0–3), and the label which represents the actual option was chosen by the user. After analyzing the results obtained, we decided to filter the training data as there was too much noise. Therefore, we calculated the post and question embeddings and used only the top 50 posts more similar to each question as training data in the fine-tuning process.

To run both fine-tunings, we used a batch size of eight (four, when fine-tuning with the seven options dataset), a maximum sequence length of 320, a learning rate of 5×10^{-5}, two epochs, and two gradient accumulation steps.

To carry out inference, we feed the model with every post from each user combined with each question from the BDI-II questionnaire with all its options. As a result, we will receive the model's confidence on each option for each pair of post-questions. Given that confidence, we can extract the inferred answer for each paired user-question by selecting the option with the most appearances.

In this phase, we used the following parameters: a batch size of 48, a maximum sequence length of 320, and a minimum option probability of 0.4 (0.2 for seven options questions). We subtract 0.01 from the minimum probability if no posts achieve that minimum probability.

Finally, to facilitate the whole inference process, we built another dataset using the test data from the 2020 task and following the same approach as explained before.

3. Results and Discussion

In Table 1, we show the results obtained following the explained approach.

Table 1. Results of our model, along with the best baselines of eRisk 2020. Bold values correspond to the best result for each metric.

Model	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
BioInfo@UAVR [10]	**38.30**	69.21	76.01	30.00
ILab [11]	37.07	**69.41**	81.70	27.14
Prhlt-Upv [12]	34.56	67.44	80.63	**35.71**
Relai [13]	36.39	68.32	**83.15**	34.29
MCQA Model	25.03	57.76	75.58	31.43

On the one hand, we can see in the results table that we get good results in both ADODL and DCHR metrics. However, on the other hand, the results obtained in AHR and ACR metrics were poor compared with the best results of the 2020 task.

Inspecting the model's answers to the BDI-II, we could see that it overestimates depression severity. This is likely because both the train and test data are still noisy. With this in mind, in future work, we plan to design more effective data filtering processes on both the training and test data.

4. Conclusions

In this article, we studied the application of pre-trained multiple-choice question answering models to automatically estimate the depression severity of users on social media. The results obtained are promising and a good starting point to continue researching this type of model.

Acknowledgments: This work was supported by projects RTI2018-093336-B-C22 (MCIU & ERDF), GPC ED431B 2019/03 (Xunta de Galicia & ERDF) and CITIC, which is financial supported by Consellería de Educación, Universidade e Formación Profesional of the Xunta de Galicia through the ERDF (80%) and Secretaría Xeral de Universidades (20%), (Ref ED431G 2019/01).

References

1. World Health Organization. Health & Environmental Research Online (HERO). In *Proceedings of the Preamble to the Constitution of the World Health Organization as Adopted by the International Health Conference, New York, NY, USA, 19–22 June 1946*; World Health Organization: Geneva, Switzerland, 1948.
2. Beck, A.T.; Steer, R.A.; Brown, G.K. *Beck Depression Inventory (BDI–II)*; Pearson: London, UK, 1996; Volume 10.
3. De Choudhury, M.; Counts, S.; Horvitz, E. Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference, Paris, France, 2–4 May 2013; pp. 47–56.
4. Xu, K.; Tin, J.; Kim, J. A BERT based model for Multiple-Choice Reading Comprehension. *Passages* **2019**, *6*, 362.
5. Losada, D.E.; Crestani, F.; Parapar, J. eRisk 2020: Self-harm and depression challenges. In Proceedings of the European Conference on Information Retrieval, Lisbon, Portugal, 14–17 April 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 557–563.
6. Losada, D.E.; Crestani, F.; Parapar, J. Overview of erisk 2019 early risk prediction on the internet. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Lugano, Switzerland, 9–12 September 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 340–357.
7. Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. Race: Large-scale reading comprehension dataset from examinations. *arXiv* **2017**, arXiv:1704.04683.
8. Zellers, R.; Bisk, Y.; Schwartz, R.; Choi, Y. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv* **2018**, arXiv:1808.05326.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *CoRR* **2017**, *2017*, 5998–6008.
10. Oliveira, L. BioInfo@ UAVR at eRisk 2020: On the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. In Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece, 22–25 September 2020; Volume 2696; CLEF (Working Notes).

11. Martínez-Castaño, R.; Htait, A.; Azzopardi, L.; Moshfeghi, Y. Early risk detection of self-harm and depression severity using BERT-based transformers: iLab at CLEF eRisk 2020. In Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece, 22–25 September 2020; Volume 2696; CLEF (Working Notes).
12. Uban, A.S.; Rosso, P. Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece, 22–25 September 2020; CLEF (Working Notes); Sun SITE Central Europe: Aachen, Germany, 2020; Volume 2696, pp. 1–12.
13. Maupomé, D.; Armstrong, M.D.; Belbahar, R.M.; Alezot, J.; Balassiano, R.; Queudot, M.; Mosser, S.; Meurs, M.J. Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. In Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece, 22–25 September 2020; Volume 2696; CLEF (Working Notes).