FilipN@LT-EDI-ACL2022-Detecting signs of Depression from Social Media: Examining the use of summarization methods as data augmentation for text classification

Filip Nilsson and György Kovács
EISLAB Machine Learning
Lulea University of Technology
filnil-8@student.ltu.se, gyorgy.kovacs@ltu.se

Abstract

Depression is a common mental disorder that severely affects the quality of life, and can lead to suicide. When diagnosed in time, mild, moderate, and even severe depression can be treated. This is why it is vital to detect signs of depression in time. One possibility for this is the use of text classification models on social media posts. Transformers have achieved state-of-the-art performance on a variety of similar text classification tasks. One drawback, however, is that when the dataset is imbalanced, the performance of these models may be negatively affected. Because of this, in this paper, we examine the effect of balancing a depression detection dataset using data augmentation. In particular, we use abstractive summarization techniques for data augmentation. We examine the effect of this method on the LT-EDI-ACL2022 task. Our results show that when increasing the multiplicity of the minority classes to the right degree, this data augmentation method can in fact improve classification scores on the task.

1 Introduction

The number of people suffering from depression has been steadily increasing since the 1990s (of Health Metrics and Evaluation, 2019), therefore it is essential that we find an efficient method to identify this on the internet. Over the past few years, transformers have taken over the field of Natural Language Processing (NLP) and achieved state-of-the-art results on various problems (Wolf et al., 2020).

Some classification problems in machine learning deal with the problem of class imbalance. In this paper, we examine the effect different degrees of data augmentation have on the performance of transformer models on a text classification task. The method of data augmentation is done using abstractive summarizations. Our data augmentation is done by first generating summarizations for each of the training examples and then balance the dataset using these generated summarizations. For this, first we discuss the related literature in Section 2. Then in Section 3 we briefly describe the dataset used. This will be followed by the description of our methods in Section 4, after which we discuss our results in Section 5, then end the paper with our conclusions and plans for future work in Section 6.

2 Related work

Data augmentation is nothing new to the field of NLP, it is one of the standard approaches when improving the results of a model. There are many different approaches to data augmentation and the NLP survey (Feng et al., 2021) puts these methods into three different categories rule-based, example interpolation and model-based techniques. The latter is the approach that we focus on in this paper in which we use the T5 (Raffel et al., 2019) model to summarize the original posts. There are multiple different tasks that data augmentation can aim to solve such as mitigating bias, fixing class imbalance and few-shot learning. Yet in this paper we solely focus on fixing class imbalance.

The area of data augmentation for fine-tuning transformers is limited and is still being explored (Feng et al., 2021). Yet some research has been done such as GenAug (Feng et al., 2020) which describes methods to use data augmentation to fine-tune text generators. GenAug focuses on character-level synthetic noise and keyword replacement as augmentation methods for fine-tuning. Although this data augmentation is done for text generation with GPT-2 (Radford et al., 2019) and not for a sentiment-analysis task. The (Kumar et al., 2020) paper shows that there are effective ways to use data augmentation methods to fine-tune transformers to achieve better results on abstractive summarization.

Using transformers as a data augmentation method has been done previously in papers such as (Sabry et al., 2022) where they used DialoGPT (Zhang et al., 2019) to generate new training data

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 283 - 286
May 27, 2022 ©2022 Association for Computational Linguistics

and effectively double the training data. The augmented dataset was then used to fine-tune a T5 model where they showed a positive effect on the results.

Previous work using transformers to detect depression in social media posts has been done by (Martínez-Castaño et al., 2020). Where they used the BERT transformer to analyze the risk a user was of self-harming themselves by classifying social media posts from that user.

3 Data

The dataset was provided by the organizers of LT-EDI-ACL2022 (Sampath et al., 2022) in the competition and it contains social media posts from different users, categorized as severe, moderate and not depression. The dataset was created and annotated by the methods described in (Kayalvizhi and Thenmozhi, 2022). These posts differ greatly from BERTs training data, the dataset is not very large and very imbalanced (some classes are largely underrepresented). Therefore a good dataset to study the effect of data augmentation.

We were provided a training set, validation set and a test set. Labels for the test set however, have not been published yet. Hence this paper is based solely on the results from the training and validation set, therefore our validation set is used as a test set, further references in this paper to a test set is the validation set from the LT-EDI-ACL2022 competition. Table 1 shows our datasets.

Class	Training	Test
Severe	901	360
Moderate	6019	2306
Not Depression	1971	1830
All	8891	4496

Table 1: The number of labels for the two datasets.

4 Methodology

The pipeline used in this paper consists of two major parts, the first part performs the data augmentation by summarizing the training dataset, the second part is the classification of the levels of depression. Our approach to solving the problem of detecting depression in social media posts is to fine-tune a multiclass BERT transformer on our dataset using different degrees of data augmentation. The following section will describe in-depth how this is done. To ensure repeatability, our code is shared in a Github[1] repository.

4.1 Preprocessing

Minimal preprocessing was done on the dataset, URLs were removed and we used Huggingface base pre-trained BERT tokenizer[2] trained on WordPiece.

4.2 Data Augmentation

In NLP two major approaches to summarization has evolved, namely extractive and abstractive summarization. Extractive is when the model receives a text as input and has to select the best concatenation of sentences that best summarizes that input. Whereas in abstractive summarization the model has to generate the summarization by itself.

In this paper we chose to use Google's T5 (Raffel et al., 2019) which is a text-to-text transformer pre-trained on the c4 dataset[3] then fine-tuned for abstractive summarization. Even though the c4 dataset differs from our dataset T5 still achieves state of the art results in summarization and hence was used for the task of summarization in this task. We chose to use Huggingface implementation of t5-base[4]. T5 is trained on a maximum sequence length of 512 tokens, that is a limitation of the model since it cannot always take the entire post as input.

As seen in Table 2 the summarizations that T5 produces do not always include the most important part of the sentiment in the social media post. In Table 2 the 30 token summary of the severe example includes suicide attempts which can be vital for the classification while the 10 token summary misses that.

We chose a specific length by which t5-base should generate its summarizations. These lengths were between 10 and 50 tokens and went through all the training examples in the underrepresented classes. Which were randomly sampled when balancing respective class for the training dataset.

The data augmentation was done in five different degrees of how balanced the underrepresented classes were. This balancing can be seen in Table 3, in the 0 example we have the training dataset.

[1] https://github.com/flippe3/DSDSM_augmentation

[2] https://ai.googleblog.com/2021/12/a-fast-wordpiece-tokenization-system.html

[3] https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

[4] https://huggingface.co/t5-base

Type	Severe	Not depression
Original	I'm really struggling : So I don't know how to start things like this, So I'll start with basics. I'm 16yo, diagnosed depression at 14yo. Since then, my life is total mess. I've already been to two different psychologists, both of them said that i'm kind of unfixable, but that might be due to my young age.	Hows everyone doing on this new years eve? : I know as well as everyone new years makes you think back and regret and wonder why you're still here one more year. Well i'm spending new years alone at home and I'm just wondering how everyone is doing and if you need a place to vent or talk this can be a safe place for you.
30 token	i'm 16yo, diagnosed depression at 14yo. since start of 2019, my life got even worse. i've had 6 suicide attempts, all of them ended up with getting yelled at	i'm spending new years alone at home and wondering how everyone is doing. if you need a place to vent or talk this can
10 token	i'm 16yo, diagnosed depression at 14yo. i'm	i'm spending new years alone at home and

Table 2: Two examples of different summarizations over different classes.

Class	0	25	50	75	100
Severe	901	1505	3010	4515	6019
Moderate	6019	6019	6019	6019	6019
Not Depression	1971	1971	3010	4515	6019

Table 3: The different degrees of balancing.

In the 50 example we can see that the two underrepresented classes have been augmented by the summarization examples, the 50 represents that both of the classes now are at least 50 percent as large as the largest class.

4.3 Classification

To measure how well our model preforms we chose to use Googles's BERT (Devlin et al., 2019), with the base configuration that has 12 layers and 110 million parameters. BERT is a bidirectional transformer trained for language modeling. We chose to use BERT as the underlying model as it has become the standard in transformers when fine-tuned on downstream tasks. We used the Huggingface implementation of bert-base-uncased[5] for easier experimentation. The fine-tuning was done using a multiclass labelling version of BERT. We used weighted Adam optimizer and a linear scheduler as that generated the best results.

5 Experiments and results

The following subsection describes the experiments and result that were produced. BERT ran four epochs of fine-tuning and the best Macro F1-score was chosen to represent the result for that model. The models were trained on a shared DGX-1 cluster with 8 × 32GB Nvidia V100 GPUs.

5.1 Results

To evaluate the proposed data augmentation method, we applied it on the multi class classification task of the LT-EDI-ACL2022 challenge. In the competition we placed 31th using a method of classification that preformed similarly to the model presented in this paper. Before writing this paper however we changed our model to BERT and added the data augmentation. Our results on the validation set distributed with the challenge are outlined in Table 4. As can be seen in Table 4, although augmenting the data to the point of a completely balanced dataset improves the recall, it is at the cost of a lower precision. However, when selecting the degree of balancing carefully, one can improve the recall without a significant negative effect on precision. Unfortunately, here, we were not able to add results on the test set, however, upon the release of test labels, we would amend our table with classification scores attained on the test set too.

Score	0	25	50	75	100
Macro F1	0.50	0.52	0.52	0.52	0.49
Macro Recall	0.50	0.52	0.54	0.51	0.52
Macro Precision	0.57	0.57	0.57	0.55	0.56

Table 4: F1-Macro scores on five different degrees of data augmentation.

[5] https://huggingface.co/bert-base-uncased

6 Conclusions and Future Work

We examined a method of using an abstractive summarization model, T5 to do data augmentation. This was done before fine-tuning a BERT transformer on the dataset which was balanced to different degrees. We found that with the right degree of augmentation, the proposed method improved the performance of the BERT model on the task of detecting signs of depression. One technique that can be examined for further improvement of the proposed model is splitting the posts longer than 512 tokens into several parts, summarizing the parts individually, and creating the final summary by concatenating the individual summaries. Future work for data augmentation for fine-tuning transformers could be done by comparing the result using an extractive summarization method such as MatchSum (Zhong et al., 2020). Our method was examined on one dataset, future work should use this technique of data augmentation to fine-tune for different domains on multiple datasets. There are also different methods of augmentation other than summarization that future work should examine and compare. In this paper we used BERT for our classification, future work should use other transformer models to see how they perform using our augmentation method.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*.

Steven Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020. *CEUR Workshop Proceedings*, 2696. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020. urn:nbn:de:0074-2696-0; Early Risk Prediction on the Internet : CLEF workshop, eRisk at CLEF ; Conference date: 22-09-2020 Through 25-09-2020.

Institute of Health Metrics and Evaluation. 2019. Gbd results tool.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovacs, Foteini Liwicki, and Marcus Liwicki. 2022. Hat5: Hate language identification using text-to-text transfer transformer.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.