Method based on transformers for the early suicide detection task

Ilicea Anca Stefania
Department of Natural Language Processing, University of Bucharest
anca.ilicea@s.unibuc.ro

Abstract

This report presents an end-to-end approach for classifying, from given text, if an individual is suicidal or just suffering from a light depression using a modern natural language processing technique, based on transformers. My approach uses a tokenizer to preprocess the data, so the text inputs will be prepared for the model, this ensuring that all samples have the maximum length that the model can take, either by padding or truncating them. Furthermore, I will present some of the dataset's disadvantages, errors in the labeling process and the results of an alternative strategy in which the labels were changed.

1 Introduction

Depression is a disease that is affecting more and more people nowadays. Detecting it early could be life-saving, allowing patients suffering from depression to receive the appropriate medicine, as soon as possible. However, distinguishing between a text written by a person suffering from depression and a text written by a person suffering from a more severe form of depression, being on the verge of committing suicide, is a difficult task due to the lack of sufficiently large and correctly labeled datasets. Not only the dataset would be the problem, but also the fact that this differentiation can hardly be done even by a specialist, given the fact that he would have a single text at his disposal. An example of texts belonging to these two categories are presented in Table 1.

To begin, I took the ground truth of an existing dataset (SDCNL) [1] as the starting point for the work. Specifically, I worked with the data precisely as it was provided, assuming that the text examples are labeled correctly, and I tried to achieve the best results by determining if a text represents depression or a deeper disease that can lead to suicide.

The next part summarizes the associated work, and the three sections that follow will draw an end-to-end approach to the problem, including dataset specifics, results, analysis, issues with the original dataset, and ultimately a conclusion and some ideas for future work.

2 Related work

As a main starting point I followed the paper named 'Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction' which was written by Ayaan Haque, Viraaj Reddi and Tyler Giallanza in the summer of 2021. This paper not only addresses various models based on transformers and convolutional neural networks for analysis of the results, but also proposes an interesting method for correcting mislabeled texts based on two unsupervised models (*KMeans* and *Gaussian Model Mixture*).

The best accuracy obtained by the authors of this paper for the dataset with original labels was 72.24 percentage using the GUSE model with a fully-dense neural network (guse-dense), while the state-of-the-art for this task is represented by the solution in which the labels were corrected. GUSE is a transformer trained and optimized for greater-than-word length text and it returns a 512 × 1 dimensional vector.

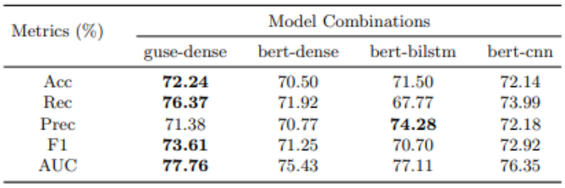
Thus, the highest accuracy is 93.68 percentage for the same guse-dense model, but with labels predicted by the unsupervised methods. The results can be seen in the Figure 2.

Although, when it comes to the simple clas-

[1] https://github.com/ayaanzhaque/SDCNL

Input text	**label**
'I'm never really depressed over stuff and my mind is clear.'	0
'I wish I was someone else. I wish I wasn't so broken.'	1

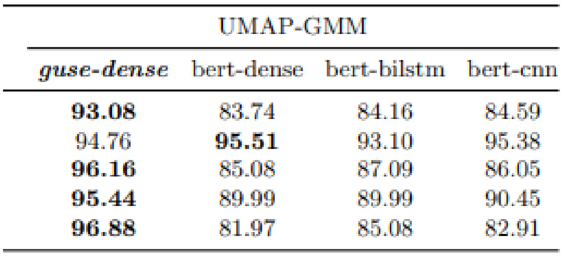
Table 1: Suicide and non-suicide texts.



Metrics (%)	Model Combinations			
	guse-dense	bert-dense	bert-bilstm	bert-cnn
Acc	**72.24**	70.50	71.50	72.14
Rec	**76.37**	71.92	67.77	73.99
Prec	71.38	70.77	**74.28**	72.18
F1	**73.61**	71.25	70.70	72.92
AUC	**77.76**	75.43	77.11	76.35

Figure 1: Performance of the four best combinations of embedding models and classifiers.

sification between depression and a healthy mind, more research can be discussed. In the current, case the paper presented above is the only one so far studying this case.



UMAP-GMM			
guse-dense	bert-dense	bert-bilstm	bert-cnn
93.08	83.74	84.16	84.59
94.76	**95.51**	93.10	95.38
96.16	85.08	87.09	86.05
95.44	89.99	89.99	90.45
96.88	81.97	85.08	82.91

Figure 2: Final classification performance after using the label correction method.

3 Dataset

Many existing approaches for detecting suicidal thoughts rely on data from sources such as surveys, Electronic Health Records (EHRs), and suicide notes. However, research has shown that people who suffer from depression tend to express their feelings more through written text, preferably through an anonymous account, rather than communicating negative feelings effectively, for example, with friends.

Well, the authors of the dataset relied on this idea when building it, so the dataset contains about 1500 training data and 400 test messages, all representing real Reddit posts. Reddit could be described as an online social media platform in which users can create, alternate and use discardable accounts to ensure privacy and anonymity while posting their thoughts.

Although the existing dataset is a good starting point for this classification problem, in order to have better quality results and have the opportunity to try other methods based on convolutional neural networks build from scratch, a much larger dataset would be needed. In this case, its labels should be assigned by a specialist in the field of psychology.

Split. Due to the fact that this available dataset is not so sizeable, it was devided into train and validation datasets with a 8 : 2 ratio (80 percentage for train, 20 percentage for validation).

4 Method

In order to build the final solution, I used a tokenizer from the *Huggingface* library for the bert-uncased-model to preprocess the text. These tokenized data were passed on to a bert-base-uncased model. BERT is an acronym that stands for Bidirectional Encoder Representations from Transformers. By conditioning on both left and right context in all layers, the model is aimed to pretrain deep bidirectional representations from unlabeled text. As a result, the pre-trained BERT model may be fine-tuned with just one extra output layer to generate state-of-the-art models for a wide range of tasks, such as text categorization, without requiring significant task-specific architectural changes.[2]

The model was trained over **5** epochs with a batch-size of **8** and a learning rate of **5e-5**, all these parameters being chosen after multiple attempts. For example, I realized that for a lower learning rate, the loss did not decrease during training, respectively the accuracy did not increase. I utilized the Huggingface library with Pytorch for all transformer-based trainings. The overall accuracy obtained by it was **74.4 percent**.

[2] https://arxiv.org/pdf/1810.04805.pdf

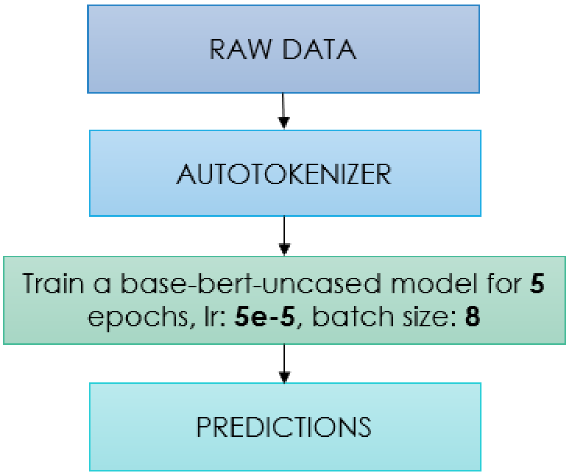

RAW DATA
AUTOTOKENIZER
Train a base-bert-uncased model for **5** epochs, lr: **5e-5**, batch size: **8**
PREDICTIONS

Figure 3: **End to end architecture.** I used Autotokenizer to extract tokens from text and the BERT base uncased model to map those tokens to a probability distribution over is-suicide class

4.1 Baseline

To build a baseline for the current problem I thought of using a classic model for data training. In this case, the first step was the preprocessing of the text. Saying that, I turned all the data into lowercase, I removed multiple whitespaces, hashtags, mentions, links, numbers, punctuation, emoticons, and I also performed a lemmatization of the text. Having this pre-processed texts available, I further chose to use a word representation (of bag of word type) in order to prepare the data for the model. Finally, I trained the data using not only a Support Vector Machine model with linear kernel and C = 1 for both preprocessed and raw data, but also a KNN Classifier with a n-neighbors parameter of 7. The results were pretty good, as you can see in Table 2.

5 Results

In the figures below you can see both the evolution of the loss and the evolution of the accuracy during the training of the BERT type model.

I noticed that the value of the loss increases during training and after researching I realized that a main cause of this increase could be the value of the learning rate being too high, but it turned out that a lower value would lead to a poorer performance. Specifically, the resulting accuracy for a learning rate of 4e-4 was 50.9 percent. However, the final accuracy obtained for the final parameters specified in the Method section was 74.4 percent for the task with the original labels.

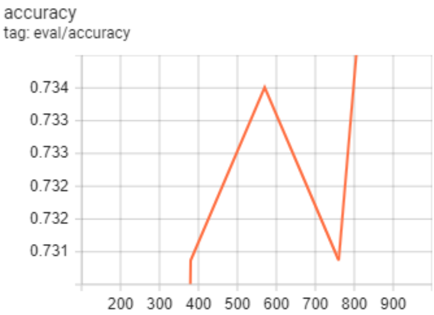


Figure 4: **Accuracy.** Accuracy evolution during the training of the BERT type model for 5 epochs

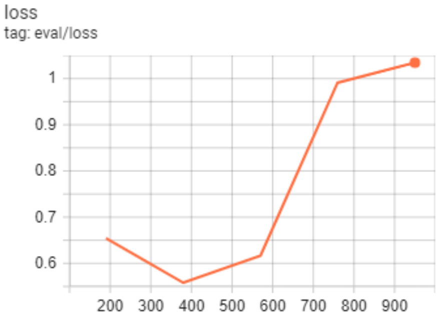


Figure 5: **Loss.** Loss evolution during the training of the BERT type model for 5 epochs

6 Conclusions and future work

Although the results using BERT base model for the data with the original labels were better than those presented in the paper "Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction", I consider that the addition of an unsupervised method for re-labeling the texts together with the chosen model could lead to an even better accuracy.

Moreover, I consider that the application of the two methods on a more comprehensive data set that would also contain posts from Reddint but labeled together with a specialist in the field of psychology could lead to much better results.

Model	Accuracy
base-bert-uncased	74.41 percent
SVM preprocessed data	66.49 percent
SVM raw data	64.37 percent
KNN preprocessed data	56.20 percent
KNN raw data	51.97 percent

Table 2: Results with different models.

References

[1] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova (2019) *Article*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[2] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza (2021) *Article*, Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction

[3] Jay Alammar (2018) *Article*, The Illustrated Transformer

[4] Katz, C., Bolton, J. Sareen , Te prevalence rates of suicide are likely underestimated worldwide: why it matters. Soc. Psychiatry Psychiatr. Epidemiol. 51, 125–127 (2016).

[5] Preventing Suicide: A global imperative. World Health Organisation (2014).

[6] De Choudhury, M., De, S.: Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 8 (2014) 2

[7] Yin Zhang, Rong Jin, Zhi-Hua Zhou. Understanding bag-of-words model: A statistical framework

[8] Theodoros Evgeniou, Massimiliano Pontil. Support Vector Machines: Theory and Applications

[9] Gongde Guo, Hui Wang, David A. Bell, Yaxin Bi. KNN Model-Based Approach in Classification

[10] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., Huang, Z.: Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems (2020) 1, 2

[11] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proc. of the Association for Computational Linguistics. pp. 142–150 (2011) 6