Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210118

Self-Harm Detection for Mental Health Chatbots

Saahil DESHPANDE[a,1] and Jim WARREN[a]
[a] *University of Auckland, Auckland, New Zealand*

Abstract. Chatbots potentially address deficits in availability of the traditional health workforce and could help to stem concerning rates of youth mental health issues including high suicide rates. While chatbots have shown some positive results in helping people cope with mental health issues, there are yet deep concerns regarding such chatbots in terms of their ability to identify emergency situations and act accordingly. Risk of suicide/self-harm is one such concern which we have addressed in this project. A chatbot decides its response based on the text input from the user and must correctly recognize the significance of a given input. We have designed a self-harm classifier which could use the user's response to the chatbot and predict whether the response indicates intent for self-harm. With the difficulty to access confidential counselling data, we looked for alternate data sources and found Twitter and Reddit to provide data similar to what we would expect to get from a chatbot user. We trained a sentiment analysis classifier on Twitter data and a self-harm classifier on the Reddit data. We combined the results of the two models to improve the model performance. We got the best results from a LSTM-RNN classifier using BERT encoding. The best model accuracy achieved was 92.13%. We tested the model on new data from Reddit and got an impressive result with an accuracy of 97%. Such a model is promising for future embedding in mental health chatbots to improve their safety through accurate detection of self-harm talk by users.

Keywords. Chatbot, Self-harm, Mental health, sentiment analysis, LSTM, BERT

1. Introduction

The increasing rate of suicide in youth is an alarming problem that New Zealand is facing. Mental health experts have called for immediate action to bring the situation under control. The statistics since 2010 show that the rate of suicide in youth has been larger than the overall rate of suicide in the country [1]. This indicates that the people in the age range of 15 to 24 years have a larger probability to commit suicide or cause harm to themselves. In 2019 New Zealand had the highest mortality rate for youth between the age of 10 to 24, around 35 deaths per 100,000 [2]. These alarming rates of suicide motivate the need to search for solutions to this problem, and to the problem of providing further mental health support generally.

One such solution that people have come up with is that of a chatbot to supplement the traditional mental health workforce Use of such technology has been shown to improve mood [3-4]. There has been public criticism, however, in that widely

[1] Corresponding Author, Saahil Deshpande, University of Auckland, Auckland CBD, Auckland 1010, New Zealand; E-mail: sdes343@aucklanduni.ac.nz.

deployed mental health chatbots have been demonstrated to fail to identify obvious signs of sexual abuse, eating disorders, drug use and risk of self-harm [5]. These shortcomings raise concerns regarding the recommendations of the chatbots for the youth.

We intend to address the issue of identifying risk of self-harm in term of text inputs by users which imply the users have suicidal thoughts or a tendency to harm themselves. This is an issue that requires immediate help in which case a chatbot should not go through its regular routine but should escalate to the part of the program where the chatbot starts providing resources to the youth that help them deal with self-harm thoughts or decides if human intervention is required. The research question that we address in this paper is if we can build a self-harm classifier which can accurately identify the texts indicating self-harm. If we can train such a classifier, we wish to implement it on existing chatbots.

2. Methodology

Our plan was to build a self-harm classifier in three steps. The first step was to train a sentiment analysis classifier to model the emotion of the text. Next, we wanted to train the self-harm classifier and finally we wanted to integrate the results from the sentiment analysis classifier into the self-harm classifier. The purpose of doing so was to account for the emotion of the text while classifying the texts as self-harm, as the emotion plays an important role.

The ideal data for training the models is confidential patient data and is not easily available. We used alternate sources of data which would be the closest representation of the texts a chatbot may receive from the user. The sentiment analysis was done on a labelled twitter dataset, sentiment140 [6]. The dataset is balanced with equal number of positive and negative sentiments. For the self-harm classifier, the data was scraped from Reddit [7]. Reddit segregates data in various subreddits. We used the suicidewatch subreddit which provided us with data close to that we would expect from the chatbot to indicate risk of self-harm. We used videogames, soccer and games subreddits to get the data which did not indicate risk of self-harm. Finally, the doemsticviolence subreddit was scraped for data with both positive as well as negative instances of risk of self-harm. The dataset was manually labelled and the number of positive and negative instances of risk of self-harm were kept equal to balance the dataset.

The models were trained in Python version 3.7.7. Computational models used in our analysis included Random Forest [8], Long short-term memory (LSTM) networks [9] and Bidirectional Encoder Representations from Transformers (BERT) in LSTM networks [10]. Random Forests are tree-based ensemble machine learning methods which can be used both for regression and classification. We used the Random Forest classifier through the scikit-learn library in Python. LSTM networks are a special type of Recurrent Neural Networks, which were designed to avoid long-term dependencies. We build the LSTM networks using the Keras library in Python which is an interface for the TensorFlow library. BERT is an encoder and works different from the traditional encoder in Keras. It takes into consideration the context in which a word is used in a sentence. This means that the same words may have different encodings based on the preceding and proceeding words. We used the bert library in Python along

with the LSTM network to train the models. In each of the cases, we performed hyperparameter tuning in order to improve the model performance.

3. Results

Sentiment Analysis on Twitter data

The Twitter dataset consisted of 1.6 billion labelled tweets. Data were distributed in an 80:20 train/test proportion. We used two approaches to model this data. First we implemented a LSTM network. The training data was converted into numeric vectors by an embedding layer and fed to the network which was trained in 20 iterations. Accuracy achieved on the test set was 75.58%. The precision of the model was 0.77, the recall was 0.76 and the f1-score was 0.77. The model gave an equivalent number of true negatives and true positives implying that it did not tend to overclassify in one direction, but the accuracy left ample room for improvement. The second approach was using a tree-based algorithm. We applied the Random Forest algorithm and used 1000 trees to train the model on the training set. This approached improved accuracy from the neural network model, taking it up to 80.14%. The precision of the model was 0.75, recall was 0.90 and f1-score was 0.82. The model showed improvement in accuracy but it had a higher tendency to classify the outcome as a negative sentiment.

Self-harm classifier for the Reddit data

The challenge with the self-harm classifier was to be able to correctly identify the text which was similar to the ones indicating risk of self-harm but did not do so. These texts were scraped from the domesticviolence subreddit. They were associated with a strong negative sentiment but did not generally indicate self-harm. We created two datasets, dataset 1 and dataset 2. Dataset 1 combined posts from suicidewatch, games, videogames and soccer to create the training set and all the posts from domesticviolence made up the test set. Dataset 2 combined post from all five subreddits and then used an 80:20 split to create the training and test set respectively. The purpose of doing so was to check if the model trained on dataset 1 can classify the posts from domesticviolence correctly. If not, we would have to train the model on the domesticviolence posts to make it a better classifier. We used a similar modelling approach for the data as we did for sentiment analysis. Using LSTM and the training set of dataset 1, training for 20 iterations resulted in an accuracy of 85.59% on the test set. However, the test set was heavily unbalanced with only 2 instances with indication of risk of self-harm, so accuracy was not a good metric to judge the model. The precision for the model was 1, the recall was 0.67 and the f1-score was 0.8.

The Random Forest algorithm was applied using 2000 trees to train the data on the training set of dataset 1. The model had an accuracy of 80% on the test set. The precision for the model was 1, recall was 0.66 and f1-score was 0.8. In both cases a large number of texts were incorrectly classified as indicating self-harm. Neither of the models performed well enough indicating that we needed to train the models on the domestic violence posts. We retrained the models on dataset 2 again using both algorithms. The LSTM model had an accuracy of 86.43% on the test set. The precision was 0.86, the recall was 0.86 and the f1-score was 0.86. The model did not show tendency to overclassify one class. The Random Forest model had an accuracy of 80.5%. The precision was 0.77, the recall was 0.77 and the f1-score was 0.76. The models did improve but were still not good enough to be implemented in a chatbot.

Combining results of Sentiment analysis with the self-harm classifier

In a bid to improve the accuracy of the trained models we used the results of the sentiment analysis classifier to train the self-harm classifier. The sentiment analysis classifier was trained on the Twitter data using the Random Forest algorithm. This classifier then predicted the sentiment for the Reddit data. The Reddit data along with its corresponding sentiment was given as the input to a LSTM network to train the self-harm classifier. After 20 iterations the model had an accuracy of 92.77% on the training data and 86.69% on the test data.

Self-harm classifier using BERT encoder

The previous algorithms left considerable room for improvement and we decided to use newer and more complex methods encoding techniques. BERT has provided some outstanding results in the recent years on several Natural language Processing projects [11]. We decided to use the encoder on the dataset 2. The model trained on the training set for 20 iterations and achieved an accuracy of 96.45% on the training set and 92.13% on the test set. We tested the model on more recent data scraped from Reddit (12442 posts scraped from 7th July, 2020 to 14th October, 2020) and it achieved an accuracy of 97% on this data. This result was close to what we had hoped for to have a practical application of the model in a chatbot.

4. Discussion

We were able to identify relevant data sources which were used to train the sentiment analysis classifier and the self-harm classifier. The Twitter dataset was used to train the sentiment analysis classifier which achieved a maximum test accuracy of 80.14% using the Random Forest algorithm. The accuracy was not satisfactorily high and left considerable unexplained variance in the data. The Reddit data was used to create the two datasets, dataset 1 and dataset 2. Dataset 1 exposed the issues a model would have in correctly identifying posts similar to indicating risk of self-harm. This issue emphasized the need of using the domesticviolence posts for the purpose of training the models. Training the models on dataset 2 achieved a maximum accuracy of 86.43%. This was an improvement from previous results, but there was still variance which needed to be explained. Combining the results of sentiment analysis with the self-harm classifier resulted in a classifier accuracy of 86.89% on the test set. There was a small improvement in the accuracy, but it was still not what we hoped for. Finally, we used a BERT encoder to improve the model performance and this achieved an accuracy of 92.13%. We tested the model on newer data to check for its consistency and it resulted in an accuracy of 97%. The exact reason as to why the model performs so much better on the newer data is unknown, but the results are promising as the accuracy did not decrease.

The results from our experiment were largely as expected. One exception was the sentiment analysis classifier not working as well as we expected. This may have been due to the length of each post being only 140 characters. Also, the self-harm classifier did not work well with dataset 1, but the performance improved when dataset 2 was used. Further, we believed that combining the results of the sentiment analysis with the self-harm classifier would improve the results but that was not the case. This might have happened as the sentiment analysis classifier itself was not good enough or the self-harm classifier was already taking into account the sentiment of posts. The BERT

encoder had a better performance than all other models, as it has on several previous occasions [12].

Our results provide a promising solution to improve the safety of existing mental health chatbots. The classifier can easily be integrated in chatbots and we intend to start with the HeadStrong chatbot [13] developed at the University of Auckland. The classifier can work alongside a chatbot's usual intent recognizer as a resource to indicate self-harm text. The chatbot can then respond appropriatelya to this signal, such as entering a clarification dialog and directing the user to a human-staffed help line.

Although we achieved satisfactory results to answer our research question, there are still some limitations to the paper. The data sources we used were assumed to be representative of the data users would input to a chatbot and the performance of the model on the actual data is unknown. Platforms such as Google Dialogflow [14] and Raza [15] can be used to implement a self-harm classifier and we have not evaluated the performance of our model against these platforms. Finally, the sentiment analysis classifier has room for improvement and a stronger classifier may improve the performance of the self-harm classifier.

To conclude, given the difficulty to access conversations between therapists and patients, we found suitable data sources, and trained a model which may be usable in mental health chatbots to detect risk of self-harm.

References

[1] New Zealand Ministry of Health. (2012). Suicide Facts: Deaths and Intentional Self-Harm Hospitalisations 2010.
[2] Henry, D. (2019). New Zealand suicides in 2018-19 highest since records began, Available at: https://www.nzherald.co.nz/nz/new-zealand-suicides-in-2018-19-highest-since-records-began/KIZY3N5C5ICQ2LL6BRWBENRXEE/, Accessed Nov 11, 2020.
[3] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456-464.B.
[4] Inkster, S. Sarda, V. Subramanian. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth.2018;6(11):e12106.
[5] White, G. Child advice chatbots fail to spot sexual abuse. 2018, Available at https://www.bbc.com/news/technology-46507900?, Accessed Nov 20, 2020.
[6] Kazanova, MM. Sentiment140 dataset with 1.6 million tweets. 2017, Available at: https://www.kaggle.com/kazanova/sentiment140, Accessed May 27, 25020.
[7] Ji S, Yu CP, Fung, S, Pan S, Long G.Supervised Learning for Suicidal Ideation Detection in Online User Content. Complexity (New York, N.Y.), 2018. p.1-10.
[8] A complete guide to the Random Forest algorithm, Availabe at https://builtin.com/data-science/random-forest-algorithm., Accessed Nov 22, 2020.
[9] Sak H, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling.2014, Availabe at: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/43905.pdf, Accessed Nov. 20, 2020.
[10] Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805. 2018
[11] González-Carvajal S, Garrido-Merchán EC. Comparing BERT against traditional machine learning text classification.2018. Available at https://arxiv.org/abs/2005.13012.
[12] Sun C, Qiu X, Xu Y, Huang X. How to Fine-Tune BERT for Text Classification? 2019, Avalable at https://arxiv.org/abs/1905.05583.
[13] Holt-Quick C, Warren J, Stasiak K, Williams R, Christie G, Hetrick S, Hopkins S, Cargo T, Merry S. A chatbot architecture for promoting youth resilience.2020. arXiv Preprint arXiv 2005.07355
[14] Google Dialogflow. Available at https://cloud.google.com/dialogflow/cx/docs, Accessed Nov 22, 2020.
[15] Rasa, Availble at https://rasa.com/, Accessed Nov 22, 2020.