Detecting Intentional Self-Harm Reported in Clinical Text Using a Deep Learning Model

Jihad S. Obeid, MD¹, Brian E. Bunnell, PhD², Jennifer Dahne, PhD¹, Tami Crawford, PhD¹, Lewis Frey, PhD¹

¹Medical University of South Carolina, Charleston, SC; ²University of South Florida, Tampa, FL

Abstract

Suicide is an important public health concern in the United States. There is active research in establishing predictive models for suicide to help direct patients to proper care. Here we examine a novel approach for using a deep learning classifier to identify patients with suicidal behavior and prediction of future suicidal attempts.

Introduction

Suicide ranks among the leading causes of death in the United States (US)¹. More than 100 individuals in the U.S. die by suicide each day, on average. Suicidal ideation (i.e., having thoughts about harming oneself or taking one’s own life) and suicidal/self-harm behavior (i.e., harming oneself or attempting to take one’s own life) are relatively common risk factors for suicide. Better understanding predictors of suicidal behavior can help the development of approaches to preventing patient suicides. There has been significant work in the area of utilizing machine learning approaches to help address the limitations of traditional statistical models in the prediction of suicidal behavior and attempts². Several researchers have explored the use of electronic health record (EHR)-based models³, Others have focused on using natural language processing and linguistics-driven prediction models⁴,⁵. In this presentation, we explore the utility of a deep learning approach for a) automated detection of concurrent self-harm International Classification of Diseases (ICD) code assignment based on clinical text; and b) the prediction of future suicide attempt and intentional self-harm based on ICD labeled encounters within the EHR.

Methods

This project was approved by the IRB at the Medical University of South Carolina (MUSC) under protocol Pro00087416. We extracted clinical text notes from the MUSC EHR for 835 patients with ICD codes for suicide attempt and intentional self-harm (ISH) as defined in a National Health Statistics Report from the Centers for Disease Control and Prevention in the US⁶, and 1670 control cases who never had any ISH ICD codes within our EHR. We first trained a number of algorithms on clinical notes using the ISH codes as a positive label. Clinical notes associate with the first visit with ISH codes (on average 5 notes per patient) were concatenated chronologically and truncated at 8000 words. The following models were tested: Naïve Bayes classifier (NBC); single decision tree (SDT) with a maximum depth of 20; Random Forrest (RF) with 201 trees and the number of variables randomly sampled as candidates at each split (mtry)=150; Support Vector Machines (SVM) Type 1 with a radial basis kernel; a simple Multilayer Perceptron (MLP) with 64-node input and hidden layers; and finally a deep learning model with a randomly initialized word embedding layer (200 dimensions per word), a convolutional neural network (CNN) layer with multi-filter sizes (3, 4 and 5), 200 nodes each, with global maxpooling, followed by a merge tensor, a fully connected 200 node layer, then a single sigmoid activation output node. A dropout rate of 0.2 was used after both the embedding layer and the last dense layer. Text pre-processing included lower casing, and punctuation removal. In addition, stop-word removal and stemming was used for bag-of-words models, i.e. all the models except the CNN. The vectors for the CNN were generated using tokenization, followed by sequence padding to ensure that all sequences have the same length. Thirty percent of the data were held out as a test set and 70% was used for training and cross validation. We used R version 3.6.1 for constructing the machine learning pipelines, and Keras and TensorFlow for the deep learning models.

After establishing the best performing model, we trained the model on clinical text prior to the dates of the first ISH ICD code assignments (the index visit). We selected patients and controls who had visits with dates ranging from 180 to 30 days prior to the index visit. In order to test future predictive power, we used patients with index visits prior to 2018 for training, and patients with index visits in 2018 or later for testing. The training set included 480 patients and 645 controls, and the test set had 106 patients and 106 controls. Again, controls never had ISH codes throughout the span of the EHR.
Results
The results for the training and testing on the notes concurrent with ISH codes are shown in table 1. Since the CNN was the best performer in detecting ISH in clinical notes concurrent with ICD codes, it was used for the second part, namely prediction of future ISH. The results of the CNN for predicting future ISH codes on the test set were as follows: AUC of 0.868 (with a 95% CI of 0.82-0.915), precision: 0.869, recall: 0.689 and F1: 0.768.

Table 1. The results for the training and testing on the notes concurrent with ISH ICD codes.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (95% CI)</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.929 (0.909-0.949)</td>
<td>0.818</td>
<td>0.896</td>
<td>0.855</td>
</tr>
<tr>
<td>DT</td>
<td>0.927 (0.903-0.951)</td>
<td>0.932</td>
<td>0.884</td>
<td>0.908</td>
</tr>
<tr>
<td>RF</td>
<td>0.973 (0.96-0.986)</td>
<td>0.876</td>
<td>0.904</td>
<td>0.890</td>
</tr>
<tr>
<td>SVM</td>
<td>0.959 (0.944-0.975)</td>
<td>0.912</td>
<td>0.788</td>
<td>0.845</td>
</tr>
<tr>
<td>MLP</td>
<td>0.966 (0.953-0.979)</td>
<td>0.884</td>
<td>0.888</td>
<td>0.886</td>
</tr>
<tr>
<td>CNN</td>
<td>0.988 (0.981-0.996)</td>
<td>0.950</td>
<td>0.916</td>
<td>0.933</td>
</tr>
</tbody>
</table>

Discussion
Most of the models showed a relatively good performance when detecting initial ISH ICD code assignments within clinical notes associated with concurrent hospital visits. This is likely due to a strong signal within concurrent notes, and is likely associated with a high fidelity of ICD code attribution specifically for ISH. This has yet to be established by manual chart reviews (as a gold standard). The deep learning model (CNN) was the best performer with an area under the receiver operating characteristics curve (AUC) approaching 99%. When applied to predicting future ISH code assignment for the first time in a patient chart based on clinical notes preceding those visits by anywhere from 6 months to 30 days prior, the AUC for the CNN was 87%, with modest recall and precision. Expectedly, this is significantly lower than prediction from concurrent clinical notes. There are several limitations to our work so far. We intend to perform chart reviews on a number of patients in our data set to use as a gold standard for both training and testing of the models. Our model currently only addresses features within clinical text. Other clinical information could be added to the model e.g. associated demographics, comorbidities and risk factors e.g. codes for depression or substance use. With respect to suicide prediction, EHR data alone may not provide the full picture. Ideally, our data should be linked with statewide cause of death data, which should yield an improved predictive power.

Nevertheless, our results are competitive with results from other models reported in the literature, and indicate that some patients with potential suicidal behavior or ideation are being missed, or not flagged with relevant ICD codes. Improving the precision of these algorithms could lead to better follow up and care by mental health professionals for patients who are at risk. Future work will include manual chart reviews to provide verification of the accuracies for both ISH ICD codes and deep learning algorithms.

References