Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1303–1310
Marseille, 11–16 May 2020

Using Deep Neural Networks with Intra- and Inter-Sentence Context to Classify Suicidal Behaviour

Xingyi Song[1], Johnny Downs[2], Sumithra Velupillai[2], Rachel Holden[2], Maxim Kikoler[2], Kalina Bontcheva[1], Rina Dutta[2], Angus Roberts[2]
[1]Department of Computer Science, The University of Sheffield
Sheffield, UK
{x.song, k.bontcheva}@sheffield.ac.uk

[2]Maudsley Biomedical Research Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London
London, UK
{johnny.downs, sumithra.velupillai, rina.dutta, angus.roberts}@kcl.ac.uk

Abstract

Identifying statements related to suicidal behaviour in psychiatric electronic health records (EHRs) is an important step when modeling that behaviour, and when assessing suicide risk. We apply a deep neural network based classification model with a lightweight context encoder, to classify sentence level suicidal behaviour in EHRs. We show that incorporating information from sentences to left and right of the target sentence significantly improves classification accuracy. Our approach achieved the best performance when classifying suicidal behaviour in Autism Spectrum Disorder patient records. The results could have implications for suicidality research and clinical surveillance.

Keywords: Suicide Detection, Neural Network, Contextual Classification, Document Encoding

1. Introduction

Suicide is the second leading cause of death among 15 to 29 year-olds, causing about 800,000 deaths per year worldwide (WHO, 2018). Mental health problems are a major risk factor for suicide attempts. For example, among the people who attempted (or completed) suicide in the US in 2010, approximately 44% were diagnosed with a mental health problem, and 31% were receiving mental health care (Parks et al., 2014). The ubiquity of Electronic Health Records (EHRs) in many countries, and the potential for the reuse of information contained within them, has led to attempts to model and predict patient suicide risk from the content of those records. In building such risk models, it is necessary to extract information about suicidal behaviour from the EHR. In common with much of the information in the mental health record, this is generally recorded in the free text, unstructured portion of the records.

When applying natural language processing to EHRs, the EHR can be analysed at different levels of granularity (Velupillai et al., 2018). Considering these levels of granularity for detection of suicidal behaviour, previous approaches can be categorised into **patient level** - a set of EHRs that belong to a certain patient who has exhibited or exhibits suicidal behaviour (Barak-Corren et al., 2017; Tran et al., 2015; Walsh et al., 2017; Fernandes et al., 2018; Choi et al., 2018; Cook et al., 2016); **mention level** - word or sub-sentences that mention suicidal thoughts or behaviour (Haerian et al., 2012; Anderson et al., 2015; Downs et al., 2017; Gkotsis et al., 2016b); and combinations of these into **document and patient level** - aggregating mention level annotations to derive document- and patient-level labels (Velupillai et al., 2019).

To the best of our knowledge, none of these previous approaches have addressed this problem on a sentence level. There are several benefits to modeling the problem of classifying suicidal information in EHRs on a sentence level:

- Sentences are relatively shorter than patient/document level, so researchers/clinicians can easily verify or identify the suicidal thought.
- Compared with the mention level, data labelling is relatively cheap, requiring an answer to a yes/uncertain/no question, rather than labelling a particular span of words.
- Sentence level suicidal information can be transferable to either document or patient level.

Considering the above benefits and filling the gap of sentence level suicidal behaviour research, in this paper we treat suicidal behaviour extraction as a sentence classification problem.

In theory, sentence level suicidal behaviour extraction can be achieved by any text classification algorithm. However, openly available annotated datasets are very scarce. Moreover, most previous approaches have employed mention level annotations, but with this approach contextual information can be lost. Whilst the context of a suicidal behaviour mention may be intra-sentential, i.e. within the same sentence, it may also be inter-sentential, spanning adjacent sentences. Consider the following example:

Example 1 1. He says that he wants to jump out of the window. 2. He feels life is not worth continuing.

Example 2 1. He says that he wants to jump out of the window. 2. He wants to escape from his family.

Sentence 1 in both Example 1 and 2 are identical. When combined with context (Sentence 2), Example 1 expresses a clear suicidal thought, whereas Example 2 does not necessarily convey suicidality.

In order to address the importance of context, we propose a sentence level suicidal behaviour classification approach based on a C-LSTM-CNN (Song et al., 2018) algorithm. **C** stands for context. We encode intra-sentence context using a bi-directional LSTM. We encode the surrounding inter-sentence context using a Fixed Size Ordinally Forgetting Encoding (FOFE (Zhang et al., 2015)) based algorithm. Compared with previous LSTM based context encodings (Lee and Dernoncourt, 2016), a FOFE can be generated before training, to massively reduce computational cost in both training and application.

This reduction in computational cost satisfies an important requirement that our method be usable in the typically compute resource constrained environments of hospitals, where additionally, the ethically sensitive nature of the data means that it cannot be processed elsewhere.

Whilst we present a mental health use case for our method, the importance of inter-sentence context and of computationally efficient deep learning are not restricted to this domain, as discussed in (Song et al., 2018). Other contributions of this paper are: (1) Mapping the suicidal behaviour data presented in (Downs et al., 2017) from the mention level to the sentence level. (2) The first deep neural network sentence level suicidal behaviour extraction work for EHRs, to the best of our knowledge. (3) A systematic comparison of several strong baselines for EHR suicidal behaviour detection.

2. Related Work

Early EHR suicide-related information detection works were mostly rule based approaches using external knowledge bases (Haerian et al., 2012), lists of key words (Anderson et al., 2015), or patterns (Gkotsis et al., 2016b; Fernandes et al., 2018). These rule based approaches are relatively cheap to build, as very little training data is required. However, these approaches are less robust than machine learning, and not transferable to different languages, or even EHRs with different language styles.

Shallow machine learning with human selected features are now the main method used in suicidal behaviour detection. Recent approaches include naive Bayes classifiers (Barak-Corren et al., 2017) and multi-layer perceptron (Choi et al., 2018; Bhat and Goldman-Mellor, 2017) using human labelled features; random forest (Walsh et al., 2017) over patient metadata and history features; N-gram based linear regression (Cook et al., 2016); Support Vector Machine(SVM) over hand picked (Bittar et al., 2019) or restricted Boltzmann machine encoded features (Tran et al., 2015). More recently, (Metzger et al., 2017) has compared seven different learning algorithms [1] in suicidal attempt classification, with nine different features.

Surprisingly, we have not been able to find any modern deep learning [2] based research on suicidal behaviour detection, to date. This may be because of the challenges in data access, where access to EHR data is typically restricted due to governance regulations. However, in non-restricted data (e.g. social media data), deep learning approaches have dominated work on detecting suicide-related information in recent years. Approaches have included CNN (Shing et al., 2018; Gaur et al., 2019), LSTM (Ji et al., 2018), attention LSTM (Coppersmith et al., 2018), combined CNN and LSTM (Sawhney et al., 2018). Most recently, (Matero et al., 2019) applied context word embeddings (BERT) to improved performance.

3. C-LSTM-CNN

The C-LSTM-CNN architecture is shown in Figure 1. It takes three inputs: 1. the **focus sentence** - the sentence we are aiming to classify; 2. **left context** - all document text to the left of the focus sentence; 3. **right context** - all document text to the right of the focus sentence. The architecture contains five major parts, represented in five different colors in Figure 1:

1. Word embedding – purple
2. Bi-directional LSTM – pink
3. Multiple window CNN – yellow
4. Context encoder – green
5. Softmax classifier – blue

For details of the C-LSTM-CNN architecture please refer our previous paper (Song et al., 2018). We briefly describe its motivation here.

The input words of the focus sentence are first transformed into vector space representations in order to capture the semantic representation of the words. In this paper, we use the Word2Vec(w2v) embedding model (Mikolov et al., 2013). The bi-directional LSTM layer (Hochreiter and Schmidhuber, 1997) is used to enrich the word vector representation with sentence level sequential information. This is followed by CNN with max-pooling layers (Kim, 2014), which extract local features at specific points from the LSTM outputs.

In addition to processing the focus sentence with the LSTM and CNN layers, the input words of the left and right context are encoded by an adapted FOFE encoder. All sentences prior to the focus are considered part of the left context, and all sentences following the focus to be part of the right context. Each sentence i in the context can be represented as $S_i = \{x_0, x_1, ...x_U\}$, a sequence of U vector representations x_u, one for each word in the sentence. The context representation z for S can then be calculated recursively as:

$$\begin{cases} z_u = x_u & , u = 0 \\ z_u = \alpha \cdot z_{u-1} + x_u & , u > 0 \end{cases} \tag{1}$$

In this paper, the FOFE encoder is applied in two hierarchical steps. First we encode each sentence S_i in the left/right

[1] The algorithms are: random forest, naive Bayes, support vector machines, predictive association rules, decision trees, multi-layer perceptron, and logistic regression

[2] Neural Network approach without human selected features

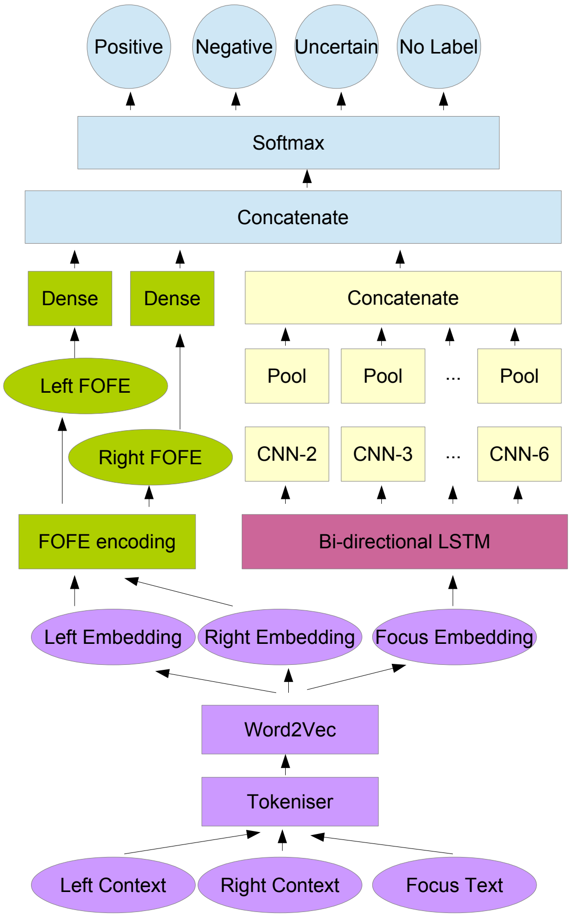

Figure 1: Structure of the C-LSTM-CNN model

context into a sentence level FOFE embedding z_i^{sent}, with a forgetting factor α_{sent} that is slowly decreased as we move further from the focus sentence.

$$z_{sent\_i} = FOFE_{sent}(S_i, \alpha_{sent}) = \sum_{u=0}^{U} \alpha_{sent}^{(U-u)} x_u \qquad (2)$$

After first FOFE encoder, we group all left side encoded sentences as $Z_{left\_sent} = \{z_{l\_0}, z_{l\_1}, ...z_{l\_I}\}$, and all right side encoded sentences as $Z_{right\_sent} = \{z_{r\_0}, z_{r\_1}, ...z_{r\_I}\}$.

The grouped left and right context FOFE embeddings are then themselves encoded into one context embedding for left context, and one for right context, using a rapidly decreasing α_{cont}.

$$z_{left}^{cont} = FOFE_{cont}(Z_{left\_sent}, \alpha_{cont}) = \sum_{i=0}^{I} \alpha_{cont}^{(I-i)} z_{l\_i} \qquad (3)$$

$$z_{right}^{cont} = FOFE_{cont}(Z_{right\_sent}, \alpha_{cont}) = \sum_{i=0}^{I} \alpha_{cont}^{(I-i)} z_{r\_i} \qquad (4)$$

Finally, a softmax layer takes the LSTM-CNN output and the FOFE context outputs, and combines them in to a multi-class classifier.

4. Materials and Methods

4.1. Data and Labeling

We use a revised version of the corpus described in (Downs et al., 2017)[3], which contains free text documents from the EHRs of adolescent patients who have an Autism Spectrum Disorder (ASD) and have been referred to the South London and Maudsley NHS Foundation Trust (SLaM). These records were extracted from the Clinical Record Interactive Search (CRIS) system (Perera et al., 2016), a resource with de-identified EHRs which allows data retrieval for secondary data analysis, approved by the Oxfordshire Research Ethics Committee C (reference 08/H0606/71+5).

	No Label	SR-Pos	SR-Neg	Uncertain
Sentence	327,051	6514	2993	1090
Document	8	2911	1557	442
Patient	0	331	100	68

Table 1: Counts of labels in sentences, documents, and patients. Label names are described in the text.

The corpus is described in Table 1, and below. The corpus consists of documents from 499 ASD patients containing at least one pre-defined term related to suicidal behaviour [4]. Suicidal behaviour mentions in each document were independently annotated by one of two domain experts. Mentions were loosely defined: the annotators were asked to label any explicit mention of suicidality in the text, marking each mention as *suicidality risk positive* (SR-Pos, i.e. a patient with suicidality risk), *suicidality risk negative* (SR-Neg, i.e. not a patient with suicidality risk) or *uncertain*. In total, 4,918 documents were annotated, containing 6697 *SR-Pos*, 3,701 *SR-Neg* and 1,097 *uncertain* mentions.

From these mention-level labels, labels were generated at the sentence, document and patient level. For sentence labels, each document was split into sentences using the GATE NLP toolkit (Cunningham et al., 2011), and sentences were labeled using the following rules:

1. If a sentence contains only one suicidality-related mention (Case 1 illustrated in Figure 2), then the sentence is given the same label as the mention.

[3] The current corpus contains one patient less and seven documents more than described in (Downs et al., 2017).

[4] The suicide-related terms were: 'suicide','kill herself', 'kill himself', 'kill themselves', 'kill myself', 'take his own life', 'take her own life', 'take their own life', 'end his own life', 'end her own life', 'end their own life', 'want to die', 'were dead'. These were identified in the documents using the tool described in (Downs et al., 2017)

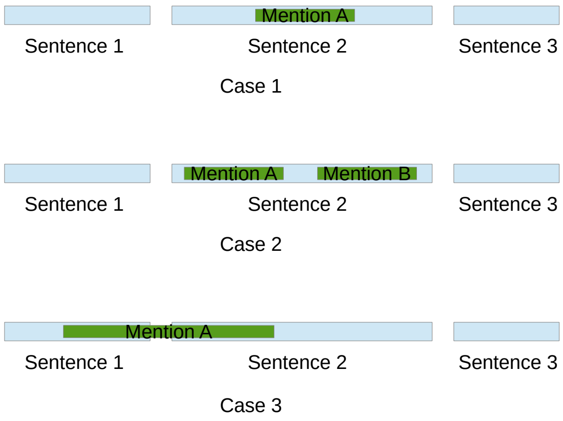

Figure 2: Possible cases when generating sentence labels from mention labels (refer to text and Algorithm 1

2. If a sentence contains more then one mention (Case 2 illustrated in Figure 2), then we label the sentence using Algorithm 1.

3. If a mention is split across two or more sentences (Case 3 illustrated in Figure 2), we combine those sentences and treat them as a single sentence in subsequent experiments. We give this combined sentence the same label as the mention.

4. If the sentence contains no mention, then we label the sentence as 'No label'.

Algorithm 1 Multi-mention sentence labeling algorithm

```
if mentions in the sentence have different labels then
    if any mention contains SR-Pos label then
        sentence labeled as SR-Pos
    else if any mention contains Uncertain label then
        sentence labeled as Uncertain
    else
        sentence labeled as SR-Neg
    end if
else
    sentence assigned same label as any mention
end if
```

Label counts are shown in Table 1. Only 34 sentences contain more than two different mentions. We manually inspected these sentence individually, to verify the label correctness. [5].

For document and patient level labelling, we adapted the patient level process described in (Downs et al., 2017)[6]. If there are any SR-Pos sentences in a document, then the document is labelled as SR-Pos. If no SR-Pos sentences occurred in the document, but there are sentences labelled Uncertain, then the document is labelled Uncertain. Otherwise, the document is labelled SR-Neg. An identical process was followed to percolate document labels to the patient level. Final counts are shown in Table 1.

4.2. Experimental setup

We compared C-LSTM-CNN to three baselines: (1) a rule based tool designed to detect affirmation and negation of suicide related information (Gkotsis et al., 2016a), which we refer to as NegTool; (2) two statistical machine learning approaches - Support Vector Machine (SVM) and Maximum Entropy (ME) both with a bag-of-words feature representation; (3) State-of-the-art deep neural network models - CNN, LSTM and LSTM-CNN, all with tokens represented as Word2Vec embedidngs. The CNN, LSTM and LSTM-CNN architectures are subsets of full C-LSTM-CNN, and use the same hyper-parameters.

Two sets of experiments were performed: **four class experiments** - which aimed to classify each sentence into four different classes, as detailed in Section 4.1.; and **two class experiments** - which only classify SR-Pos and SR-Neg classes, and ignore the other two classes. This two class experiment is included because the NegTool baseline only contains rules for these two classes.

Five-fold cross validation was used for evaluation, rather than ten-fold. This was because: (1) experiments were conducted in a computationally restricted environment (2); there are only 68 patients with the Uncertain label in the data set, and so ten folds gives a high chance of some test splits having no patients with this label. The rule based NegTool does not require training, but was tested with the same folds as all other systems.

For the shallow machine learning models, features were the top 2000 bag of words based on word count. SVM is computationally complex when trained with large numbers of instances and dimensions, which can cause difficulties in a resource constrained environment, such as a hospital. Therefore, in four class SVM and MaxEnt training we randomly selected up to 2500 instance from each class the training sample. With SVM and MaxEnt, we applied a One-vs-Rest strategy to model the four class experiments as binary classifiers.

For neural network models, we trained each fold with 50 epochs and minibatch size of 64 using the Adamax (Kingma and Ba, 2015) optimization algorithm. To deal with label imbalance in the data, class weights w_i for class i were set proportional to $\max(f_i)/f_i$ where f_i is the frequency of class i.

We used the Word2Vec Skip gram embeddings (Mikolov et al., 2013). The word embedding dimensions were set to 50, which is sufficient for NLP tasks according to (Lai et al., 2016). For the LSTM, 64 hidden units were used. For the CNN, layers for kernel sizes 2 to 6 were included in the network, and 64 features were used for each. For the context, 64 fully connected perceptrons were used in the projection dense layer. The forgetting factor was $\alpha_{cont} = 0.9$ and $\alpha_{sent} = 1.0$

[5] (Downs et al., 2017)'s experiment did not consider the uncertain label.

[6] (Downs et al., 2017) directly transform the mention level (rather than the sentence level) to the document level, and hence use a slightly different approach.

5. Results

The results of the four class experiments are shown in Table 2 and Table 3. Table 2 shows the mean accuracy of sentence level, document level and patient level classifiers over the cross validation folds. In general, deep neural models have better performance than shallow models. All deep models are able to achieve accuracy over 97 %. LSTM has better performance over CNN on the sentence and document level, but worse than CNN on the patient level. When LSTM and CNN are combined, the performance improves over either on their own, by about 4%. By adding context, C-LSTM-CNN consistently improves accuracy over all three classification levels.

Model	Sentence	Document	Patient
SVM	85.89 (0.971)	61.53 (1.105)	63.70 (1.587)
MaxEnt	77.60 (2.284)	59.68 (1.058)	63.41 (1.180)
CNN only	97.98 (0.257)	80.81 (1.372)	80.81 (3.215)
LSTM only	98.12 (0.272)	81.92 (3.024)	80.17 (3.273)
LSTM-CNN	98.61 (0.121)	85.28 (1.269)	84.47 (2.179)
C-LSTM-CNN	**98.71 (0.139)**	**85.64 (1.424)**	**85.31 (2.877)**

Table 2: Four class mean accuracy. Highest values are marked as bold, standard deviations in parentheses.

Table 3 shows the average F-measure for each class. We ignore the No Label F-measure on document and patient level, because there are no patient without labels, and because there are only 8 'No Label' documents.

The overall performance is consistent with the accuracy results shown in Table 2. C-LSTM-CNN remains the best model in all classes and all levels of labelling. In the sentence level, C-LSTM-CNN improves over CNN alone by 8.36 in the SR-Pos class; 10.42, in the SR-Neg class; and 5.91 in the 'Uncertain' class. When compared to LSTM-CNN, the C-LSTM-CNN shows an improved F-measure of more than 1 across all classes.

We note that for the Uncertain label, SVM obtains an F-measure of 0 at the sentence level, but 2.30 at the Document level. This is because false positive Uncertain sentences can belong to true positive Uncertain document, by chance. This justifies the measurement of the sentence level performance alongside the more clinically meaningful document and patient level performance.

The results also show that the shallow models tend to be heavily biased towards the majority 'No Label' class, resulting in low F-measure in the other three classes for both SVM and MaxEnt in the four class task. This may be because: (1) Both shallow models use a binary classification approach, which is less suited to a multi-class task. (2) the hidden layers in deep neural networks can be treated as a feature generation step, transforming the surface level features (e.g. words) to deep features (e.g. output of LSTM). However, shallow models do not contain these steps, and thus require manual feature selection. (3) Training of SVM requires solution of a quadratic programming problem, which does not scale well to large data sets.

For both SVM and MaxEnt, F-measure increases for the SR-Pos label, but drops for SR-Neg, when moving from sentence to document level classification. Both models have around 70% recall for the SR-Pos and SR-Neg classes, but low precision. When creating document level labels from sentence level labels, the existence of any SR-Pos sentence in the document will cause the document to be classified as an SR-Pos document. Therefore, high SR-Pos recall will increase the chance that a correct SR-Pos label is applied to a document (the average SVM document level SR-Pos recall is 95.70). Whilst document level metrics may be important from a clinical use-case point of view, the more granular sentence level also needs to be considered to understand the performance of the underlying algorithms.

Table 4 and Table 5 show the results of the two class experiments. The training and testing data are filtered to leave only the SR-Pos and SR-Neg classes. At the sentence level, machine learned models have better performance than the rule based NegTool. Shallow models perform better than in the 4 class experiment, with both models able to achieve an accuracy above 83%.

C-LSTM-CNN has the best performance at both the document and patient level, but slightly worse (0.13 for SR-Pos and 0.03 for SR-Neg) than LSTM-CNN at the sentence level. This may because: the two class task is much easier than the four class task, and it is difficult to improve over LSTM-CNN which is already achieving very high accuracy.

6. Discussion and Future Work

We introduce a sentence level suicidality detection method using C-LSTM-CNN that combines the strength of LSTM and CNN with a light weight context encoder. We apply this method on a dataset of EHR notes annotated with mention-level suicide-related information (positive, negative and uncertain), and aggregate these to form sentence level labels.

C-LSTM-CNN shows consistently better results than all baselines: rule based, shallow learning and non-contextual deep learning algorithms. We also include classification of the challenging 'Uncertain' label, where our proposed approach also performs well. This approach could be applied to other similar EHR classification tasks.

The model architecture could be further simplified as Context-CNN for situations where computational power is further limited.

The work described has several limitations that will be considered in the future. The SVM and Maximum entropy algorithms were only used with bag-of-words features, and the experiment could be further extended to use word embedding features for these algorithms. The corpus used was restricted to young people with ASD: future work needs to consider generalisability, and construction of a broader training corpus. Additionally, the original annotations were designed for a mention-level task. Our approach to aggregate these to a sentence level might not reflect the distribution of labels that would have existed if the annotation task was designed differently. Annotating suicide-related information is a challenging task in general, and defining relevant labels is not trivial. We plan to look into alternative ways of designing the annotation task to better incorporate contextual information, such as by only focusing on

Sentence	No Label (327,051)	SR-Pos (6,514)	SR-Neg (2993)	Uncertain (1090)
SVM	92.54 (0.560)	17.04 (1.629)	50.07 (1.448)	0.00 (0)
MaxEnt	87.39 (1.495)	22.69 (1.824)	20.26 (3.068)	3.94 (2.369)
CNN only	99.14 (0.119)	69.98 (2.277)	72.45 (3.331)	47.90 (1.859)
LSTM only	99.21 (0.148)	71.56 (3.180)	79.01 (2.277)	42.92 (9.509)
LSTM-CNN	99.43 (0.052)	77.05 (0.974)	81.59 (2.304)	53.14 (3.907)
C-LSTM-CNN	**99.48 (0.071)**	**78.34 (1.064)**	**82.87 (1.074)**	**53.81 (2.414)**

Document	No Label (8)	SR-Pos (2911)	SR-Neg (1557)	Uncertain (442)
SVM	-	74.87 (0.594)	21.36 (7.640)	2.30 (4.609)
MaxEnt	-	75.59 (0.713)	9.57 (1.935)	8.92 (3.330)
CNN only	-	87.02 (1.086)	77.46 (2.234)	45.78 (6.058)
LSTM only	-	87.75 (1.915)	79.65 (5.349)	48.56 (10.563)
LSTM-CNN	-	90.25 (0.886)	84.80 (1.964)	55.37 (6.419)
C-LSTM-CNN	-	**90.39 (0.949)**	**85.45 (1.476)**	**56.03 (7.592)**

Patient	No Label (0)	SR-Pos (331)	SR-Neg (100)	Uncertain (68)
SVM	-	78.23 (1.080)	12.55 (6.244)	0.00 (0)
MaxEnt	-	78.18 (0.943)	4.80 (3.879)	4.51 (5.564)
CNN only	-	88.01 (2.049)	71.68 (6.485)	44.16 (1.490)
LSTM only	-	87.53 (2.237)	70.00 (6.441)	48.91 (1.142)
LSTM-CNN	-	90.89 (1.338)	80.33 (2.635)	52.98 (1.013)
C-LSTM-CNN	-	**90.97 (1.844)**	**82.11 (3.036)**	**56.46 (1.140)**

Table 3: Four class mean F-measure. Highest values are marked as bold, standard deviations in parentheses.

Model	Sentence	Document	Patient
NegTool	68.68 (1.558)	86.80 (0.542)	87.28 (1.851)
SVM	83.18 (1.118)	84.41 (1.177)	85.58 (2.392)
MaxEnt	87.59 (1.252)	89.94 (0.672)	91.17 (1.591)
CNN only	92.96 (0.506)	94.13 (0.605)	93.92 (1.159)
LSTM only	93.55 (0.340)	94.70 (0.619)	94.63 (1.152)
LSTM-CNN	**94.55 (0.813)**	95.49 (0.643)	95.75 (1.392)
C-LSTM-CNN	94.43 (0.655)	**95.68 (0.687)**	**96.42 (1.314)**

Table 4: Two classes Average test accuracies, best values are marked as bold, standard deviations in parentheses

a sentence or even document-level label without employing heuristics deriving from mention-level annotations.

7. Conclusion

In conclusion, C-LSTM-CNN combines the strength of LSTM and CNN with a light weight context encoder that can consider both intra-sentence and inter-sentence context. C-LSTM-CNN can be used to develop a multi-class sentence level suicidal behaviour detection method that scales well with large amounts of input data. This method not only improves suicidal behaviour detection accuracy over rule based and traditional shallow based machine learning algorithms, it also consistently improves accuracy over state-of-art deep neural network models. In addition to demonstrating the value of incorporating wide context in a sentence classification use case, a high-quality suicide behaviour classifier could provide important information both for the epidemiological study of suicidality, and for day to day use in clinical surveillance tools.

Sentence	SR-Pos	SR-Neg
NegTool	73.08 (1.208)	62.54 (2.257)
SVM	88.76 (0.873)	66.55 (1.610)
MaxEnt	90.86 (0.894)	80.62 (2.256)
CNN only	94.95 (0.379)	88.32 (1.125)
LSTM only	95.32 (0.302)	89.56 (0.540)
LSTM-CNN	**96.07 (0.605)**	**91.10 (1.323)**
C-LSTM-CNN	95.94 (0.470)	91.07 (1.197)

Document	SR-Pos	SR-Neg
NegTool	89.50 (0.498)	82.22 (0.684)
SVM	88.96 (0.846)	74.24 (2.047)
MaxEnt	91.64 (0.483)	86.01 (1.107)
CNN only	95.47 (0.411)	91.67 (1.075)
LSTM only	95.88 (0.465)	92.56 (0.929)
LSTM-CNN	96.49 (0.477)	93.67 (0.977)
C-LSTM-CNN	**96.62 (0.526)**	**94.02 (0.996)**

Patient	SR-Pos	SR-Neg
NegTool	90.70 (1.496)	79.77 (2.545)
SVM	90.22 (1.551)	72.47 (5.179)
MaxEnt	93.64 (1.205)	85.51 (2.327)
CNN only	95.58 (0.866)	90.16 (1.832)
LSTM only	96.08 (1.078)	91.43 (2.476)
LSTM-CNN	96.89 (1.010)	93.23 (2.265)
C-LSTM-CNN	**97.37 (0.963)**	**94.36 (2.072)**

Table 5: Two class Average test F-measure, best values are marked as bold, standard deviations in parentheses

8. Availability

The corpus and annotations are part of the CRIS case register at the South London and Maudsley NHS Foundation Trust. CRIS access is controlled by a strict information governance framework that includes both project approval and researcher approval[7].

The source code can be obtained from https://bitbucket.org/deansong/contextlstmcnn/ under the MIT License

9. Acknowledgements

JD is supported by a National Institute of Health Research (NIHR) Clinician Scientist Award (CS-2018-18-ST2-014) and has received support from a MRC Clinical Research Training Fellowship (MR/L017105/1) and Psychiatry Research Trust Peggy Pollak Research Fellowship in Developmental Psychiatry. RD is supported by a Clinician Scientist Fellowship (project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences. AR is supported by the Maudsley Charity and by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SV and AR are supported by the NIHR Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London. RD and SV declare previous research funding received from Janssen. AR declares current research funding from Takeda.

This work uses data provided by patients and collected by the NHS as part of their care and support.

10. References

Anderson, H. D., Pace, W. D., Brandt, E., Nielsen, R. D., Allen, R. R., Libby, A. M., West, D. R., and Valuck, R. J. (2015). Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med*, 28(1):65–71, February.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., and Reis, B. Y. (2017). Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *American Journal of Psychiatry*, 174(2):154–162.

Bhat, H. S. and Goldman-Mellor, S. J. (2017). Predicting adolescent suicide attempts with neural networks. *arXiv preprint arXiv:1711.10057*.

Bittar, A., Velupillai, S., Roberts, A., and Dutta, R., (2019). *Text Classification to Inform Suicide Risk Assessment in Electronic Health Records*. 4.

Choi, S. B., Lee, W., Yoon, J.-H., Won, J.-U., and Kim, D. W. (2018). Ten-year prediction of suicide death using cox regression and machine learning in a nationwide retrospective cohort study in south korea. *Journal of affective disorders*, 231:8–14.

Cook, B. L., Progovac, A. M., Chen, P., Mullin, B., Hou, S., and Baca-Garcia, E. (2016). Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid. *Computational and mathematical methods in medicine*, 2016.

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., et al. (2011). Developing language processing components with gate version 7 (a user guide). *University of Sheffield, Department of Computer Science*.

Downs, J. M., Velupillai, S., Gkotsis, G., Holden, R., Kikoler, M., Dean, H., Fernandes, A. C., and Dutta, R. (2017). Detection of suicidality in adolescents with autism spectrum disorders: Developing a natural language processing approach for use in electronic health records. *AMIA Annual Symposium Proceedings*.

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., and Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):7426.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R., and Pathak, J. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, WWW '19, pages 514–525, New York, NY, USA. ACM.

Gkotsis, G., Velupillai, S., Oellrich, A., Dea, H., Liakata, M., and Dutta, R. (2016a). Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Computational Linguistics and Clinical Psychology 2016*.

Gkotsis, G., Velupillai, S., Oellrich, A., Dean, H., Liakata, M., and Dutta, R. (2016b). Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105, San Diego, CA, USA, June. Association for Computational Linguistics.

Haerian, K., Salmasian, H., and Friedman, C. (2012). Methods for identifying suicide or suicidal ideation in EHRs. In *AMIA Annual Symposium Proceedings*, pages 1244–1253. American Medical Informatics Association.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ji, S., Yu, C. P., Fung, S.-f., Pan, S., and Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *The 3rd International Conference for Learning Representations*.

Lai, S., Liu, K., He, S., and Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, Nov.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *Proceedings of the 2016 Conference of the*

[7] See `https://www.slam.nhs.uk/research/cris` for details.

North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., and Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Metzger, M.-H., Tvardik, N., Gicquel, Q., Bouvry, C., Poulet, E., and Potinet-Pagliaroli, V. (2017). Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study. *International journal of methods in psychiatric research*, 26(2):e1522.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Parks, S. E., Johnson, L. L., McDaniel, D. D., and Gladden, M. (2014). Surveillance for violent deaths â national violent death reporting system, 16 states, 2010. *Morbidity and Mortality Weekly Report: Surveillance Summaries*, 63(1):1–33.

Perera, G., Broadbent, M., Callard, F., Chang, C.-K., Downs, J., Dutta, R., Fernandes, A., Hayes, R. D., Henderson, M., Jackson, R., Jewell, A., Kadra, G., Little, R., Pritchard, M., Shetty, H., Tulloch, A., and Stewart, R. (2016). Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open*, 6(3).

Sawhney, R., Manchanda, P., Mathur, P., Shah, R., and Singh, R. (2018). Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175.

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., and Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Song, X., Petrak, J., and Roberts, A. (2018). A deep neural network sentence level classification method with context information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 900–904.

Tran, T., Nguyen, T. D., Phung, D., and Venkatesh, S. (2015). Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of biomedical informatics*, 54:96–105.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., d Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., and Dutta, R. (2018). Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future adv ances. *Journal of Biomedical Informatics*, 88:11 – 19.

Velupillai, S., Epstein, S., Bittar, A., Stephenson, T., Dutta, R., and Downs, J. (2019). Identifying Suicidal Adolescents from Mental Health Records Using Natural Language Processing. *Studies in Health Technology and Informatics*, 264:413–417, August.

Walsh, C. G., Ribeiro, J. D., and Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*, 5(3):457–469, June.

WHO. (2018). Suicide fact sheet, reviewed January 2018. Online, January.

Zhang, S., Jiang, H., Xu, M., Hou, J., and Dai, L. (2015). The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.