HHS Public Access

Author manuscript

Proc Conf. Author manuscript; available in PMC 2023 February 25.

Published in final edited form as:

Proc Conf. 2022 July ; 2022: 1029–1040. doi:10.18653/v1/2022.naacl-main.75.

ScAN: Suicide Attempt and Ideation Events Dataset

Bhanu Pratap Singh Rawat[1,*], **Samuel Kovaly**[1,#], **Wilfred R. Pigeon**[2,3,†], **Hong Yu**[1,3,4,‡]

[1]CICS, UMass-Amherst,

[2]University of Rochester,

[3]U.S. Department of Veterans Affairs,

[4]Center of Biomedical and Health Research in Data Sciences

Abstract

Suicide is an important public health concern and one of the leading causes of death worldwide. Suicidal behaviors, including suicide attempts (SA) and suicide ideations (SI), are leading risk factors for death by suicide. Information related to patients' previous and current SA and SI are frequently documented in the electronic health record (EHR) notes. Accurate detection of such documentation may help improve surveillance and predictions of patients' suicidal behaviors and alert medical professionals for suicide prevention efforts. In this study, we first built **S**uicide **A**ttempt and Ideatio**n** Events (ScAN) dataset, a subset of the publicly available MIMIC III dataset spanning over 12k+ EHR notes with 19k+ annotated SA and SI events information. The annotations also contain attributes such as method of suicide attempt. We also provide a strong baseline model ScANER (**S**ui**c**ide **A**ttempt and Ideatio**n** **E**vents **R**etreiver), a multi-task RoBERTa-based model with a *retrieval module* to extract all the relevant suicidal behavioral evidences from EHR notes of an hospital-stay and, and a *prediction module* to identify the type of suicidal behavior (SA and SI) concluded during the patient's stay at the hospital. ScANER achieved a macro-weighted F1-score of 0.83 for identifying suicidal behavioral evidences and a macro F1-score of 0.78 and 0.60 for classification of SA and SI for the patient's hospital-stay, respectively. ScAN and ScANER are publicly available[1].

1 Introduction

For decades, suicide has been one of the leading causes of death (CBHSQ, 2020). The suicide rate in the United States increased from 10.5 per 100, 000 in 1999 to 14.2 in 2018, a 35% increase (Hedegaard et al., 2020). Globally, 740, 000 people commit suicide each year. The rates of suicidal behaviors, suicide attempt (SA) and suicide ideation (SI), are much higher (WHO, 2021).

A prior study shows that a large proportion of suicide victims sought care well before their death (Kessler et al., 2020). Suicidal behaviors, including SA and SI are recorded by

[1]The annotations, code and the models are availble at https://github.com/bsinghpratap/ScAN.

[*] brawat@umass.edu,. [#] skovaly@umass.edu,. [†] wilfred_pigeon@urmc.rochester.edu,. [‡] hong_yu@uml.edu .

[2] https://github.com/medspacy/medspacy

clinicians in electronic health records (EHRs). This knowledge can in turn help clinicians assess risk of suicide and make prevention efforts (Jensen et al., 2012). The diagnostic ICD codes include suicidality codes for both SA and SI. However a study has shown that ICD codes can only capture 3% SI events, while 97% of SI events are described in notes (Anderson et al., 2015). In addition, of patients described with SA in their EHR notes, only 19% had the corresponding ICD codes (Anderson et al., 2015). Therefore, it is important to develop natural language processing (NLP) approaches to capture such important suicidality information.

Researchers have developed NLP approaches to detect SA and SI from EHR notes (Metzger et al., 2017; Downs et al., 2017; Fernandes et al., 2018; Cusick et al., 2021). These studies either used rule-based approaches (Downs et al., 2017; Fernandes et al., 2018; Cusick et al., 2021) or built the SA and SI identification models on a small set (Metzger et al., 2017) or private set (Bhat and Goldman-Mellor, 2017; Tran et al., 2013; Haerian et al., 2012) of EHR notes. It is also difficult to compare the results of those studies as they varied in EHR data, data curation, as well as NLP models, which were not made available to the public.

In this study, we present ScAN: **S**uicide **A**ttempt and Ideatio**n** Events Dataset, a publicly available EHR dataset that is a subset of the MIMIC III data (Johnson et al., 2016). ScAN contains 19, 690 expert-annotated SA and SI events with their attributes (e.g., methods for SA) over 12, 759 EHR notes. Specifically, experts annotated suicidality evidence or sentences relevant to SA and SI events during a patient's stay at the healthcare facility, an example of SA annotations is shown in Fig 1. The evidences were put together to assess whether the patient has an SA or SI event.

We also present ScANER (**S**ui**c**ide **A**ttempt and Ideatio**n** **E**vents **R**etriever), a RoBERTa-based NLP model that is built on a multi-task learning framework for retrieving evidences from the EHRs and then predicting a patient's SA or SI event using the complete set of EHR notes from the hospital stay using a multi-head attention model. We focus on the prediction of SA and SI using all the EHR notes during a patient's stay because for the whole duration, multiple EHR notes and note types are generated, including *admission* notes, *nursing* notes, and *discharge summary* notes. Suicidal information are described in multiple notes throughout the stay. For example, a patient was admitted to the hospital with opioid overdose. It was documented initially in the admission note as an SA, but later dismissed as an accident after physician's evaluation. In another example, an opioid overdose admission was first documented as an accident on admission, but later documented to be an SA event after clinical assessment. Both ScAN and ScANER capture SA and SI information at the hospital-stay level. ScANER is able to retrieve suicidal evidences from EHR notes with a macro-weighted F1-score of 0.83 and is able to predict SA and SI with a macro F1-score of 0.78 and 0.60, respectively. Our annotation guidelines, ScAN, and ScANER system will be made publicly available, making ScAN a benchmark EHR dataset for SA and SI events detection. We will release the training and evaluations splits used in this study for benchmarking new models.

2 Related Works

Efforts on detecting SA and SI within EHRs have been explored in recent years. Most work used rule-based or traditional machine learning-based approaches. In one study, experts created hand-crafted rules from mentions of suicidality (both SA and SI) and then used the rules to identify suicidality as positive, negative, or unknown in a document (Downs et al., 2017). The rule-based approaches are limited by their scalability. In another study, structured and unstructured EHRs were used to classify at the hospital-stay level as SA, SI, or no mention of suicidal behavior (Metzger et al., 2017). The training data consisted of only 112 SA, 49 SI and 322 unrelated examples. In contrast, ScAN comprises of 697 hospital-stays with more than 19,000 suicidal event examples over 12,759 clinical notes. Only traditional machine learning models such as random forest (Breiman, 2001) were explored. In contrast, ScANER was built on the state-of-the-art self-attention based model.

Hybrid approaches have also been developed to identify SA at the hospital-stay level (Fernandes et al., 2018). In that study, a post-processing heuristic rule-based filter (e.g., removing negated events) was applied to the machine-learning-based classifier (a SVM (Cortes and Vapnik, 1995) classifier) to reduce false positives. Training and evaluation were done also on relatively small datasets (500 for training and 500 for evaluation).

Finally, weakly supervised approaches have been developed to identify SI from EHRs (Cusick et al., 2021). In that study, authors used ICD codes to identify 200 patients with SI and then obtained EHR notes of those patients (6, 588). This EHR note dataset was then used as the ‘current’ SI training data. The remaining 400 patients were labelled as ‘potential’ SI and their 12, 227 EHRs were also labelled the same. Authors used multiple statistical machine learning models and one deep learning model: convolutional neural network. (Bhat and Goldman-Mellor, 2017) also used feed-forward neural networks to predict suicide attempts over 500k unique patients but the EHR data for this study is not publicly available. (Ji et al., 2020) surveyed multiple studies where the researchers worked on private datasets (Tran et al., 2013; Haerian et al., 2012) for suicide attempt and ideation prediction. Whereas in our study, in contrast to using the ICD codes which has considerable errors, domain experts chart-reviewed a large, publicly available set of EHRs for SI and SA, along with their attributes (e.g., positive or negative SA, SI and the type of self-harm such as asphyxiation and overdose).

3 Dataset

In this section, we introduce ScAN (**S**ui**c**ide **A**ttempt and Ideatio**n** Events Dataset) and describe it’s data collection and annotation process. We also discuss some examples from ScAN along with basic dataset statistics.

3.1 Dataset collection

For annotation, we selected the notes from the MIMIC-III (Johnson et al., 2016) dataset, which consists of the de-identified EHR data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012 (Johnson et al., 2016). The data includes notes, diagnostic codes, medical history, demographics, lab

measurements among many other record types. We chose MIMIC-III because it is publicly available under a data use agreement and allows clinical studies to be easily reproduced and compared.

The diagnostic ICD codes for the patients are provided at hospital-stay level in MIMIC with admission identification numbers (HADM_ID in MIMIC database). We first filtered the hospital stays that had ICD codes associated with suicide and overdose. This resulted in 697 hospital-stays for 669 unique patients. For each stay, multiple de-identified notes such as nursing notes, physician notes, and discharge summaries are available. For the selected 697 hospital-stays we extracted a total of 12, 759 notes. Each medical note contains multiple sections about a patient such as family and medical history, assessment and plan, and discharge instructions. We extracted different sections from these clinical notes using MedSpaCy's[2] `clinical_sectionizer` and filtered the relevant sections from these clinical notes for annotation. The extensive list of these sections is provided in Appendix A.

3.2 Annotation Process

The aim was to annotate all instances of SA and SI documented in the medical notes as defined by Center of Disease Control and Prevention (CDC) (Hedegaard et al., 2020). The filtered 12, 759 notes were annotated by a trained annotator under the supervision of a senior physician. Each note consisted of instances of SA, SI, both or none. The senior physician randomly annotated 330 notes and had a 100% agreement with the annotator on hospital-stay level annotation and 85% agreement on sentence-level annotations. After adjudication between the senior physician and the annotator, the disagreements were discussed and adjusted by the annotator.

Suicide Attempt (SA): The annotator labelled all the sentences with a mention of SA. Some hospital stays could represent multiple types of SA, such as in Fig. 1, where 'tried to hang himself' is labelled as a *positive* SA and 'Tylenol overdose' is labelled as *unsure* since the overdose was never confirmed as an SA event elsewhere in the medical notes of the patient's hospital-stay. The label *unsure* is used when it is not clearly documented if a self-harm was an SA event or not. The *negative* instance, example shown in Fig. 2, is a sentence that confirms that the self-harm, an "accidental overdose", responsible for the patient's hospital-stay is not an SA event. In this work, we only focused on suicidal self-harm and not non-suicidal self-harm (Crosby et al., 2011).

Further sub-categories are also provided for an SA annotation in the form of the ICD label group: a.) T36-T50: Poisoning by drugs, medications and biological substances b.) T51-T65: Toxic effects on non-medical substances c.) T71: Asphyxiation or suffocation and d.) X71-X83: Drowning, firearm, explosive material, jumping from a high place, crashing motor vehicles, other specified means.

Suicide Ideation (SI): SI is defined as any mention and/or indication of wanting to take one's own life or harm oneself. Similar to SA, any sentence with a mention of SI was

[2]These splits would be released as part of ScAN for benchmarking models.

labelled within the patient's notes. A SI annotation could be labeled as *positive* or *negative*, an example for each label is shown in Fig. 3.

A sentence without SA or SI annotation would be considered as a *neutral-SA* or *neutral-SI* sentence respectively. Sentence level annotations provide more visibility to a medical expert for the hospital-stay level annotation.

3.3 Dataset statistics

ScAN consists of 19, 690 unique evidence annotations for the suicide relevant sections of 12, 759 EHRs of 697 patient hospital-stays. There are a total of 17, 723 annotations for SA events and 1, 967 annotations for SI events. The distribution for both SA and SI events is provided in Table 1.

4 Methodology

In this section, we introduce *ScANER* (**S**ui**c**ide **A**ttempt and Ideatio**n** **E**vents **R**etreiver): a strong baseline model for our dataset. ScANER consists of two sub-modules: (1) An *evidence retriever module* that extracts the evidences related to SA and SI and (2) A *predictor module* that predicts SA or SI label for the patient's hospital-stay using the evidences extracted by the retriever module.

4.1 Evidence Retriever

Problem Formulation: Given an input clinical note, the model extracts the evidences (one or more sentences) related to SA or SI (SA-SI) from the note. This is a binary classification problem where given a text snippet the model predicts whether it has an evidence for SA-SI or not. We learn this task at paragraph level where the input is a set of 20 consecutive sentences because the local surrounding context provides additional important information (Yang et al., 2021; Rawat et al., 2019). A paragraph was labeled as *evidence no*, if all the sentences in that paragraph are *neutral-SA* and *neutral-SI*. If there was at least one SA-SI sentence, it was considered an *evidence yes*. As the number of non-evidence sentences significantly outsized the evidence sentences, we decided to use an overlapping window of 5 sentences between the paragraphs to build more evidence paragraphs. The distribution of the paragraphs, across all evidence, SA and SI labels for train, validation, and test set is provided in Table 2. We segregated the train and test set such that any patient observed by the *retriever module* during training was not seen in the test set. This is important as there are patients who had multiple hospital-stays in ScAN.

Proposed Model: Transformer (Vaswani et al., 2017) based language models (Devlin et al., 2018; Liu et al., 2019) have shown great performance for a broad range of NLP classification tasks. Hence, to extract the evidence paragraphs we trained a RoBERTa (Liu et al., 2019) based model. It has been previously shown that the domain-adapted versions of the pre-trained language models, such as clinicalBERT (Alsentzer et al., 2019) or BioBERT (Lee et al., 2020), work better than their base versions. So, we further pre-trained the RoBERTa-base model over the MIMIC dataset to create a clinical version of RoBERTa model, hereby referenced as medRoBERTa. During our initial exploration, we experimented with

clinicalBERT and BioBERT but found that medRoBERTa consistently outperformed both models. medRoBERTa achieved an overall F1-score of 0.88 whereas both clinicalBERT and BioBERT achieved an overall F1-score of 0.85. Our hospital-level SA and SI predictor would work with any encoder-based evidence retriever model.

Multi-task Learning: We trained medRoBERTa in a multi-task learning setting where along with learning the evidence classification task, the model also learns two auxiliary tasks: (a.) Identifying the label for SA between *positive*, *negative*, *unsure* and *neutral-SA* and, (b.) Identifying the label for SI between *positive*, *negative* and *neutral-SI*. The training loss ($L(\theta)$) for our evidence retriever model was formulated as:

$$L(\theta) = L_{evi} + \alpha * L_{SA} + \beta * L_{SI} \quad (1)$$

Where L_{evi} is the negative log likelihood loss for evidence classification, L_{SA} and L_{SI} are SA and SI prediction losses respectively, and α and β are the weights for the auxiliary tasks' losses. The distribution of labels across all the three tasks is highly skewed, hence, we applied the following techniques to learn an efficient and robust model.

- Weighted log loss was used in both main task and auxiliary tasks. The total loss for each task was calculated as the weighted sum of loss according to the *label* of the input paragraph. Log weighing helps smooth the weights for highly unbalanced classes. The weight for each class was calculated using:

$$w_{l,t} = \begin{cases} 1.0 & \text{if}(w_{l,t} < 1.0) \\ log(\gamma * N_t / N_{l,t}) \end{cases}$$

  Where N_t is the count of all training paragraphs for the task t and $N_{l,t}$ is the count of paragraphs with label l for the task t and $w_{l,t}$ is the calculated weight for those paragraphs. We tuned γ as a hyper-parameter. All training hyper-parameters for our best model are provided in Appendix B.

- We also employed different sampling techniques (Youssef, 1999), up and down sampling, to help our model learn from an imbalanced dataset. After sweeping for different sampling combinations as hyper-parameters, we found that down-sampling the *no*-evidence paragraphs by 10% resulted in the best performance.

- The *negative* label of SA is severely under-represented in ScAN making it difficult for the model to learn useful patterns from such instances, refer Table 2. After discussion with the experts, we decided to group the instances of *negative* and *unsure* together and label them as *neg_unsure* because for both groups the general psych outcome is to let the patient leave after the hospital-stay as there is no solid evidence defining whether the self-harm was a SA event.

4.2 Hospital-stay level SA and SI Predictor

Problem Formulation—Given all the clinical notes of a patient during the the hospital stay, the model predicts the label for SA (*positive*, *neg_unsure* and *neutral-SA*) and SI

(*positive*, *negative* and *neutral-SI*). The *prediction module* uses the evidence paragraphs extracted by the *retriever module*.

Robust Finetuning—The retriever module is not perfect and can extract false positives. This results in extracting irrelevant paragraphs, with evidence label *No*, along with evidence paragraphs for a hospital-stay with SA or SI and extracting irrelevant paragraphs as evidences for a hospital-stay with both SA and SI marked as *neutral*. To tackle such situations and train a robust model, we applied three techniques:

- For a hospital-stay with a non-*neutral* label for SA or SI, during training we added some noise in the form of irrelevant paragraphs (a paragraph with no SA or SI annotation) from the notes to the set of actual evidence paragraphs for the input. An irrelevant paragraph from a clinical note was sampled with a probability of 0.05. This forced the predictor module to learn effectively even with noisy inputs.
- For a *neutral* hospital-stay with no evidence paragraphs, we randomly chose X unique irrelevant paragraphs from the notes. X was sampled from the distribution of number of evidence paragraphs of the non-*neutral* hospital-stays. This prevented the leaking of any information to the prediction module during training by keeping the distribution of number of input paragraphs the same across *neutral* and non-*neutral* instances.
- Since these hospital-stays were extracted using the ICD codes related to suicide and overdose, the data is quite skewed with only 102 *neutral* events from a total of 697 hospital-stays. Whereas in a real-world scenario, *neutral* hospital-stays would be much higher than non-*neutral* ones. Hence, to facilitate a balanced learning of the predictor module we introduced 1, 800 *neutral* hospital-stays from the MIMIC dataset. The distribution for SA and SI at hospital-stay level is provided in Table 3.

Proposed Model—The paragraphs extracted using the *retriever module* for a patient's hospital-stay were provided as an input to the predictor module. We used a multi-head attention model to predict the SA and SI label for a hospital-stay as self-attention based models have proved to be quite effective for a lot of prediction tasks in machine learning (Devlin et al., 2018; Cao et al., 2020; Hoogi et al., 2019).

We encoded the extracted paragraphs ($[p_1, p_2 \ldots p_n]$) using the retriever module, medRoBERTa, to get a vector representation of 768 dimensions for each of the paragraphs ($[v_1, v_2 \ldots v_n]$). Training the *retriever module* on auxiliary tasks of predicting SA and SI helped align these paragraph representations for SA and SI prediction. Then, we added a prediction vector (v_0) along with all the vector representations of the paragraphs to get $\mathcal{V} = [v_0, v_1, v_2 \ldots v_n]$. We passed $\mathcal{V}$ through our multi-head attention model to get the hidden representations $\mathcal{H} = [h_0, h_1 \ldots h_n]$. We then passed h_0 through a SA inference layer and SI inference layer to predict the labels. During the whole training process, the weights of the retriever module were frozen whereas v_0 was a learnable vector initialised as an embedding in the multi-head attention model. We used a separate v_0 prediction vector

so that it could retain the information from all the other paragraph representations for hospital-stay level prediction similar to how [CLS] is utilized in different transformer-based models for sequence prediction (Devlin et al., 2018; Liu et al., 2019). We tuned the number of layers and number of attention heads of our prediction module as hyper-parameters and achieved the best performance using a 2-layer and 3-attention head model. Our complete ScANER model is illustrated in Fig 4.

5 Results and Discussion

Since the labels for both the retriever and prediction task are imbalanced, we used macro-weighted precision, recall, and F1-score to evaluate the overall performance of our models. Macro-weighted metrics provide better model insights across all labels.

Evidence Retriever Performance

Our multi-task learning model achieved a F1-score of 0.83 for extracting positive evidence paragraphs and an F1-score of 0.88 overall. The retriever model has higher recall than precision for the positive evidence paragraphs (0.87 > 0.79), SA (0.74 > 0.71), and SI (0.62 > 0.46) events, as shown in Table 4. In healthcare, there is an incentive to maximize recall over precision (Watson and McKinstry, 2009). As mentioned in §4.2, ScANER was trained with added noisy paragraphs and is therefore robust to the extracted evidence paragraphs if they contain some false positives.

The *retriever module* achieves an overall F1-score of 0.63 for SA prediction and 0.64 for SI prediction at paragraph-level. The performance for positive SA and SI evidence is much higher than the performance for *neg_unsure* SA and *negative* SI. We looked at the confusion matrices for SA and SI paragraph-level prediction and found that largely ScANER made mistakes between *positive* and *neg_unsure* labels for SA prediction and between *positive* and *negative* labels for SI prediction (refer Appendix C). The poor performance in SA for *neg_unsure* evidence prediction is mainly due to data sparsity where the *neg_unsure* cases are only 1743; in contrast, the positive cases are 4-fold higher. Similarly, for SI the *positive* cases are 1.4 times higher than the *negative* cases.

Hospital-stay level Prediction Performance

Our multi-head attention model is able to achieve an overall macro F1-score of 0.78 for SA prediction and 0.60 for SI prediction, as shown in Table 5. For SA, the *prediction module* achieves a recall of 0.93 for the positive label. After analysing the confusion matrix, the model largely predicts a *positive* label for the visits with *neg_unsure* label, as shown in Table 6. The poor performance for *neg_unsure* is largely because of its small representation in the training set of ScAN, 54 negative cases as compared to 377 positive and 1, 381 neutral instances. In our future work, we plan to expand ScAN with more instances of *negative* SA events.

For SI, the prediction module achieves an overall F1-score of 0.60 with a precision of 0.63 and recall of 0.66. The model has a high recall for *neutral-SI* and *positive* but the *positive* label has a low precision of 0.49. After analysing the test set, we observed that a lot of patient hospital-stays with *negative* labels are getting wrongly predicted as *positive*,

as shown in Table 6. After doing error analysis for hospital-stays with *negative* labels, we observed that a lot of extracted evidence paragraphs contain information that suggests that the patient had SI before the SA but does not have SI anymore during the hospital-stay. As shown in the example in Fig 5, the past SI is an explanation for the SA but then the patient does not have any further SI during the hospital-stay. This suggests that period assertions for these annotations are quite important and we aim to add period assertion property in our future work by further annotating ScAN.

6 Conclusion

In this paper, we introduce ScAN: a publicly available suicide attempt (SA) and ideation (SI) events dataset that consists of 12, 759 EHR notes with 19, 960 unique evidence annotations for suicidal behavior. To our knowledge, this is the largest and publicly available dataset for SA and SI, an important resource for suicidal behaviors research. We also provide a strong RoBERTa baseline model for the dataset: ScANER (SA and SI retriever) which consists of two sub-modules: (a.) an *evidence retriever module* that extracts all the relevant evidence paragraphs from the patient's notes and (b.) a *prediction module* that evaluates the extracted evidence paragraphs and predicts the SA and SI event label for the patient's stay at the hospital. ScAN and ScANER could help extract suicidal behavior in patients for suicide surveillance and predictions, leading to potentially early intervention and prevention efforts by medical professionals.

A: Selected Clinical Sections

The sections selected for annotations after using `clinical_sectionizer` are enumerated below:

1. Allergys
2. Case Management
3. Consult
4. Discharge Summary
5. Family history
6. General
7. HIV Screening
8. Labs and Studies
9. Medication
10. Nursing
11. Nursing/other
12. Nutrition
13. Observation and Plan
14. Past Medical History

15. Patient Instructions
16. Physical Exam
17. Physician
18. Present Illness
19. Problem List
20. Radiology
21. Rehab Services
22. Respiratory
23. Sexual and Social History
24. Social Work

B: Hyper-parameter Settings

All the hyper-parameter settings for both modules of ScANER are provided in Table 7.

Table 7:

Hyper-parameter setting for both retriever and prediction module of ScANER.

Evidence Retriever Module			
Learning Rate	Warmup steps	Optimizer	Adam ϵ
2e-5	2,000	Adam	1e-8
γ	α	β	
2.5	1,1	1,5	
Hospital Stay SA-SI Prediction Module			
Attention Heads	Attention Layers	Learning Rate	Warmup steps
3	2	1e-3	1,200
Optimizer	Adam ϵ		
Adam	1e-8		

C: Confusion matrices

The confusion matrices for SA and SI prediction at paragraph level is provided in Table 8.

Table 8:

Confusion matrices for the predictions on the test set of evidence retriever.

Paragraph SA Prediction			
	Positive	Neg_Unsure	Neutral-SA
Positive	1,804	285	344
Neg_Unsure	253	118	80

Neutral-SA	472	204	7,314
Paragraph SI Prediction			
	Positive	Negative	Neutral-SI
Positive	206	69	56
Negative	71	87	31
Neutral-SI	170	73	10,111

References

Alsentzer Emily, Murphy John R, Boag Willie, Weng Wei-Hung, Jin Di, Naumann Tristan, and McDermott Matthew. 2019. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323.

Anderson Heather D, Pace Wilson D, Brandt Elias, Nielsen Rodney D, Allen Richard R, Libby Anne M, West David R, and Valuck Robert J. 2015. Monitoring suicidal patients in primary care using electronic health records. The Journal of the American Board of Family Medicine, 28(1):65–71. [PubMed: 25567824]

Bhat Harish S and Goldman-Mellor Sidra J. 2017. Predicting adolescent suicide attempts with neural networks. arXiv preprint arXiv:1711.10057.

Breiman Leo. 2001. Random forests. Machine learning, 45(1):5–32.

Cao Ran, Fang Leyuan, Lu Ting, and He Nanjun. 2020. Self-attention-based deep feature fusion for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 18(1):43–47.

Center for Behavioral Health Statistics Quality CBHSQ. 2020. 2019 national survey on drug use and health (nsduh): Methodological summary and definitions. Substance Abuse and Mental Health Services Administration.

Cortes Corinna and Vapnik Vladimir. 1995. Support-vector networks. Machine learning, 20(3):273–297.

Crosby Alex, Ortega LaVonne, and Melanson Cindi. 2011. Self-directed violence surveillance; uniform definitions and recommended data elements.

Cusick Marika, Adekkanattu Prakash, Campion Thomas R Jr, Sholle Evan T, Myers Annie, Banerjee Samprit, Alexopoulos George, Wang Yanshan, and Pathak Jyotishman. 2021. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. Journal of psychiatric research, 136:95–102. [PubMed: 33581461]

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Downs Johnny, Velupillai Sumithra, George Gkotsis, Holden Rachel, Kikoler Maxim, Dean Harry, Fernandes Andrea, and Dutta Rina. 2017. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. In AMIA annual symposium proceedings, volume 2017, page 641. American Medical Informatics Association. [PubMed: 29854129]

Fernandes Andrea C, Dutta Rina, Velupillai Sumithra, Sanyal Jyoti, Stewart Robert, and Chandran David. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific reports, 8(1):1–10. [PubMed: 29311619]

Haerian Krystl, Salmasian Hojjat, and Friedman Carol. 2012. Methods for identifying suicide or suicidal ideation in ehrs. In AMIA annual symposium proceedings, volume 2012, page 1244. American Medical Informatics Association. [PubMed: 23304402]

Hedegaard Holly, Curtin Sally C, and Warner Margaret. 2020. Increase in suicide mortality in the united states, 1999–2018.

Hoogi Assaf, Wilcox Brian, Gupta Yachee, and Rubin Daniel L. 2019. Self-attention capsule networks for image classification. arXiv preprint arXiv:1904.12483.

Jensen Peter B, Jensen Lars J, and Brunak Søren. 2012. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6):395–405.

Ji Shaoxiong, Pan Shirui, Li Xue, Cambria Erik, Long Guodong, and Huang Zi. 2020. Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems.

Johnson Alistair EW, Pollard Tom J, Shen Lu, H Lehman Li-Wei, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, and Mark Roger G. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9.

Kessler Ronald C, Bossarte Robert M, Luedtke Alex, Zaslavsky Alan M, and Zubizarreta Jose R. 2020. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Molecular psychiatry, 25(1):168–179. [PubMed: 31570777]

Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, So Chan Ho, and Kang Jaewoo. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240. [PubMed: 31501885]

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, Zettlemoyer Luke, and Stoyanov Veselin. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Metzger Marie-Hélène, Tvardik Nastassia, Gicquel Quentin, Bouvry Côme, Poulet Emmanuel, and Potinet-Pagliaroli Véronique. 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. [PubMed: 27634457]

Rawat Bhanu Pratap Singh, Li Fei, and Yu Hong. 2019. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2547–2555.

Tran Truyen, Phung Dinh, Luo Wei, Harvey Richard, Berk Michael, and Venkatesh Svetha. 2013. An integrated framework for suicide risk prediction. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1410–1418.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Aidan N Gomez Łukasz Kaiser, and Polosukhin Illia. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Watson Philip WB and McKinstry Brian. 2009. A systematic review of interventions to improve recall of medical advice in healthcare consultations. Journal of the Royal Society of Medicine, 102(6):235–243. [PubMed: 19531618]

WHO. 2021. Suicide fact sheet: https://www.who.int/news-room/fact-sheets/detail/suicide.

Yang Baosong, Wang Longyue, Wong Derek F, Shi Shuming, and Tu Zhaopeng. 2021. Context-aware self-attention networks for natural language processing. Neurocomputing, 458:157–169.

Youssef Abdou. 1999. Image downsampling and up-sampling methods. National Institute of Standards and Technology.

Context:
[*Name*] is a 22 year old male with a history of angry and impulsive behavior who is transferred from an outside hospital s/p Tylenol overdose. [*Name*] reports that he and his girlfriend broke up last Wednesday, and that he subsequently went on an alcohol and cocaine binge lasting from Thursday to Saturday. He has used alcohol and cocaine regularly in the past, but he denies having had a binge of this quantity or duration before. On Saturday night, [**Name**] told his father that he had tried to hang himself at a nearby park, but the rope had broken.

Annotations:
Two instances of suicide attempts are annotated in the paragraph.
1. [*Name*] ... overdose: Annotated for suicide attempt and is assigned 'unsure' category as there is no definite documentation that it is a suicide attempt.
2. On Sat ... broken: This is also annotated for suicide attempt and is assigned the category: 'T71' (asphyxiation, hanging).

Figure 1:
An example of *positive* and *unsure* evidence annotations for SA in an EHR note.

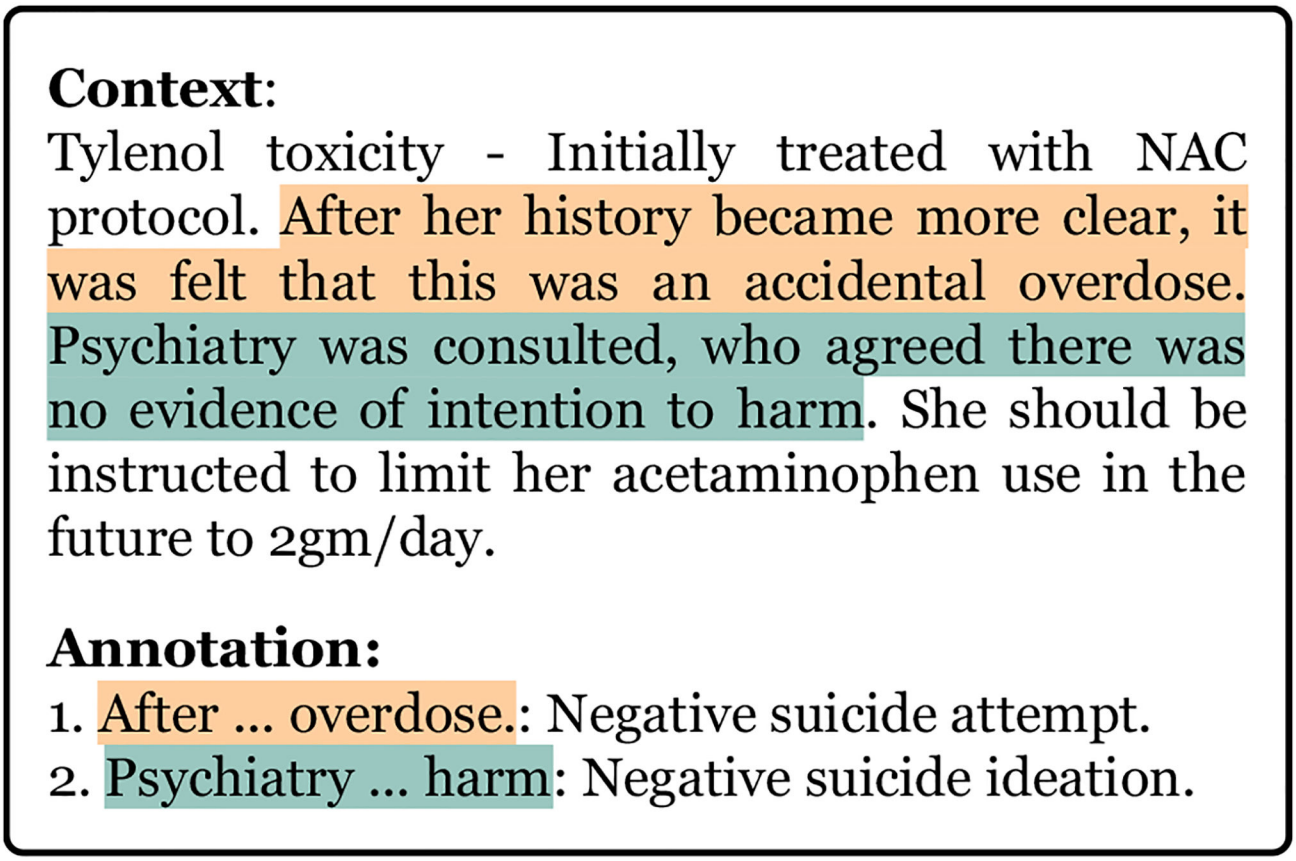

Figure 2:
An example with *negative* SA and *negative* SI annotations.

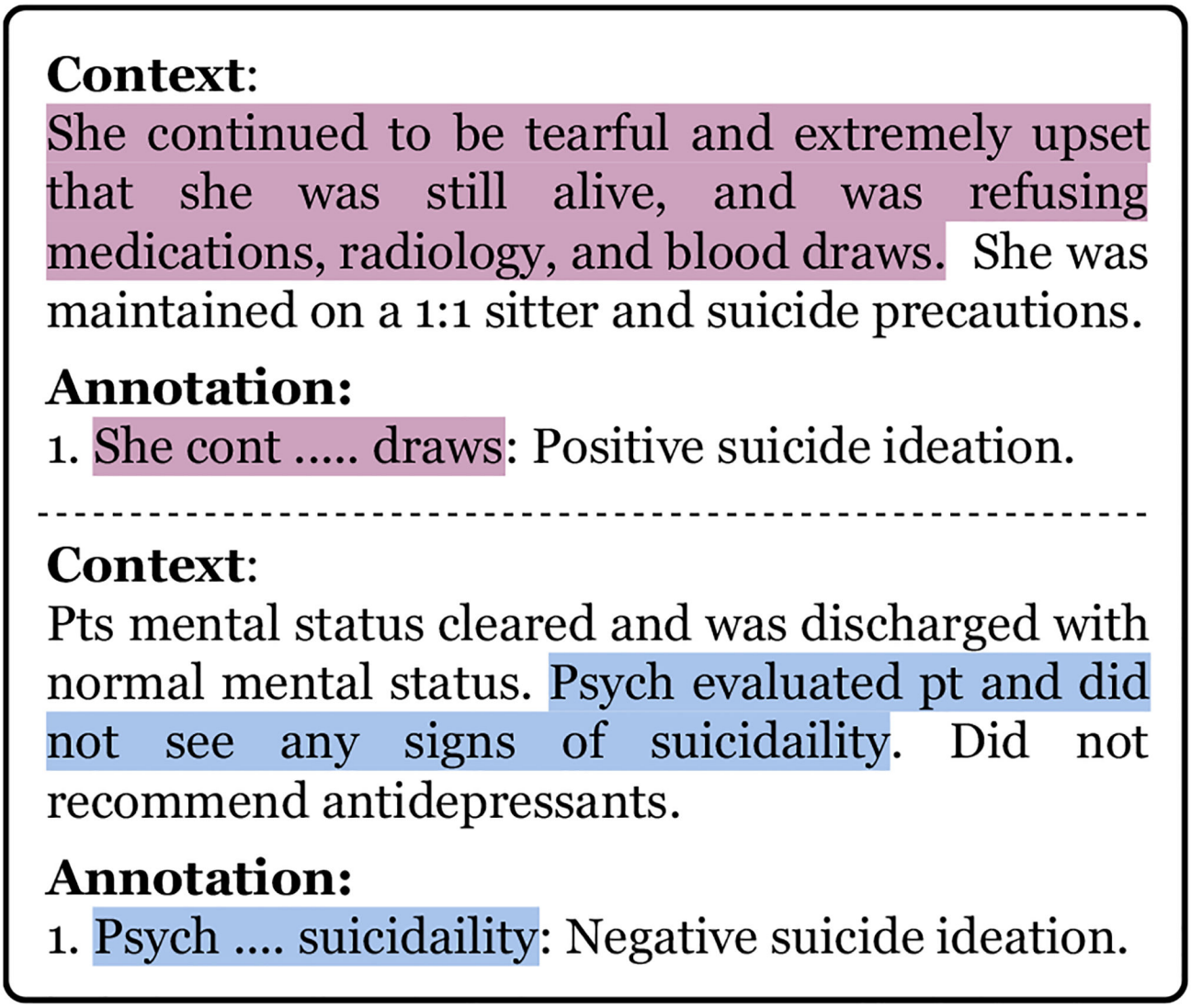

Figure 3:
Examples of *positive* and *negative* SI annotations.

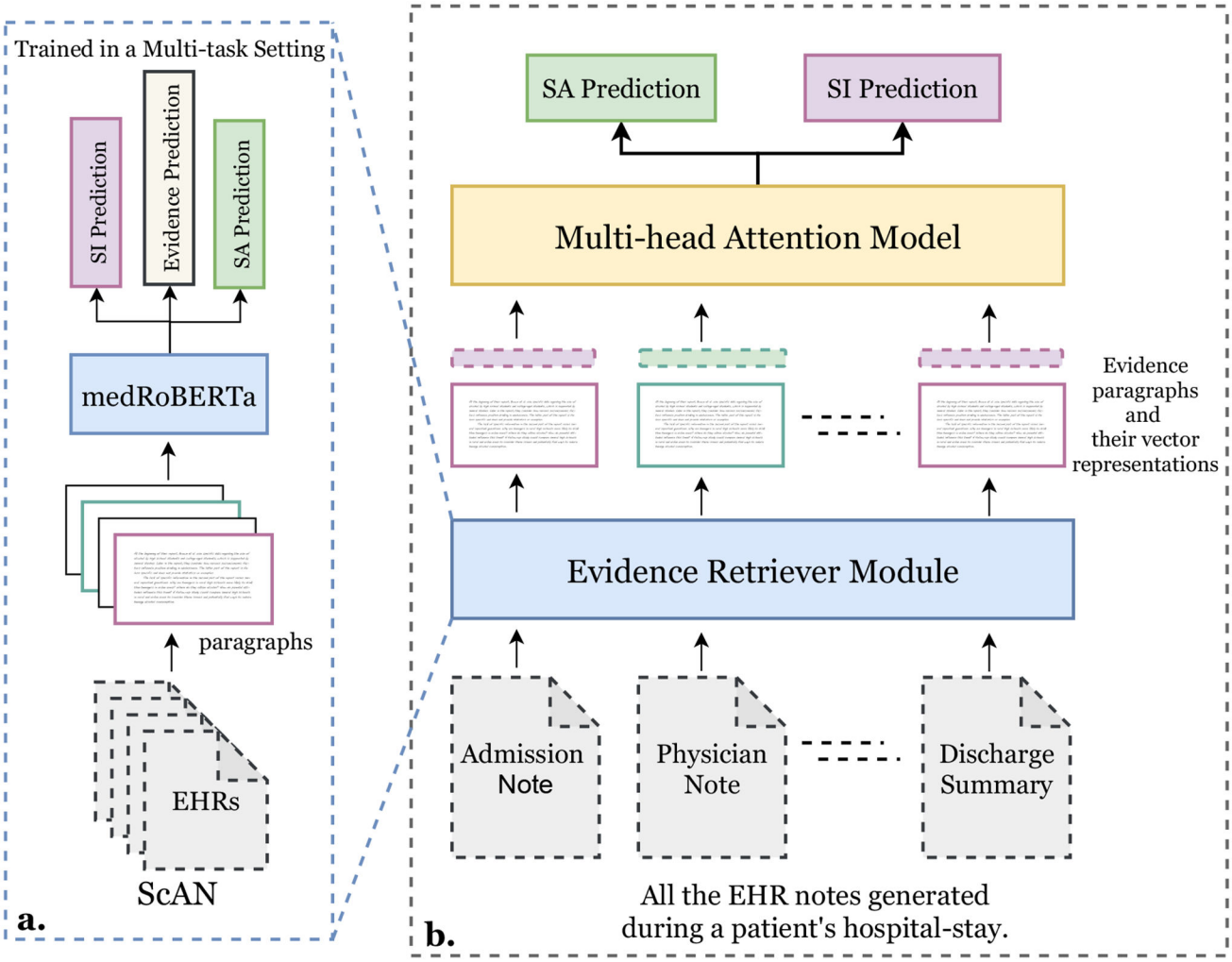

Figure 4:

ScANER (**Sui**cide **A**ttempt and Ideatio**n** **E**vents **R**etreiver) consists of two sub-modules: (a.) Evidence retriever module extracts evidence paragraphs from all EHR notes. We trained the module using all annotated paragraphs from ScAN. (b.) Prediction module predicts the SA and SI label for a patient using the evidence paragraphs extracted by the *retriever module* from EHR notes during the patient's hospital-stay.

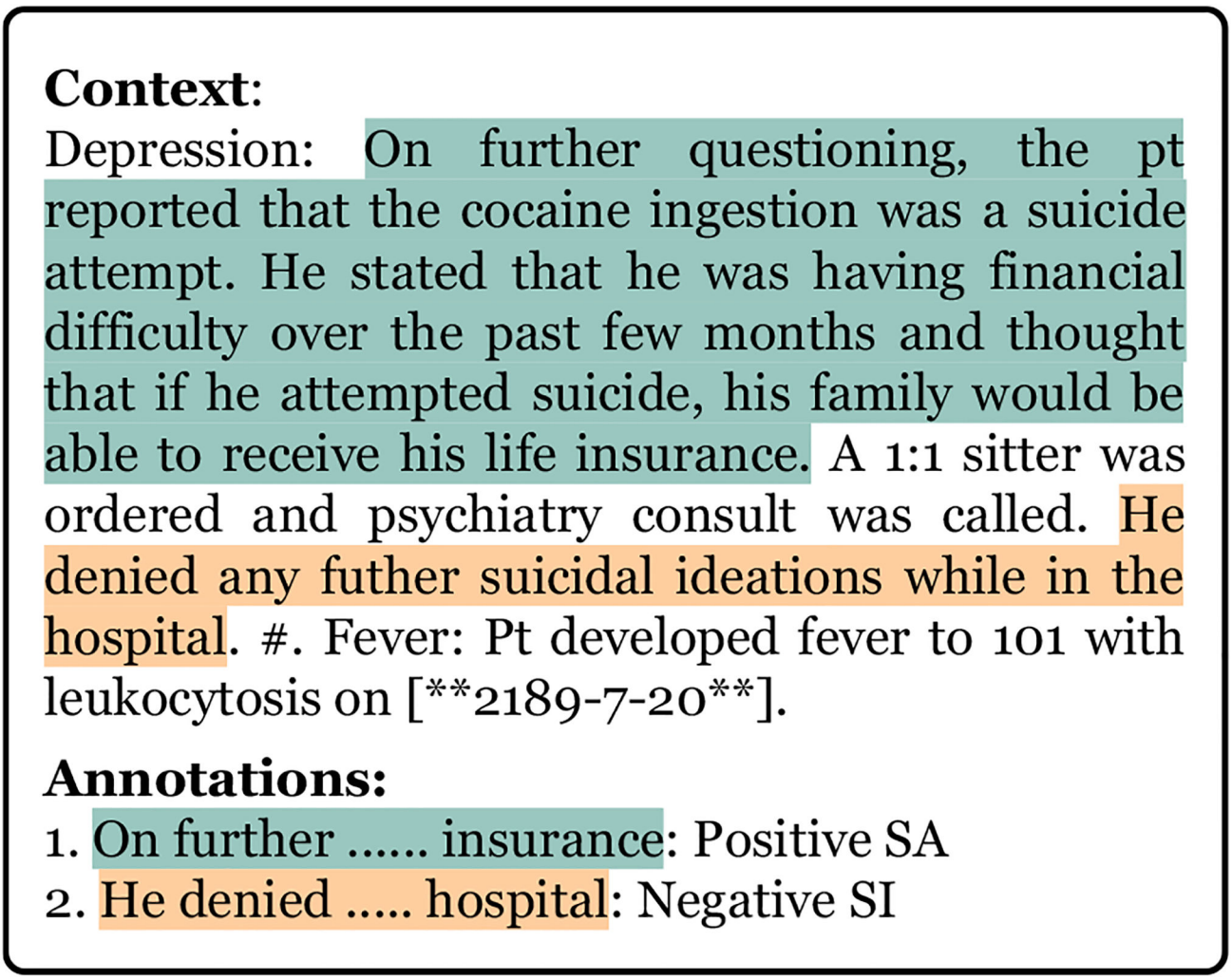

Figure 5:
An instance for which ScANER incorrectly predicted a *negative* hospital-level SI as *positive*.

Table 1:

Distribution of unique annotations at the patient, hospital-stay and notes level in ScAN.

General Statistics	Patients 669	Hospital-stays 697	Notes 12, 759
Suicide Attempt	Positive 14, 815	Negative 170	Unsure 2, 738
Suicide Ideation	Positive 1,167	Negative 800	

Table 2:

Distribution of evidences at paragraph level in ScAN for train, validation and test sets. A paragraph was considered an *evidence*, labeled as *Yes*, if it had at least one sentence annotated as SA or SI. A *No* evidence paragraph was both *Neutral-SA* and *Neutral-SI*.

Evidence	**Train**	**Validation**	**Test**
Yes	9,880	1,803	3,038
No	30,133	4,864	7,836
Suicide Attempt (SA)			
Positive	7,597	1,474	2,433
Negative	136	36	20
Unsure	1,607	216	431
Neutral-SA	30,673	4,941	7,990
Suicide Ideation (SI)			
Positive	928	153	331
Negative	654	107	189
Neutral-SI	38,431	6,407	10,354

Table 3:

Distribution of SA and SI at hospital-stay level in training, validation and testing set.

Suicide Attempt	**Positive**	**Neg_Unsure**	**Neutral-SA**
Train	377	54	1, 381
Val	50	10	189
Test	91	19	326
Suicide Ideation	**Positive**	**Negative**	**Neutral-SI**
Train	377	214	1, 521
Val	45	28	208
Test	44	35	357

Table 4:

Paragraph level performance of the *evidence retriever module*. The overall evaluation metrics (precision, recall and F1-score) are macro-weighted. Evidence prediction is the main task whereas SA and SI prediction are auxiliary tasks and help the model align the vector representations of the paragraphs for the *hospital-stay level suicidal behavior prediction*.

Paragraph Evidence Prediction				***Paragraph SA Prediction***				***Paragraph SI Prediction***			
Evidence	**P**	**R**	**F**	**Labels**	**P**	**R**	**F**	**Labels**	**P**	**R**	**F**
Yes	0.79	0.87	0.83	Positive	0.71	0.74	0.73	Positive	0.46	0.62	0.53
No	0.95	0.91	0.93	Neg_Unsure	0.19	0.26	0.22	Negative	0.38	0.46	0.42
-	-	-	-	Neutral-SA	0.95	0.92	0.93	Neutral-SI	0.98	0.99	0.98
Overall	0.87	0.89	0.88	Overall	0.62	0.64	0.63	Overall	0.61	0.69	0.64

P: Precision, R: Recall and F: F1-score.

Table 5:

Hospital-stay level SA and SI prediction performance of ScANER.

Hospital-stay SA Prediction			
Labels	**Precision**	**Recall**	**F1-score**
Positive	0,81	0.93	0.87
Neg_Unsure	0,48	0.58	0.52
Neutral-SA	0,98	0.93	0.96
Overall	0,76	0.81	0.78
Hospital-stay SI Prediction			
Labels	**Precision**	**Recall**	**F1-score**
Positive	0,49	0.93	0,65
Negative	0,40	0.11	0.18
Neutral-SI	0,99	0.95	0.97
Overall	0,63	0.66	0.60

Table 6:

Confusion matrices for SA and SI prediction at hospital-stay level.

Hospital-stay SA Prediction			
	Positive	**Neg_Unsure**	**Neutral-SA**
Positive	85	4	2
Neg_Unsure	5	11	3
Neutral-SA	15	8	303
Hospital-stay SI Prediction			
	Positive	**Negative**	**Neutral-SI**
Positive	41	2	1
Negative	27	4	4
Neutral-SI	15	4	338