World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:17, No:8, 2023

COVID_ICU_BERT: A Fine-tuned Language Model for COVID-19 Intensive Care Unit Clinical Notes

Shahad Nagoor, Lucy Hederman, Kevin Koidl, Annalina Caputo

Abstract—Doctors' notes reflect their impressions, attitudes, clinical sense, and opinions about patients' conditions and progress, and other information that is essential for doctors' daily clinical decisions. Despite their value, clinical notes are insufficiently researched within the language processing community. Automatically extracting information from unstructured text data is known to be a difficult task as opposed to dealing with structured information such as physiological vital signs, images and laboratory results. The aim of this research is to investigate how Natural Language Processing (NLP) techniques and machine learning techniques applied to clinician notes can assist in doctors' decision making in Intensive Care Unit (ICU) for coronavirus disease 2019 (COVID-19) patients. The hypothesis is that clinical outcomes like survival or mortality can be useful to influence the judgement of clinical sentiment in ICU clinical notes. This paper presents two contributions: first, we introduce *COVID_ICU_BERT*, a fine-tuned version of a clinical transformer model that can reliably predict clinical sentiment for notes of COVID patients in ICU. We train the model on clinical notes for COVID-19 patients, ones not previously seen by Bio_ClinicalBERT or Bio_Discharge_Summary_BERT. The model which was based on Bio_ClinicalBERT achieves higher predictive accuracy than the one based on Bio_Discharge_Summary_BERT (Acc 93.33%, AUC 0.98, and Precision 0.96). Second, we perform data augmentation using clinical contextual word embedding that is based on a pre-trained clinical model to balance the samples in each class in the data (survived vs. deceased patients). Data augmentation improves the accuracy of prediction slightly (Acc 96.67%, AUC 0.98, and Precision 0.92).

Keywords—BERT fine-tuning, clinical sentiment, COVID-19, data augmentation.

I. Introduction

CLINICAL notes include information that is not recorded elsewhere in patients' health record such as patients' medication reactions, patients' attitude, or doctors' social interactions with patients' families. This information is essential for doctors to make their daily clinical decisions. Monitoring patients' attitude, their reaction to medications, and their daily progress can be especially valuable in the context of critical care as patients need close and frequent monitoring and their conditions can change or deteriorate quickly. Failing to recognize the patient's holistic condition accurately and comprehensively could result in inaccurate medical plans or mistakes that incur high costs [1]–[4]. Unintentionally inaccurate decision-making or medical mistakes could happen as doctors have narrow time to review patients' chart, intense work environment, and exchange shifts in treating patients. Despite their value, clinical notes are insufficiently researched [5], [6]. This is because notes are unstructured as opposed to structured information such as physiological vital signs, images and laboratory results. In addition, variations in language expressions, diversity in style of note takers, as well the normal ambiguity that characterises natural language, make the process and representation of clinical notes a complex task. Processing unstructured text is difficult [5]. It requires intense pre-processing as well as manual feature engineering and sometimes mapping to ontologies for semantic interpretation [7], [8]. Therefore, most clinical studies utilize structured health data and few target clinical notes. In addition, providing NLP experts access to clinical notes is often unfeasible given the challenges posed by the need for data protection.

Shahad Nagoor, Lucy Hederman, Kevin Koidl, are Annalina Caputo with Trinity College Dublin, School of Computer Science and Statistics, Ireland (e-mail: nagoors@tcd.ie, HEDERMAN@tcd.ie, KOIDLK@tcd.ie, annalina.caputo@dcu.ie).

II. Motivation

With the advancement of machine learning techniques and NLP capabilities, the analysis of clinical text is becoming more feasible. Existing neural language models are becoming reasonably capable of achieving multiple language tasks such as question answering, identifying entities, translation, and summarization. Moreover, modern NLP classifiers in the state of the art are pre-trained and released to public use but may require refinement to adjust to new tasks and new data. In fact, there has been a paradigm shift in machine/deep learning where the standard has become to transfer learning from existing pre-trained models to target tasks [9]. One example is the BERT model which was released in 2018 [10]. Successive versions of BERT have been proposed, each of which adds to its training and adapts its internal parameters by reinforcing its learning capabilities using datasets previously unseen by the model. However, there has been a lack of classifiers in the medical domain due to data protection acts and difficulty of accessing data for NLP experts. In response to this lack, two recent customizations of BERT, namely Bio_ClinicalBERT [11], and Bio_Discharge_Summary_BERT [11], were trained on clinical text extracted from a dataset named MIMIC [12] which allowed them to familiarize themselves with textual knowledge in clinical language and facilitate adaption for downstream language-based classification or prediction. Both clinical models are based on Bio_BERT and were trained and fine-tuned on multiple down stream tasks including clinical named entity recognition (NER) and clinical natural language inference (NLI). The study demonstrated improvement in the tasks using clinical-specific training in contrast to training on general or non-specific domains. Therefore, the goal of the study was to produce clinical-specific embeddings and clinical pre-trained models which are suitable for adaption and tuning at much lower training cost compared to starting

from scratch (training both models was computationally expensive consuming 17-18 days of computational runtime as well as major GPU, CPU, and memory resources) [11]. Bio_ClinicalBERT was trained on all notes in the data whereas Bio_Discharge_Summary_BERT was trained on discharge summaries. A notable limitation recognised by the authors of Bio_ClinicalBERT and Bio_Discharge_Summary_BERT study is that the data (MIMIC-III) [12] used to train these models are from a single hospital and healthcare practices significantly differ from one institution to another. Therefore, the authors emphasize the importance of using notes from other institutions in future work.

A. Redefining Sentiment Analysis for Clinical Context

Classification of clinical notes poses a computational challenge given the domain-specific nature of clinical text. A particular challenge in medical domain is assessing how positive or negative is the patient's status [13]. Assessing the polarity is known in NLP as sentiment analysis. However, in medical domain, generic approaches of sentiment analysis cannot function effectively as medical text is more objective, and composed of concrete language including clinical terms, jargon words, and abbreviations with very few sentimental, subjective words. In addition, assessing positivity or negativity in clinical narratives is highly dependent on medical events and outcomes of treatments. Therefore, there have been arguments for altering the interpretation of sentiment analysis for medical context and introducing the concept of clinical sentiment [13], [14]. Unlike regular sentiment analysis which targets sentimental words (lexicons) in text (e.g. happy, sad, joyful,...etc), clinical sentiment is driven by clinical events, interventions, treatments and outcomes. In particular, a positive note may reflect successful intervention, improvement or progress in patient health status, good impact of a medication, and also desirable outcomes such as discharge and survival. On the other hand, a negative note may reflect failed intervention, regress in patient health status, adverse medication reaction, and undesirable outcomes such as mortality and readmission.

B. The Challenge of Labeling in Clinical Sentiment

Another challenge to developing clinical sentiment models is the requirement for annotation and labeling of individual notes for input to the learning process. Models in supervised learning learn from samples presented in the form of {{sample1, label1}, {sample2, label2}, ...etc} and this requires intensive manual work by experts, such as physicians in the case of medical domain. In this study, we propose to label notes depending on two outcomes: mortality and survival. We assume that mortality reflects that a patient has been in deteriorating health condition and deterioration signs will be recorded in patient clinical notes and as such we label these notes as negative. On the other hand, we assume that survival reflects that a patient has been recovering and recovery signs will be recorded in patient clinical notes and as such we label these notes as positive. We aim to test if this is feasible considering that other recent research proved that sentiments analysis results on clinical notes for patients in ICU were correlated with their mortality and survival [15], [16]. Labeling is explained further in section *IV.C*. Some research in the literature tried to address the the problem of manual labeling by assigning pseudo-labels. Reference [9] used pseudo-labeling but their labeling strategy is to create a dataset for the target task with a distribution that resembles the distribution of the data previously used in pre-training the model. Although their method did not use direct labels for the samples, their results suggested that the use of pseudo-labels can be effective in clinical data.

C. No Standard Fine-tuning Strategy

There is no standard fine-tuning strategy; the process of tuning pre-trained models can be performed in multiple ways. Reference [17] experimented on multiple trials of fine-tuning for text classification and found that despite the effective performance of BERT in many NLP tasks, its full capabilities have been partially unexplored. There has been little research studying the performance of BERT on target tasks or investigating factors for improving performance on them, e.g. [18]. Therefore, fine-tuning methods are still subject to experiments, investigations, and development.

D. Transferring Existing Models to COVID-19 Data

Many studies presented in literature that attempt to apply NLP techniques on clinical text were mainly focusing on notes extracted from MIMIC dataset. "MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012" [12]. Because MIMIC-III data only go until 2012, it does not cover recent medical issues like coronavirus disease 2019 (COVID-19). It is worth mentioning here that both Bio_ClinicalBERT and Bio_Discharge_Summary_BERT were trained on MIMIC-III clinical notes and so their learnt vocabularies, and embedding did not include COVID related terms. Therefore, their performance on tasks may differ on COVID clinical notes since they may include previously unseen vocabularies and context. It is known that exposing the models to new and custom datasets can boost their performance in language tasks as it increases models' generalizability and predictability by educating the models for new writing styles, and new terminologies in the domain they are used for. Therefore fine-tuning these models on COVID datasets will potentially improve their performance on linguistic tasks applied to COVID related texts. We base our work on these models and raise the following questions:

- Can transfer learning of the pre-trained models cinincalBert and Bio_Discharge_Summary_BERT applied to doctors' clinical COVID notes predict clinical sentiment for COVID patients in ICU?
- Can data augmentation for the given dataset improve the predictability for these models?

III. Related Work

A. Sentiment Analysis of Clinical Text

NLP, and more specifically sentiment analysis techniques, have been applied to clinical text only in more recent years. According to [19], sentiment in clinical context has been limited. Some target patient reviews in social media and patients' suicide notes. Others study the correlation of sentiment scores in nursing notes and discharge summaries with patients mortality, and other studies focus on providing a descriptive comparison between nursing and radiology clinical narratives. The preservation of patients' privacy and confidentiality has limited the availability of open clinical text for training and test purposes, hampering the applicability of ML techniques to this domain. MIMIC-III dataset is an example of very few datasets that allow NLP researchers to have access to ICU clinical notes. Therefore, only a few recent studies [20], [15], and [16] have applied sentiment analysis on clinical notes; [20] studied the association of sentiment scores derived from nursing notes with the in-hospital 28-day mortality of sepsis patients, whereas [16] and [15] studied the same association for 30-day mortality.

Whereas many MIMIC-based studies used nursing notes, only a few used all notes, or the admission/discharge summaries, depending on the purpose of studies. In this study we aim to produce COVID_ICU_BERT to perform clinical sentiment analysis for COVID patients using clinical notes. Moreover, the existing methods use general language models which are not necessarily able to sustain their performance on clinical notes as they were trained on non-medical data. When describing a patient state, the physician may make use of language that bears some sentiment, or implicit polarity, reflection of the untold opinion of the doctor, something that if captured by an algorithm can be used to detect early signs of specific patient complications or disease. Bio_ClinicalBERT and Bio_Discharge_Summary_BERT are first attempts to address this problem, but they are not general enough being trained on only one dataset with a very specific time span, missing for example all the references that have been generated by COVID-19 pandemic, and being trained for specific target prediction tasks. Therefore, in this paper, we perform transfer learning on these models to specific medical domain for the purpose of clinical sentiment prediction.

B. Prediction in COVID-19

Several studies have focused on prediction tasks related to COVID-19, for example [21], [22] focused on predicting severity level of COVID-19. Reference [23] proposed a novel feature selection methodology for prognosis of COVID-19. Other studies focused on mortality prediction [24]–[28]. Reference [29] classifies WHO patients' reports, which include clinical notes, using four findings/labels including COVID-19. Despite the existence of many models for forecasting survival and mortality for COVID-19, the use of clinical notes in the mentioned prediction tasks for COVID-19 related research is limited.

IV. Data and Method

A. Data

The publicly available dataset used in this study is clinical notes attached and related to *chest X-ray* and *CT images* of patients infected with or suspected of COVID-19 or other pneumonia causes. It was mainly collected from publications, but some samples were collected independently from hospitals and physicians. Data include images, as well as other supplemental metadata. Details about the project and data collection can be found in the project repository [30], [31].

The original dataset consists of 950 samples. It has multiple fields which describe the images and patients' conditions. Clinical notes are provided, mostly in the form of summaries of patients' hospitalization and discharge condition. Each patient record is associated with death or survival. Other fields include incubation status, oxygen saturation and counts of leukocyte, neutrophil, lymphocyte. We split the data to form training set, validation set and test set. The test set is 20% and the validation set is 25% of the training set. The training samples extracted from the dataset are imbalanced; 71 are positive (correspond to patients who survived) and 21 are negative (correspond to patients who deceased). Imbalanced data classes can result in a biased model where prediction can be high for the majority class and poor for the minority class. Therefore, we apply data augmentation to even the samples distribution using contextual word embeddings for creating new samples with replaced synonyms. Data augmentation is explained in section *IV.D*.

B. Notes Pre-processing

The data is not limited to COVID-19 cases and contains other pathologies and pneumonia causes like SARS, ARDS, Influenza, and Lipoid. As we are interested only in COVID-19, the samples with non-COVID findings were removed, reducing the sample size to 584. There were no duplicate notes, though some notes reference the same patient and share the same introduction of patient history and status and differ only in the last lines indicating different time/day of examination. For example, four notes were found for patient id 178. The first two are reported in Fig. 2 in the Appendix.

Notes with no survival information were removed since this information is necessary for the annotation step as explained in Section IV *C*. The resulting sample size after this step is 155 notes. In addition to the above, and in order to set up the notes for the learning process, we removed deceased words like ("died", "terminal", "death", "passed away") from the notes as these provide direct and easy-to-predict outcomes for the model. Similarly, the phrases "full recovery" and "complete recovery" were removed. We applied a few preparation steps to convert the notes into the appropriate input for Bio_ClinicalBERT and Bio_Discharge_Summary_BERT models. The steps are as follows:

- Tokenization to split the text into subword units using BERT specific tokenizer provided by the transformers library. This step is important as the BERT model was trained on certain vocabulary sets and the tokenizer must identify and handle tokens previously unseen by the model.

- Indexing the tokenized vocabulary by token ids to look up the embedding corresponding to these words.
- Adding special tags at the beginning [CLS] and end [SEP] of sentences. CLS is placed at start of sentence and refers to classification as it will be later used in the classification step and SEP refers to sentence separator at sentence end.
- Finally, truncating and padding with zeros all sentences to a specific maximum length. An example of a note before and after applying these steps is shown in Fig. 3 in the Appendix.

C. Labeling

The prediction model is built using supervised learning which requires labeling of individual notes. As introduced earlier in Subection *B*, we propose a labeling scheme depending on survival and mortality information. We label notes for patients who died as *neg* and notes for patients who survived as *pos*. This labeling is based on the assumption that patients who survived have reflection of recovery, presumably recorded in their notes; and patients who died have some reflection of deterioration which contributed to their death and are presumably recorded in their medical notes. An example of a note for a patient who survived is shown in Fig. 4 while an example of a note for a patient who died is shown in Fig. 5 in the Appendix.

D. Method

In this study, two versions of BERT model [10] are used and modified to address the proposed purpose: 1. Bio_ClinicalBERT; 2. Bio_Discharge_Summary_BERT.

We use clinical notes from COVID-chest-xray dataset to perform fine-tuning on both Bio_ClinicalBERT and Bio_Discharge_Summary_BERT and monitor their performance on predicting patients' mortality and survival sentiment. Pre-trained models in all BERT family have embedding space which contain numerical representations (vectors) for every word presented to the model during pre-training. These are contextually dependent vectors (i.e., the vector for a single word e.g., the word *bank* in the sentence: I have a *bank* account is different than the vector for the same word in the sentence: I stand by the river *bank*). This is because these words bear different meaning depending on the context they present in.

Fine-tuning modifies the embedding space. Reference [18] studied how fine-tuning changes BERT and they conclude that in classification tasks, the model increases the distances between samples associated with different labels/categories. So, in the case of this study, the model should maximize the distance between clinical notes for people who survived vs. clinical notes for people who died.

With respect to the model's architecture, the original architecture of both models is retained. On top of the original architecture, we added an additional untrained dense layer of neurons to form the classifier and we train the new model. Given clinical notes as input, the clinical sentiment of the notes is predicted using the final output layer which takes the classification representation from the token [CLS] and calculates probabilities of classes as in:

$$p(c|h) = softmax(Wh) \tag{1}$$

where W is matrix of parameters for binary classification, h corresponds to the hidden state in the first token that has the features of the sequence in the pre-processed text to feed a classifier, and *softmax* is a function to a probability distribution of the possible classes. In fine-tuning, both W and other BERT parameters are adjusted by maximizing the log-probability of the correct class labels [17]. The fine-tuning of the model is done by the cross-entropy loss function. Fig. 1 shows an overview of the proposed method. We evaluate the models

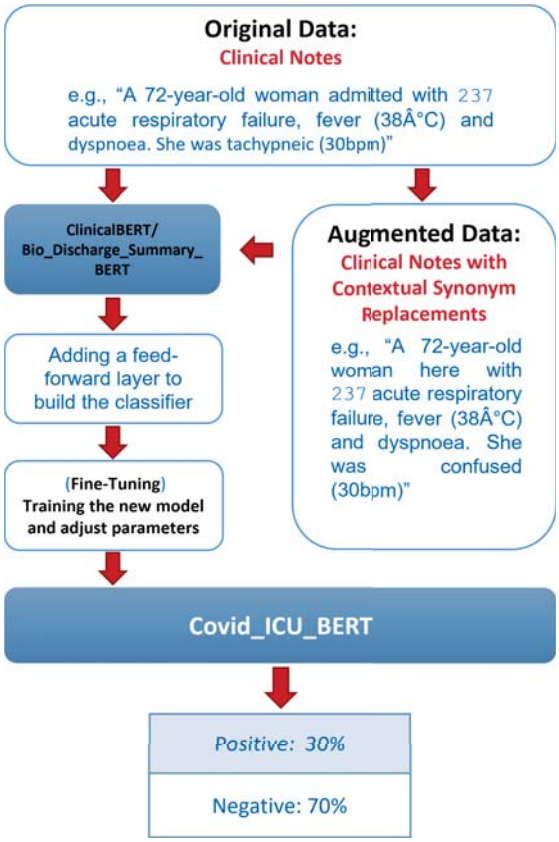

Fig. 1 A diagram for the proposed methodology (examples from [30], [31])

using Accuracy, Area Under The Curve (AUC), and Precision.

1) Fine-Tuning: We set random seeds to 42 prior to the fine-tuning process for reproducibility. We use Adam with decoupled weight decay (AdamW) as it has been reported to provide better performance than standard Adam according to [32]. We empirically use a learning rate of 1.e-4 (we attempt on smaller learning rates like 5.e-5 and 1.e-5 but those were less effective) and a batch size of 32 (due to no availability of GPUs). We compute the loss after each training epoch on the validation set, and the average loss over the entire training data at the end of training. The results are summarized in table 1 (upper section). The sample size for both models is 155 as indicated earlier. Both models are trained for 4 epochs.

2) Data Augmentation: The training samples extracted from the dataset are imbalanced; 71 are positive (correspond to patients who survived) and 21 are negative (correspond to

patients who deceased). We perform data augmentation using synonym replacement to balance the dataset and increase the number of negative samples. We applied contextual word embedding augmentation using *emilyalsentzer/Bio_ClinicalBERT* as this model is familiar with medical text unlike general English-based models like *bert-base-uncased*. Contextual augmentation, first proposed by Kobayashi [33], is a technique that utilizes a bidirectional method to read text and substitute words with their synonyms based on the given context of the sentence. The technique has shown improvements in text classification tasks in Kobayashi's experimental results. In our study, we choose to augment text in a proportion of 10% and performed the augmentation on the 50 negative samples. The results of training both models on the augmented data are summarized in Table I (lower section).

TABLE I
ACCURACY, AUC, AND PRECISION FOR TWO IMPLEMENTATIONS OF COVID_ICU_BERT: 1-FINE-TUNING BIO_CLINICALBERT AND 2-FINE-TUNING BIO_DISCHARGE_SUMMARY_BERT ON A TESTSET OF CLINICAL NOTES EXTRACTED FROM COVID CHESTXRAY DATASET

Model	AUC	Acc	Prec
Bio_ClinicalBERT	0.98	93.33%	0.96
Bio_Discharge_Summary_BERT	0.97	86.67%	0.95
Augmented Data			
Bio_ClinicalBERT	0.98	96.67%	0.92
Bio_Discharge_Summary_BERT	0.85	90.00%	0.96

E. Discussion and Limitations

Based on the results, the performance of both models on the given dataset is similar. The accuracy of the models is relatively high. This reflects that the models were to a reasonable extent able to distinguish a note for a patient who died from a note for a patient who survived. Some of these notes have repeated text, making the notes easier to classify. Another factor which might contribute to the accuracy is that the given clinical notes contain clear language and a complete summary of a patient trajectory in the hospital, which might facilitate the classification task for the models. Therefore, it will be worthwhile to examine the models' performance and do a second fine-tuning step on more fine-grained pieces of information broken down in multiple notes like daily progress notes. An example of a note that was not correctly classified in the test dataset is shown in Fig. 6 in the appendix. This note was misclassified possibly because it was relatively short and does not provide much descriptive information about the patient, such as vitals, blood pressure, or ventilation information, which was mentioned in others. The quality of learned representations of text depends on the text shown to the model. We expect that increasing the data sample and size will improve the representation and classification further. Since alarm fatigue is a known challenge in medicine, and high precision is required in support systems adopted in healthcare [6], we also consider comparing precision of both models. We evaluated our models using a confusion matrix to check their performance, especially on positive predictive values. Both models have high precision (around 0.96 for Bio_ClinicalBERT, and 0.95 for Bio_Discharge_Summary_BERT) and in the case of augmented data, the averages are 0.92, and 0.96 for Bio_ClinicalBERT and Bio_Discharge_Summary_BERT respectively.

Another limitation of this clinical sentiment analysis is that it is operating at the level of complete clinical notes rather than studying the internal aspects of smaller events in clinical notes; a note may include fluctuations in clinical polarity and this fluctuation might be missed by the proposed labeling scheme. We plan to address this by involving medical experts who can identify the detailed polarity inside the notes as described in future work in Section V.

V. FUTURE WORK

The dataset used in this study is formed as summarised information about patient hospitalisation; it did not include consistent timing information for all samples which would allow for time-dependent analysis. With a time series of clinical notes we could use a time-dependent labeling scheme. For example, notes in days/time close to death may contain more negative events compared to others in days away from mortality point. We plan to develop clinical sentiment models on daily progress notes from another dataset as a next step. We expect that predicting the sentiment may be more challenging on daily notes as these reflect daily events as opposed to summary notes which reflect the whole trajectory of a patient in ICU. The following points summarize the future plan:

- Collaborate with medical annotators to identify entities reflecting progress and entities of deterioration for specialised evaluation of the clinical sentiment. We expect that physicians label complete notes by choosing a label that reflects the patient status. e.g., label 1 is "stable", when a note indicates that the patient is OK or has made some progress and label 2 is "unstable", when a note indicates that the patient's condition has deteriorated.
- Try other augmentation techniques e.g. WordNet for vocabulary synonym replacement and compare the results with contextualized word embedding.
- Consider different fine-tuning strategies: further pre-train both models on masked language modeling task and the next sentence prediction tasks prior to fine-tuning, or apply multi-step fine-tuning. We may also consider changing truncation methods to set and observe the effect of different max sequence length for BERT.

VI. CONCLUSION

In this study, we investigated clinical sentiment of clinical notes for COVID patients in critical care. We apply transfer learning to produce a language model which perceives and distinguishes the linguistic features written in clinical notes for COVID patients and determine notes' polarity, where polarity is determined based on patient survival information. We hypothesized that clinical outcomes like survival or mortality can be useful to influence the clinical sentiment in ICU. We also applied contextual word embedding to augment the training data and balance the classes to avoid biased learning. Based on our results, clinical sentiment for COVID-19 clinical notes influenced by mortality and survival outcomes has the potential to recognize valuable polarity signals from clinical notes. This recognition will allow for more advanced clinical and favorable decision support in ICU like improving detection of early signs of deterioration or mortality.

APPENDIX
SAMPLE OF NOTES

Note 1: "A 72-year-old woman admitted with acute respiratory failure, fever (38Â°C) and dyspnoea. She was tachypneic (30bpm), with lymphopenia and low oxygen saturation (SpO2 85%, PAFI<250). Patient presented to the emergency department two days earlier with fever (up to 38.6Â°C), dry cough, odynophagia and general malaise. She was discharged from hospital because she did not present alarm criteria at that time. The patient required mechanical ventilation and was admitted to intensive care. During her stay in ICU, poor evolution to respiratory distress syndrome and to multi-organic failure. The patient passed away 24 hours later. Figure 1 (at admission). Ill-defined bilateral alveolar consolidation with peripheral distribution" [30], [31].

Note 2: "A 72-year-old woman admitted with acute respiratory failure, fever (38Â°C) and dyspnoea. She was tachypneic (30bpm), with lymphopenia and low oxygen saturation (SpO2 85%, PAFI<250). Patient presented to the emergency department two days earlier with fever (up to 38.6Â°C), dry cough, odynophagia and general malaise. She was discharged from hospital because she did not present alarm criteria at that time. The patient required mechanical ventilation and was admitted to intensive care. During her stay in ICU, poor evolution to respiratory distress syndrome and to multi-organic failure. The patient passed away 24 hours later. Figure 2 (4 hours later). Radiological worsening, with affectation of lower lobes. Endotracheal tube and central venous line were required" [30], [31].

Fig. 2 The first two notes (among four) for an individual patient

Original: "Extubated since the prior radiograph. Partial regression of the diffuse lungs changes, however air bronchograms are now evident in both upper lobes" [30], [31]. **Token IDs**: [101, 4252, 25098, 2913, 1290, 1103, 2988, 2070, 15241, 119, 7597, 1231, 24032, 1104, 1103, 4267, 3101, 5613, 8682, 2607, 117, 1649, 1586, 9304, 1320, 8401, 12139, 1116, 1132, 1208, 10238, 1107, 1241, 3105, 27645, 119, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Fig. 3 An example of a note before and after applying pre-processing steps for clinical BERT models

"84M with COPD, HFpEF, and BPH with recurrent UTIs. Was in his USOH and recovering from recent admission for UTI and subsequent stay in short-term rehab 2 weeks ago. Found unresponsive by wife at home with labored breathing. Afebrile, hypoxic and tachycardic in the field, arrived to ED on non-rebreather satting well but altered and in respiratory distress, intubated for airway protection. CBC/BMP completely unremarkable aside from anion gap of 19 and leukocytosis 16. CXR (image 1) showed some questionable linear opacities compared to recent prior, and there was concern for infection given #COVID-19 epidemic and respiratory distress. However further labs revealed POC trop 2.16, BNP 4K, VBG pH 7.37, lactate 4.3. EKG showed new RBBB and S1Q3T3 pattern (image 2, right).Echo (image 2, left) showed severely dilated RV, apical hypokinesis and McConnell's sign. PERT code was activated and pt was taken for stat CTA (image 3) revealing massive #Pulmonaryembolism (?saddle) but predominately occluding the entire R side. Systemic thrombolysis was considered, however, pt had known meningioma, which showed interval growth on stat CT head, raising concern for bleed. Within 4h of presentation, pt was taken to IR suite for embolectomy, which was successful at restoring flow to entire R lung field (image 4). Involvement of L pulmonary artery was non-occlusive and not intervened on. After the procedure, pt was taken for LE dopplers (image 5), which identified residual clot burden in the L profunda femoral vein. The following day, pt was extubated to room air and made a full recovery" [30], [31].

Fig. 4 An example of a note (prior to pre-processing) for a patient who survived

"79-year-old woman who presented with chest pain, cough, and fever for 3 days. Coronavirus disease (COVID-19) had recently been diagnosed in two of her household members. Patient developed acute respiratory distress syndrome within subsequent few days and died 11 days after admission. (Courtesy of Song F, Shanghai Public Health Clinical Center, Shanghai, China). show ground glass opacification (GGO) on day 1" [30], [31].

Fig. 5 An example of a note (prior to pre-processing) for a patient who deceased

"In Case 3, X-ray showed patchy high-density shadows in both lungs, which were more prominent in the lower lobes (C1) and worsened during the couple of days before (C2). Case 3 exhibited focal interstitial thickening (Fig. 3c)" [30], [31].

Fig. 6 An example of a note that was not correctly classified in the test dataset

REFERENCES

[1] I. P. Lynch, P. E. Roberts, J. R. Keebler, O. Guttman, and P. E. Greilich, "Error Detection and Reporting in the Intensive Care Unit: Progress, Barriers, and Future Direction," *Current Anesthesiology Reports*, vol. 7, no. 3, pp. 310–319, Sep. 2017. [Online]. Available: https://doi.org/10.1007/s40140-017-0228-3

[2] W. G. Johnson, T. A. Brennan, J. P. Newhouse, L. L. Leape, A. G. Lawthers, H. H. Hiatt, and P. C. Weiler, "The economic consequences of medical injuries. Implications for a no-fault insurance plan," *JAMA*, vol. 267, no. 18, pp. 2487–2492, May 1992.

[3] E. J. Thomas, D. M. Studdert, J. P. Newhouse, B. I. Zbar, K. M. Howard, E. J. Williams, and T. A. Brennan, "Costs of medical injuries in Utah and Colorado," *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, vol. 36, no. 3, pp. 255–264, 1999.

[4] L. L. Leape, T. A. Brennan, N. Laird, A. G. Lawthers, A. R. Localio, B. A. Barnes, L. Hebert, J. P. Newhouse, P. C. Weiler, and H. Hiatt, "The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II," *The New England Journal of Medicine*, vol. 324, no. 6, pp. 377–384, Feb. 1991.

[5] H.-J. Kong, "Managing Unstructured Big Data in Healthcare System," *Healthcare Informatics Research*, vol. 25, no. 1, pp. 1–2, Jan. 2019, publisher: Korean Society of Medical Informatics. [Online]. Available: http://e-hir.org/journal/view.php?id=10.4258/hir.2019.25.1.1

[6] K. Huang, A. Singh, S. Chen, E. Moseley, C.-Y. Deng, N. George, and C. Lindvall, "Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, 2020, pp. 94–100. [Online]. Available: https://www.aclweb.org/anthology/2020.clinicalnlp-1.11

[7] S. N. Kasthurirathne, B. E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin, and S. J. Grannis, "Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection," *Journal of Biomedical Informatics*, vol. 60, pp. 145–152, Apr. 2016.

[8] S. Nuthakki, S. Neela, J. W. Gichoya, and S. Purkayastha, "Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks," *arXiv:1912.12397 [cs]*, Dec. 2019, arXiv: 1912.12397. [Online]. Available: http://arxiv.org/abs/1912.12397

[9] Y. Wang, K. Verspoor, and T. Baldwin, "Learning from Unlabelled Data for Clinical Semantic Textual Similarity," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, 2020, pp. 227–233. [Online]. Available: https://www.aclweb.org/anthology/2020.clinicalnlp-1.25

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[11] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly Available Clinical BERT Embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909

[12] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/sdata201635

[13] Y. Deng, T. Declerck, P. Lendvai, and K. Denecke, "The Generation of a Corpus for Clinical Sentiment Analysis," in *The Semantic Web*, ser. Lecture Notes in Computer Science, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, and C. Lange, Eds. Cham: Springer International Publishing, 2016, pp. 311–324.

[14] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, May 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365715000299

[15] I. E. R. Waudby-Smith, N. Tran, J. A. Dubin, and J. Lee, "Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients," *PloS One*, vol. 13, no. 6, p. e0198687, 2018.

[16] Y. Zou, J. Wang, Z. Lei, Y. Zhang, and W. Wang, "Sentiment Analysis for Necessary Preview of 30-Day Mortality in Sepsis Patients and the Control Strategies," *Journal of Healthcare Engineering*, vol. 2021, p. 1713363, 2021.

[17] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" in *Chinese Computational Linguistics*, ser. Lecture Notes in Computer Science, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham: Springer International Publishing, 2019, pp. 194–206.

[18] Y. Zhou and V. Srikumar, "A Closer Look at How Fine-tuning Changes BERT," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 1046–1061. [Online]. Available: https://aclanthology.org/2022.acl-long.75

[19] G. E. Weissman, L. H. Ungar, M. O. Harhay, K. R. Courtright, and S. D. Halpern, "Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness," *Journal of Biomedical Informatics*, vol. 89, pp. 114–121, Jan. 2019.

[20] Q. Gao, D. Wang, P. Sun, X. Luan, and W. Wang, "Sentiment Analysis Based on the Nursing Notes on In-Hospital 28-Day Mortality of Sepsis Patients Utilizing the MIMIC-III Database," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 3440778, 2021.

[21] M. Abbaspour Onari, S. Yousefi, M. Rabieepour, A. Alizadeh, and M. Jahangoshai Rezaee, "A medical decision support system for predicting the severity level of COVID-19," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 2037–2051, Aug. 2021. [Online]. Available: https://link.springer.com/10.1007/s40747-021-00312-1

[22] M. Chieregato, F. Frangiamore, M. Morassi, C. Baresi, S. Nici, C. Bassetti, C. Bnà, and M. Galelli, "A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data," *Scientific Reports*, vol. 12, no. 1, p. 4329, Mar. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-07890-1

[23] O. Kocadagli, A. Baygul, N. Gokmen, S. Incir, and C. Aktan, "Clinical prognosis evaluation of COVID-19 patients: An interpretable hybrid machine learning approach," *Current Research in Translational Medicine*, vol. 70, no. 1, p. 103319, Jan. 2022.

[24] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, and Y. Yuan, "An interpretable mortality prediction model for COVID-19 patients," *Nature Machine Intelligence*, vol. 2, no. 5, pp. 283–288, May 2020, number: 5 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-020-0180-7

[25] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan, K. W. Johnson, M. Bicak, E. Klang, Y. J. Kwon, A. Costa, S. Zhao, R. Miotto, A. W. Charney, E. Böttinger, Z. A. Fayad, G. N. Nadkarni, F. Wang, and B. S. Glicksberg, "Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach," *JMIR Medical Informatics*, vol. 9, no. 1, p. e24207, Jan. 2021. [Online]. Available: http://medinform.jmir.org/2021/1/e24207/

[26] A. Karthikeyan, A. Garg, P. K. Vinod, and U. D. Priyakumar, "Machine Learning Based Clinical Decision Support System for Early COVID-19 Mortality Prediction," *Frontiers in Public Health*, vol. 9, p. 626697, 2021.

[27] J. Berenguer, A. M. Borobia, P. Ryan, J. Rodríguez-Baño, J. M. Bellón, I. Jarrín, J. Carratalà, J. Pachón, A. J. Carcas, M. Yllescas, and J. R. Arribas, "Development and validation of a prediction model for 30-day mortality in hospitalised patients with COVID-19: the COVID-19 SEIMC score," *Thorax*, vol. 76, no. 9, pp. 920–929, Sep. 2021, publisher: BMJ Publishing Group Ltd Section: Respiratory infection. [Online]. Available: https://thorax.bmj.com/content/76/9/920

[28] P. Schwab, A. Mehrjou, S. Parbhoo, L. A. Celi, J. Hetzel, M. Hofer, B. Schölkopf, and S. Bauer, "Real-time prediction of COVID-19 related mortality using electronic health records," *Nature Communications*, vol. 12, no. 1, p. 1058, Feb. 2021.

[29] A. M. U. D. Khanday, S. T. Rabani, Q. Khan, N. Rouf, and M. M. U. Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *International journal of information technology : an official journal of Bharati Vidyapeeth's Institute of Computer Applications and Management*, 2020.

[30] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[31] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image

data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[32] I. Loshchilov and F. Hutter, "DECOUPLED WEIGHT DECAY REGULARIZATION," *The International Conference on Learning Representations, ICLR 2019*, p. 18, 2019.

[33] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," May 2018, arXiv:1805.06201 [cs]. [Online]. Available: http://arxiv.org/abs/1805.06201