Comparison of Neural Language Modeling Pipelines for Outcome Prediction From Unstructured Medical Text Notes

CHERUBIN MUGISHA®, (Member, IEEE), AND INCHEON PAIK®, (Senior Member, IEEE)

School of Computer Science and Engineering, The University of Aizu Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima 965-8580, Japan
Corresponding author: Incheon Paik (paikic@u-aizu.ac.jp)

ABSTRACT Machine learning techniques and algorithm-based approaches are becoming more and more vital to support clinical decision-making. In the medical area, natural language processing (NLP) techniques have shown the ability to extract useful information from electronic health records. On the one hand, statistic, semantic, and contextualized word embedding-based models and on the other hand preprocessing approaches are the keys to a better representation of a document. Using narratives from the Intensive Care Unit, we elaborated a comparison of the most used methods and preprocessing approaches to tackle an outcome prediction problem and guide researchers into NLP pipelines in the medical area. We used real data from Medical Information Mart for Intensive Care-III (MIMIC-III). We selected all notes related to patients with pneumonia. We conducted a deep analysis on text preprocessing tasks producing three datasets: raw data with minor preprocessing, meticulous preprocessing, and extreme preprocessing filtering only medical-related terminologies using Named Entity Recognition algorithms. We then used these three sets in five models, of which two are based on the traditional noncontextual word embedding techniques and three use contextualized word embedding based on a transformer. We demonstrated that transformer-based models outperform other word embedding models and a profound preprocessing yielded an accuracy of 98.2 F1-score. These results show the highly competitive ability of NLP predictive models against other models that use medical data. With an appropriate NLP pipeline, the information contained in medical narratives can be used to draw up a patient profile, and admission notes can help to ascertain a mortality risk of a patient admitted to the Intensive Care Unit.

INDEX TERMS BERT, medical notes, NER, NLP, outcome prediction, pre-trained language models, text preprocessing, word embedding.

I. INTRODUCTION

Pneumonia is an infectious disease of the lungs affecting alveoli and caused by bacteria, fungi, or viruses. Pneumonia can range in seriousness from mild to life-threatening. It remains the commonest infective reason for admission to intensive care as well as being the most common secondary infection acquired while in the Intensive Care Unit (ICU) [1], [2].

Electronic Health Records (EHRs) are health-related information on an individual created in a health care organization. EHR systems contain structured data such as demographics, vital signs, laboratory test results, medications, and procedures. They also have unstructured medical or nonmedical data in a free format such as imaging reports or care-provider notes [3]. In medical assessment, it is common and practical to use all types of data to understand the status of a patient or to predict his outcome. However, for caregivers, medical notes are of paramount importance.

Within a hospitalization or a clinical visit, a patient might have several note documents which can constitute a rich and long clinical history. Clinical notes provide a deep understanding of a patient’s illness because they describe symptoms, clinical history, reasons for admission, and details of any intervention made by a multidisciplinary team [4]. With the medical texts representing 80% of the EHR data [5], admission constitute an extensive informative source used by doctors to draw a patient’s profile within the first 24 hours of admission. It is then crucial to be able to use his history and admission description to predict what is likely to happen during his stay.
In recent years, machine learning algorithms have been increasingly used to predict the outcome by using structured or unstructured medical data. However, using free-text notes to achieve such tasks may encounter a lot of challenges. Even though there are standards [6] when taking medical notes, most of the texts are biased by internal and conventional writing methods that make generalization harder for resulting models in a different environment. In addition, medical text notes can be too long to be handled by conventional natural language processing (NLP) models. Consequently, achieving good results requires a better algorithm that preprocesses the data and models the determinants of health condition for an overall understanding of a patient’s status.

Although there are different studies on medical text classification, not many have demonstrated a clear statistical comparison of NLP pipelines to guide researchers on the selection of methods ensuring the best results.

Our main insight was that admission narrative notes have the potential to predict outcome only if, in the NLP pipeline, we can find the best combination of preprocessing methods, document representation, and learning models.

The question of this research is what combination of NLP models and preprocessing methods are appropriate to unlock the information from medical narratives. The present study aims to use medical pneumonia patients notes, written by a multidisciplinary team of care-providers, to investigate among the dynamic word embeddings and static models to assess and compare their performance on the outcome prediction of an ICU hospitalization. We evaluate the resulting models using admission notes taken within 24 hours.

II. RELATED WORK
A. TRADITION LINEAR MODELS
Prediction of prognosis to inform decision-making in the ICU has a long history. Traditionally, statistical methods were widely used to evaluate the survival rate of a patient using domain experts features. Linear models such as Kaplan–Meier (KM) estimator were the most popular combination with Cox proportional hazards regression to handle regression problems [7]. Using advanced methods based on logistic regression, Simplified Acute Physiology Score (SAPS) and the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) demonstrated a clear improvement in assessing the disease severity of a hospitalized patient. However, those models use predetermined features, which greatly limits their applicability in a real situation.

In recent years, those traditional methods have been surpassed by more modern and accurate algorithms mainly using data-driven models based on machine learning architectures. Since the linguistic string project [8] in analyzing clinical documents, most of the work has been conducted around general medical management, treatment, test and results, patient state, and patient behavior using medical text.

However, it has been more challenging to demonstrate the real usability of nonmedical data. The nature of medical text data requires a combination of steps to unlock the information embedded in a clinical text (e.g., disease, treatment, patient status) by transforming the text into structured medical data. The automation of this process and the efficiency of models to perform tasks such as clinical text classification has been investigated [4].

B. ML MODELS AND TEXT PREPROCESSING
Currently, ML techniques have shown a major improvement for prediction in the general domain and particularly in the medical domain. They can perform better using either structured and unstructured data or even both through an ensemble of machine learning processes [3], [9]. An NLP task starts with a preprocessing stage, to extract useful information and structure the raw text into a format that can make use of automated computing power.

A review [10] conducted on 67 publications from 2000 to 2015 has shown that extracting information from the EHR narratives can improve case detection for classification tasks. While this process can use different techniques of information extraction, 67% of the studies incorporated rule-based, 24% used keywords, and only 9% include machine learning in their approach. However, this trend has changed and a recent review [11] shows that machine learning-based methods are more used. The medical data transformation stage is challenging for messy medical notes because preprocessing can clean out significant information that is clinically important to predict the outcome accurately [12], [13].

To process medical text, Authors have been using the Unified Medical Language System (UMLS) in order to reduce ambiguity from abbreviations and conventional annotations. However, according to Liu et al. [14], 31% of UMLS abbreviations have multiple meanings. This can be resolved by computing the proximity of the abbreviation to its expanded form and replacing it with the most suitable term. This abbreviation disambiguation pipeline has brought a lot of controversy among the community and led to the creation of several resources for different data.

In NLP, deep learning methods such as Long Short-Term Memory (LSTM) and its variants use a preprocessing pipeline that includes a filtering process based on predefined controlled vocabulary terms before transforming data into training vectors [15], [16]. In this recent study [17], the authors propose an online medical pre-diagnosis support in which semantic and sequential features are extracted from a patient’s inputs using a CNN-RNN-based architecture model to predict a diagnosis.

Geraci et al. proposed a neural network to extract phenotype information from electronic medical record (EMR) text notes [18]. Their goal was to identify suitable candidates for medical research using doctors’ narratives within a supervised learning process. They extract useful information through a Document Term Matrix (DTM) using the TF-IDF algorithm. Their results show that NLP can help to identify criteria for models to perform better for a task such as classification. Wang et al. [19] illustrated a paradigm of clinical text
classification using deep representation and weak supervision. Their work demonstrated that it could be possible to use a deep neural network like CNN and outperform traditional NLP rule-based algorithms. Their approach also compared the importance of using word embeddings over count vector algorithms such as TF-IDF. However, their method has limitations because the model is trained from scratch; it requires a lot of training data. The input size is also dictated by the word embedding methods, which are usually not suitable for long text like medical narratives.

Authors have tried to tackle this problem in recent literature by using more sophisticated NLP models. For example, model like BERT (Bidirectional Encoder Representations from Transformers) [20] have shown impressive results from an architecture of a multilayer encoder models, to learn words and documents representation more deeply. In the medical area, researchers have been trying to leverage the knowledge from general documents to pretrained specific-domain models for higher performance for medical-related tasks [21], [22].

Despite that evolution in NLP, we still have many publications that use either advanced machine learning models or archaic models. Although most of these studies claim to have achieved the best scores in various tasks by using very different approaches, we can only wonder if there is no room for improvement. To our knowledge, no other research trying to elucidate a fair comparison of those methods has been published.

III. MATERIAL AND METHODS

Prediction task has been a great topic for academic research. Finding a correlation between the massive amount of clinical data and the outcome has the potential to understand life better and the factors involved in its end. In the ICU, time could be the determinant, and to obtain more valuable and easily interpretable information, medical notes are a good alternative to identifying key problems of a patient when other sources are not available.

Text narratives contain a concise description of a patient that can inform a caregiver about the status of the patient as soon as he is admitted to the ICU. Nonetheless, those narratives could incorporate more information than necessary, such as duplications of the structured data or repetition from different contributors, making it hard to be modeled for prognosis or prediction.

Machine learning and NLP have shown an incredible ability to learn from data; no matter how messy they are, there is always a potential to obtain an output from a model. However, building up a consistent and valuable model requires a deep understanding of the data to know what preprocessing steps are needed and what model architecture is more suitable for those data.

A. STUDY WORKFLOW DIAGRAM

This research was conducted in several steps consisting of three main processes illustrated in Fig. 1. A summarized workflow for the study is as follows:

1) Step 1: Data sampling and selection from a large dataset of MIMIC notes.
2) Step 2: Data preprocessing through a light(A), thorough(B), and extreme(C) cleaning performed by extracting medical entities, producing three separate datasets.
3) Step 3: Use of various embeddings to train different NLP classifier models.

Detailed descriptions of each of these steps are provided in the following sections.

B. DATA DESCRIPTION

To initiate this research, we need to use real medical text to challenge our approach and models to real-life data. We extracted our narratives from the Medical Information Mart for Intensive Care-III (MIMIC-III) [23] data using SQL queries. MIMIC is a publicly available multiparameter monitoring system used in the ICU for 11 years. It contains structured information such as physiological medical data and unstructured data such as text notes taken by different healthcare actors. The overall mortality of patients in the MIMIC-III database is 23.2%. To narrow our scope of research, we had to find a disease with enough data and a higher mortality rate for balanced learning for our models. Therefore, using only the admission diagnostic instead of ICD9 codes, “Pneumonia” as the main disease comes first with 2059 cases and a mortality rate of 29%. These patients have a total of 85085 notes taken by physicians, nurses, radiologists, and nutritionists, making an average of 41 notes per admission. Our cohort of patients comprised only adults, all over 15 years old.

We associated with each sequence of notes a label from the patient outcome. This binary sequence labeling considered all notes for discharged patients as a negative class represented by “0” and for those who deceased in hospital as a positive class represented by “1”. In addition, in case a patient has multiple admissions, they were considered as new. We, therefore, included the new admission data and used admission ID instead of patient ID to corroborate our queries. All these exploratory data analyses were done using python libraries and their statistics are reported in Table 1, and more details on the data sets sampling are given in section IV-D.

C. DATA CLEANING

Data cleaning refers to steps that we took to standardize our data and to remove text and characters that are not relevant to be left with a clean text dataset that is ready to be analyzed.
Authors have suggested many methods of text data cleaning. Most of the NLP cleaning tasks are based on basic rules such as converting text to lower case, regular expression and word replacement, punctuation, and nonalphanumeric character removal. Advanced preprocessing will include more tasks such as stop-words and tokenization, stemming and lemmatization, word tagging, or Named Entity Recognition (NER). All of these will depend on the data, whether it has a dictionary or not, or simply the NLP task we want to perform. For data privacy, an illustration sample on the cleaning results is provided as a reference in Fig. 2.

To create our three datasets and analyze in detail the impact of each cleaning approach, especially for medical notes, we proceed as follows:

1) MINOR CLEANING
This cleaning task follows the basic rules of the NLP cleaning task utilizing the natural language toolkit (NLTK). For case sensitivity, we converted all text into lower case and used regular expressions to remove punctuation extra white spaces, line breaks, and nonregular expressions. In addition, We utilized the Stop-words dictionary to filter out irrelevant entities.

2) THOROUGH CLEANING
To take our cleaning process even further, and harmonize clinical abbreviations and acronyms, we manually built up a matching dictionary of 80 terms. Using the UMLS [24], [25] with its metathesaurus inventory, we selected the top used acronyms in medical notes presented in these studies [26], [27] and added predominant risk factors for pneumonia such as acute respiratory distress syndrome (ARDS) and acute respiratory failure (ARF) [28], [29]. We also removed de-identification characters and harmonized typos and conventional spellings (e.g., pt, dr, W/O). These two cleansing processes are suitable for the emergent bidirectional models because their tokenization uses a word-piece technique and does not need a deep cleaning.
3) NAMED ENTITY RECOGNITION

Narratives can be very long and full of information that it could be necessary to filter out domain-related data. NER has shown the ability to process data semantically by identifying and categorizing key information (entities) in text [30]–[32]. Traditionally, dictionary NER-based models have been used for text data mining, and recently, deep learning-based models have shown outstanding progress leveraging pretrained language models. For our case, we addressed this step as sentence-level biomedical information extraction tasks. The biomedical language representation model for biomedical text mining (BioBERT) [21] is a domain-specific language model that has been trained on medical text data. BioBERT NER (BERN) [33] is one of its modules for recognizing biomedical entities and discovering new entities. We used BERN to extract entities related to disease, drugs/chemicals, genes/proteins, and species. However, the resulting entities are independent of each other, so they can only be used by nonsequential models and their association would be considered as correlated features.

IV. MODELS

In this study, we used several methods, from the traditional to the recent NLP models. NLP in the medical area has been using count-vector-based models, word-embedding-based models, and transformers-based models. To make a fair comparison, we conducted this study using all of them as we prepared data accordingly.

A. FEATURE EXTRACTION

To convert the text into a numeric format, comprehensible by computers, the narratives need to be encoded. We used various encoding techniques such as:

1) BAG OF WORDS (BOW)

BOW is a statistical representation of words and sentences and their compositionality [34]. Boosted by the success of text classification [34], [35], BOW has become one of the most used methods to classify text and documents using keywords. Although this method can help to classify text, it represents words in a singular dimension vector. It does not carry any semantic or syntactic meanings. To use these methods, we hypothesized that discriminative terms could be highlighted by a term frequency counter. Counter Vectorizer is a low-level one-hot encoder that transforms a given text into a vector based on the frequency of each term that occurs in the entire document. This representation can be efficient if each class has particular discerning words. Term Frequency Inverse Document Frequency (TFIDF) measures the relevancy of a given term by multiplying its frequency by the logged inverse document frequency of that term across the entire corpus.

\[ x_{ij} = TF_{ij} \times \log(J/|TF_{ij}|) \] (1)

where \( TF_{ij} \) is the term frequency and \( J \) the number of documents in the corpus. TFIDF is more efficient since it normalizes the count by scaling up rare terms and diminishes the weight of frequent words like "patient" in our case. We utilized Count vectorizer and TFIDF encoder for dataset where the order of words are ignored or broken (NER) and we additionally varied the vocabulary size between 1000 and 5000 words.

2) GloVe

To take advantage of the multidimensions of text data, we must use a model that vectorizes the text from a large number of precise syntactic and semantic word relationships. Global Vectors for Word Representation(GloVe) [36] represents words as real-valued vectors in a vector space of relatively low dimensions compared with its vocabulary size. This means that words will be close in a vector space only if their semantic and syntactic meanings are also relatively close and vice versa. In contrast with precedent embeddings such as Word to Vector(Word2Vec) [37], frequency of co-occurrences within context windows is vital as semantic information and should be carried on. In our case, we assumed that frequent co-occurrence words determine the outcome; besides, global corpus statistics are already incorporated in the embeddings. We downloaded the pre-trained word vectors and we used it as our word embedding to train a sequence model.

3) BERT EMBEDDINGS

BERT is a contextualized word vector representation. BERT embeddings create different vectors for a word used in different contexts. It utilizes the transformer encoder to represent a word in a higher-dimensional space, capturing relations between distant words more efficiently than traditional bidirectional encoders. Using a vocabulary size of more than 30000 tokens, BERT can encode any words or subwords using its position in the input sequence. A text representation by BERT prepends to each sequence a [CLS] token that can be used for a classification task [20]. Utilizing subwords has an advantage especially in biomedical text because we can encode more properly uncommon medical terminologies.
B. LEARNING MODELS

1) RECURRENT NETWORKS

In deep learning, a recurrent neural network (RNN) is a traditional reference for time series data such as sound and monitoring data in medical scenarios [38]. Conventional RNNs are slow and for long sequences, backpropagation in time tends to either vanish or explode [39]. Variants of RNN such as LSTMs have been developed to overcome RNN issues for long sequential inputs such as text but also to learn a bidirectional dependency via an attention mechanism using BiLSTM [40]. These models use an encoder part for a classification task where several recurrent cells handle each element as an input vector and propagate it forward.

\[ h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \]  

(2)

Hidden states \( h_t \) are computed by applying some weights \( w^{(hh)} \) on the previous input vector \( x_t \) where \( i \) is the order of the input words. For the output, a decoder part will calculate a vector where each value will represent a probability score for each class.

\[ y_t = \text{SoftMax}(W^y h_t) \]  

(3)

Here, \( h_t \) represents the output of the encoder for an input \( i \) and \( W^y \) is the respective weight applied by the decoder before feeding to a SoftMax function [41].

2) TRANSFORMER

RNNs and LSTMs are slow because data need to be passed sequentially. Transformer-based models can use the advancement of the computation technology by parallelizing the process and learning faster. Leveraging the knowledge from models like BERT, authors proposed multiple variants of language models dedicated to medical text. Among them, for this study, we specifically utilized ClinicalBERT and BioBERT. Even though these models are domain-specific based models for biomedical, they were pretrained on different data. BioBERT initialized its weights from BERT and used PubMed abstracts and Central full-text articles while ClinicalBERT leveraged BioBERT weights and pretrained using MIMIC medical notes. To use these models, we leveraged their weights and fine-tuned each to predict an outcome through a classification. As for a classification task, even though the [CLS] token can be used alone for a classification, the authors of [20] recommend trying different approaches. For our case, we averaged the four last layers by vector-wise summation to obtain a sentence embedding vector as an input to train a logistic regression classifier computing a binary probability.

\[ P(\text{outcome} = 1|h_n) = \text{ArgMax}(W h_n), \]  

(4)

where \( h_n \) is the averaged output of \( n \) hidden layers and \( W \) is the parameter matrix of the classifier.

C. EXPERIMENT DESIGN

Using free-text narratives from the MIMIC-III database, we propose a model that predicts a binary outcome utilizing the NLP process from the cleaning stage to the prediction. As shown in Fig. 1, our experiment was conducted in three main steps. The first is data gathering and selection. As we described, only patients with pneumonia as the primary disease were chosen for our experiment.

The second step is related to data preprocessing. To prepare data for models and improve its quality, this study proposes three cleaning methods that lead to three different datasets. We generated sets with minor cleaning, thorough cleaning, and medical entities extraction (NER). For simplicity, we will call these sets A, B, and C.

The third step is about optimizing the NLP machine learning models. This research demonstrates the performances of different machine learning algorithms to use static and contextualized word embeddings.

1) STATIC WORD EMBEDDINGS

They map each word to a single vector. Moreover, these vectors are dense and have much lower dimensionality than the size of the vocabulary. For this reason, we utilize two different vectorizations; a very simple document vectorization using count vectorization and TF-IDF. Such models ignore the meaning and context of a word in a document, for example, the word “pneumonia” will have the same value as “cancer” as long as they have the same number of occurrences in the document. These methods should then be applied to a bag of words that do not have any sequential relationship such as entities extracted by NER methods. The idea is to use medical terms with low frequency by n-gram (\( n = 1 \)) vectors to achieve better discrimination for our classification. We passed the resulting vectors to a logistic regression model for a classification. We set the maximum number of iterations to 5000, the solver to liblinear and kept other parameters to their default values. As for the more advanced static word embedding method, we conducted this study with Global vectors for word representation (GloVe) [36] that has been pretrained on a Wikipedia dataset and has approximately 6000 words represented each by a vector size of 300. These pretrained word embeddings vectors include the sequential dimension, which can be well learned by sequential models such as LSTM and BiLSTM.

2) DYNAMIC WORD EMBEDDINGS

To analyze the importance of contextualized word embeddings for medical free text, we relied on 12 layers of language representation models (BERT, BioBERT, and ClinicalBERT). To handle long narratives, we shrunk them into small sentences of 380 words, to leave some room for the tokenization process that will output 512 tokens.

D. EXPERIMENT SETTING

To evaluate the effectiveness of using text notes to predict outcomes, we hypothesized that we should make predictions...
TABLE 2. Evaluation and comparison of BOW-based models. A logistic regression classifier used a unigram representation from count-vectorizer and TF-IDF of 1000 and 5000 vocabulary size, respectively.

|                | Accuracy | Precision | Recall | F1-Score |
|----------------|----------|-----------|--------|----------|
| **B dataset**  |          |           |        |          |
| CountVec-1000  | 0.51     | 0.743     | 0.532  | 0.621    |
| CountVec-5000  | 0.59     | 0.798     | 0.63   | 0.71     |
| TF-IDF-1000    | 0.657    | 0.761     | 0.804  | 0.789    |
| TF-IDF-5000    | 0.778    | 0.784     | 0.928  | 0.852    |
| **C dataset**  |          |           |        |          |
| CountVec-1000  | 0.521    | 0.777     | 0.567  | 0.656    |
| CountVec-5000  | 0.62     | 0.837     | 0.656  | 0.736    |
| TF-IDF-1000    | 0.697    | 0.802     | 0.827  | 0.815    |
| TF-IDF-5000    | 0.801    | 0.805     | 0.993  | 0.889    |

as soon as a patient is admitted to the ICU. We then decided to sample our test set by utilizing only admission notes. Within the database, no tags were available to determine admission narratives among others. Using SQL queries, we filtered from the database on this criterion: an admission note = Unique per admission ID & taken by nurse & the first taken within 24 hours. This was sampled as a test set and we divided the rest into 90% for training and 10% for the validation set in order to have a separate validation set. The training and validation sets comprised progress, nursing, and procedure notes. The contextualized encoding has a limitation in terms of sentence lengths. However, medical narratives are usually too long, without any indication of which part contains the most useful information. This constrained us to truncate each long note to a size of 380 words to be used in all embeddings. That transformation changed our dataset from 85,085 long notes to 1,101,524 notes of a maximum size of 380 words. On average, each note produced almost 12 small consecutive notes, which we labeled as their original narratives. The distribution of our datasets will be described later in the results (3). Although we trained these notes separately, we averaged the predicted classes to calculate the loss.

V. RESULTS

For a fair comparison, given the low mortality rate in the data (29%), we calculated the accuracy in terms of sensitivity and specificity respectively by recall and precision metrics. As an overall evaluation metric, we reported the F1-scores and balanced accuracy \( \frac{1}{2} \left( \frac{TP}{P} \times \frac{TN}{N} \right) \) as well as Matthews correlation coefficient (MCC) within the Table 3. In contrast, we evaluated each model with different datasets resulting from our cleaning methods. This evaluation was conducted on the aforementioned test set made from admission notes. The scores demonstrate how well each model with a particular preprocessing can predict the outcome using the information described by admission notes.

From the very basic BOW, Table 2 shows that vectorization from NER leads to better results than a thorough cleaning process by all metrics. Between Count-vectorizer and TF-IDF, the latter performs better with accuracy, recall, and F1-score of 0.801, 0.993, and 0.889, respectively.

Table 3 shows a deep comparison of the static and contextualized embeddings as well as the accuracy of models to predict the outcome. For LSTM and BiLSTM, we performed a k = 10-fold cross-validation, using the training set and we trained both models for 10 epochs. Contextualized embeddings demonstrated a higher ability to understand the medical narratives than the static embeddings with a difference of 6% between their respective best scores. For static embedding, BiLSTM shows a better F1-score of 92.01% using the B dataset, however, using independent entities from the C dataset, BOW outperforms LSTM with an F1-score of 88.9% from TF-IDF against 54.6% from LSTM.

A. PERFORMANCE ON THE BEST-PERFORMING MODELS

Fig. 3 reports values obtained from the training of contextualized word representation using different word-piece tokenizations and embeddings. Extensive training of 50 epochs on the B dataset shows that BioBERT and ClinicalBERT have a more stable logarithmic training loss curve while BERT needs more training epochs. This is also illustrated by the validation (Fig. 3b) and the test (Fig. 3c), where BioBERT and ClinicalBERT performed similarly but BERT needed more than 20 epochs to gain stability.

However, even though BERT was pretrained on the general corpus, they all showed good discriminatory power as described in Table 3 with F1-scores of 98.2%, 97.4%, and 98.2%. Although these scores seem to be close, the precision scores of 98.1%, 96.7%, and 97.4% show a clearly inferior ability for BioBERT to handle unbalanced data. The best-performing model made use of the B dataset demonstrating the importance of a deeper cleansing process before training a contextualized model. Improvements of 9.79%, 7.7%, and 4.3% for MCC scores were observed among BERT, BioBERT, and ClinicalBERT, respectively.

B. ADDITIONAL ANALYSIS

Word and document embedding are starting points to represent any knowledge behind the input text. To understand how each model and embedding have a different interpretation of the medical notes, we tried to elucidate that contrast with vector similarity. Once we have vectorized our narratives, statistical similarity methods can be used. However, on one hand, multidimensional vectorization such as BERT cannot be properly reshaped into two dimensions without losing their important information. On the other hand, the lack of semantic and contextual information for BOW-based models constitutes a handicap to initiate any clustering behavior from the beginning of the NLP pipeline.

As shown in Fig. 4, using cosine similarity distance, it was clear that there is no evidence of cluster between features representing the two classes. With random samples, we ordered six positive and six negative notes and we tested the effect of using cosine similarity to look for any similarity between our two classes. None of the noncontextualized models demonstrated such ability. However, using Uniform Manifold Approximation and Projection (UMAP), we reduced the dimension of contextualized embedding on 100 notes for each
TABLE 3. Results from global vectors and BERT-based vectorization.

| Models | Data size | Training size | Test size | Tokenization | Embeddings | MCC | Acc. | Precision | Recall | F1-score |
|--------|-----------|---------------|-----------|--------------|------------|-----|------|-----------|--------|----------|
| LSTM   | A 1101524 | 2239          | GloVe 300 | GloVe 300    | -          | 89.04 | 80.7 | 58.2      | 70.1   |
|        | B 1101524 | 2239          | GloVe 300 | GloVe 300    | -          | 92.5  | 89.3 | 67.7      | 77.0   |
|        | C 1101524 | 2239          | GloVe 300 | GloVe 300    | -          | 84.8  | 52.7 | 40.5      | 54.6   |
| BiLSTM | A 1101524 | 2239          | GloVe 300 | GloVe 300    | -          | 93.2  | 81.4 | 80.6      | 81.3   |
|        | B 1101524 | 2239          | GloVe 300 | GloVe 300    | -          | 97.05 | 93.14| 90.91     | 92.01  |
| BERT   | A 1101524 | 2239          | BERT Wordpiece | BERT       | -          | 88.01 | 96.07| 90.6      | 90.6   |
|        | B 1101524 | 2239          | BERT Wordpiece | BERT       | -          | 97.8  | 99.3 | 98.3      | 98.2   |
| BioBERT| A 1101524 | 2239          | BioBERT Wordpiece | BioBERT   | -          | 89.4  | 99   | 86.2      | 91.7   |
|        | B 1101524 | 2239          | BioBERT Wordpiece | BioBERT   | -          | 97.1  | 99.1 | 96.7      | 97.4   |
| ClinicalBERT | A 1101524 | 2239          | BioBERT Wordpiece | ClinicalBERT | - | 93.5  | 96.9 | 96.2      | 92.7   |
|        | B 1101524 | 2239          | BioBERT Wordpiece | ClinicalBERT | - | 97.8  | 99.3 | 97.4      | 98.2   |

FIGURE 3. BERT-based models training. Each model was trained for 50 epochs using a validation set of 20% to control the over-fitting, this shows a comparison of the training, validation, and accuracy of BERT, BioBERT, and ClinicalBERT.

(a) Training (b) Validation (c) Val. accuracy

FIGURE 4. Cosine similarity for noncontextualized models. This shows a sample of 12 inputs, where the first 6 are notes of discharged patients and others are for deceased patients.

(a) Count-vectorizer cosine similarity (b) TF-IDF cosine similarity (c) GloVe 300 cosine similarity

VI. DISCUSSION

With this research, pneumonia patients were selected among ICU EHR data. Through the NLP pipeline, we aim to demonstrate the ability to predict outcomes using the narratives taken during a patient’s stay by comparing the existing NLP approaches. Cleaning thoroughly improves the contextual understanding of the inputs as demonstrated by ClinicalBERT with an MCC score of 4.3%.

BERT-based domain-specific models can perform slightly better than the general BERT but the difference resides mainly on the convergence time between the training and the validation process. In case we want to use independent entities, such as NER, BOW models are more suitable to handle prediction tasks because only term frequency over inverse document frequency is more relevant. Nonetheless, the preprocessing methods may change with a different EHR, that utilizes a different notation. Besides the performance of a prediction model, it should be interpretable. For our case, the accuracy of the prediction should be quantified by diagnosis, drugs, bioinformation, or even demographics. However, the high dimensionality of modern NLP models contains abstract class and the Fig. 5 shows clear clusters of the two classes before the learning and classification processes.

To analyze the certainty of advanced models through the prediction probabilities, we utilized the same random 200 samples. We extracted the logits from the last layer of each model, before being transferred to the activation function. Fig. 6 shows a distribution set of unnormalized scores $\theta$ from the three models, and each class is represented by 100 consecutive samples. It demonstrates that BioBERT and ClinicalBERT have a constant better distance score between the two classes than BERT. This means that if a score is close to 0, the probability for that sequence to fall under a certain class is around 0.5 which can be translated as a low confidence score for that prediction.

VI. DISCUSSION

With this research, pneumonia patients were selected among ICU EHR data. Through the NLP pipeline, we aim to demonstrate the ability to predict outcomes using the narratives taken during a patient’s stay by comparing the existing NLP approaches. Cleaning thoroughly improves the contextual understanding of the inputs as demonstrated by ClinicalBERT with an MCC score of 4.3%.

BERT-based domain-specific models can perform slightly better than the general BERT but the difference resides mainly on the convergence time between the training and the validation process. In case we want to use independent entities, such as NER, BOW models are more suitable to handle prediction tasks because only term frequency over inverse document frequency is more relevant. Nonetheless, the preprocessing methods may change with a different EHR, that utilizes a different notation. Besides the performance of a prediction model, it should be interpretable. For our case, the accuracy of the prediction should be quantified by diagnosis, drugs, bioinformation, or even demographics. However, the high dimensionality of modern NLP models contains abstract
FIGURE 5. Contextualized embeddings clustering. This figures show that dimension reduction using Uniform Manifold Approximation and Projection (UMAP) can easily find similarities between embedding vectors from the two classes.

FIGURE 6. Logits from 200 input samples represented by $y = \theta_{pos} - \theta_{neg}$. On the left are sentences from deceased patients while on the right are for discharged patients. The evaluation is on the $y$-axis. This shows the polarity of each model’s output.

features for which we have no words or mental concepts. To visualize which feature or medical terminology activated the model along with the input text does not necessary have a meaning for us. Therefore, we judged that interpreting the outcomes from the narratives is beyond the scope of this research.

There are also some limitations of this study. First, we limited this analysis to pneumonia patients who stayed in the ICU, and our prediction test used admission notes. NLP models require a lot of data to be generalizable and avoid over-fitting the models. Therefore, our guarantee for reproducibility using different domain data is limited.

A second limitation is common to EHR-driven prediction models for supervised learning. It is rare to have sufficient balance between classes, however for our case, the minority class represented by 29% of the data did not give an alarming false-negative for high-dimensional models and its precision score was as high as the majority class.

VII. CONCLUSION

Through this paper, we present a deep comparison of Neural Language modeling pipelines for outcome prediction from medical text notes using pneumonia patients. We compare the performance of medical notes preprocessing, their representation as well as a supervised learning mechanism to predict outcomes of ICU admissions. We demonstrated that text preprocessing is of paramount importance as the first step of the pipeline. A light preprocessing will not achieve results as good as a deeper processing. Replacing medical jargon and abbreviation terms to harmonize the data will have a high positive impact. For example, changing “dx PE” to “diagnosis Pulmonary Embolism” will allow models to add more weight to each of those tokens as related to pneumonia and appears multiple times. However, extreme processing such as NER will limit the applicable models besides cutting off some useful information from the text. The choice of the embeddings depends mostly on the input type, size, and domain. Current NLP models, utilizing transformers as the fundamental structure, understand the medical text better at the expense of prediction interpretability. A meticulous data cleaning, and subword level representation from a medical domain embedding and a fine-tuned transformer-based model yielded better optimal results.

REFERENCES

[1] A. C. Morris, “Management of pneumonia in intensive care,” J. Emerg. Crit. Care Med., vol. 2, p. 101, Dec. 2018. [Online]. Available: https://jeccm.amegroups.com/article/view/4830
[2] I. Rudan, K. O’Brien, H. Nair, L. Liu, E. Theodoratou, S. Qazi, I. Lukisc, C. Walker, R. Black, H. Campbell, and G. Reference, “Epidemiology and etiology of childhood pneumonia in 2010: Estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries,” J Glob Health, vol. 3, May 2013, Art. no. 10401.
[3] C. Mugisha and I. Paik, “Pneumonia outcome prediction using structured and unstructured data from EHR,” in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Dec. 2020, pp. 2640–2646.
[4] I. Li, J. Pan, J. Goldwasser, N. Verma, W. P. Wong, M. Y. Nuzumulah, B. Rosand, Y. Li, M. Zhang, D. Chang, R. A. Taylor, H. M. Krumholz, and D. Radev, “Neural natural language processing for unstructured data in electronic health records: A review,” 2021, arXiv:2107.02975.
[5] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis,” IEEE J. Biomed. Health Inform., vol. 22, no. 5, pp. 1589–1604, Sep. 2018.
[6] R. Mann and J. Williams, “Standards in medical record keeping,” Clin. Med. London, U.K., vol. 3, pp. 329–332, Jul. 2003.
[7] N. R. Council, Biomedical Models and Resources: Current Needs Future Opportunities. Washington, DC, USA: National Academies Press, 1998. [Online]. Available: https://www.nap.edu/catalog/6066/biomedical-models-and-resources-current-needs-and-future-opportunities
[8] J. L. Schlossberg, “Book review: Medical language processing: Computer management of narrative data by naomi sager, carol friedman, and margaret S. Lyman (Addison–Wesley 1987),” ACM SIGCHI Bull., vol. 20, no. 1, pp. 70–71, Jul. 1988, doi: 10.1145/49103.1046397.

[9] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, and H. M. El-Bakay, “Improving patient mortality prediction: An improved patient-specific stacking ensemble model,” IEEE Access, vol. 8, pp. 133541–133564, 2020.

[10] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, “Extracting information from the text of electronic medical records to improve case detection: A systematic review,” J. Amer. Med. Inform. Assoc., vol. 23, no. 5, pp. 1007–1015, Sep. 2016.

[11] S. Sheikhalishahi, R. Motto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, “Natural language processing of clinical notes on chronic diseases: Systematic review,” JMIIR Med. Informat., vol. 7, no. 2, Apr. 2019, Art. no. e12239.

[12] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, “Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review,” J. Amer. Med. Inform. Assoc., vol. 24, no. 1, pp. 198–208, Jan. 2017.

[13] V. Osmani, L. Li, M. Danieletto, B. Glicksberg, J. Dudley, and O. Mayora, “Automatic processing of electronic medical records using deep learning,” in Proc. 12th EAI Int. Conf. Pervasive Comput. Technol. Healthcare, May 2018, pp. 251–257.

[14] H. Liu, Y. Lussier, and C. Friedman, “A study of abbreviations in the UMLS,” in Proc. AMIA. Annu. Symp. AMIA Symp., Feb. 2001, pp. 393–397.

[15] G. Maragatham and S. Devi, “LSTM model for prediction of heart failure in big data,” J. Med. Syst., vol. 43, no. 5, pp. 1–13, May 2019.

[16] R. Ailsaad, Q. Malluhi, I. Janahi, and S. Boughorbel, “Interpreting patient-specific risk prediction using contextual decomposition of BiLSTMs: Application to children with asthma,” BMC Med. Inform. Decis. Making, vol. 19, no. 1, pp. 1–11, Dec. 2019.

[17] X. Zhou, Y. Li, and W. Liang, “CNN-RNN based intelligent recommenders for online medical pre-diagnosis support,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 18, no. 3, pp. 912–921, May 2021.

[18] J. Geraci, P. Wilansky, V. de Luca, A. Roy, J. L. Kennedy, and J. Strauss, “Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression,” Evidence Based Mental Health, vol. 20, no. 3, pp. 83–87, Aug. 2017.

[19] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, “A clinical text classification paradigm using weak supervision and deep representation,” BMC Med. Inform. Decis. Making, vol. 19, no. 1, pp. 1–13, Dec. 2019.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[22] K. Huang, J. Altaosar, and R. Ranganath, “ClinicalBERT: Modeling clinical notes and predicting hospital readmission,” 2019, arXiv:1904.05342.

[23] A. Johnson, T. Pollard, L. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark, “MIMIC-III, a freely accessible critical care database,” Sci. Data, vol. 3, May 2016, Art. no. 160035.

[24] O. Bodenreider, “The unified medical language system (UMLS): Integrating biomedical terminology,” Nucleic acids Res., vol. 32, p. D267, Feb. 2004.

[25] L. V. Grossman, E. G. Mitchell, G. Hripcsak, C. Weng, and D. K. Vawdrey, “Patient-specific stacking ensemble model,” in Proc. AMIA Annu. Symp. Proc., 2014, pp. 1532–1543.

[26] R. S. Puttini, A. A. Toffanello, R. M. Chaim, G. Alves, J. M. P. Rotzsch, E. O. Carvalho, E. Ishikawa, A. P. Araujo, and E. C. Oliveira, “Semantic framework for electronic health records,” in Proc. IEEE 11th Int. Conf. Semantic Technol. Comput. (ICSC), Jan. 2017, pp. 334–337.

[27] H. Sun, K. Depraetere, J. D. Roo, G. Mels, B. D. Vloed, M. Twagirumukiza, and D. Colaert, “Semantic processing of EHR data for clinical research,” J. Biomed. Informat., vol. 58, pp. 247–259, Dec. 2015.

[28] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Data processing and text mining technologies on electronic medical records: A review,” J. Healthcare Eng., vol. 2018, pp. 1–9, Apr. 2018.

[29] D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, “A neural named entity recognition and multi-type normalization tool for biomedical text mining,” IEEE Access, vol. 7, pp. 73729–73740, 2019.

[30] Y. Zhang, R. Jin, and Z. Zhou, “Understanding bag-of-words model: A statistical framework,” Int. J. Mach. Learn. Cybern., vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[31] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in Proc. Eur. Conf. Mach. Learn. Springer, 1998, pp. 137–142.

[32] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1532–1543.

[33] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proc. Adv. Neural Inf. Process. Syst., vol. 26, Dec. 2013, pp. 1–9.

[34] E. Choi, T. Bahadori, and J. Sun, “Doctor AI: Predicting clinical events via recurrent neural networks,” Nov. 2015, arXiv:1511.05942.

[35] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[36] P. Zhou, W. Shi, Z. Tian, Q. B. Li, H. Hao, and B. Xu, “Attention-based bidirectional long-short memory networks for relation classification,” in Proc. ACL, Berlin, Germany, Aug. 2016, pp. 207–212. [Online]. Available: https://www.aclweb.org/anthology/P16-2034

[37] I. Sutskever, O. Vinyals, and Q. Le, “Sequence to sequence learning with neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 4, Sep. 2014, pp. 3104–3112.

CHERUBIN MUGISHA (Member, IEEE) received the bachelor’s degree in computer science from the Université Lumière de Bujumbura, Burundi, in 2013, and the M.Sc. degree in computer science from The University of Aizu, Japan, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include algorithms for multimodal machine learning methods integrating NLP, structured and unstructured data and its application to medical data. He is the President of the IEEE University of Aizu Student Branch.

INCHEON PAIK (Senior Member, IEEE) received the M.E. and Ph.D. degrees in electronic engineering from Korea University, in 1987 and 1992, respectively. He is currently a Professor with The University of Aizu, Japan. His research interests include semantic web, web services and their composition, data mining, deep learning, and big data science infrastructure. He is a member of ACM, IEICE, and IPSJ.