A Comparison of Feature Extraction Models in NLP to Predict Patient Mortality Using the MIMIC-III Dataset

Elizabeth Garrison, Sai Shi, Tanuja Sajja

Professor Wang

May 4, 2021

Abstract

The goal of this project is to compare natural language processing (NLP) techniques in extracting information from the MIMIC III dataset in order to predict patient mortality. In this project, we compare the following NLP models: Bag of Words (BoW), term frequency-inverse document frequency (TF-IDF), long short-term memory (LSTM), and bidirectional long short-term memory (BiLSTM). We also study the use of pretrained word embedding layers and self-attention layers in our models. We find that the BiLSTM model performs the best for classification of mortality in our dataset.

Introduction

Natural language processing is an interdisciplinary subfield of computer science and linguistics. Research in the field of NLP aims to explore how computers can understand and manipulate natural language text or data in order to perform specific tasks [1]. The early NLP techniques were created based on Chomsky's Linguistic Theory which assumed that grammar rules were universal. However, this approach had its set of limitations which were caused mainly by the fact that there are exceptions to all rules and the ambiguity of natural language [10]. Because of the limitations in early NLP techniques, there was a quick rise in statistical NLP models which are the models that are most frequently used today [9]. Through supervised or unsupervised learning models, we can study patterns in natural language text.

One of the main uses of NLP is to support the automation of tasks in sectors of work that require human coding. In the medical field, NLP can support many tasks such as automation with coding, billing, and diagnostics [2]. In recent years, healthcare systems have also slowly moved towards adopting electronic health records. Although many hospitals have already started using electronic health records, the lack of data integration leads to data only being used for immediate patient care, despite this data having the potential to provide insight into understanding illnesses and illness progression. The MIMIC-III dataset aims to solve this issue by making electronic health record data openly available to researchers, which contains de-identified health-related data associated with forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [4].

Specifically, information contained in clinical notes, can also lead to the potential prediction of disease prognosis, therefore supporting medical personnel in decision making regarding treatment at various stages of disease. In clinical notes, there is a wealth of data such as demographic information, medical imaging, medical testing, as well as relationships outlined between symptoms, treatment, and outcome of treatment [3]. Coding information and pattern recognition can be extremely time-consuming for the human researcher and medical personnel that use this information to make treatment decisions and to diagnose patients.

With NLP models such as bag of words, TF-IDF and Bidirectional LSTM, features can be extracted from the clinical notes to build models that better support the medical field and patient care.

Related Work:

There have been many natural language processing models that have been implemented to extract data from a variety of sources. These implementations have largely benefitted the medical field by building models that extract data from clinical notes. In 2016, Yoav Goldberg published a paper in the *Journal of Artificial Intelligence* that provides a tutorial for the different available neural network models from an NLP perspective. This paper contains helpful background information on how the existing models work and how they can be applied to specific NLP problems [1]. Marafino et al. developed a sparse classifier to predict ICU mortality risk based on nursing notes. Text features from the notes that were strongly associated with mortality were classified. In this study, researchers used the MIMIC-II database to build the model [14]. We seek to extend their work by using the MIMIC-III database [5].

In addition to these studies, Ju et al. built a model using SVM that can efficiently recognize patterns in large datasets in order to perform Named Entity Recognition (NER) in biomedical texts. When the model was tested with the GENIA corpus dataset it performed with a precision rate of 84.24%. This paper provides information on SVM and how it can be applied to build a model with a high precision rate [6]. Singh et al. used NLP to automate the extraction of codes from unstructured clinical notes. They successfully identified the top ten and fifty most common diagnoses and procedures found in the MIMIC-III database. BERT was used to tune the language model and ultimately achieved an accuracy of 87.07%. This paper gives background information on how to generalize the knowledge discovery process used in this paper to other clinical notes [2].

Hasan et al. explored the idea of combining traditional NLP techniques with word and sentence embeddings such as LSTM in order to improve the relation extraction. They found that combining the techniques, created models that significantly outperformed other baseline models when tested using the same datasets [3]. Chalapathy et al. also used the bidirectional LSTM with CRF decoding in order to identify and classify concepts from patient clinical records into predefined categories. The paper provides helpful information on how they implemented the BiLSTM-CRF framework to complete the task of concept extraction from clinical notes [7].

In addition to these works, Liu et al. proposed an attention-based BiLSTM with a convolution layer for text classification. The combination of these three methods allows this proposed method to outperform other currently used classification methods in accuracy. Although this paper did not focus on clinical notes, the information on how this model is implemented for text documents will be helpful for us when testing a BiLSTM network to predict patient mortality [8].

Methods:

For our models, we used Python and the following packages: Pandas, NLTK, Keras, and SciKit Learn. We used a 70/30 train/test split for our models, with 20% of train data used for validation and hyperparameter tuning of our models.

Data Preprocessing
Data were obtained from the MIMIC-III database. We imported the MIMIC-III .csv file "Patient Clinical Notes" to a Pandas Dataframe, and then removed all features except subject id, Clinical Notes (text), and mortality label. Clinical notes for this dataset were from patients located in the ICU of the Beth Israel Deaconess Medical Center between 2001 and 2012. We shuffled the data, and randomly sampled 10,000 examples for our model. We removed stopwords using the standard NLTK package, and added additional stopwords of "Admission," "Discharge," and "Date," as these words were contained in all of the notes for all patients.

Bag of Words
BoW is a model that is used to extract specific features from text and stores these features for downstream use. BoW contains information on the vocabulary of the known word and the occurrences of the known word. It disregards any information about the order or structure of the words in the text document. This model creates fixed-length vectors from text by counting how many times each word appears in the text. The features generated using the BoW model can then be used to train other machine learning algorithms such as SVM in order to accomplish the wanted NLP tasks [11], [12].

TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

TF-IDF is a statistical measure that assigns a weight to each word based on its relevance. This model calculates the score of each word by combining two metrics. The first is the term frequency in a document. This means that if a word appears more times in a document it will be assigned a higher weight. The second metric uses the inverse of document frequency. The more times a word appears in a set of documents, the lower its weight is. Then finally multiplying both of these metrics assigns a score to each word in the text as shown in the formula above. These scores can then be combined with machine learning algorithms (such as SVM) to perform NLP tasks[5], [13].

SVM

Support Vector Machine (SVM) is a Machine Learning algorithm that was originally proposed for binary classification. The objective of the SVM is to find a hyperplane in an N-dimensional space, where N is the number of features, that distinctly classifies the data points. Multiple choices of hyperplanes exist and SVM is aimed at finding the plane that has the maximum margin, i.e. the maximum distance between data points of both classes [6]. In addition to linear classification, SVM is also capable of performing non-linear classification efficiently, using what is called ‘kernel trick’, which implicitly maps their inputs into high-dimensional feature spaces [6]. SVM with target-dependent and target-independent feature engineering methods were applied and proven to be effective in many NLP tasks[6].

LSTM

Among the neural network models that have been applied, recurrent neural networks and specifically the Long Short-Term Memory (LSTM) model is the most popular and promising one, due to its ability to preserve information from input that has passed through it using the hidden state. LSTM networks are a unique kind of recurrent neural network (RNN) [14]. Traditional RNNs have the problem of being unable to connect information to the present task when the gap is too large. LSTM networks are neural networks that are capable of learning such long-term dependencies. They are able to preserve information for long durations of time and are able to recall that information to predict future words based on context. LSTMs take in three pieces of information at each time step, current input data, short-term memory from the last cell (hidden state), and long-term memory. Then the cells use different trained gates to determine which information should be retained, passed through, and discarded.[14]

BiLSTM

Although LSTM networks solve the problem of long-term retention of information, it comes with its own limitations. In order to understand a word, we need not just the previous word, but also the future word, which together provide a better understanding of the sentence. Bidirectional LSTM overcomes the limitations of regular LSTM by using all available input information in the past and future of a specific time frame, and has been proven to be more effective than unidirectional ones in many NLP applications like speech recognition. Instead of running an LSTM only in the forward mode starting from the first token, we start another one from the last token running backward. BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm [7], [8].

Word Embedding

One way to improve model performance in NLP is using pre-trained word embedding matrices as an embedding layer for the LSTM and BiLSTM networks. Pre-trained word embeddings are the embeddings learned in one task that are used for solving another similar task. The reason we use pre-trained word embeddings is because learning embeddings from scratch is very challenging due to the sparsity of training data and large number of trainable parameters. For our models, we used the GloVe (Global Vectors for Word Representation) in one model as the embedding layer [15], and fastText in another model [18] for comparison. We chose these pre-trained word embedding matrices due to their success in previous work [3, 7,15, 18].

Self-Attention

Our final BiLSTM model was tested using a self-attention layer as part of the neural network. The attention mechanism focuses on the information outputted by the hidden layers of BiLSTM. We chose this last model to test based on the success from Liu et. al [8]. The intuition behind the attention mechanism is that one word often 'attends' to other words in the same sentence differently. For example, given the sentence 'she is eating a green apple', when we see 'eating', we expect to see a food word very soon, hence 'eating' should pay higher attention to 'apple', compared to 'green', even though 'green' is closer to 'eating'. Traditional RNN models, such as LSTM, suffer from its incapability of remembering long sentences due to the fixed-length context vector design, hence we need a better mechanism to learn the long-range dependencies within the paragraph. This is also applicable to our dataset, since many clinical notes are lengthy, complicated, and contain long-range dependencies based on our observation. The attention model computes an attention score for each input by deriving the key, query, and value from inputs using neural networks, which shows how attended each query is against the keys.

Results

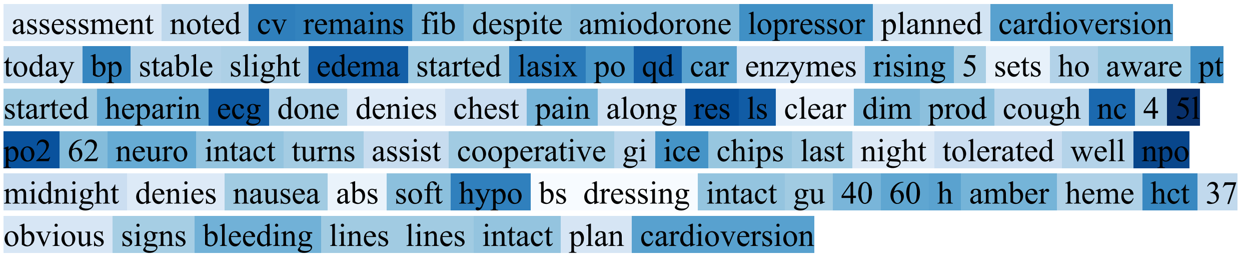
Models	**Accuracy**	**Precision**	**Recall**	**F1**
BoW + SVM	0.73	0.69	0.68	0.68
TF-IDF + SVM	0.75	0.69	0.71	0.70
LSTM (scratch)	0.71	0.66	0.65	0.65
LSTM (GloVe)	0.73	0.66	0.78	0.71
LSTM (Fasttext)	0.73	0.66	0.75	0.70
BiLSTM (GloVe)	0.72	0.64	**0.81**	0.71
BiLSTM (Fasttext)	0.73	0.68	0.67	0.68
BiLSTM + Attention	**0.75**	**0.71**	0.80	**0.75**

From these results, we observe that the best model for accuracy, precision and F1 for predicting patient mortality is the BiLSTM with attention model. BiLSTM Fasttext scored the highest in recall.

It is also important to note that the TF-IDF model scored the same accuracy as the BiLSTM + Attention model. This model also compiled the quickest, as compared to LSTM and BiLSTM models which took several hours to train. This illustrates that machine learning models can potentially perform as well as deep learning models, especially with smaller datasets. It is possible that the results we obtained could be improved with a larger sample size, and potentially different pre-trained word embeddings that are specifically trained for the medical domain.

We also present the following examples from the BiLSTM+Attention model. These examples show where the model accurately predicted mortality to match the ground truth of mortality label of the patient. In these examples, words are shown from light to dark blue, where dark blue represents a word that has more importance according to the attention model.

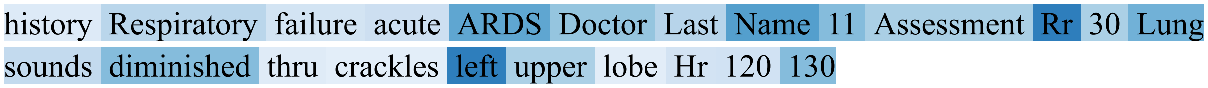
Example 1:



assessment noted cv remains fib despite amiodorone lopressor planned cardioversion today bp stable slight edema started lasix po qd car enzymes rising 5 sets ho aware pt started heparin ecg done denies chest pain along res ls clear dim prod cough nc 4 5l po2 62 neuro intact turns assist cooperative gi ice chips last night tolerated well npo midnight denies nausea abs soft hypo bs dressing intact gu 40 60 h amber heme hct 37 obvious signs bleeding lines lines intact plan cardioversion

Attention: edema (swelling caused by excess fluid trapped in your body's tissues), 45l po2 (blood oxygen, less than 40 is severe), ecg, cardioversion is a medical procedure that restores a normal heart rhythm in people with certain types of abnormal heartbeats (arrhythmias).

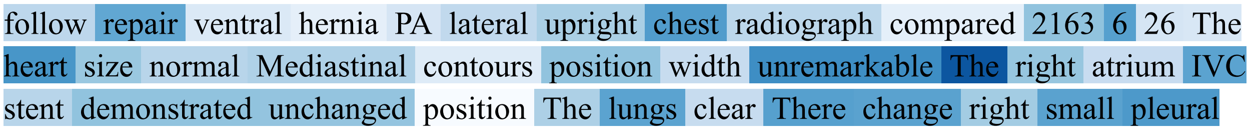
Example 2:



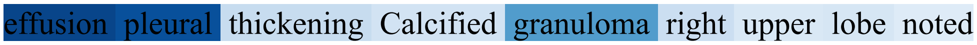
history Respiratory failure acute ARDS Doctor Last Name 11 Assessment Rr 30 Lung sounds diminished thru crackles left upper lobe Hr 120 130

Attention: Respiratory, ARDS (Acute respiratory distress syndrome), lung, diminished.

Example 3:



follow repair ventral hernia PA lateral upright chest radiograph compared 2163 6 26 The heart size normal Mediastinal contours position width unremarkable The right atrium IVC stent demonstrated unchanged position The lungs clear There change right small pleural



Attention: pleural effusion, “water on the lungs,” pleural thickening, calcified granuloma--infection of lung.

Discussion

Prior to starting the experiment, we hoped that the BiLSTM NLP models would perform the best in extracting relevant information from the MIMIC-III dataset in order to predict patient mortality. This was the initial prediction because of BiLSTMs ability to consider both past and future words based on context instead of the unidirectional approach of the other models. BiLSTM NLP model showed the best performance in many previous studies [7]. In our study, we found that the BiLSTM models performed better than BoW, TF-IDF, and LSTM models.

Implementing an attention mechanism to the model improved the performance. An attention mechanism can improve performance of models by storing all relevant information and using that information when needed depending on the context of the prediction [15].

There were some limitations. Because of the nature of the MIMIC-III dataset, the models could have benefited from a larger sample size. Using a larger sample size could have led to an improved performance with the LSTM and BiLSTM models. However, it was not feasible for us to go higher than 10,000 examples as we had limited resources and time to run the models.

Conclusion

Clinical natural language processing is a subfield of NLP that has gained more popularity in recent years. It has the ability to drastically improve patient diagnosis, treatment, and care. Because of the complex nature of clinical notes, NLP models that can be used to accurately predict the needed information are still a work in progress. As can be seen by the results of this experiment, the accuracy in predictions was not above 75%. Other relevant studies have shown that BiLSTM models can reach precision levels up to 85% [7].

The results of this project add to the current available information about NLP techniques that can be used with clinical notes extraction. It shows a comparison of multiple different techniques that all perform relatively well in the assigned task. With further improvements of NLP models tailored to clinical notes extraction, MIMIC-III and future medical information databases will play an important role in patient care. Future research should test different word embeddings that are pre-trained from domain specific resources, and larger sample sizes should be used when improving NLP models.

References.

[1] Goldberg, Yoav (2016). "A Primer on Neural Network Models for Natural Language Processing". Journal of Artificial Intelligence Research. 57: 345–420. arXiv:1807.10854. doi:10.1613/jair.4992. S2CID 8273530.

[2] Singh, A. K., Guntu, M., Bhimireddy, A. R., Gichoya, J. W., & Purkayastha, S. (2020). Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes. arXiv preprint arXiv:2003.07507.

[3] F. Hasan, A. Roy and S. Pan, "Integrating Text Embedding with Traditional NLP Features for Clinical Relation Extraction," 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 2020, pp. 418-425, doi: 10.1109/ICTAI50040.2020.00072.

[4] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635 (Links to an external site.)

[5] Marafino, B. J., Boscardin, W. J., & Dudley, R. A. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. Journal of biomedical informatics, 54, 114-120.

[6] Z. Ju, J. Wang and F. Zhu, "Named Entity Recognition from Biomedical Text Using SVM," *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, Wuhan, China, 2011, pp. 1-4, doi: 10.1109/icbbe.2011.5779984.

[7] Chalapathy, Raghavendra, Ehsan Zare Borzeshi, and Massimo Piccardi. "Bidirectional LSTM-CRF for clinical concept extraction." *arXiv preprint arXiv:1611.08373* (2016).

[8] Liu, Gang, and Jiabao Guo. "Bidirectional LSTM with attention mechanism and convolutional layer for text classification." *Neurocomputing* 337 (2019): 325-338.

[9] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, *Journal of the American Medical Informatics Association*, Volume 18, Issue 5, September 2011, Pages 544–551, https://doi.org/10.1136/amiajnl-2011-000464

[10] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

[11] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework." *International Journal of Machine Learning and Cybernetics* 1.1-4 (2010): 43-52.

[12] Wang, Jin, et al. "Bag-of-words representation for biomedical time series classification." *Biomedical Signal Processing and Control* 8.6 (2013): 634-644.

[13] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." *Proceedings of the first instructional conference on machine learning*. Vol. 242. No. 1. 2003.

[14] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." *Physica D: Nonlinear Phenomena* 404 (2020): 132306.

[15] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[16] Vu, Thanh, Dat Quoc Nguyen, and Anthony Nguyen. "A Label Attention Model for ICD Coding from Clinical Text." *arXiv preprint arXiv:2007.06351* (2020).

[17] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[18] Agibetov, A., Blagec, K., Xu, H. et al. Fast and scalable neural embedding models for biomedical sentence classification. BMC Bioinformatics 19, 541 (2018). https://doi.org/10.1186/s12859-018-2496-4