

# Understanding patient complaint characteristics using contextual clinical BERT embeddings

Budhaditya Saha, Sanal Lisboa, Shameek Ghosh

Medius Health, Sydney, Australia

{aditya.saha, sanal.lisboa, shameek.ghosh}@mediushealth.org

**Abstract**—In clinical conversational applications, extracted entities tend to capture the main subject of a patient’s complaint, namely symptoms or diseases. However, they mostly fail to recognize the characterizations of a complaint such as the time, the onset, and the severity. For example, if the input is “I have a headache and it is extreme”, state-of-the-art models only recognize the main symptom entity - *headache*, but ignore the severity factor of *extreme*, that characterises *headache*. In this paper, we design a two-stage approach to detect the characterizations of entities like symptoms presented by general users in contexts where they would describe their symptoms to a clinician. We use Word2Vec and BERT to encode clinical text given by the patients. We transform the output and re-frame the task as a multi-label classification problem. Finally, we combine the processed encodings with the Linear Discriminant Analysis (LDA) algorithm to classify the characterizations of the main entity. Experimental results demonstrate that our method achieves 40-50% improvement in the accuracy over the state-of-the-art models.

## I. INTRODUCTION

Clinical Named Entity Recognition systems based on neural networks [1] [2] are trained to detect entities in text. In the clinical domain, there are different types of inter-related entities. Existing systems lack the ability to detect these relations because these systems are not trained to understand the context in a text. For example, in the text “I have severe headache and nausea”, the parent entities are *headache* and *nausea*. The child of *headache* is *severe*. Existing systems may detect the three entities but they are unable to predict if they are related.

A relationship prediction mechanism is required to link parent and child entities in a text. Such techniques are useful in applications like clinical conversational chat platforms [3], which predict disease differentials based on the symptoms entered. Here, the quality of predicted disease differentials depends on the accuracy of identified clinical information ( and their characteristics) in the text. The input text may contain two main components (a) clinical entities or *parent* entities and (b) the characterization of the clinical entities or *children* entities. Table I shows example of duration, severity, onset and frequency onset characterizations respectively.

The clinical named entity recognition model has been researched extensively [1], [2], [4]. The state-of-the-art clinical entity recognition methods [5], [6] mostly recognizes the parent entity in the text. The two most popular clinical entity recognition tools are METAMAP [5] and Amazon Medical

Comprehend [6]. The METAMAP framework recognizes a medical concept, whereas the Amazon Medical Comprehend service predicts the named entities in clinical texts. While these tools can separately detect the parent entity and child entities, they are not able to predict the relationship or context. In the example discussed earlier, the Amazon medical comprehend predicts the *headache* and *nausea* as a *parent* entities and *severe* as a characterization. Similar outcomes can also be found for the METAMAP. But they fail to recognize that the *severe* is related to *headache*. The main reason behind this failure is that the data modeling method ignores the contextual information that denote the relation between the neighboring words.

In this paper we build a solution to recognize the time, onset and severity characterization of a *parent* entity in user input. In a clinical chatbot, the *parent* entities are mostly symptoms or diseases. To achieve this, we seek to capture the contextual information in an input text and convert this information into a vector space representation. These vector space model will effectively map contextually related texts close to each other and unrelated texts far away from each other. For example, sentences like *I have a pain in the head for hours* and *I have got a headache since morning* have similar contextual information for a target entity *headache*, hence, they will be placed close to each into the vector space. Similarly, examples like *I am having continuous headache* and *I get headache infrequently* will be mapped far away from each other, as confirmed by our experiments.

Here, we propose a framework to understand the language of the clinical text and predict the time, onset and severity characterization of a *parent* entity. For language understanding, we use state-of-the-art, deep learning models designed for natural language processing tasks. In our model, we have fine-tuned these models for the clinical text. These models take text as an input and output a continuous vector representation. This outcome is processed and the downstream task is framed as a multi-label classification problem. Finally, the intermediate representation is fed to a classifier. We have applied the model on the dataset curated from potential users of chatbot Quro [3]. We compare the outcome of this model with other popular state-of-the-art [5], [6] architectures. The performance evaluation shows that our model archives up to 50% higher accuracy, 40% higher precision, 40% higher recall, and 30% higher F1 score compared to state-of-the-art.

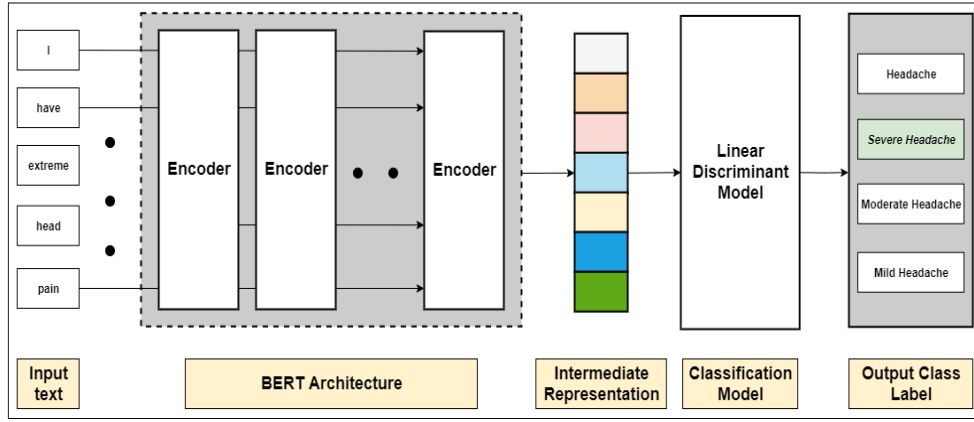


Figure 1: A schematic diagram of the proposed time-severity-onset recognition model

User Input in a chatbot	Parent Entity	Children type	Children Class
I have a <i>headache</i> for the last <i>2 months</i> .	Symptom (Headache)	Duration (time)	Months
She is having a <i>headache</i> since last <i>five days</i> .			Days
<i>headache</i> lasted for <i>several hours</i> .			Hours
I'm having <i>headache</i> from <i>few minutes</i> .			Minutes

(a) Duration (time)

User Input in a chatbot	Parent Entity	Children type	Children Class
<i>Pain</i> is <i>extreme</i> in my <i>head</i>	Symptom (Headache)	Severity	Severe
I am having a <i>moderate headache</i>			Moderate
I am having a <i>slight pain in head</i>			Mild

(b) Severity

User Input in a chatbot	Parent Entity	Children type	Children Class
my <i>headache</i> starts <i>abruptly</i>	Symptom (Headache)	Onset	Sudden
<i>gradual pain in my head</i>			Gradual

(c) Onset

User Input in a chatbot	Parent Entity	Children type	Children Class
I am having a <i>constant pain in head</i>	Symptom (e.g. Headache)	frequency (time)	Continuous
I usually get <i>pain in head occasionally</i>			ON - OFF

(d) Frequency (time)

Table I: Examples of duration, severity, onset, and frequency characterizations respectively. The color represents the type of the entities (*blue* for parent and *red* for children).

## II. CLINICAL DIALOGUE DATASET

The clinical conversation data is curated from text inputs of potential users of Quro [3]. The Quro bot is an AI-driven clinical conversational platform that orchestrates the patient throughout the primary health care journey. We have recorded around 2000 instances of unlabeled text which contains *parent* symptoms with one or more *time-severity-onset* factors for

each parent. We use a semi-supervised technique to label these unlabeled text instances. The semi-supervised technique detects keywords in the text and labels them. These labels are then reviewed by clinicians to reduce the noise. Table I shows recorded data with annotations.

The time factor has two components ie. duration and frequency. There are 4 characteristics (hours, days, weeks, months) of duration component and 2 characteristics (on-off, continuous) of frequency component as shown in Table Ia and Table Id. If user text is *I have been having regular back pain since the last 3 days*, the word *regular* in text signifies *continuous* characteristic of frequency component; *last 3 days* signifies the *days* characteristic of duration component.

The severity factor has 3 characteristics ie. severe, mild and moderate as shown in Table Ib. For example if the user text is *I have extreme headache* the word *extreme* in text signifies *severe* characteristic.

The onset factor has 2 characteristics ie. sudden and gradual as shown in Table Ic. For example if the user text is *my headache starts abruptly* the word *abruptly* in text signifies *sudden* characteristic.

## III. PROPOSED APPROACH

The proposed framework has two main components: (a) A text encoder, and (b) a classification model. In our architecture, a tokenized sentence is given as an input to the encoder which maps the sentence to a continuous vector representation. The dimensions of the vector representation are reduced because most encoders produce high dimensional vector representations. Then, the classification model maps the vector representation to a child class. Fig. 1 depicts a schematic diagram of the proposed model which uses BERT as the encoder and LDA as the classifier. Before the classifier is applied the task is framed as a multi label classification problem. In the following sections, we will describe the two components in detail.

### A. Text Encoding

Word embeddings are vectors that represent words in a text in the semantic space. Similarity between words can be found using a distance measure such as cosine similarity. The Word2Vec is a deep learning model which attempts to create high quality word embeddings [7]. These word embeddings capture context form the training corpus and as such embeddings from general text cannot be used in highly specific context.

Bidirectional Encoder Representations from Transformer (BERT) [8] is a viable solution to the context problem. BERT model reads a text input sequentially from left-to-right and right-to-left to learn the contextual relationship amongst the words and embed this learning into a low dimensional continuous vector space. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

In our framework, we can use Word2Vec to encode words in sentence and combine the word vectors. Alternatively, we can use BERT encoder model to get a continuous vector representation of the entire input text. These models can be pre-trained on large corpus of text and used in a different context. However the accuracy would depend on the relevance and quality of the corpus.

### B. Sentence Classification

We use LDA for sentence classification. Consider,  $C$  number of children class denoted by  $\mathcal{L}$ , where  $\mathcal{L} = (\ell_1, \ell_2, \dots, \ell_m)$ , the LDA [9] model maps the  $D$  dimensional vector representation  $\mathbf{h} \in \mathcal{R}^D$  to a class label in  $\mathcal{L}$ . Here, the training data can be expressed as  $(\mathbf{h}_i, \ell_i)$  where  $i \in 1, \dots, N$ ,  $N$  is the number of training samples. The number of vectors in class  $\ell_i$  is denoted by  $n_i$ , thus  $N = \sum n_i$ . The LDA tries to find an optimal hyperplane such that the separability between two classes is maximized. The hyperplane is computed by minimizing the within class distance and maximizing the between class distance simultaneously. The within class ( $\mathbf{H}_w$ ) and between class ( $\mathbf{H}_b$ ) scatter matrices are defined as

$$\mathbf{H}_b = \sum_{i=1}^C (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

$$\mathbf{H}_w = \sum_{i=1}^C \sum_{\mathbf{h} \in \ell_i} (\mathbf{h} - \mathbf{m}_i)(\mathbf{h} - \mathbf{m}_i)^T$$

where  $\mathbf{m}_i$  denotes the class mean of  $i^{th}$  class and  $\mathbf{m}$  is the global mean of samples  $\{\mathbf{h}_i\}_{i=1}^N$ . The LDA model learns the hyperplane by optimizing the fisher criterion as

$$J(\mathbf{W}) = \max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{H}_b \mathbf{W}}{\mathbf{W}^T \mathbf{H}_w \mathbf{W}} \quad (1)$$

where  $\mathbf{W}$  is a parameter of the hyperplane. The equation (1) can also be modified as  $\mathbf{H}_b \mathbf{W} = \lambda \mathbf{H}_w \mathbf{W}$  which turns to a generalized eigenvalue problem with eigenvectors  $\mathbf{W}$  and eigenvalue  $\lambda$ . The optimal hyperplane is spanned by the eigenvectors in  $\mathbf{W}$ .

Model	Accuracy	Precision	Recall	F1-score
BERT+ LDA	<b>0.642</b>	<b>0.9245</b>	<b>0.8376</b>	<b>0.8789</b>
Amazon Medical	0.1714	0.5982	0.3418	0.4351
METAMAP	0.3	0.6838	0.4701	0.5441

Table II: Performance evaluation of the proposed BERT+LDA model with METAMAP and Amazon medical comprehend

## IV. EXPERIMENTS

We evaluate the performance of the proposed BERT+LDA model into two different stages. In the first stage, we compare the performance of the The BERT + LDA model the with state-of-the-art UMLS concept recognition tool METAMAP [5] and clinical named entity recognition model Amazon Comprehend Medical [6]. In the second stage, we compare the BERT+LDA against Word2Vec+LDA. In both architectures we use Principal Coefficient Analysis (PCA) [10] for dimensionality reduction of text vectors. We use Chain Classifier [10] to frame the task as multi label classification.

### A. Experimental Setup

We use curated clinical text to train and test our models. All of the text instances contain at least the parent entity and some contain a combination of time-onset-severity characteristics. We use instances for the symptom “headache” in the following experiments. Randomly sampled 80% of the data is used for training and 20% is used for evaluation.

The fine tuned BERT model has 24 layers and 1024 neurons per layer, activation function used is Gaussian Error Linear Units(GELU) and the vocabulary size of 30522. LDA model is optimized using a off-the-shelf Singular value decomposition (SVD) solver where convergence limit is set to 1.0e-7. The Word2Vec [7] model is pre trained on Google News dataset and contains 3 million words with each word having a vector of 300 dimensions.

### B. Performance Metric

We use the following metrics:

- 1)  $accuracy = \frac{\text{correct predictions}}{\text{sample size}}$
- 2)  $precision = \frac{TP}{TP+FP}$
- 3)  $recall = \frac{TP}{TP+FN}$
- 4)  $F_1 score = 2 \times \frac{precision * recall}{precision + recall}$

Where TP is True Positive, FP is False Positive and FN is False Negative.

### C. Experimental Results

Table II shows the comparison of the proposed BERT + LDA model with the METAMAP and Amazon medical comprehend respectively. The BERT+LDA model is superior than the Amazon medical comprehend about 3.8 times on accuracy, 1.6 times on precision, 2.4 times on recall and 2 times on  $F_1$  score respectively. A similar superior result can be observed against the METAMAP as well.

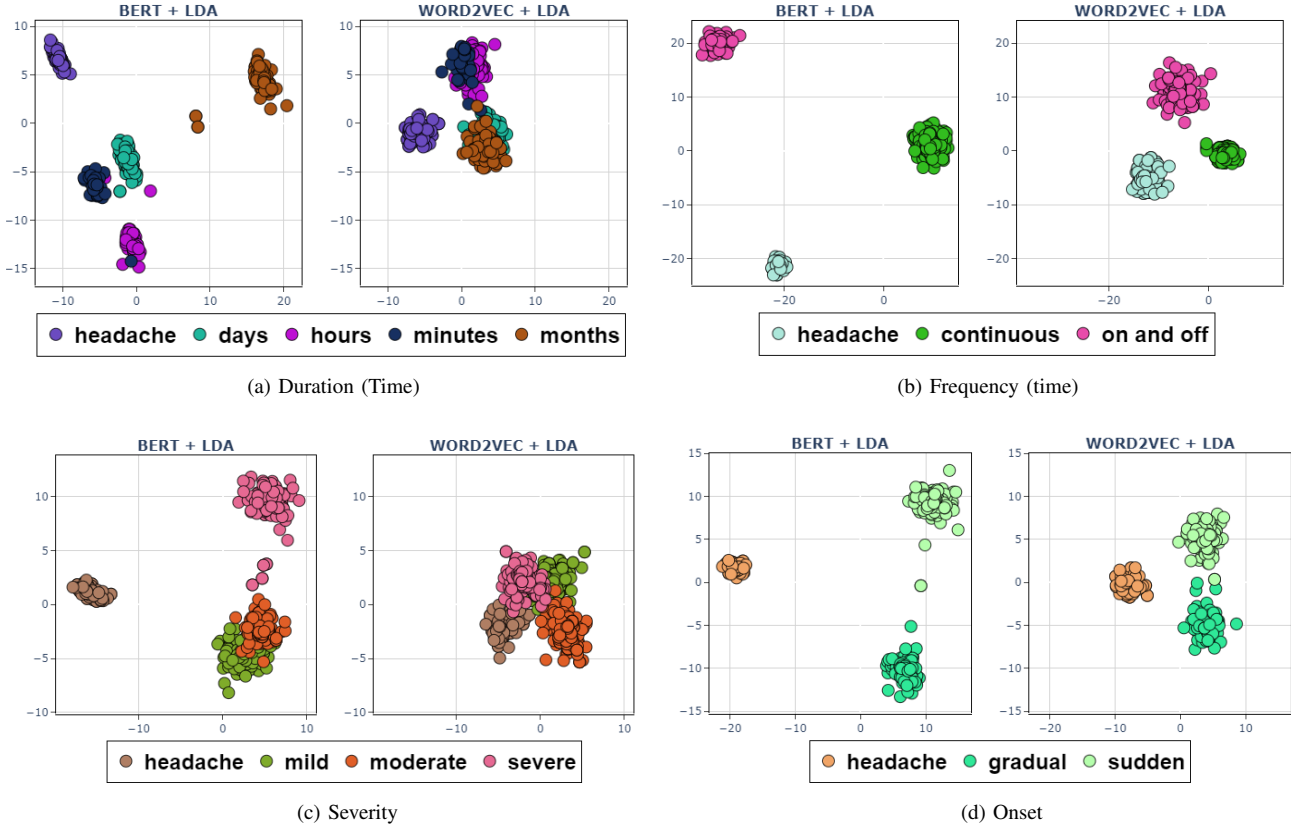


Figure 2: Clusters formed by the BERT+LDA and Word2Vec+LDA models.

Model	Contextual information	Accuracy	Precession	Recall	F1-score
WORD2VEC + LDA	NO	0.0215	0.6842	0.2532	0.3696
BERT+ LDA	YES	<b>0.6428</b>	<b>0.9245</b>	<b>0.8376</b>	<b>0.8789</b>

Table III: Performance comparison of word2vec and BERT

Table III compares the performance of the BERT + LDA and Word2vec + LDA model respectively. The BERT + LDA is 32 times better on accuracy than the Word2vec + LDA model. In this experiment, the precession, recall and  $F_1$ -score is improved for BERT+LDA by a factor of 1.35, 3.32 and 2.4 times in comparison to the Word2vec + LDA respectively.

In summary, the BERT + LDA performs superior in comparison to the other entity recognition model and popular word embedding model the Word2vec. This is because the BERT encodes the text capturing the context of the entire text while Word2vec has a vector for each word in the text. When embedding the entire text using Word2Vec, word vector for each word in the text is averaged and in doing so the sequence of words is not taken into consideration.

To check the performance of the BERT + LDA model qualitatively, we plot the output of the 2 dimensional vectors

generated by the LDA model. Fig. 2 shows how the related text is represented closer in vector space. For the Duration (time), the clusters of the BERT+LDA model are linearly separable, however, the clusters of the Word2Vec+LDA are overlapped. For the frequency (time) and the onset, the clusters in both models are clearly separable. For severity, the clusters in both models are partially overlapped, however, it seems that the configuration of the clusters in the BERT+LDA model is better than the Word2Vec+LDA model. Overall, the BERT+LDA model is superior to the Word2Vec+LDA model.

## V. CONCLUSION

Existing clinical named entity recognition models are designed to predict the parent entity (eg. *headache*) in a text input. However, these models fail to recognize the time-onset-severity characterization of a parent entity (e.g. *days* or *months*, *sudden* or *gradual*, *severe* or *mild*). In this paper, we have proposed a model which is a combination of a language understanding framework and a classification method to predict both the parent entity and the time-onset-severity characterization of the parent entity. The proposed model successfully exploits the contextual information of the parent entity to predict its time-onset-severity characterizations. The proposed model has shown a superior performance against the

state-of-the-art clinical named entity recognition frameworks METAMAP and Amazon medical comprehend.

## REFERENCES

- [1] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu, "Entity recognition from clinical texts via recurrent neural network," *BMC medical informatics and decision making*, vol. 17, 2017.
- [2] M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispace: Fast and robust models for biomedical natural language processing," *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019. [Online]. Available: <http://dx.doi.org/10.18653/v1/W19-5034>
- [3] S. Ghosh, S. Bhatia, and A. Bhatia, "Quro: Facilitating user symptom check using a personalised chatbot-oriented dialogue system," *Stud Health Technol Inform*, vol. 252, 2018.
- [4] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics*, vol. 46, 2013.
- [5] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, vol. 1, 2006.
- [6] P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel, "Comprehend medical: a named entity recognition and relationship extraction web service," *arXiv preprint arXiv:1910.07419*, 2019.
- [7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [9] S. Zhang and T. Sim, "Discriminant subspace analysis: A fukunaga-koonz approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007.
- [10] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.