Published in final edited form as:

IEEE J Biomed Health Inform. 2022 April ; 26(4): 1737–1748. doi:10.1109/JBHI.2021.3123192.

A Deep Language Model for Symptom Extraction from Clinical Text and Its Application to Extract COVID-19 symptoms from Social Media

Xiao Luo [Member, IEEE],
Department of Computer Information Technology, IUPUI, Indianapolis, IN 46202 USA

Priyanka Gandhi,
Department of Computer Information Science, IUPUI, Indianapolis, IN 46202 USA

Susan Storey,
School of Nursing, Indiana University, Indianapolis, IN 46202 USA

Kun Huang [Senior Member, IEEE]
Indiana University School of Medicine, Indianapolis, IN 46202 USA

Abstract

Patients experience various symptoms when they have either acute or chronic diseases or undergo some treatments for diseases. Symptoms are often indicators of the severity of the disease and the need for hospitalization. Symptoms are often described in free text written as clinical notes in the Electronic Health Records (EHR) and are not integrated with other clinical factors for disease prediction and healthcare outcome management. In this research, we propose a novel deep language model to extract patient-reported symptoms from clinical text. The deep language model integrates syntactic and semantic analysis for symptom extraction and identifies the actual symptoms reported by patients and conditional or negation symptoms. The deep language model can extract both complex and straightforward symptom expressions. We used a real-world clinical notes dataset to evaluate our model and demonstrated that our model achieves superior performance compared to three other state-of-the-art symptom extraction models. We extensively analyzed our model to illustrate its effectiveness by examining each component's contribution to the model. Finally, we applied our model on a COVID-19 tweets data set to extract COVID-19 symptoms. The results show that our model can identify all the symptoms suggested by CDC ahead of their timeline and many rare symptoms.

Keywords

Natural Language Processing; Symptom Extraction; Deep Language Model; COVID-19; Social Media

luo25@iupui.edu .

I. Introduction

Patients with chronic or acute disease experience a variety of symptoms as a result of their disease or the treatments for their disease [1], [2]. Research shows that the majority of patients (76%) are admitted to the hospital due to symptoms [3]. Unrelieved symptoms can have deleterious effects on patient outcomes (e.g., functional status, mood states, and quality of life) [4]. Experiencing symptoms, minor to severe, prompts millions of patients to visit their healthcare providers each year. Among hospital admission of the cancer patients, the chief complaints include dyspnoea(15.7%), pain(15.2%), and neurological symptoms(14.5%) [5]. Neuropsychiatric symptoms (NPSs) affect almost all individuals with dementia (97%) over the course of the disease and associate with early institutionalization [6]. During the COVID-19 pandemic, symptoms are indicators for severity of disease [7] and the needs of hospitalization [8]. Hence, it is critical to identify symptoms and link them to disease prediction and healthcare outcome management.

In the typical electronic health record (EHR) system, the structured data includes demographic information, allergies, immunizations, diagnosis, medications, procedures, etc. Besides the structured data, clinicians record a large amount of information in clinical notes or progress notes. The clinical notes are narrative text that describes patients' symptoms, the rationale for the treatments, advice on taking certain medications, etc. Very few researchers use EHR clinical text data to extract symptoms for symptom management, risk assessment, and clinical outcome prediction. There is no ontology or existing standard vocabulary for symptoms. Another challenge to automate the symptoms extraction is data inconsistency. Clinicians write symptoms in different expressions or abbreviations. Because of the varieties of representations and large amounts of clinical notes, advanced algorithms need to be developed to extract the symptom terms and phrases. If symptoms can be extracted from clinical text, better precision healthcare plans can be built by linking them to the structured data.

With traditional text mining and Natural Language Processing (NLP) techniques, domain experts need to specify the symptom keywords to identify the most relevant text from the EHR. The keyword search approach often ignores the detailed context and the semantic similarities between words and phrases. Although Unified Medical Language System (UMLS) MetaMap [9], cTAKES [10], and other existing tools can extract symptoms, most of them have limitations in recognizing complex symptom representations. For example, only "pain" instead of "pain in nail" is extracted from the sentence "Patient reported pain in nail" if MetaMap or cTAKES is used to identify terms that belong to the semantic category sign or symptoms. In recent years, NLP techniques for health informatics, such as drug adverse events detection [11]-[13], and medical entity and entity-relationship recognition [14]-[16], deep learning-based NLP algorithms performed better than traditional NLP algorithms. However, no robust deep learning model has been created to extract symptoms from clinical text. This research develops an integrated deep learning language model to extract patient-reported symptoms from clinical text and apply it to extract COVID-19 symptom mentions from social media.

Our research objective is to extract patient-reported symptoms, which means if a patient reports “cough” during an encounter, this is categorized as a patient-reported symptom. However, if a clinician observes “cough”, or records “cough” as a diagnosis during an encounter, this would not be categorized as a patient-reported symptom. The reason is that an observed symptom can be interpreted differently by clinicians. Symptoms can be described within any clinical note section or a chunk of narrative text, including the present illness history. For example, “Patient reported lower back pain 1 week ago, but resolved after taking medication”. In order to extract the symptoms that are clinically insightful and actionable, an NLP system needs not only to identify the “true” symptoms cases, such as direct symptom mention or complaint statement; but also need to determine the “not true” symptoms, such as negation statements or conditional statements (“Return to the hospital if pain getting worse.”), etc., as these statements do not describe the symptoms patient is experiencing. These “not true” cases make the symptom extraction more difficult. Another challenge is that clinicians often use incomplete short sentences in clinical notes to record patients’ conditions, such as “Take this medication when needed”, etc. Therefore, it requires a system or model beyond traditional keyword searches or matches to incorporate the context information for practical clinical use.

The deep learning language model developed in this research consists of two components. Both components are based on the notion of named entity recognition (NER) [17]. One component - DeepSymptom utilizes the state-of-the-art language model BERT [18] with a conditional random field (CRF) [19] layer to exam the semantic relations between words for symptom extraction, which has been applied to perceive various medical entities from biomedical corpus [20]-[22]. The other component - SynSymptom, analyzes the syntactic dependency between words by feeding the syntactic feature representation of each word into a multilayer perceptron (MLP) network. We compared our language model against different baseline models and showed that our model works better than the baseline models. We also evaluated the generalizability of our model on a dataset of COVID-19 tweets. Our method can successfully identify COVID-19 symptoms mentioned on the Center for Disease Control (CDC) website. Also, we identify rare symptoms, such as “blurred vision” [23], which are discussed in the medical literature.

The main contributions of this research include:

- Develop an integrated deep language model for symptom extraction from the clinical text. The integrated deep language model has a component to explicitly consider the syntactic dependencies between words for symptom extraction and a component to consider the semantic relations for symptom extraction,
- Compare our model against the baseline models for symptom extraction and demonstrate that our model performs better than the baseline models,
- Evaluate the generalizability of our model on processing social media text for symptoms extraction, and
- Demonstrate that our model can extract typical COVID-19 symptoms listed by CDC and rare symptoms mentioned in the literature.

II. Related Work

A. Symptom Extraction from Clinical Text

Various researchers applied NLP and machine learning techniques to extract symptoms from the clinical notes of the EHR for clinical research. Many of them were designed to extract symptoms of a specific disease or disorder. Vijayakrishnan et al. [24] developed an NLP pipeline to identify signs and symptoms of heart failure (HF) patients. The NLP pipeline took a rule-based approach that first use Framingham HF diagnostic criteria to identify the major and minor symptoms of HF, then rules including noun phrase recognition, negation identification, and disambiguation were used to extract the symptoms. Jackson et al. [25] developed an NLP analysis model to capture critical symptoms of severe mental illness from clinical text. They first classified sentences in the clinical notes into two categories: catatonic symptoms and negative symptoms using a keywords-based approach. A team of psychiatrists defined the keyword lexicon of the symptoms. Gundlapalli et al. [26] investigated an NLP algorithm to extract urinary symptoms to detect indwelling urinary catheters. They first defined a set of lexicons for urinary symptoms. Then, the rule-based NLP algorithms were used to detect the positive and negative cases. Divita et al. [27] developed an NLP pipeline to extract symptoms from the clinical notes. The pipeline started with a set of human-annotated symptoms extracted from the consumer health vocabulary, then used UMLS Metathesaurus to expand the vocabulary. Although UMLS Metathesaurus can be used to identify symptoms, the limitation is that it does not include the different expressions of the same symptom.

B. Name Entity Recognition (NER) for Clinical Concept Extraction

Many studies have achieved good results on clinical entity recognition from text through applying NER techniques. Jonnalagadda et al. [28] proposed the semantic conditional random fields (CRF) model using CRF as a sequential discriminant classifier that analyzes the correlation between the words to improve the accuracy of NER. Fu et al. [29] explored different methods that can be applied to text before and after the CRF for clinical concept extraction. One important step is to convert the abbreviations into standard expressions before feeding the text into the CRF. Boag et al. [30] proposed CliNER which is a lightweight system for clinical concept extraction. CliNER first used a linear-chain CRF to identify concept boundaries and then used support vector machines to classify the identified concepts to various clinical types. Boag et al. also expanded their work to utilize the LSTM model for the same task [31]. Wu et al. [32] utilized a Deep Neural Network (DNN) working on embeddings to predict four different clinical entities within the clinical text. The results show that the DNN based approach works better than the traditional CRF approach. Chalapathy et al. [33] employed the BiLSTM model with the CRF layer for clinical concept recognition. The Glove model was used to generate word embeddings to feed into the model for the labeling process of the word sequence. Their results showed that BiLSTM with CRF worked better than the compared methods. Steinkamp et al. [34] employed BiLSTM with CRF for symptom extraction and compared it with a BERT-based approach. They achieved 86.4% F1 on their data set. Yang et al. [35] compared different BERT-based architectures for extracting various types of clinical concepts using 3 public datasets. Si et al. [36] explored three different embedding methods with a BiLSTM architecture for different biomedical concept extraction. Li et al. [37] first pre-trained the BERT model on the unlabeled Chinese clinical records, then adopted BiLSTM with CRF layer for Chinese clinical named entity recognition. Based on the literature, the neural network-based approach with the word embeddings can automatically learn the relations between words, thereby avoiding complex feature design phases, which sometimes involve professionals specifying rules. The neural network-based NER models have shown better performances compared to traditional methods on clinical concept extraction.

To our knowledge, this work is the first to investigate a novel NER model to automate the patient-reported symptom extraction from clinical text. Our NER model is the first to incorporate the syntactic dependencies between words for symptom extraction.

III. Methods

Our method is a NER model, which assigns a label to each word to identify symptom words. BIOE labels are commonly used for sequence labeling for the NER tasks. B labels the beginning word of a symptom, I labels the interior words of a symptom, O labels the non-symptom words, and E labels the ending word of a symptom. For example, given a sentence “The patient complains neck pain.”, words ‘The’,’patient’, ‘complains’ are labeled as O, word ‘neck’ is labeled as B indicating the beginning of a symptom, and word ‘pain’ is labeled as E indicating the end of a symptom.

Figure 1 shows the overall model architecture of our method. An input sentence passes through a syntactic analysis component - SynSymptom and a semantic analysis component - DeepSymptom. The final decision is made by integrating the output of both components. The following sections explain the details of each component and the integration process.

SynSymptom takes a sentence as a word sequence and feeds it into a multilayer perceptron (MLP) network. The syntactic information of each word and syntactic dependencies between the words are explicitly encoded into the word representation. So that it can capture the syntactic word relations in the long symptom expression, including the body location and severity of the symptom, such as ‘mild left shoulder pain’. Where ‘shoulder pain’ is a noun phrase with compound dependency between ‘shoulder’ and ‘pain’, ‘mild’ is an adjective modifier to ‘pain’, and ‘left’ is an adjective modifier to ‘shoulder’. The DeepSymptom is a transformer-based name entity recognition (NER) model designed to extract symptoms as entities. Although the transformer-based NER model shows more advantages than the typical recurrent neural network-based models on the NER task, they both have limitations on extracting long symptom expression. Hence, the SynSymptom is designed to overcome this limitation.

A. SynSymptom

SynSymptom utilizes a multilayer perceptron (MLP) network and works on the one-hot syntactic embedding representations of the words in a sentence. The syntactic embedding explicitly encodes the dependencies between the word and the neighbor words in the sentence and the part-of-speech (POS) tags of the word and the neighbor words. The MLP network predicts the NER tag of each word. There are several fully connected hidden layers between the input and the output layer. The left part of Figure 1 shows the model architecture.

Syntactic parsing is a method by which sentences are tokenized, POS of the words are tagged. The sentence is also converted into a dependency tree structure that exhibits the associations among tokenized words administered by syntax standards. The syntactic dependencies between words are generated along with the POS tagging. Figure 2 shows an example of the dependency tree. A dependency tree is a directed acyclic graph with nodes representing the words $S\{w_0, w_1, \ldots, w_n\}$ and edges representing the dependencies $E\{e_1, e_2, \ldots, e_3\}$ between words. Each word has a Part-Of-Speed (POS) tag, such as proper noun (NN), verb (VBD), adjective (JJ), etc. Each edge has a property that specifies the dependency between two words, such as adverb modifier, conjunct, etc. If the word modifies another word, it is an outgoing edge. Likewise, if another word modifies the word, then there is an incoming edge. In this work, Stanford CoreNLP [38] is used to generate the dependency tree.

Based on the syntactic dependency tree, we generate the one-hot embedding V with m dimension to present each word in a sentence. For each word, we consider the POS tags of the word itself and the l neighbor words to its left and right, as well as all outgoing and incoming dependency edges. m is the total types of POS tag and dependency edge in the corpus. In V, the entries corresponding to the POS tags of the word, the neighbors of the word, and all edge types connected to the word are set to be 1. The rest of the entries are set to be 0. Table I shows an example of generated one-hot embedding for each word in the sentence shown in Figure 3. In this example, only one neighbor word to the left and right of each word is considered. The word ‘caught’ has a left neighbor word ‘James’ and a right neighbor word ‘cold’, and five outgoing dependency edges. So the corresponding entries, such as ‘NNP’, ‘VBD’, ‘JJ’, ‘out_obj’ and ‘out_nobj’ etc., are set to be 1, and the rest of the entries are set to be 0. The advantage of this encoding is that it considers the POS tagging, syntactic dependencies between the words explicitly to identify the words that are part of the symptoms.

After generating the syntactic one-hot embedding of each word in the whole corpus, the MLP takes each one-hot embedding and predicts the probability of the NER tag. The ReLu activation function is used to produce the output of each layer and forward it to the next layer. The softmax function of the output layer is used to predict the BIOE tag.

B. DeepSymptom

Other than the syntactic analysis for symptom extraction, we also investigate the most recent language model - BERT [18] for symptom extraction. BERT has been used for NER tasks in various domains [37], [39]-[41]. In this research, the DeepSymptom is composed of a BERT model with a token-level classifier on top followed by a Linear-Chain CRF, as shown in Figure 1. Given input sequence with n words, BERT first converts it into m tokens, then outputs an encoded token sequence with dimension H. The classification layer then classifies the token representations to classes of tags, i.e., $\mathbb{R}^H \mapsto \mathbb{R}^K$, where K is the number of tags and depends on the tagging scheme of NER. The output scores of the classification model are then fed to the Conditional Random Field (CRF) layer. The CRF layer is implemented to understand the context of the documents, and the neighboring values influence the prediction of the current value. It is observed that when CRF is combined with a baseline NER model, good confidence and efficiency are achieved [42]. Given an input sequence x_1, x_2, x_3,…, x_n in X, the CRF model of tag sequence as output y_1, y_2, y_3, , …, y_n in Y. By training the model parameters, the CRF model predicts the conditional probability of Y using the equation 1. The model calculates the conditional probability through normalization factor $Z(x)$, eigenfunctions specified on transfer feature t_k and state feature s_1. The values λ_k and μ_1 are the weights assigned to t_k and s_1 respectively. If the characteristic condition is satisfied, then the transfer feature and state feature values are 1; otherwise, it is 0.

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,\mathrm{k}} \lambda_\mathrm{k} t_\mathrm{k}\left(y_{i-1}, y_i, x, i\right) + \sum_{i,1} \mu_1 s_1\left(y_i, x, i\right)\right) \tag{1}$$

The BERT model is trained with the CRF layer jointly during training. In the experimental section, we also compare results without the CRF layer, in which the BERT model is optimized by minimizing the cross-entropy loss. Based on our comparison of three different BERT models on NER. The BioBERT performs better than the others, so it is used with the CRF layer for prediction.

C. Integration of SynSymptom and DeepSymptom

To integrate the symptom extraction output from both models, we utilize an n-gram based approach. Based on the initial evaluation of both models, it is found that DeepSymptom works better on extended symptom expressions that have more than three words, such as “difficulty in memorizing words”. Whereas DeepSymptom works much better on short symptom expressions that have three or less than three words. Hence, the integration process first checks the length of the extracted symptoms by words. If the symptoms identified by DeepSymptom are 1-, 2- or 3-grams, it is added to the final extracted symptom list. If the symptoms identified by SynSymptom or DeepSymptom have more than three words (3+ grams), it is added to the final extracted symptom list. If any extracted symptom is a substring of another extracted symptom, it is removed from the final list. For example, if ‘pain in arm’ and ‘pain’ are both in the final list, ‘pain’ is removed. Algorithm 1 shows the detailed process of the integration.

Algorithm 1: Model Integration

```
Sentence Symptoms Symptoms = ∅;
$S_1$ = DeepSymptom(Sentence);
$S_2$ = SynSymptom(Sentence);
$S_a = S_1 \cup S_2$;
for $s\ in\ S_1$ do
    if $(length(s) < \ = 3)$ then
        $Symptoms = Symptoms \cup \{s\}$;
for $s\ in\ S_a$ do
    if $(length(s) > 3)$ then
        if $s\ is\ not\ a\ substring\ of\ s' \in S_a$ then
            $Symptoms = Symptoms \cup \{s\}$;
```

Since SynSymptom and DeepSymptom components can be trained in parallel and then integrated after the models are trained, the efficiency of the system is determined by the efficiency of each component. The efficiency of SynSymptom is higher than DeepSymptom since it only utilizes MLP. The DeepSymptom has similar efficiency as the base BERT with a CRF layer for NER tasks. Hence, the efficiency of our model is comparable to the typical deep learning-based NER models.

IV. Experimental Setting and Results

This section presents the baseline methods, the creation and annotation of the dataset, and the experimental results.

A. Dataset Creation and Annotation

Our initial dataset consisted of 5000 clinical notes randomly extracted from different types of clinical notes, including medical history, physical exam, and consulting notes, of patients diagnosed with breast cancer, colorectal cancer, or respiratory diseases. The patient illness sections from the 5000 clinical notes are extracted. It is worth noting that not all clinical notes have the patient illness section. Because the clinical notes are not well structured, and some expressions or sentences do not end with proper punctuation. We selected sentences with more than five words and less than one hundred words from the extracted patient illness sections. In total, we selected 1,693 sentences for annotation.

Two annotators labeled the data. The Cohen’s kappa value was calculated to measure inter-rater reliability. The Cohen’s kappa value between the two annotators is 0.89. Conflicts were resolved using a consensus vote from a third annotator. In this research, we considered the positive symptoms relevant to patients as (1) direct symptom mention (“Pt here c/o pain above and into the armpit on her left side.”) and (2) complaint statement (“Feeling tired or poorly since the last PM”). The following situations are considered as negative or “not true” symptom mentions: (1) negation statements (“Not feeling tired or poorly, not tiring easily, and no lethargy”) (2) conditional statements (“If you have chest pain, return to the hospital”), (3) statements for the medication (“Take Tylenol as needed for fever”), (4) resolved symptoms (“had some chest tightness overnight resolved with neb treatment”) (5) statements that only describe the existence of symptom-related event (“he is allergic to aspirin-that it causes itchy spots on his head,”), as these statements do not describe the symptoms patient experiencing. We noticed that the resolved symptoms mainly cause the inconsistency between the two annotators.

After annotation, 755 sentences among the 1,693 sentences have no or “not true” symptoms, and 938 sentences have one or more positive symptom mentions relevant to the patient. The distribution of the symptoms based on n-grams is shown in Figure 3. In our data set, there are more symptoms in 1- and 2-grams, fewer symptoms in 3- and more than 3-grams.

B. Baseline Methods

In the domain of medical informatics, there are existing tools that can be used to identify specific symptoms from the clinical text, such as UMLS MetaMap [9]. On the other hand, other existing NER models can also be applied to extraction symptoms, such as BiLSTM with CRF layer, etc. In this research, we compare our model with the following baseline methods.

- **UMLS MetaMap** [9]: The UMLS Metamap is an NLP tool that uses various sources to categorize the phrases or terms in the text to different semantic types. MetaMap are often used to extract information from clinical notes and biomedical text [43]-[45]. The UMLS MetaMap can also identify the abbreviations of medical terminologies, detect negation, and be configured to use word sense disambiguation (WSD). In this research, three semantic types - Sign or Symptom, Physiologic Function, and Mental or Behavioral Dysfunction are used to identify the symptoms from the text, which means any term tagged by MetaMap to one of these three semantic types with no negation detection are extracted as symptoms.

- **BiLSTM+CRF** [46]: As mentioned in the related work, BiLSTM is capable of classifying data. When it is combined with a CRF layer, a strong performance is observed on the task of NER. Much recent work compared BiLSTM with CRF against other models for clinical concept extraction [33], [47]. Although it has not been applied to symptom extraction, we investigate it as another baseline model for comparison. The BioWord2Vec [48] is used to generate word embeddings to feed into the model. BioWord2Vec has been pre-trained on 10M PubMed abstracts.

- **BERT** [18]: The BERT model itself can also be used for NER. Much previous research demonstrated that BERT performs well on a variety of natural language tasks in the clinical domain, including concept extraction [35], [36]. BERT uses WordPiece tokenization which could split a single word into multiple tokens [18]. If one token of a word is predicted as a symptom, the word is considered a positive prediction. The base BERT model was pre-trained by various research groups using PubMed data and/or clinical notes of MIMIC-III

[49]. The BioBERT [50] published in 2020 is a pre-trained BERT model using English Wikipedia, BooksCorpus, PubMed abstracts (PubMed), and PubMed Central full-text articles (PMC). The ClinicalBERT [39] published in 2019 is initialized from base BERT and pre-trained using all note types of MIMIC-III. The BlueBERT [51] published in 2019 is pretrained using PubMed abstracts and clinical notes of MIMIC-III. We compared our model against these different BERT models.

C. Evaluation Metric

Various matching techniques can be used to evaluate the performance of a NER model. In this research, we considered the exact match and relaxed match metrics.

- **Exact match** considers the prediction is correct if the model prediction matches all words of each symptom annotated in a sentence. We calculated the precision and recall for an exact match based on the extracted symptoms from each sentence.
- **Relaxed match** considers a partial match between the annotated symptoms and predictions and provides partial matching scores. It is similar to the fragment match in the related research about biomedical named entity recognition [52] [53]. For the relaxed match, all extracted symptoms are converted into words. Then, we calculated the word-based precision and recall.

Given the sentence “Patient has joint pain in past few weeks”, the annotated symptom is “joint pain”. If only “pain” is extracted by a model, both precision and recall are 0 when an exact match metric is used. When the relaxed match metric is used, recall is 0.5, and precision is 1.

D. HyperParameter Setting

We split our dataset into training, validation, and test sets with percentage values as 80%, 10%, and 10%, respectively. For SynSymptom, the number of layers for the MLP is set to be 3. There are 200, 100, and 40 neurons in each layer, respectively. For DeepSymptom, we used the BioBERT [50] model since our preliminary results show that it works better than ClinicalBERT and BlueBERT on symptom extraction.

E. Symptom Extraction Performance

We compared our symptom extraction model to the baselines mentioned above in Tables II. Our model outperforms all the previous baselines in all precision, recall, and F1-measure performance regarding both exact match and relaxed match evaluation metrics.

These results confirm that UMLS MetaMap has the lowest F1 value since it performs based on UMLS metathesaurus, a biomedical thesaurus organized by concepts and links to similar names from nearly 200 different vocabularies. The vocabulary-based approach has limitations on recognizing symptoms expressions that are not included in the vocabulary. For example, the symptom ‘loss of appetite’ is recognized by MetaMap as a term that belongs to the semantic category ‘Disease or Syndrome’ and ‘Findings’ instead of ‘Sign and Symptoms'. The deep language models perform better than the UMLS MetaMap, with the advantages of identifying new symptoms. BiLSTM with CRF performs better than the different BERT-only based models. When a relaxed match metric is used, the BiLSTM with CRF performs closely to our model. However, our model has advantages in identifying the complete symptom representations (e,g,'numbness in fingertips and toes') instead of partial representations (e,g, 'numbness'). The performances of using different pre-trained BERT models show that BioBERT works better than the ClinicalBERT and BlueBERT on symptom extraction. Comparing the two components of our models with the baselines, SynSymptom performs better than UMLS MetaMap and performs similarly to the BERT-only based models. DeepSymptom works better than all the other models based on the overall recall, precision, and F1. After further analysis, we found that SynSymptom works better on those long and complex symptoms since it utilizes the syntactic dependencies between words in a sentence.

Table III shows various symptoms that can be extracted by our model and baseline models. The ✓ means the exact match metric is used for the extraction. These examples show that through integrating syntactic and semantic analysis, our model can recognize the long symptom expressions and symptoms that are not in the UMLS metathesaurus. The BiLSTM with CRF layer and BioBERT model can also identify symptoms with up to 3 or 4 words. Nevertheless, they cannot recognize the symptoms with more than four words. The results explain that because the percentage of the symptoms with more than three words in our dataset is low, the overall performance of our model and BiLSTM with CRF are close.

We also investigated the cases that our model cannot correctly recognize. We found that our model has a limitation on recognizing some abbreviation symptoms and hyphenated words relevant to symptoms and could not identify some 'not true' cases. For example, given a sentence "recently discharged from WMH 7/16 after being hospitalized for the same symptoms who re-presents with N/V/ abd pain", the labeled symptoms are 'N/V/' and 'abd pain'. Our model can only recognize 'pain' as a symptom. The abbreviations 'N/V/' for 'Nause/Vomitting/' and 'adb' for 'abdominal' cannot be recognized. One reason is that the words 'N/V/' and 'abd' are not included in our training data. The other reason is that the complex compound relations between 'N/V/', 'abd' and 'pain' are not recognized by the SynSymptom because there is a very minimum number of such cases in the training data. Similarly, the symptom 'left-sided shoulder pain' in the sentence 'The patient presents to the pain clinic with a history of left-sided shoulder pain which radiates to the left upper lateral arm.' is not extracted by our model. Only the 'shoulder pain' is extracted. The syntactic dependency between 'left-sided' and 'should pain' is not identified by SynSymptom, since there is punctuation dependency between the hyphen and word 'sided'. Given a 'not true' case, the 'nausea' is not annotated as a symptom in sentence 'Nausea (ACTIVE) ICD9 787.02 resolved seen in ED for UTI- treated.' since it is resolved. However, our model extracted it. In this case, there is no sentence structure and direct syntactic dependency between the 'nausea' and 'resolved'. It is not identified as a 'not true' case by our model. We further categorized the test data into three categories: 'not true' symptom, true symptom, and no symptom. The precision and recall of the 'not true' category are 0.56 and 0.65, the precision and recall of the true category are 0.93 and 0.89, and the precision and recall of the no symptom category are 0.83 and 0.83. The quantitative analysis shows that our model has some limitations on identifying the ‘not true’ cases.

F. Model Analysis

Our model has both syntactic and semantic components - SynSymptom and DeepSymptom. We evaluated the effect of each component. We disabled either SynSymptom or DeepSymptom to extract symptoms of various lengths based on the number of words (n-grams). Table IV shows the performances of both components on n-grams using the exact match metric on the test data and the comparison with the baseline models. Instead comparing all different BERT models, we selected BioBERT for this comparison, since it shows slightly better performance than ClinicalBERT and BlueBERT (shown in Table II). The results show that transformer-based models work better than the BiLSTM with CRF and SynSymptom on 1-, 2-, and 3-grams. This analysis shows that DeepSymptom performs better than SynSymptom on short symptom expression, such as 1-, 2-, 3-grams. When the symptom expression gets longer, such as 4- and 5-grams, SynSymptom works better than DeepSymptom. The reason is that the SynSymptom considers syntactic dependencies between words. This analysis confirms the contributions of the components. However, because our data set does not have a large amount of 4+ grams symptoms, the current design of the integration of the components is taking predictions from both components equally.

V. COVID-19 Symptom Extraction from Social Media

The World Health Organization (WHO) declared COVID-19 a global pandemic started on March 11, 2020 [54]. Individuals who contracted the COVID-19 could have mild or severe symptoms leading to hospitalization or death [55]. Some research [56] suggested that symptoms could be used as a screening tool to identify people with potential mild cases who could be recommended to self-isolate. Other research showed that self-reported symptoms could be used to predict potential COVID-19 [57]. Various research investigated self-reported symptoms of COVID-19 using social media data [58]-[61]. These previous researchers either used regular expression or predefined keywords to extract symptoms or classified tweets with symptom mentions. To our best knowledge, none of the existing research investigated a deep language model to extract symptoms from tweets.

A. Dataset

Individuals and organizations use Twitter to distribute information [62], [63] or express opinions or feelings [64]. During the pandemic outbreak of COVID-19, Twitter has been a helpful resource for researchers and people to understand the global health crisis [65]-[67].

This research employed the proposed deep language model to extract symptom mentions from the COVID-19 related tweets published between eight weeks from March 15, 2020, to May 09, 2020. The COVID-19 Twitter chatter dataset published by Banda et al. [68] is used. The tweets were collected using a set of COVID-19 related keywords: “COVD19”, “CoronavirusPandemic”, “COVID-19”, “2019nCoV”, “CoronaOutbreak”, “coronavirus”, “WuhanVirus”. The clean version of the COVID-19 Twitter chatter dataset is used. Since only tweet IDs are included in the data set, we used the Twitter API and tweet IDs to extract the actual tweets. In total, there are 2,841,624 tweets extracted from the period. The total number of tweets by each week is shown in Figure 4.

B. Results

We applied our trained model to all tweets for symptom extraction. Then, we sampled 200 tweets to be labeled by two annotators. The Cohen's kappa value between the two annotators is 0.78. Then, a third annotator is involved in solving the inconsistency between the annotation. We found that most of the inconsistency is caused by the cases whether they describe the symptom a patient experiences. For example, given 'There s a lot of anxiety and uncertainty going around with the pandemic so we want to make informed as easy and stress free as possible #COVID19.', one annotator labeled 'anxiety' as the symptom. In contrast, the other annotator labeled none symptom in this tweet since it does not seem to be a symptom experienced by the tweet writer.

Among the 200 tweets, 49 are labeled with symptoms. The rest are labeled as without symptoms or 'not true' symptoms. We applied our model to the 200 sample tweets to understand the model performance on social media data. When the relaxed match metric is used, the precision, recall, and F1 on the tweets with symptoms are 0.82, 0.9, and 0.86, respectively. We found that our model cannot recognize some of the symptoms mentioned in the language used in social media. For example, given a tweet 'It's so scary all this with I can't even believe it I have fever and headache my nose drop also afraid that I have corona #Corona #COVID19.', the labeled symptoms are 'fever', 'headache', and 'nose drop'. Our model can extract 'fever and headache', but not 'nose drop'. The symptom 'nose drop' is not typically written in the clinical notes.

Using our model, we found that 51,439 tweets have symptom mentions, as shown in Figure 4. At the beginning of the pandemic, there were a lot more tweets that mention symptoms. We think it is because of many unknowns about the disease, and people talk more about symptoms. The number of tweets about COVID-19 increased significantly from the end of March 2020 but then dropped to around 5000 per week. Examples of tweets with symptom mentions (highlighted in orange) extracted by our model are presented in Figure 5. Some of the symptoms are suggested by the CDC [69] or mentioned in the literature. We extracted all symptoms and analyzed the symptoms trends over the eight weeks.

Figure 6 shows the percentage of each symptom listed by the CDC by each day. Since one tweet can contain multiple symptoms, the added sum of the percentages is more than 1. Users can mention the same symptom in various ways. For example, "muscle ache" can be mentioned as "muscle pain". We first generated semantic embeddings for all extracted symptoms using universal sentence encoder [70], then used the CDC symptoms as seed terms to identify the various symptom representations. If a CDC symptom is defined as two terms using the word 'or', it is treated as multiple seed symptoms. For example, the symptom "fever or chill" is treated as two seed symptoms "fever" and "chill". Cosine similarity is used to identify the additional symptom representations to the seed symptoms. If the cosine similarity between an extracted symptom and a CDC seed symptom is larger than threshold θ, the extracted symptom is counted as the corresponding CDC symptom. After a few experiments on θ, we set θ to be 0.8. For example, the cosine similarity between extracted symptom “higher fever” and CDC symptom “fever” is larger than 0.8, “higher fever” is counted as “fever”. Based on Figure 6, symptoms of “cough” and “fever or chill” are the top ones mentioned in the tweets at the beginning of the pandemic till the beginning of April. The symptoms “muscle or body ache”, “shortness of breath or difficulty breathing” and “sore throat” are also mentioned by many users at the beginning of the pandemic, although “muscle or body ache” and “sore throat” are suggested by CDC in late April [61]. Starting from the beginning of April, many tweets mention “loss of taste or smell” added by CDC in late April. Although the symptoms “fatigue”, “diarrhea”, “congestion or running nose”, “nausea or vomiting” are additional COVID-19 symptoms suggested by CDC in May [61], all were mentioned in tweets since the beginning of the pandemic.

Other than the CDC symptoms, we investigated the top frequent symptoms, not on the CDC list. After using semantic embeddings to group the symptoms with similar meanings, we listed nine frequent ones in Figure 7. Many of these symptoms are more relevant to mental health or illness, reflecting the prevalence of mental illness symptoms during the pandemic, which needs to be addressed by the community during and post-pandemic [71], [72]. Through extraction all the symptoms from the tweets, we also found some less frequent symptoms, such as “blurred vision” [23], “hearing loss” [73], and “hair loss” [74] mentioned together with the symptoms listed by CDC.

The results show that our deep language model can be applied to social media text for symptom extraction. The longitudinal social media data can provide an overview of the real-time symptoms mentioned by the public [75]. Our method can also identify the typical and rare symptoms noted by the public before CDC suggests them.

VI. Conclusions, limitations and Future Work

This research proposed and evaluated a novel deep language model to extract patient report symptoms from clinical text and social media. This new language model contains two components: one analyzes the syntactic dependency between words for symptom extraction - SynSymptom, and the other analyzes the semantic relations between words for symptom extraction - DeepSymptom. Two components are integrated based on the n-grams of the extracted symptoms. We illustrated the effectiveness of the symptom extraction model by comparing it with different baseline NER models applicable for symptom extraction. our model outperformed the baselines. The model analysis shows the contributions of each component of our model. Finally, we applied our model to extract symptoms from a COVID-19 tweets data set. The results showed that our language model could successfully identify both prevalent and rare symptoms mentioned by the public, and some symptoms are identified before suggested by CDC.

The limitations of our study include: (1) The size of our annotated clinical notes is limited. it could not cover all various complex symptoms. (2) Although we sampled a number of tweets to validate the performance of our model on social media data, it is not feasible to validate all the tweets with symptoms identified by our model. Other than the symptoms mentioned in the literature, the rare symptoms still need human validation.

Future work includes enriching training data to include various complex symptom representations to improve our model performance further and link our symptom extraction models with other clinical attributes to assist disease prediction and clinical decision support.

Acknowledgment

The authors acknowledge Regenstrief Institute, Inc.'s participation in this project.

This work was supported by the grants from NIH (1R15GM139094-01A1) (PI Luo), the Oncology Nursing Foundation (RE01) (PI Storey), and the Indiana University Simon Cancer Center – Cancer Prevention and Control Pilot Grant (PI Storey).

References

[1]. Molarius A and Janson S, "Self-rated health, chronic diseases, and symptoms among middle-aged and elderly men and women," Journal of clinical epidemiology, vol. 55, no. 4, pp. 364–370, 2002. [PubMed: 11927204]

[2]. Carfì A, Bernabei R, Landi F et al. , "Persistent symptoms in patients after acute covid-19," Jama, vol. 324, no. 6, pp. 603–605, 2020. [PubMed: 32644129]

[3]. Felix HC, Seaberg B, Bursac Z, Thostenson J, and Stewart MK, "Why do patients keep coming back? results of a readmitted patient survey," Social work in health care, vol. 54, no. 1, pp. 1–15, 2015. [PubMed: 25588093]

[4]. Miaskowski C, "Symptom clusters: establishing the link between clinical practice and symptom management research," 2006.

[5]. Numico G, Cristofano A, Mozzicafreddo A, Cursio OE, Franco P, Courthod G, Trogu A, Malossi A, Cucchi M, Sirotovà Z et al. , "Hospital admission of cancer patients: avoidable practice or necessary care?" PloS one, vol. 10, no. 3, p. e0120827, 2015. [PubMed: 25812117]

[6]. Lanctôt KL, Amatniek J, Ancoli-Israel S, Arnold SE, Ballard C, Cohen-Mansfield J, Ismail Z, Lyketsos C, Miller DS, Musiek E et al. , "Neuropsychiatric signs and symptoms of alzheimer's disease: New treatment paradigms," Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 3, no. 3, pp. 440–449, 2017. [PubMed: 29067350]

[7]. Li J, Chen Z, Nie Y, Ma Y, Guo Q, and Dai X, "Identification of symptoms prognostic of covid-19 severity: multivariate data analysis of a case series in henan province," Journal of medical Internet research, vol. 22, no. 6, p. e19636, 2020. [PubMed: 32544071]

[8]. Díaz LA, García-Salum T, Fuentes-López E, Ferrés M, Medina RA, and Riquelme A, "Symptom profiles and risk factors for hospitalization in patients with sars-cov-2 and covid-19: A large cohort from south america," Gastroenterology, vol. 159, no. 3, pp. 1148–1150, 2020. [PubMed: 32437750]

[9]. Aronson AR, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001, p. 17.

[10]. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, and Chute CG, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010. [PubMed: 20819853]

[11]. Li F, Liu W, and Yu H, "Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning," JMIR medical informatics, vol. 6, no. 4, p. e12159, 2018. [PubMed: 30478023]

[12]. Emadzadeh E, Sarker A, Nikfarjam A, and Gonzalez G, "Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology," in AMIA Annual Symposium Proceedings, vol. 2017. American Medical Informatics Association, 2017, p. 679.

[13]. Dandala B, Joopudi V, and Devarakonda M, “Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks,” Drug safety, vol. 42, no. 1, pp. 135–146, 2019. [PubMed: 30649738]

[14]. Armengol-Estapé J, Soares F, Marimon M, and Krallinger M, “Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts,” Genomics & informatics, vol. 17, no. 2, 2019.

[15]. Cai X, Dong S, and Hu J, “A deep learning model incorporating part of speech and self-matching attention for named entity recognition of chinese electronic medical records,” BMC medical informatics and decision making, vol. 19, no. 2, pp. 101–109, 2019. [PubMed: 31138219]

[16]. Suárez-Paniagua V, Zavala RMR, Segura-Bedmar I, and Martínez P, “A two-stage deep learning approach for extracting entities and relationships from medical texts,” Journal of biomedical informatics, vol. 99, p. 103285, 2019. [PubMed: 31546016]

[17]. Nadeau D and Sekine S, “A survey of named entity recognition and classification,” Lingvisticae Investigationes, vol. 30, no. 1, pp. 3–26, 2007.

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

[19]. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, and Torr PH, “Conditional random fields as recurrent neural networks,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.

[20]. Tanabe L, Xie N, Thom LH, Matten W, and Wilbur WJ, “Genetag: a tagged corpus for gene/protein named entity recognition,” BMC bioinformatics, vol. 6, no. 1, pp. 1–7, 2005. [PubMed: 15631638]

[21]. Zeng D, Sun C, Lin L, and Liu B, “Lstm-crf for drug-named entity recognition,” Entropy, vol. 19, no. 6, p. 283, 2017.

[22]. Wang X, Yang C, and Guan R, “A comparative study for biomedical named entity recognition,” International Journal of Machine Learning and Cybernetics, vol. 9, no. 3, pp. 373–382, 2018.

[23]. Million M, Lagier J-C, Gautret P, Colson P, Fournier P-E, Amrane S, Hocquart M, Mailhe M, Esteves-Vieira V, Doudier B et al. , “Early treatment of covid-19 patients with hydroxychloroquine and azithromycin: A retrospective analysis of 1061 cases in marseille, france,” Travel medicine and infectious disease, vol. 35, p. 101738, 2020. [PubMed: 32387409]

[24]. Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, Williams BA, Defilippi C, Ebadollahi S, and Stewart WF, “Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record,” Journal of cardiac failure, vol. 20, no. 7, pp. 459–464, 2014. [PubMed: 24709663]

[25]. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, Roberts A, Dobson RJ, and Stewart R, “Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (criscode) project,” BMJ open, vol. 7, no. 1, 2017.

[26]. Gundlapalli AV, Divita G, Redd A, Carter ME, Ko D, Rubin M, Samore M, Strymish J, Krein S, Gupta K et al. , “Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing,” Journal of biomedical informatics, vol. 71, pp. S39–S45, 2017.

[27]. Divita G, Luo G, Tran L, Workman TE, Gundlapalli AV, and Samore MH, “General symptom extraction from va electronic medical notes.” Studies in health technology and informatics, vol. 245, p. 356, 2017. [PubMed: 29295115]

[28]. Jonnalagadda S, Cohen T, Wu S, and Gonzalez G, “Enhancing clinical concept extraction with distributional semantics,” Journal of biomedical informatics, vol. 45, no. 1, pp. 129–140, 2012. [PubMed: 22085698]

[29]. Fu X and Ananiadou S, “Improving the extraction of clinical concepts from clinical records,” Proceedings of BioTxtM14, pp. 47–53, 2014.

[30]. Boag W, Wacome K, Naumann T, and Rumshisky A, “Cliner: A lightweight tool for clinical named entity recognition,” AMIA joint summits on clinical research informatics (poster), 2015.

[31]. Boag W, Sergeeva E, Kulshreshtha S, Szolovits P, Rumshisky A, and Naumann T, “Cliner 2.0: Accessible and accurate clinical concept extraction,” arXiv preprint arXiv:1803.02245, 2018.

[32]. Wu Y, Jiang M, Lei J, and Xu H, “Named entity recognition in chinese clinical text using deep neural network,” Studies in health technology and informatics, vol. 216, p. 624, 2015. [PubMed: 26262126]

[33]. Chalapathy R, Borzeshi EZ, and Piccardi M, “Bidirectional lstm-crf for clinical concept extraction,” arXiv preprint arXiv:1611.08373, 2016.

[34]. Steinkamp JM, Bala W, Sharma A, and Kantrowitz JJ, “Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes,” Journal of biomedical informatics, vol. 102, p. 103354, 2020. [PubMed: 31838210]

[35]. Yang X, Bian J, Hogan WR, and Wu Y, “Clinical concept extraction using transformers,” Journal of the American Medical Informatics Association, 2020.

[36]. Si Y, Wang J, Xu H, and Roberts K, “Enhancing clinical concept extraction with contextual embeddings,” Journal of the American Medical Informatics Association, vol. 26, no. 11, pp. 1297–1304, 2019. [PubMed: 31265066]

[37]. Li X, Zhang H, and Zhou X-H, “Chinese clinical named entity recognition with variant neural structures based on bert methods,” Journal of biomedical informatics, vol. 107, p. 103422, 2020. [PubMed: 32353595]

[38]. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, and McClosky D, “The stanford corenlp natural language processing toolkit,” in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.

[39]. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, and McDermott MBA, “Publicly available clinical bert embeddings,” 2019.

[40]. Yan H, Deng B, Li X, and Qiu X, “Tener: adapting transformer encoder for named entity recognition,” arXiv preprint arXiv:1911.04474, 2019.

[41]. Liu M, Tu Z, Wang Z, and Xu X, “Ltp: A new active learning strategy for bert-crf based named entity recognition,” 2020.

[42]. Guo H, “Accelerated continuous conditional random fields for load forecasting,” in 2016 IEEE 32nd International Conference on Data Engineering (ICDE), 2016, pp. 1492–1493.

[43]. Gandhi P, Luo X, Storey S, Zhang Z, Han Z, and Huang K, “Identifying symptom clusters in breast cancer and colorectal cancer patients using ehr data,” in Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019, pp. 405–413.

[44]. Guo D, Li M, Yu Y, Li Y, Duan G, Wu F-X, and Wang J, “Disease inference with symptom extraction and bidirectional recurrent neural network,” in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018, pp. 864–868.

[45]. Guo D, Duan G, Yu Y, Li Y, Wu F-X, and Li M, “A disease inference method based on symptom extraction and bidirectional long short term memory networks,” Methods, vol. 173, pp. 75–82, 2020. [PubMed: 31301375]

[46]. Lample G, Ballesteros M, Subramanian S, Kawakami K, and Dyer C, “Neural architectures for named entity recognition,” 2016.

[47]. Zhu H, Paschalidis IC, and Tahmasebi A, “Clinical concept extraction with contextual word embedding,” arXiv preprint arXiv:1810.10566, 2018.

[48]. Moen S and Ananiadou TSS, “Distributional semantics resources for biomedical text processing,” in Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan, 2013, pp. 39–43.

[49]. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG, “Mimic-iii, a freely accessible critical care database,” Scientific data, vol. 3, p. 160035, 2016. [PubMed: 27219127]

[50]. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, and Kang J, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020. [PubMed: 31501885]

[51]. Peng Y, Yan S, and Lu Z, “Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets,” arXiv preprint arXiv:1906.05474, 2019.

[52]. Seki K and Mostafa J, “A probabilistic model for identifying protein names and their name boundaries,” in Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003. IEEE, 2003, pp. 251–258.

[53]. Tsai RT-H, Wu S-H, Chou W-C, Lin Y-C, He D, Hsiang J, Sung T-Y, and Hsu W-L, “Various criteria in the evaluation of biomedical named entity recognition,” BMC bioinformatics, vol. 7, no. 1, pp. 1–8, 2006. [PubMed: 16393334]

[54]. Archived: WHO Timeline - COVID-19, https://www.who.int/news/item/27-04-2020-who-timeline—covid-19.

[55]. Giorgi Rossi P, Marino M, Formisano D, Venturelli F, Vicentini M, Grilli R, and R. E. C.-. W. Group, “Characteristics and outcomes of a cohort of covid-19 patients in the province of reggio emilia, italy,” PLoS One, vol. 15, no. 8, p. e0238281, 2020. [PubMed: 32853230]

[56]. Spinato G, Fabbris C, Polesel J, Cazzador D, Borsetto D, Hopkins C, and Boscolo-Rizzo P, “Alterations in smell or taste in mildly symptomatic outpatients with sars-cov-2 infection,” Jama, vol. 323, no. 20, pp. 2089–2090, 2020. [PubMed: 32320008]

[57]. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, Ganesh S, Varsavsky T, Cardoso MJ, Moustafa JSE-S et al. , “Real-time tracking of self-reported symptoms to predict potential covid-19,” Nature medicine, vol. 26, no. 7, pp. 1037–1040, 2020.

[58]. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, and Yang Y-C, “Self-reported covid-19 symptoms on twitter: an analysis and a research resource,” Journal of the American Medical Informatics Association, vol. 27, no. 8, pp. 1310–1315, 2020. [PubMed: 32620975]

[59]. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, Liang B, Cai M, and Cuomo R, “Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with covid-19 on twitter: retrospective big data infoveillance study,” JMIR public health and surveillance, vol. 6, no. 2, p. e19509, 2020. [PubMed: 32490846]

[60]. Al-Garadi MA, Yang Y-C, Lakamana S, and Sarker A, “A text classification approach for the automatic detection of twitter posts containing self-reported covid-19 symptoms,” 2020.

[61]. Guo J-W, Radloff CL, Wawrzynski SE, and Cloyes KG, “Mining twitter to explore the emergence of covid-19 symptoms,” Public Health Nursing, vol. 37, no. 6, pp. 934–940, 2020. [PubMed: 32937679]

[62]. Lerman K and Ghosh R, “Information contagion: an empirical study of the spread of news on digg and twitter social networks,” 2010.

[63]. Romero DM, Meeder B, and Kleinberg J, “Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter,” in Proceedings of the 20th International Conference on World Wide Web, ser. WWW ’11. New York, NY, USA: Association for Computing Machinery, 2011, p. 695–704. [Online]. Available: 10.1145/1963405.1963503

[64]. Murthy D, Twitter. Polity Press Cambridge, 2018.

[65]. Park HW, Park S, and Chong M, “Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea,” J Med Internet Res, vol. 22, no. 5, p. e18897, May 2020. [Online]. Available: http://www.jmir.org/2020/5/e18897/ [PubMed: 32325426]

[66]. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, and Shah Z, “Top concerns of tweeters during the covid-19 pandemic: Infoveillance study,” J Med Internet Res, vol. 22, no. 4, p. e19016, Apr. 2020. [Online]. Available: http://www.jmir.org/2020/4/e19016/ [PubMed: 32287039]

[67]. Ferrara E, “What types of covid-19 conspiracies are populated by twitter bots?” First Monday, May 2020. [Online]. Available: 10.5210/fm.v25i6.10633

[68]. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, and Chowell G, “A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration,” arXiv preprint arXiv:2004.03688, 2020.

[69]. “Center for disease control and prevention.” [Online]. Available: https://cdc.gov

[70]. Cer D, Yang Y, Kong S.-y., Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C et al. , “Universal sentence encoder,” arXiv preprint arXiv:1803.11175, 2018.

[71]. Vindegaard N and Benros ME, “Covid-19 pandemic and mental health consequences: Systematic review of the current evidence,” Brain, behavior, and immunity, vol. 89, pp. 531–542, 2020.

[72]. Cullen W, Gulati G, and Kelly B, “Mental health in the covid-19 pandemic,” QJM: An International Journal of Medicine, vol. 113, no. 5, pp. 311–312, 2020. [PubMed: 32227218]

[73]. Satar B, “Criteria for establishing an association between covid-19 and hearing loss,” American Journal of Otolaryngology, 2020.

[74]. Goren A, Vaño-Galván S, Wambier CG, McCoy J, Gomez-Zubiaur A, Moreno-Arrones OM, Shapiro J, Sinclair RD, Gold MH, Kovacevic M et al. , “A preliminary observation: Male pattern hair loss among hospitalized covid-19 patients in spain–a potential clue to the role of androgens in covid-19 severity,” Journal of cosmetic dermatology, vol. 19, no. 7, pp. 1545–1547, 2020. [PubMed: 32301221]

[75]. YoussefAgha AH, Jayawardene WP, and Lohrmann DK, “Role of social media in early warning of norovirus outbreaks: a longitudinal twitter-based infoveillance,” in Proceedings of the International Conference on Data Science (ICDATA). The Steering Committee of The World Congress in Computer Science, Computer …, 2013, p. 1.

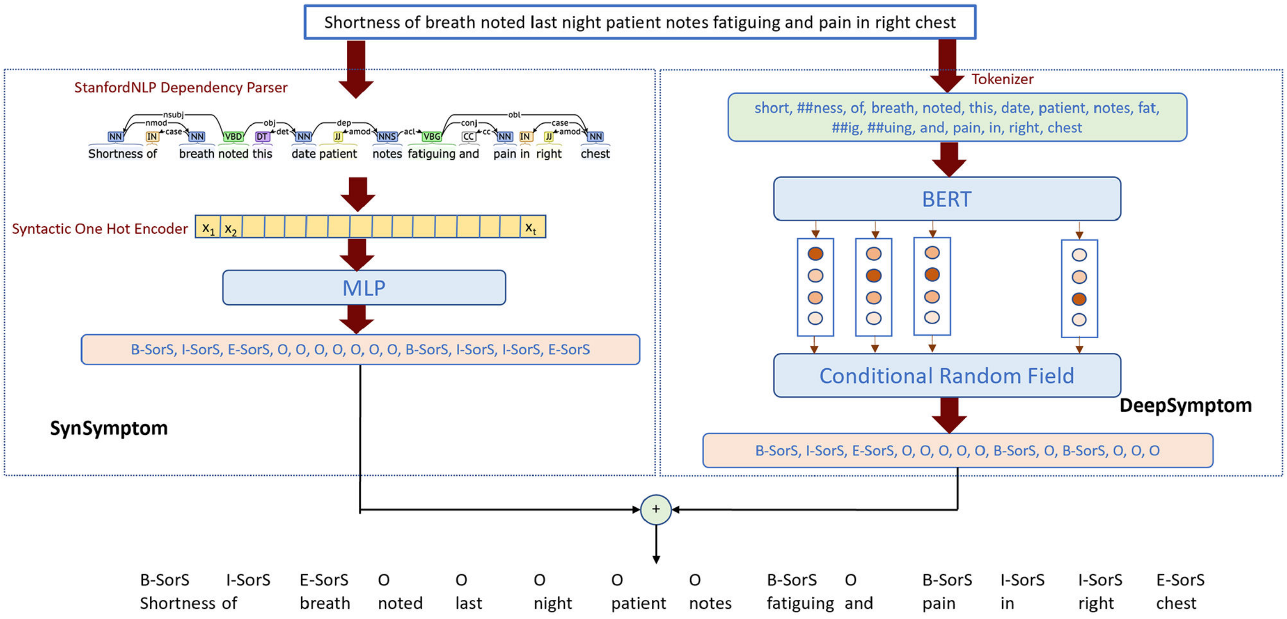

Shortness of breath noted last night patient notes fatiguing and pain in right chest
StanfordNLP Dependency Parser
nsubj
nmod
case
obj
det
dep
amod
acl
conj
cc
obl
case
amod
NN IN NN VBD DT NN JJ NNS VBG CC NN IN JJ NN
Shortness of breath noted this date patient notes fatiguing and pain in right chest
Syntactic One Hot Encoder
x_1 x_2 x_t
MLP
B-SorS, I-SorS, E-SorS, O, O, O, O, O, O, O, B-SorS, I-SorS, I-SorS, E-SorS
SynSymptom
Tokenizer
short, ##ness, of, breath, noted, this, date, patient, notes, fat, ##ig, ##uing, and, pain, in, right, chest
BERT
Conditional Random Field
DeepSymptom
B-SorS, I-SorS, E-SorS, O, O, O, O, O, B-SorS, O, B-SorS, O, O, O
+
B-SorS I-SorS E-SorS O O O O O B-SorS O B-SorS I-SorS I-SorS E-SorS
Shortness of breath noted last night patient notes fatiguing and pain in right chest

Fig. 1:

Model Architecture

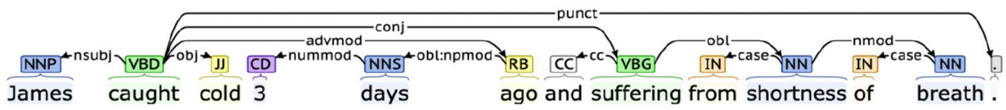


Fig. 2:
Dependency Tree

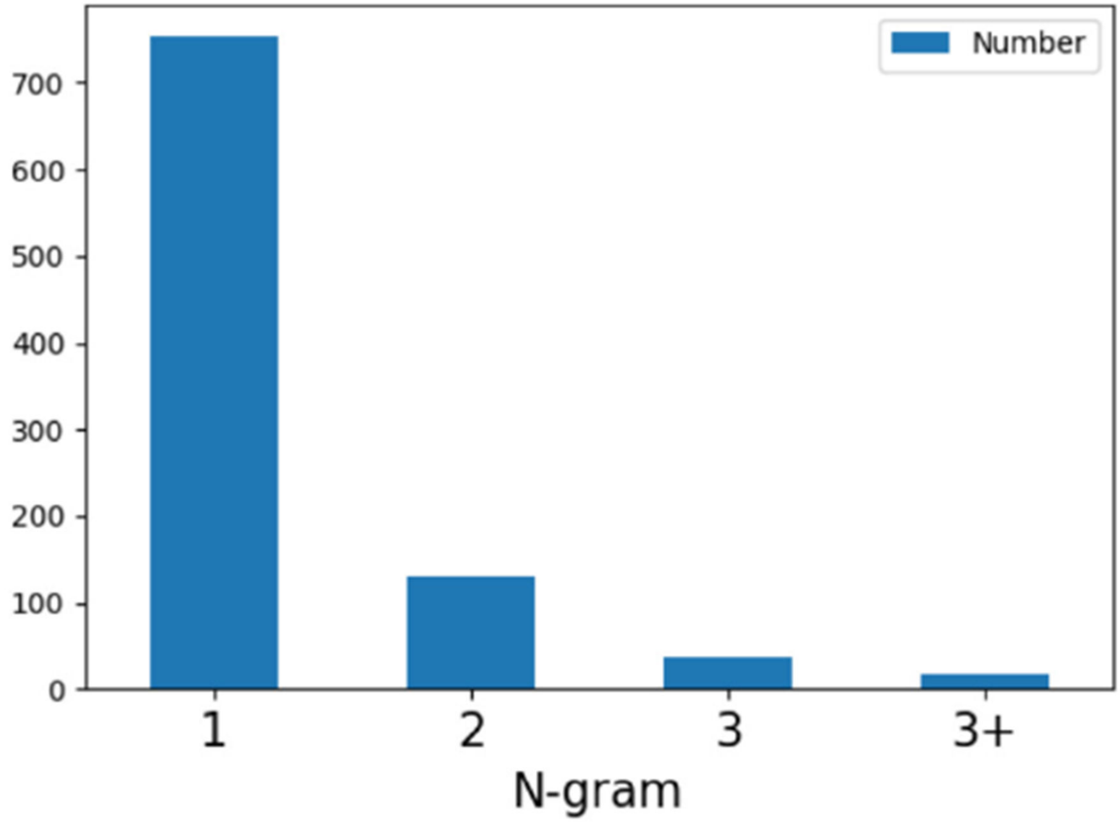


Fig. 3:
Symptom Distribution by N-grams

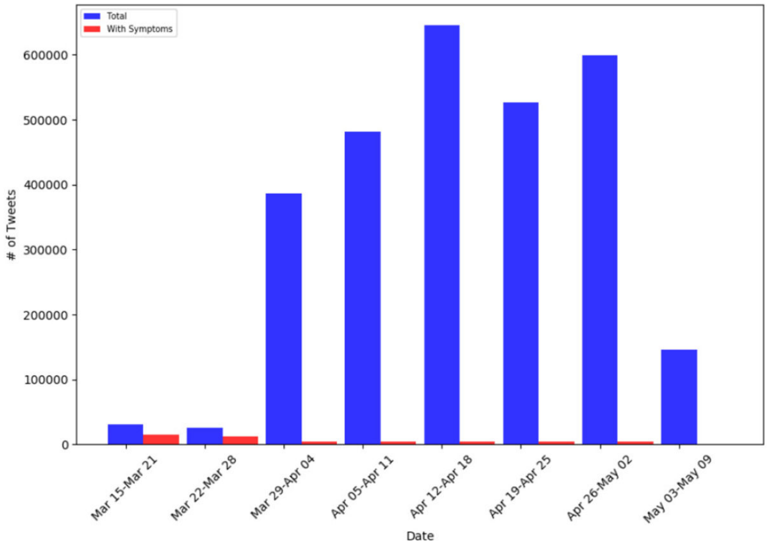
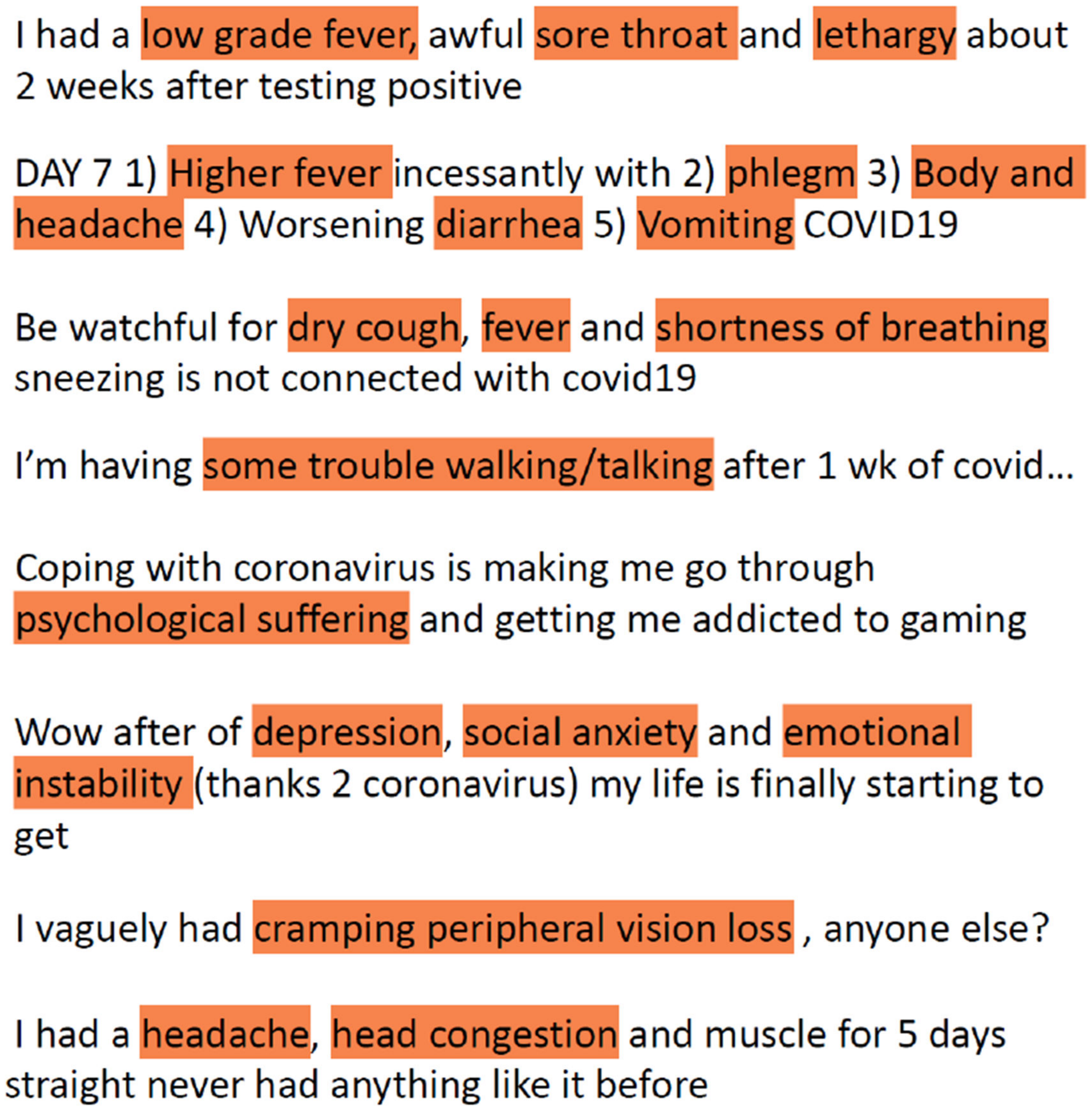


Fig. 4:

Number of Tweets by Date



I had a low grade fever, awful sore throat and lethargy about 2 weeks after testing positive

DAY 7 1) Higher fever incessantly with 2) phlegm 3) Body and headache 4) Worsening diarrhea 5) Vomiting COVID19

Be watchful for dry cough, fever and shortness of breathing sneezing is not connected with covid19

I'm having some trouble walking/talking after 1 wk of covid…

Coping with coronavirus is making me go through psychological suffering and getting me addicted to gaming

Wow after of depression, social anxiety and emotional instability (thanks 2 coronavirus) my life is finally starting to get

I vaguely had cramping peripheral vision loss , anyone else?

I had a headache, head congestion and muscle for 5 days straight never had anything like it before

Fig. 5:

Sample Tweets with Symptom Mentions

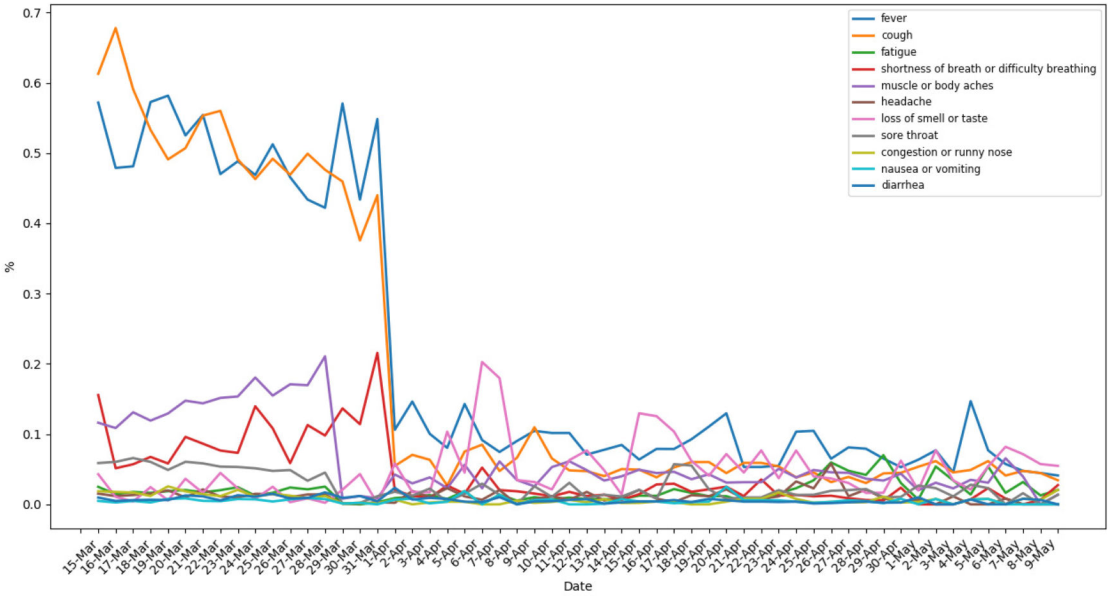


Fig. 6:

Trends of the Symptoms listed by CDC

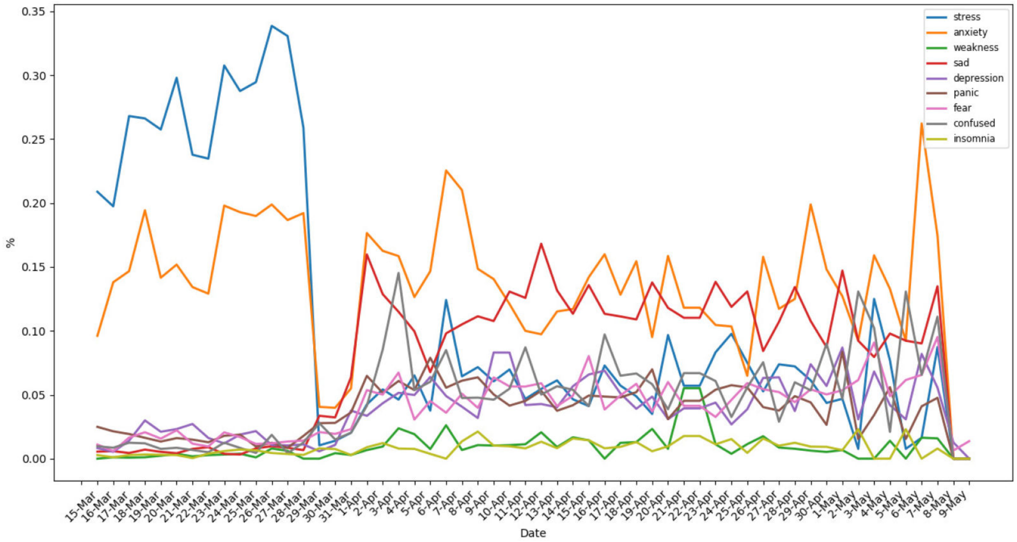


Fig. 7:

Trends of the Other Frequent Symptoms

TABLE I:

One-hot Embedding Generated based on the Dependency Tree and POS Tags

	POS Tags								**Dependency Edges**							
	NNP	**NNS**	**VBD**	**JJ**	**NN**	**IN**	**CD**	**…**	**in_obj**	**out_obj**	**out_nobj**	**in_nsubj**	**out_nmod**	**out_case**	**in_case**	**…**
James caught … shortness of breath	1	0	1	0	0	0	0	…	0	0	0	1	0	0	0	…
	1	0	1	1	0	0	0	…	0	1	1	0	0	0	0	…
	0	0	0	0	1	1	0	…	1	0	1	1	0	0	0	…
	0	0	0	0	1	1	0	…	0	0	0	0	1	0	0	…
	0	0	0	0	1	1	0	…	0	1	0	1	0	0	0	…

TABLE II:

Performance Comparison between our Model and Baseline Models

			UMLS MetaMap	BiLSTM+CRF	ClinicalBERT	BlueBERT	BioBERT	SynSymptom	DeepSymptom	Our Model
Exact Match	Training	Recall	0.48	0.84	0.79	0.78	0.86	0.82	0.91	**0.92**
		Precision	0.51	0.85	0.80	0.80	0.83	0.83	0.90	**0.90**
		F1	0.50	0.84	0.79	0.79	0.84	0.82	0.90	**0.91**
	Test	Recall	0.44	0.79	0.78	0.73	0.80	0.76	0.85	**0.86**
		Precision	0.46	0.80	0.80	0.74	0.77	0.75	0.85	**0.85**
		F1	0.50	0.80	0.79	0.74	0.79	0.75	0.85	**0.85**
Relaxed Match	Training	Recall	0.51	0.87	0.82	0.82	0.88	0.83	0.92	**0.92**
		Precision	0.54	0.89	0.84	0.85	0.87	0.86	0.93	**0.93**
		F1	0.52	0.88	0.83	0.85	0.88	0.84	0.92	**0.92**
	Test	Recall	0.49	0.86	0.81	0775	0.83	0.84	0.86	**0.86**
		Precision	0.51	0.87	0.83	0.78	0.81	0.84	0.87	**0.88**
		F1	0.50	0.86	0.82	0.76	0.82	0.84	0.86	**0.87**

TABLE III:

Examples to Demonstrate the Symptom Extraction Performances

Original Text	One of the labeled symptoms	UMLS Metamap	BiLSTM+CRF	BioBERT	Our Model
Patient feeling fatigue, having headaches with photophobia since she feels weak and also having chills.	photophobia		✓	✓	✓
She is having intermittent mild snoring plus 75 hypopneas and no apneas.	hypopneas		✓	✓	✓
The patient continues to have weight loss over the past week.	weight loss		✓	✓	✓
He presented to his PCP with weakness and loss of appetite as well as a 25-pound weight loss over the previous 6 months.	loss of appetite		✓	✓	✓
This is a patient with lethargy with nausea and vomiting and altered mental status for the past 2 days.	altered metal status	✓	✓	✓	✓
She believes that this activity increased her lower extremity discomfort related to her chemotherapy.	lower extremity discomfort			✓	✓
He presents with longstanding chest and center arm pains, which are exertional.	chest and center arm pains	✓	✓		✓
She continues to have some numbness in her fingertips and toes.	numbness in fingertips and toes				✓
Pain: 10/10 pain in lower back and abdomen at start and end of evaluation.	pain in lower back and abdomen				✓

TABLE IV:

Performance of Each Component and Baselines on N-grams

		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram+
UMLS MetaMap	Recall	0.20	0.10	0	0	0	0
	Precision	0.20	0.15	0	0	0	0
	F1	0.20	0.12	0	0	0	0
BiLSTM+CRF	Recall	0.67	0.57	0	0	0	0
	Precision	0.65	0.57	0	0	0	0
	F1	0.66	0.57	0	0	0	0
BioBERT	Recall	0.73	0.46	0.65	0	0	0
	Precision	0.71	0.50	0.65	0	0	0
	F1	0.72	0.48	0.65	0	0	0
SynSymptom	Recall	0.69	0.45	0.40	0.80	0.67	0.50
	Precision	0.62	0.47	0.40	0.80	0.67	0.50
	F1	0.65	0.46	0.40	0.80	0.67	0.50
DeepSymptom	Recall	0.77	0.55	0.80	0.20	0.50	0.25
	Precision	0.76	0.57	0.80	0.20	0.50	0.25
	F1	0.76	0.56	0.80	0.20	0.50	0.25