TRIAGE AND DIAGNOSIS OF COVID-19 FROM MEDICAL SOCIAL MEDIA

A PREPRINT

Abul Hasan
Department of Computer Science
Birkbeck, University of London
London WC1E 7HX UK
abulhasan@dcs.bbk.ac.uk

Mark Levene
Department of Computer Science
Birkbeck, University of London
London WC1E 7HX UK
mlevene@dcs.bbk.ac.uk

David Weston
Department of Computer Science
Birkbeck, University of London
London WC1E 7HX UK
dweston@dcs.bbk.ac.uk

Renate Fromson
Barnet General Hospital
Wellhouse Lane
London EN5 3DJ UK
rfromson4@gmail.com

Nicolas Koslover
Barnet General Hospital
Wellhouse Lane
London EN5 3DJ UK
nic.koslover@hotmail.co.uk

Tamara Levene
Barnet General Hospital
Wellhouse Lane
London EN5 3DJ UK
tamaralevene@gmail.com

July 16, 2021

ABSTRACT

Background: The COVID-19 pandemic has created a pressing need for integrating information from disparate sources, in order to assist decision makers. Social media is important in this respect, however, to make sense of the textual information it provides and be able to automate the processing of large amounts of data, natural language processing methods are needed. Social media posts are often noisy, yet they may provide valuable insights regarding the severity and prevalence of the disease in the population. In particular, machine learning techniques for triage and diagnosis could allow for a better understanding of what social media may offer in this respect.
Objective: This study aims to develop an end-to-end natural language processing pipeline for triage and diagnosis of COVID-19 from patient-authored social media posts, in order to provide researchers and other interested parties with additional information on the symptoms, severity and prevalence of the disease.
Materials and Methods: The text processing pipeline first extracts COVID-19 symptoms and related concepts such as severity, duration, negations, and body parts from patients' posts using conditional random fields. An unsupervised rule-based algorithm is then applied to establish relations between concepts in the next step of the pipeline. The extracted concepts and relations are subsequently used to construct two different vector representations of each post. These vectors are applied separately to build support vector machine learning models to triage patients into three categories and diagnose them for COVID-19.
Results: We report that macro- and micro-averaged F_1 scores in the range of 71-96% and 61-87%, respectively, for the triage and diagnosis of COVID-19, when the models are trained on human labelled data. Our experimental results indicate that similar performance can be achieved when the models are trained using predicted labels from concept extraction and rule-based classifiers, thus yielding end-to-end machine learning. Also, we highlight important features uncovered by our diagnostic machine learning models and compare them with the most frequent symptoms revealed in another COVID-19 dataset. In particular, we found that the most important features are not always the most frequent ones.
Conclusions: Our preliminary results show that it is possible to automatically triage and diagnose patients for COVID-19 from natural language narratives using a machine learning pipeline, in order to provide additional information on the severity and prevalence of the disease through the eyes of social media.

Key words: COVID-19, Triage and diagnosis, Medical social media, Natural Language processing, Conditional random fields, Support vector machines.

Introduction

Overview

During the ongoing coronavirus pandemic, hospitals have been continuously at risk of being overwhelmed by the number of people developing serious illness. People in the UK were advised to stay at home if they had coronavirus symptoms and to seek assistance through NHS helpline if they needed to [1]. Consequently, there is an urgent need to develop novel practical approaches to assist medical staff in the diagnosis and triage of patients. A variety of methods have been recently developed that involve natural language processing techniques, see for example [2, 3, 4]. In addition, social media search queries related to COVID-19 symptoms in China have been shown to be an effective predictor for the number of infections [5].

Herein, we propose an end-to-end *Natural Language Processing (NLP)* pipeline to automatically triage and diagnose COVID-19 cases from patient-authored medical social media posts. The triage may inform decision-makers about the severity of COVID-19, and the diagnosis could help in gauging the prevalence of infections in the population. A key concern is the production of a high-quality human labelled dataset on which to build our pipeline. In the following we give a brief overview of our pipeline and how we developed our dataset.

The first step in the pipeline is attained by developing an annotation application that detects and highlights COVID-19 related symptoms with their severity and duration in a social media post, henceforth collectively termed as *concepts*. During the second step relations between symptoms and other relevant concepts are also automatically identified and annotated. For example, *breathing hurts* is a symptom which is related to a body part *upper chest area*.

One author manually annotated our data with concepts and relations, allowing us to present posts highlighted with identified concepts and relations to three experts along with several questions, as shown in Figure 1. The first question asked the experts to triage a patient into one of the following three categories: Stay at home, Send to a GP, and/or Send to hospital. The second question asked to diagnose the likelihood of COVID-19 in a Likert Scale of 1 to 5 [6].

The three experts are foundation doctors working in the UK who were redeployed to work on COVID-19 wards during the first wave of the pandemic, between March and July 2020. Their roles involved the diagnosis and management of patients with COVID-19, including patients who were particularly unwell and required either non-invasive or invasive ventilation. There were some training sessions organised for doctors working on COVID-19 wards. However, these were only provided towards the end of the first wave, as there was initially little knowledge of the virus and how to treat it. In the hospital the doctors followed local protocols, which were adjusted as more experience was gained about the virus.

We also asked the doctors to indicate whether the highlighted text presented is sufficient in reaching their decision, in order to understand its usefulness when we incorporate them in the annotation interface. The annotations were found to be sufficient in as many as 85% of the posts, on average, as indicated by the doctors' answers to Q3 in Figure 1.

The posts labelled by the doctors were then used to construct two types of predictive machine learning model using *Support Vector Machines* (SVM) [7, 8]; see Step 4 from section Methods. The *triage models* employ hierarchical binary classifiers, which consider the risk averseness or tolerance of the doctors when making the diagnosis [9]. The *diagnostic models* first calculate the probability of a patient having COVID-19 from doctors' ratings. The probabilities are then used to construct three different decision functions for classifying *COVID* and *NO_COVID* classes; these are detailed in the Problem setting subsection in Methods.

We trained the SVM models in two different ways, first with ground truth annotations, and second using predictions from the concept and relation extraction step described above. Predictions obtained from the concept extraction step make use of *Conditional Random Fields* (CRF) [10]; see Step 1 of the Methodology Sub Section in Methods for implementation details. Relations are obtained from these predicted concepts using an unsupervised *Rule-Based* (RB) classifier [11]; see Step 2 from section Methods.

We also discuss the feature importance obtained from the constructed COVID-19 diagnostic models, and compare them with the most frequent symptoms from [4] and our dataset. We found that symptoms such as anosmia/ageusia (loss of taste and smell) rank in the top 5 most important features, whereas they do no rank in the top 5 most frequent symptoms; see Discussion. Overall, we make several contributions as follows:

1. We show that it is possible to construct machine learning models to triage and diagnose COVID-19 from patients' natural language narratives. To the best of our knowledge, no other previous work has attempted to triage or diagnose COVID-19 from social media posts.
2. We also build an end-to-end NLP pipeline by making use of automated concept and relation extraction. Our experiments show that the models built using predictions from concept and relation extraction produce similar results to those built using ground truth human concept annotation.

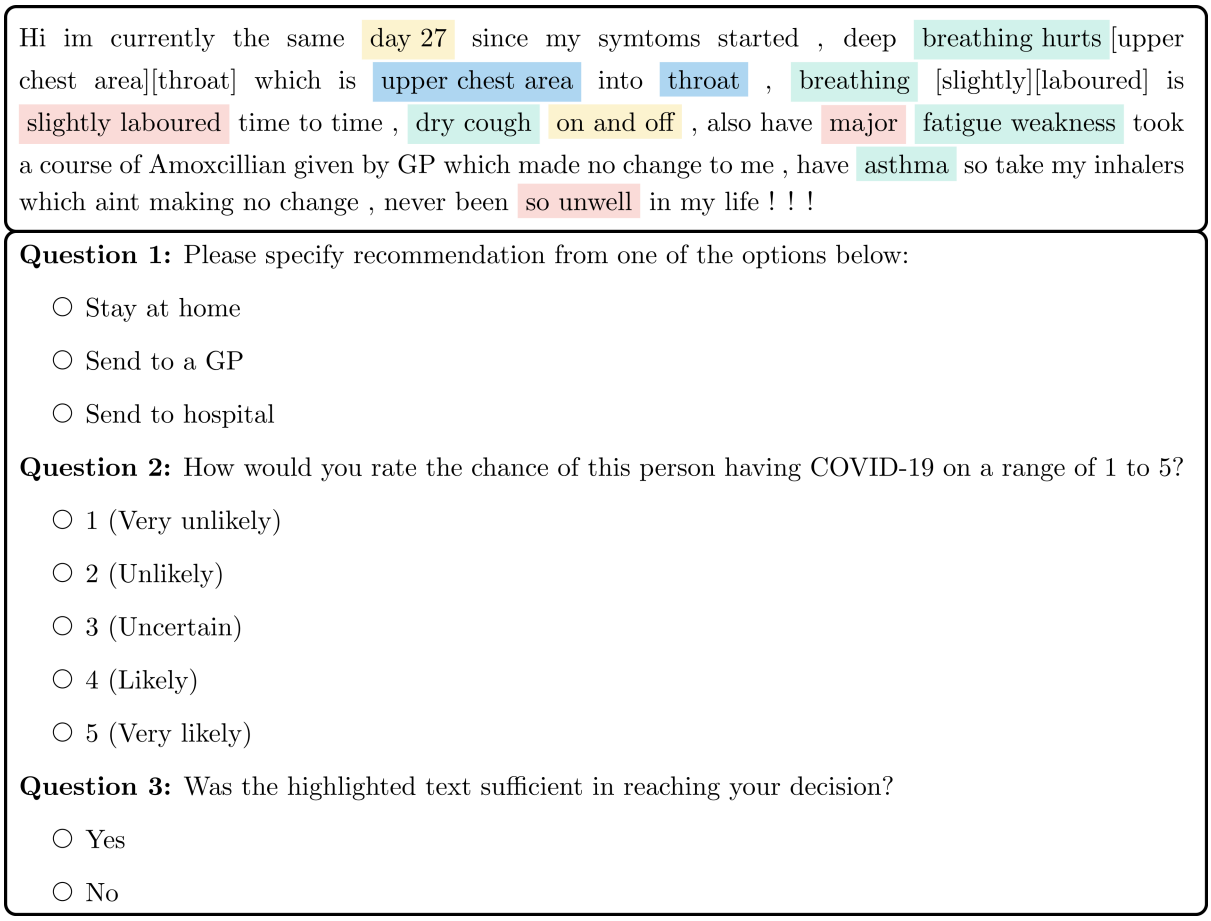

Figure 1: A patient-authored social media post is annotated with symptoms (light green), affected body parts (pale blue), duration (light yellow) and severities (pink). The phrases in the square brackets show relations between a symptom and a body part/duration/severity, when the distance was greater than 1. This annotated post was presented to three doctors to triage and diagnose the author of the post by answering *Questions 1* and *2*, respectively.

Related work

Respiratory diseases such as influenza were studied previously to determine the likelihood of symptom severity of the disease in clinical settings using Classification and Regression Trees [12]. Machine learning algorithms, such as decision trees, have shown promising results in detecting COVID-19 from blood test analyses [13].

The purpose of this study is to extract as much useful information as possible from social media, which is known to be noisy. In particular, posts from medical social media may provide an additional source of information regarding the symptom severity and prevalence of COVID-19 in the population. Here, we focus on features extracted from a textual source to triage and diagnose COVID-19. The studies related to our work deploy features obtained from online symptom tracker applications, telehealth visits, and structured and unstructured patient/doctors notes from Electronic Health Records (EHR). In general, the COVID-19 clinical prediction models can broadly be categorised into risk, diagnosis and prognosis models [14].

Since the start of the COVID-19 pandemic, a number of mobile app-based self-reported symptom tools have emerged, to track novel symptoms [15]. The mobile application in [16] applied Logistic Regression (LR) to predict percentage of probable infected cases among the total app users in the US and UK combined.

In [17], employed a portal-based COVID-19 self-triage and self-scheduling tool to segment patients into four risk categories: emergent, urgent, no-urgent and self-care. Whereas, the online telemedicine system in [18] used LR to predict low, moderate and high risk patients, by utilising demographic information, clinical symptoms, blood tests and computed tomography (CT) scan results.

In [3], developed various machine learning models to predict patients outcome from clinical, laboratory and demographic features found in EHR [19]. They reported that Gradient Boosting (XGB), Random Forest and SVM are the best performing models for predicting COVID-19 test results, and, hospital and ICU admissions for positive patients, respectively. A detailed list of clinical and laboratory features can be found in [20], where they developed predictive models for the inpatient mortality in Wuhan, using an ensemble of XGB models. Similarly, In [21], predicted mortality and critical events for patients using XGB classifiers. Finally, a critical review on various diagnostic and prognostic models of COVID-19 used in clinical settings, can be found in [14].

In [22], extracted COVID-19 symptoms from unstructured clinical notes in the EHR of patients subjected to COVID-19 PCR testing. The authors in [23] performed a statistical analysis on primary care EHR data to find longitudinal dynamics of symptoms prior to and throughout the infection. In addition, COVID-19 SignSym [24] was designed to automatically extract symptoms and related attributes from free text. Furthermore, the study in [25] utilises radiological text reports form lung CT scans to diagnose COVID-19. Similar to our approach, Lopez et al. [25] first extracted concepts using a popular medical ontology [26] and then constructed a document representation using word embeddings [27] and concept vectors [25]. However, our methodology differs from theirs with respect to the extraction of relations between concepts, and moreover, our dataset, comprising posts obtained from medical social media, is more challenging to work with, since social media posts exhibit greater heterogeneity in language than radiological text reports.

Finally, [4] published a COVID-19 symptom lexicon extracted from Twitter, which we compare our work to in the Discussion section.

Table 1: Pair-wise agreement between pairs of doctors answers for Question 1 and 2; see Figure 1 for an example.

	Question 1				Question 2		
Pair	p_o	Kappa	AC1		p_o	Kappa	AC1
AB	0.65	0.26	0.55		0.73	0.64	0.67
BC	0.63	0.14	0.53		0.73	0.64	0.67
AC	0.77	0.28	0.72		0.51	0.40	0.40

Methods

Data

We collected social media posts discussing COVID-19 medical conditions from a forum called *Patient* [28]. This a public forum that was created at the onset of the coronavirus outbreak in the UK.

We obtained permission from the site administrator to scrape publicly available posts dated between April and June, 2020. In addition, all user IDs and metadata were removed from the posts for the purpose of the study. After the posts were anonymised, and the duplicates were removed, we randomly selected 500 distinct posts. The first author annotated these posts with the classes shown in Figure 2. The class labels represent symptoms and the related concepts: (i) duration, (ii) intensifier, which increases the level of symptom severity, (iii) severity, (iv) negation, which denotes presence or absence of the symptom or severity, and (v) affected body parts. We also annotated relations between a symptom and other concepts that exist at sentence level. For example, the relation between a symptom and a severity concept is denoted as *(SYM, SEVERITY)*. The posts were then marked with concepts in different colours and the relations were placed right after the symptom in square brackets, as shown in Figure 1. Each marked post was presented to the doctors using a web application, and they were asked three questions independently; see Figure 1. We call the doctors answers to Q1 and Q2 as the COVID-19 symptom triage and diagnosis, respectively. Thus for each post we have three independent answers from three doctors, which we denote as A, B, and C, respectively; these correspond to the last three authors of the paper and have been assigned randomly.

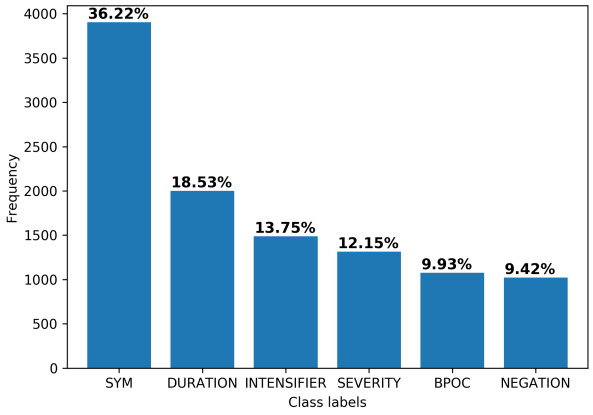

Figure 2: Frequency distribution of annotated classes/concepts from the text are shown. We also show the percentage of each class after discounting the *OTHER* labels. The average number of tokens per post is 130.17(SD = 97.83). Here, *SYM, DURATION, INTENSIFIER, SEVERITY, BPOC and NEGATION* denote symptoms, duration, intensifiers, severity, body parts and negations, respectively.

Measurement of agreement

In order to measure the agreement between the answers (recommendations and ratings) of the three doctors to Q1 and Q2 of Figure 1, we first calculated the proportion of observed agreement (p_o) as suggested by [29], who stipulate that Kappa is actually a measure of reliability rather than agreement, and observe that p_o is high in all cases as can be seen in Table 1. We note that paradoxical behaviour of Cohen's Kappa can arise when the absolute agreement (p_o) is high [30]. This may occur when there is a substantial imbalance in the marginal totals of the answers, which we have observed in the answers to Q1. Consequently, we deploy a common solution to this problem called the AC1 statistic devised by Gwet [31, 32], in addition to Cohen's Kappa.

We found that for Q1 the AC1 measure shows moderate agreement (in the middle of the moderate range) between A and B (0.55), between B and C (0.53), and substantial agreement between A and C (0.72); see [33] for benchmark scale for the strength of agreement. For Q2 it turns out that the said paradox did not occur, resulting in similar values for Kappa and AC1. The agreement between A and B (Kappa=0.64, AC1=0.67) and between B and C (Kappa=0.64, AC1=0.67) are substantial, while the agreement between A and C (Kappa=0.40, AC1=0.40) is on the boundary of fair and moderate; see Table 1.

It is important to note that COVID-19 is a novel virus, for which the doctors did not have prior experience or training before the first wave of the pandemic, and thus one would expect some difference of opinion. (We bear in mind that in our setting the doctors can only see the posts and thus cannot interact with the patients as they would in a normal scenario.) Moreover, there are probable differences in risk tolerances between the doctors, which would lead to potentially different decisions and diagnoses.

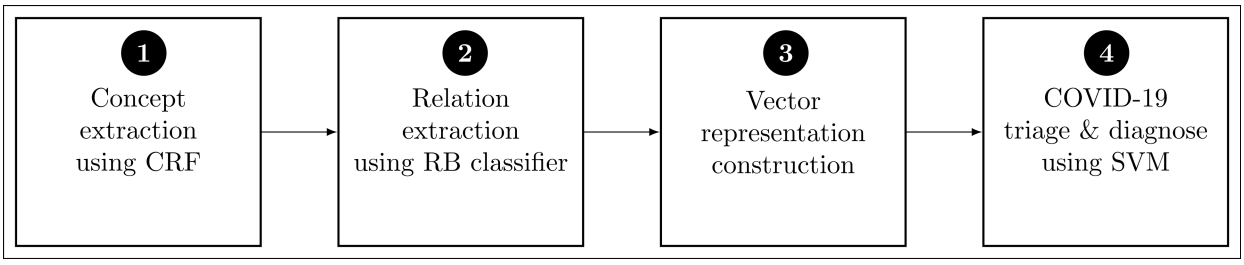

Figure 3: A block diagram of COVID-19 triage and diagnosis text processing pipeline. Here, CRF, RB classifier and SVM are acronyms for Conditional Random Fields, Rule-Based classifier and Support Vector Machine, respectively.

Problem setting

Triage classification for Question 1

We map the doctors' recommendation from Q1 to ordinal values; the options *Stay at home*, *Send to a GP*, or *Send to hospital* are transformed to the values 1, 2, and 3, respectively. In order to combine recommendations from two or more doctors, we first take their average. This result is rounded to an integer in one of two ways, either by taking the floor or the ceiling. Considering the risk attitude prevalent among medical practitioners [9], we categorise the ceiling of the average to be *risk averse*, denoted by e.g. AB(R-a), and the floor to be *risk tolerant*, denoted by e.g. AB(R-t). Thus for each patient's post, we have in total eleven recommendations from three doctors for Q1, the full enumeration can be seen in the first column of Table 3. We construct a hierarchical classification model for each of these recommendations, where the goal is to classify a post into one of the three options.

Diagnosis classification for Question 2

To diagnose whether a patient has COVID-19 from his or her post, we first estimate the probability of having the disease by normalising the rating, i.e given a rating, r, the probability of COVID-19, $Pr(COVID|r)$, which we term as the *ground truth probability* (abbreviated *GTP*), is simply:

$$Pr(COVID|r) = \frac{r-1}{4}.$$

Given our ground truth probability estimates are discrete we investigate three decision boundaries based on a threshold value of 0.5 to classify a post as follows:

LE: If $Pr(COVID|r) <= 0.5$, then *NO_COVID*, else *COVID*.

LT: If $Pr(COVID|r) < 0.5$, then *NO_COVID*, else *COVID*.

NEQ: If $Pr(COVID|r) < 0.5$ then *NO_COVID*,
else if $Pr(COVID|r) > 0.5$ then *COVID*.

Note NEQ differs to the other decisions in that we ignore those cases on the 0.5 boundary.

Methodology

A schematic of our methodology to triage and diagnose patients from their social posts is shown in Figure 3. Here, the circles denote the steps followed in the pipeline. We now detail each of these steps.

Step 1: Concept extraction

In the first step, we pre-process each patient's post by splitting it into sentences and tokens using the GATE software [34] built-in *Natural Language Processing* (NLP) pipeline. For each token in a sentence we build discrete features that signal whether the token is a member of one of the following dictionaries: (i) Symptom, (ii) Severity, (iii) Duration, (iv) Intensifier, and (v) Negation. The dictionaries were built by analysing the posts while annotating them. We also utilise the MetaMap system [26], assuming that it contains all the necessary technical terms, to map tokens to three useful semantic categories: *Sign or Symptom; Disease or Syndrome; Body Part, Organ, or Organ Component*. Due to the assumption regarding medical terms, the system does not expect any new additional terms, and thus we are justified in extracting concepts and relations in pre-processing steps. The pre-processed text is then used to build a concept extraction module to recognise the classes, shown in Figure 2, by applying a CRF [10]. A detailed description of our CRF training methodology can be found in [35]. The extracted concepts are then used for our next step to recognise the relations between concepts.

Step 2: Relation extraction

The semantic relation between a symptom and other concepts, which we formally termed as *modifiers*, is resolved using an unsupervised RB classifier algorithm. We first filter all symptom and modifier pairs from a sentence within a predefined distance and then select the closest modifier to a symptom to construct a relation. In total, we extracted five kinds of relations as follows: *(SYM, SEVERITY), (SYM, DURATION), (SYM, BPOC), (SYM, NEGATION)* and *(SYM, ?)*.

The severity modifiers are mapped to a scale of 1-5. The semantic meaning of the scale is: *very mild, mild, moderate, severe* and *very severe*, respectively. The duration modifiers are also mapped to real values in chunks of weeks. So, for example, *10 days* is mapped to the value *1.42*.

Step 3: Vector representation

Fixed length vector representations suitable as input for SVM classifiers are built as follows.

Symptom-only vector representation

Let $\langle s_0, s_1 \ldots, s_n \rangle$ be a vector of symptoms constructed from the symptom vocabulary, for our dataset the number of unique symptom words/phrases $n = 871$. To construct the vector representation for a post, we extract the concept, *SYM*, and the relation *(SYM, NEGATION)*, and set s_i to 1, 0, or -1, according to whether the symptom is present, not present, or negated, respectively.

Symptom-modifier relation vector representation

The symptom-modifier relation vector is a much larger vector than the symptom-only and comprises three appended vectors containing: (i) the absence or presence of 110 unique body parts, (ii) the absence or value of a symptom duration, and (iii) the absence, negation or value or a symptom severity.

Step 4: Triage and diagnosis

We utilise SVM classification and regression models to triage and diagnose patients' posts, respectively, from the vector representations described above. For Q1, the recommendation from a doctor or combination of doctors is the class label of the post; see section Problem setting in Methods for a description. To build a binary classifier, we first combine the *Send to a GP* and *Send to hospital* recommendations to represent a single class, *Send*. The SVM is trained to distinguish between the *Stay at home* and the *Send* options; we call this *SVM Classifier 1*. Next, the posts labelled as *Stay at home* are discarded and *SVM Classifier 2* is built utilising the remaining posts to classify the *Send to GP* and *Send to hospital* recommendations.This results in a hierarchical classifier for COVID-19 triage.

For diagnosing COVID-19 cases, we deploy a variant of SVM, called *Support Vector Regression* (SVR) [7], to estimate the probability of COVID-19. We use the GTP that is derived from answers to Q2, as the dependent variable. SVR takes as input a high dimensional feature vector such as a *symptom-only* or *symptom-modifier relation* vector representation, as described above. Classification is performed using the three decision functions, LE, LT, and NEQ, described previously.

Table 2: The concept extraction using CRF and relation extraction using RB classifier results on 3-fold cross validation.

Concept extraction using CRF				
Label	P	R	F_1	Support
SYM	0.94	0.97	0.95	1300
SEVERITY	0.80	0.79	0.79	437
BPOC	0.92	0.83	0.87	356
DURATION	0.87	0.91	0.89	667
INTENSIFIER	0.88	0.97	0.92	494
NEGATION	0.83	0.89	0.86	338
OTHER	0.99	0.98	0.98	16892
Macro-average	0.89	0.89	0.89	

Relation extraction using RB classifier						
	With stop words			Without stop words		
Distance	P	R	F_1	P	R	F_1
2	0.74	0.63	0.68	0.74	0.64	0.69
3	0.75	0.67	0.71	0.75	0.67	0.71
4	0.75	0.69	0.72	0.75	0.69	0.72
5	0.75	0.71	0.73	0.74	0.71	0.73
6	0.74	0.72	0.73	0.74	0.72	0.73
7	0.73	0.73	0.73	0.73	0.73	0.73

Results

Evaluation

We evaluate the performance of the CRF and SVM classification algorithms using the standard measures of precision (P), recall (R) and macro- and micro-averaged F_1 scores [36]. macro-averaged scores are computed by considering the score independently for each class and then taking the average, while micro-averaged scores are computed by considering all the classes together.

As our dataset is sufficiently balanced with *COVID* and *NO_COVID* classes as can be seen in Figure 5, we report micro-averaged scores for SVR classification. On the other hand, in case of concept extraction, the *Other* class dominates. So, we report the macro-averaged scores for the CRF classification results.

Table 3: Question 1: Hierarchical classification results for RBF kernel using the symptom-modifier relation vector.

Symptom-modifier relation vector													
Trained on the ground truth							Trained on the CRF predictions						
Model	SVM Classifier 1			SVM Classifier 2			SVM Classifier 1			SVM Classifier 2			
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	
A	0.82	0.91	0.86	0.73	0.95	0.83	0.81	0.89	0.85	0.72	0.91	0.80	
B	0.73	0.77	0.75	0.81	0.99	0.89	0.74	0.74	0.74	0.81	0.99	0.89	
C	0.85	0.98	0.91	—	—	—	0.85	0.96	0.90	—	—	—	
AB(R-a)	0.70	0.75	0.72	0.80	0.96	0.88	0.73	0.71	0.71	0.81	0.96	0.88	
AB(R-t)	0.84	0.96	0.89	0.85	1.00	0.92	0.84	0.94	0.88	0.84	1.00	0.92	
BC(R-a)	0.72	0.75	0.73	0.92	1.00	0.96	0.74	0.71	0.72	0.92	1.00	0.96	
BC(R-t)	0.86	0.99	0.92	—	—	—	0.88	0.98	0.93	—	—	—	
AC(R-a)	0.79	0.87	0.83	0.89	1.00	0.94	0.81	0.85	0.83	0.89	1.00	0.94	
AC(R-t)	0.88	0.98	0.93	—	—	—	0.88	0.98	0.93	—	—	—	
ABC(R-a)	0.70	0.76	0.73	0.89	0.99	0.93	0.72	0.72	0.72	0.89	1.00	0.94	
ABC(R-t)	0.88	0.99	0.93	—	—	—	0.89	0.98	0.93	—	—	—	
Symptom-only vector													
Trained on the ground truth							Trained on the CRF predictions						
Model	SVM Classifier 1			SVM Classifier 2			SVM Classifier 1			SVM Classifier 2			
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	
A	0.83	0.91	0.87	0.74	0.85	0.79	0.84	0.89	0.87	0.74	0.82	0.78	
B	0.71	0.81	0.76	0.81	0.98	0.89	0.74	0.79	0.77	0.82	0.98	0.89	
C	0.87	0.97	0.92	—	—	—	0.86	0.95	0.90	—	—	—	
AB(R-a)	0.69	0.75	0.72	0.83	0.96	0.89	0.72	0.76	0.73	0.83	0.92	0.87	
AB(R-t)	0.85	0.94	0.89	0.85	1.00	0.92	0.87	0.93	0.90	0.84	0.98	0.90	
BC(R-a)	0.71	0.79	0.75	0.92	0.99	0.95	0.72	0.78	0.75	0.92	0.99	0.95	
BC(R-t)	0.88	0.98	0.93	—	—	—	0.87	0.97	0.92	—	—	—	
AC(R-a)	0.80	0.86	0.83	0.89	1.00	0.94	0.80	0.86	0.83	0.89	1.00	0.94	
AC(R-t)	0.90	0.98	0.94	—	—	—	0.89	0.95	0.92	—	—	—	
ABC(R-a)	0.68	0.74	0.71	0.90	1.00	0.95	0.71	0.76	0.73	0.89	0.99	0.93	
ABC(R-t)	0.90	0.98	0.94	—	—	—	0.90	0.95	0.92	—	—	—	

Table 4: Question 2: Micro-averaged F_1 results for different models and decision functions. Here A, B, C are three medical doctors (abbreviated as Dr) who took part in the experiment.

Trained on the ground truth							Trained on the CRF predictions					
Model	Symptom-modifier			Symptom-only			Symptom-modifier			Symptom-only		
	LE	LT	NEQ	LE	LT	NEQ	LE	LT	NEQ	LE	LT	NEQ
A	0.72	0.61	0.78	0.70	0.59	0.74	0.68	0.64	0.76	0.50	0.79	0.74
B	0.78	0.61	0.76	0.78	0.62	0.77	0.76	0.64	0.77	0.78	0.57	0.74
C	0.87	0.75	0.87	0.88	0.75	0.87	0.86	0.75	0.87	0.87	0.74	0.86
AB	0.72	0.66	0.74	0.74	0.65	0.75	0.70	0.65	0.73	0.71	0.66	0.74
BC	0.84	0.76	0.84	0.85	0.79	0.86	0.83	0.76	0.83	0.85	0.78	0.86
AC	0.81	0.73	0.81	0.83	0.74	0.83	0.80	0.74	0.82	0.80	0.73	0.81
ABC	0.74	0.67	0.76	0.75	0.67	0.77	0.72	0.69	0.76	0.74	0.69	0.77

Experimental setup

For the CRF we report 3-fold cross validated macro-averaged results. Specifically, we trained each fold by a Python wrapper [37] for CRFsuite, see [38]. For relation extraction, we ran our unsupervised rule-based algorithm on the 500 posts and calculated the F_1 scores by varying distances considering the two cases with and without stop words.

We constructed SVM binary classifiers, *SVM Classifier 1* and *SVM Classifier 2*, using the Python wrapper for LIBSVM [39] implemented in Sklearn [40] with both Linear and Gaussian *Radial Basis Function* (RBF) kernels [8]. Similarly, the SVR [41], implemented using LIBSVM, is built with both Linear and RBF kernels. The hyperparameters ($C = 10$ for the penalty, $\gamma = 0.01$ for the RBF kernel, and $\epsilon = 0.5$ for the threshold) were discovered using grid search [40].

We simulated two cases for COVID-19 triage and diagnosis. First SVM and SVR models trained with the ground truth examine the predictive performance when they are deployed as stand-alone applications. Second, when trained with the predictions from CRF and RB classifier, they resemble an end-to-end NLP application. To get a comparable result, the models were always tested with the ground truth. As a measure of performance, we report macro and micro-averaged F_1 scores for SVM classifiers and SVR, respectively.

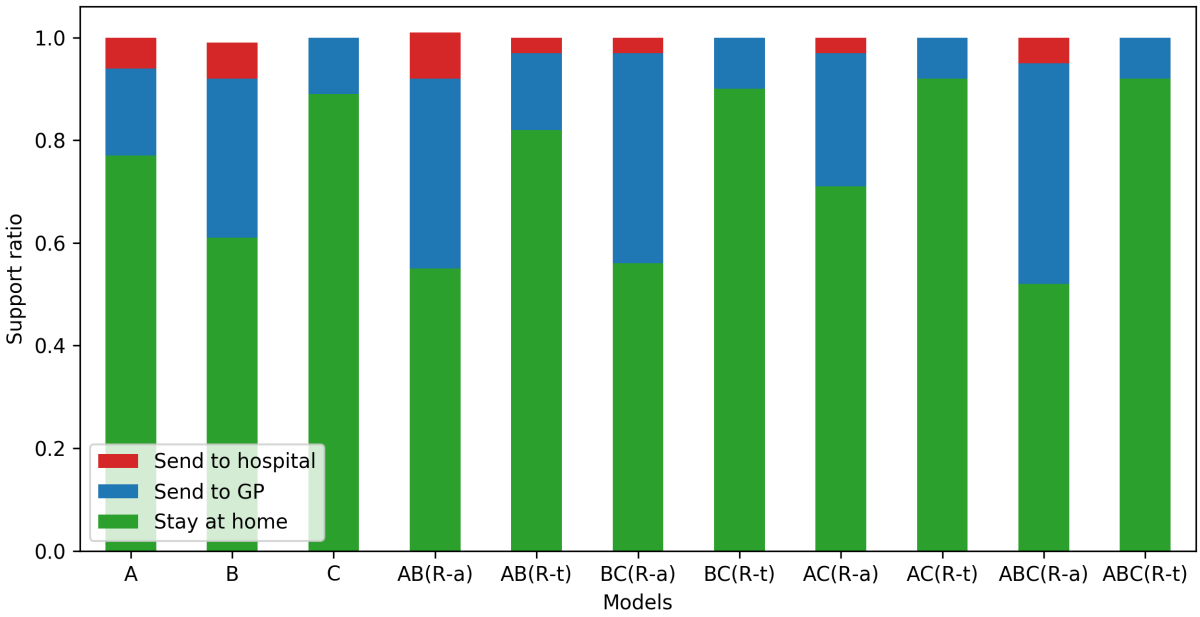

Figure 4: Support ratio of triage classes across models for Question 1 classification tasks. Absolute numbers for the *Send to hospital* class in test sets are as follows: A=10, B=12, AB(R-a)=14, AB(R-t)=5, BC(R-a)=6, CA(R-a)=5, ABC(R-a)=9; the value for the remaining models is zero.

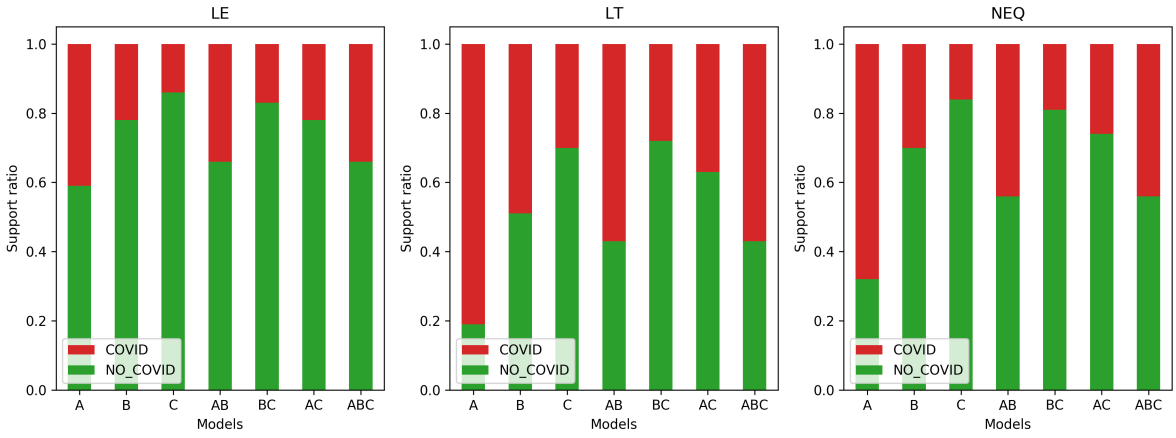

Figure 5: Support ratio of diagnosis classes across models and three decision functions for Question 2 classification tasks.

Evaluation outcomes

The concept and relation extraction phases produce excellent and very good predictive performances, respectively; see Table 2. The triage classification results from Q1 are shown in Table 3. When we trained the models with the *Symptom-modifier vector* representations from the ground truth, the results of SVM Classifier 1 and 2 are in the range of 72-93% and 83-96%, respectively. The Symptom-only vector representations produces results in the range of 71-94% and 79-95%. These results suggest that we can achieve very good predictive performance for classifying *Stay at home* and *Send*, and for *Send to a GP* and *Send to hospital*. In general, risk-tolerant models achieve better performance than the risk-averse models. However, since, in the test set, posts with the label Send to hospital are missing for some models (as can be seen from Figure 4) we cannot report them. We report macro-averaged F_1 results since Q1 is framed as a decision problem, where weights for the classes are a priori equal. The results obtained after training with CRF predictions are in similar ranges for both representations and classifiers. This is important, because it indicates that an end-to-end NLP application is likely to produce similar predictive performance.

Regarding Q2, when we trained the models with the *Symptom-modifier vector* representation from ground truth, the results of COVID-19 diagnosis are in the range of 72-87%, 61-76%, and 74-87% for the LE, LT, and NEQ decision functions, respectively; see Table 4. The Symptom-only vector representation produce results in the range of 70-88%, 59-79%, and 74-87%. In general, NEQ models perform better due to the omission of borderline cases where the GTPs are exactly 0.5. The support ratios for each model for different decision functions, is shown in Figure 5. When we trained the models with the *Symptom-modifier vector* representation from the CRF predictions, the results are in the range of 68-86%, 64-76%, and 73-87% for the LE, LT, and NEQ decision functions, respectively. This indicates that, for diagnosis as well as triage, an end-to-end NLP application is likely to perform similarly to standalone applications. Here, we report micro-averaged F_1 scores since, in our dataset, *NO_COVID* cases dominate; this largely resembles the natural distribution in the population, where people tested positive for coronavirus are relatively a low percentage in the whole population even when the prevalence of the virus is high.

Finally, we trained our models using a Linear kernel, but found that RBF dominates in most of the cases; however, Linear kernels are useful in finding feature importance [42].

Discussion

Comparison with prior work

To quantify the important predictive features in the training set, we experimented with COVID-19 diagnosis using Linear Kernel SVR regression. More specifically, we used the Symptom-only vector representation constructed from the ground truth. We summed feature weights for each s_i in $\langle s_0, s_1 \ldots, s_n \rangle$ from seven models and three decision function; see Methods. The features are then mapped to the categories found in the Twitter COVID-19 lexicon complied by [4]. The top 5 important features in our dataset are: *Cough, Anosmia/Agusia, Dyspnoea, Pyrexia and Fatigue*. In [23], quoted 4 of these symptoms as the most prevalent coronavirus symptoms, strongly correlating with our findings.

To compare our importance ranking with that of Sarker et al.'s [4] frequent categories, we compiled the corresponding frequencies of our 5 most important symptoms. Normalised weights and frequencies are then plotted in Figure 6. The top-left stacked bar chart compares our 5 most important features with Sarker's frequencies. *Cough* is the most important symptom from our dataset, where it is the 2^{nd} most frequent. *Anosmia/Aguesia* ranks 2^{nd} in our importance list, while it is 7^{th} in the most frequent list. *Pyrexia* comes 1^{st} and 4^{th} in both the frequent and importance lists, respectively.

The top-right chart in Figure 6 shows a comparison between Sarker's frequent ranking and our importance ranking. Here, we select top 5 most frequent symptoms from Sarker's frequency list and normalise them. They are: *Pyrexia, Cough, Body ache, Fatigue, and Headache*. We took the corresponding importance weights of those symptoms and plotted the stacked bar chart. Here, *Headache* ranks 22^{nd} in our importance ranking, while it is 5^{th} in the frequent ranking. We find a large difference between the two rankings, implying that the top most frequent symptoms are not necessarily the most important ones.

Next we compare our most important feature weights with our dataset's frequency ranking using the methods described above. From the bottom-left stacked bar chart of Figure 6, we observe that Anosmia/Aguesia are a relatively low in order in the frequency ranking, i.e. 11^{th}. Like Sarker's, *Cough* comes 2^{nd} in our dataset's frequency ranking.

Finally, the bottom-right chart in Figure 6 refers to the comparison between our dataset's frequency and importance rankings of the corresponding symptoms. We observe that *Anxiety* ranks 4^{th} in the most frequent list, where it is very low, i.e. 23^{rd}, in the most importance ranking.

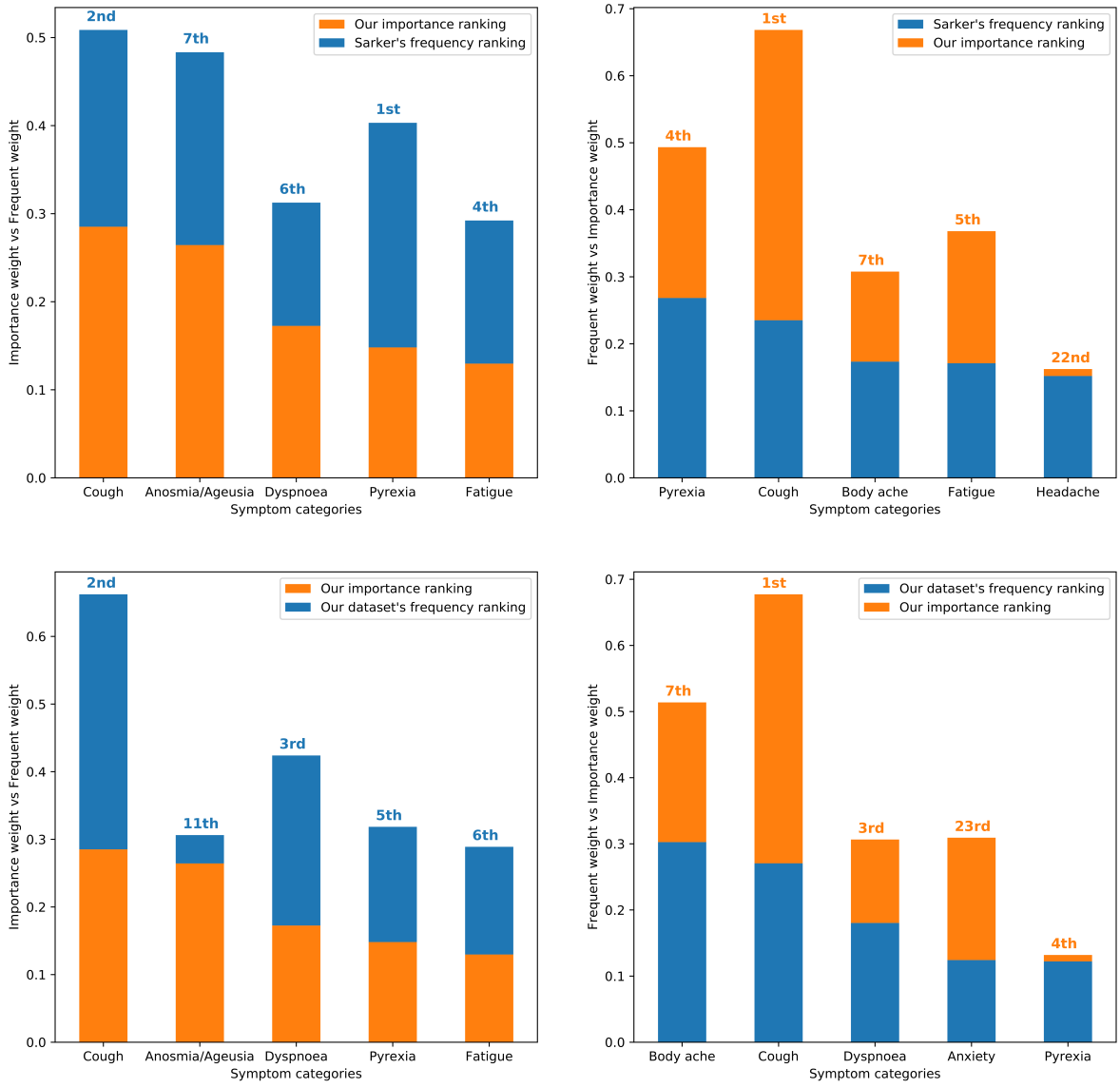

Figure 6: Feature comparison between our most important features and Sarker's most frequent symptoms (top row), and between our most important features and our most frequent symptoms (bottom row). The feature importance rankings are obtained from an SVM linear kernel using the symptom-only vector representation.

Principal findings

This study demonstrates the potential to triage and diagnose COVID-19 patients from their social media posts. We have presented a proof of concept system to predict a patient's health state by building machine learning models from their

narrative. The models are trained in two ways; using (i) ground truth labels, and (ii) predictions obtained from the NLP pipeline. Trained models are always tested on the ground truth labels. We obtained good performances in both cases which indicates that an automated NLP pipeline could be used to triage and diagnose patients from their narrative; see Evaluation outcomes in the Results section. In general, health professionals and researchers could deploy, triage models to determine the severity of COVID-19 cases in the population, and diagnostic models to gauge the prevalence of the pandemic.

Limitations

It is worth reiterating that social media posts, which are known to be noisy, are not on a par with the consultation that a patient would have with a doctor. Our manually annotated dataset has two main limitations. First having only three experts limits the quality of our labelling, although we deem this study to be a proof of concept. A larger number of experts, including more senior doctors would be beneficial in a follow-up study. The robustness of our results could be further improved by both increasing the size of our dataset and introducing posts from several alternate sources. Given that the posts come from social media, it is not clear whether the results could be used as such in a diagnostic system, without combining them with actual consultations. However, it is worth noting that medical social media such as the posts we used herein, may uncover novel information regarding COVID-19.

Conclusion

The coronavirus pandemic has drawn a spotlight on the need to develop automated processes to provide additional information to researchers, health professionals and decision-makers. Medical social media comprises a rich resource of timely information that could fit this purpose. We have demonstrated that it is possible to create an automated triage and diagnosis system, despite the heterogeneous nature of typical social media posts, that outputs of which could be used to indicate the severity and estimate the prevalence of the disease in the population.

Author Contribution

All authors were involved in the design of the work. The first author wrote the code. The first three authors drafted the article and all authors critically revised the article.

Abbreviations

CRF: conditional random fields

CT: computed tomography

EHR: electronic health records

LR: logistic regression

NLP: natural language pipeline

RBF: radial basis function

SVM: support vector machine

SVR: support vector regressions

XGB: gradient boosting

References

[1] NHS Website `https://web.archive.org/web/20200316223405/https://www.nhs.uk/conditions/coronavirus-covid-19/` Accessed: 2021-06-07

[2] Obeid J. S, Davis M, Turner M et al. An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. *J Am Med Inform Assoc*, 27(8):1321–1325, 2020.

[3] Schwab P, DuMont Schütte A, Dietz B and Bauer S. Clinical Predictive Models for COVID-19: Systematic Study. *J Med Internet Res*, 22(10):e21439, 2020.

[4] Sarker A, Lakamana S, Hogg-Bremer W et al. Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource. *J Am Med Inform Assoc*, 27(8):1310–1315, 2020.

[5] Qin L, Sun Q, Wang Y et al. Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *Int J Environ Res Public Health*, 17(7):2365, 2020.

[6] Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ*, 15(5):625–632, 2010.

[7] Drucker H, Burges C. J, Kaufman L et al. Support Vector Regression Machines. *Adv Neural Inf Process Syst*, 9:155–161, 1996.

[8] Marsland S. *Machine Learning: An Algorithmic Perspective (second ed.)*. CRC, 2014.

[9] Arrieta A, García-Prado A, Gonzalez P and Pinto-Prades J. L. Risk attitudes in medical decisions for others: An experimental approach. *Health Economics*, 26:97–113, 2017.

[10] Sutton C and McCallum A. An Introduction to Conditional Random Fields. *Found Trends Mach Learn*, 4(4):267–373, 2012.

[11] Bach N and Badaskar S. A Review of Relation Extraction. *Lit. Rev. Lang. Stat. II*, 2:1–15, 2007.

[12] Zimmerman R. K, Balasubramani G, Nowalk M. P et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC Infect Dis*, 16(1):1–11, 2016.

[13] Brinati D, Campagner A, Ferrari D et al. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *J Med Syst*, 44(8):1–12, 2020.

[14] Wynants L, Van Calster B, Collins G. S et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *British Medical Journal*, 369, 2020.

[15] Zens M, Brammertz A, Herpich J et al. App-Based Tracking of Self-Reported COVID-19 Symptoms: Analysis of Questionnaire Data. *J Med Internet Res*, 22(9):e21956, 2020.

[16] Menni C, Valdes A. M, Freidin M. B et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*, 26(7):1037–1040, 2020.

[17] Judson T. J, Odisho A. Y, Neinstein A. B et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. *J Am Med Inform Assoc.*, 27(6):860–866, 2020.

[18] Liu Y, Wang Z, Tian Y et al. A COVID-19 Risk Assessment Decision Support System for General Practitioners: Design and Development Study. *J Med Internet Res*, 22(6):e19786, 2020.

[19] Einstein Data4u.Diagnosis of COVID-19 and its clinical spectrum AI and Data Science supporting clinical decision ((from 28th Mar to 3rd Apr)). `https://www.kaggle.com/einsteindata4u/covid19`. Accessed: 2021-02-24.

[20] Wang K, Zuo P, Liu Y et al. Clinical and Laboratory Predictors of In-hospital Mortality in Patients With Coronavirus Disease-2019: A Cohort Study in Wuhan, China. *Clinical infectious diseases*, 71(16):2079–2088, 2020.

[21] Vaid A, Somani S, Russak A. J et al. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. *J Med Internet Res*, 22(11):e24018, 2020.

[22] Wagner T, Shweta F, Murugadoss K et al. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. *Elife*, 9:e58227, 2020.

[23] Mizrahi B, Shilo S, Rossman H et al. Longitudinal symptom dynamics of COVID-19 infection. *Nat Commun*, 11(1):1–10, 2020.

[24] Wang J, Abu-el Rub N, Gray J et al. COVID-19 SignSym – A fast adaptation of general clinical NLP tools to identify and normalize COVID-19 signs and symptoms to OMOP common data model. *J Am Med Inform Assoc*, 2021.

[25] López-Úbeda P, Díaz-Galiano M. C, Martín-Noguerol T et al. COVID-19 detection in radiological text reports integrating entity recognition. *Comput Biol Med*, 127:104066, 2020.

[26] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(1):267–270, 2004.

[27] Mikolov T, Sutskever I, Chen K et al. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, May 2013.

[28] Patient. `https://patient.info/forums/discuss/browse/coronavirus-covid-19--4541`. Accessed: 2021-01-18.

[29] de Vet H. C, Mokkink L. B, Terwee C. B et al. Clinicians are right not to like Cohen's κ. *British Medical Journal*, 346:f2125, 2013.

[30] Feinstein A and Cicchetti D. HIGH AGREEMENT BUT LOW KAPPA: I. THE PROBLEMS OF TWO PARADOXES. *J. Clin. Epidemiol.*, 43(6):543–549, 1990.

[31] Gwet K. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*, 61(1):29–48, 2008.

[32] Wongpakaran N, Wongpakaran T and Wedding D et al. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples.

[33] Landis J. R and Koch G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174, 1977.

BMC Med. Res. Methodol., 13(1):1–7, 2013.

[34] Cunningham H, Maynard D and Bontcheva K. *Text Processing with GATE (Version 6)*. CA: Gateway Press, 2011.

[35] Hasan A, Levene M and Weston D. Learning structured medical information from social media. *J Biomed Inform*, 110:103568, 2020.

[36] Manning C and Schutze H. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[37] Python-crfsuite. `https://python-crfsuite.readthedocs.io/en/latest/`. Accessed: 2018-03-14.

[38] Okazaki N. CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`, 2007.

[39] Chang C.-C and Lin C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol*, 2(3):1–27, 2011.

[40] Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12:2825–2830, 2011.

[41] Support Vector Machines. `https://scikit-learn.org/stable/modules/svm.html`. Accessed: 2021-01-19.

[42] Weston J, Mukherjee S, Chapelle O et al. Feature selection for SVMs. In *Advances in Neural Information Processing Systems*, 2000.