Extracting health information from social media

Extracting Health Information from Social Media

Abul Kalam Md. Rajib Hasan

Supervisors: Prof. Mark Levene
Dr. David Weston

October, 2022

This dissertation is submitted for the degree of
Doctor of Philosophy

Birkbeck, University of London
Department of Computer Science and Information Systems

Declaration

This thesis is the result of my own work, except where explicitly acknowledged in the text.

Abstract

Social media platforms with large user bases such as Twitter, Reddit, and online health forums contain a rich amount of health-related information. Despite the advances achieved in natural language processing (NLP), extracting actionable health information from social media still remains challenging.

This thesis proposes a set of methodologies that can be used to extract medical concepts and health information from social media that is related to drugs, symptoms, and side-effects. We first develop a rule-based relationship extraction system that utilises a set of dictionaries and linguistic rules in order to extract structured information from patients' posts on online health forums. We then automate the concept extraction process via; i) a supervised algorithm that has been trained with a small labelled dataset, and ii) an iterative semi-supervised algorithm capable of learning new sentences and concepts.

We test our machine-learning pipeline on a COVID-19 case study that involves patient authored social media posts. We develop a novel triage and diagnostic approach to extract symptoms, severity, and prevalence of the disease rather than to provide any actionable decisions at the individual level.

Finally, we extend our approach by investigating the potential benefit of incorporating dictionary information into a neural network architecture for natural language processing.

Acknowledgements

I would like to express my sincerest gratitude to my supervisors – Mark Levene and David Weston whose teaching, advice, feedback, and support helped me to steer through this long yet exciting journey. I am grateful to the Department of Computer Science and Information Systems and the School of Business, Economics and Informatics (BEI) for their continuous financial support during this PhD. I would like to extend gratitude to my mother Nargis, brother Shaon, and uncle Pervez for holding my family together during hard times with love, affection, and financial support. I am also indebted to Shompa for bringing up my daughter Noirita, who was born with Thalassaemia β major, in my absence. Finally, special thanks to two close friends – Manni and Shumona who have provided me with valuable emotional support throughout the journey.

Contents

List of Figures

List of Tables

Acronyms

ADE Adverse Drug Event.

ADHD Attention Deficit Hyperactivity Disorder.

ADR Adverse Drug Reactions.

BERT Bidirectional Encoder Representations from Transformers.

BiGRU Bidirectional Gated Recurrent Unit.

BiLSTM Bidirectional Long Short-Term Memory Networks.

BioBERT Bidirectional Encoder Representations from Transformers for Biomedical Text Mining.

CADEC CSIRO Adverse Drug Event Corpus.

CDC Centers for Disease Control and Prevention.

CHV Consumer Health Vocabularies.

CNN Convolutional Neural Network.

CRF Conditional Random Fields.

CT-BERT COVID-Twitter-BERT.

EHR Electronic Health Record.

ELMo Embedding from Language Model.

GCN Graph Convolution Network.

HMM Hidden Markov Models.

ILI Influenza Like Illness.

LR Logistic Regression.

n2c2 National NLP Clinical Challenges.

NHS National Health Services.

NLP Natural Language Processing.

POS Parts-of-Speech.

RB Rule-Based.

RF Random Forest.

RNN Recurrent Neural Network.

RoBERTa Robustly Optimized BERT Pretraining Approach.

SIDER Side Effect Resource.

SMM4H Social Media Mining for Health Applications.

SVM Support Vector Machines.

SVR Support Vector Regression.

UMLS Unified Medical Language System.

WNUT Workshop on Noisy User Text.

Chapter 1

Introduction

1.1 Problem settings

Patients' experiences shared on social media platforms such as Twitter, Reddit, and online health discussion forums are valuable sources of rich and timely health information. These experiences concern many diseases and health conditions, including diabetes, cancer, and mental health, to name a few. As the number of users continues to grow on social media platforms, there is an abundance of health-related knowledge contained within this domain [70]. Crucially, patients' experiences written on social media platforms provide opportunities to extract *actionable information* for health practitioners and decision-makers. Furthermore, the real time nature of posts can be particularly useful in emergency situations. A relevant and current example would be the COVID-19 pandemic. Researchers are investigating several broad categories of health applications utilising *Natural Language Processing (NLP)* including but not limited to pharmacovigilance [156, 102, 174], disease detection [30], and mental health surveillance [186]. Specifically, extracting *Adverse Drug Reactions (ADR)* [143, 90], detecting influenza epidemic [10], and monitoring mental health [186] from social media are widely studied.

In addition to social media another major source of health related text is *Electronic Health Record (EHR)* [70], which have different challenges. To process an unstructured clinical document written in natural language, e.g. a radiological report or a hospital

discharge summary, researchers have been using NLP tools specifically designed to work with formal *medical concepts* [221] defined in a medical ontology such as *Unified Medical Language System (UMLS)* [23, 12]. Several open source NLP tools exist for processing medical and biomedical documents such as MetaMap [12], cTAKES [178], and CLAMP [189]. Moreover, several benchmark datasets, for example the *National NLP Clinical Challenges (n2c2)*, exist in clinical and biomedical domains for a wide ranges of information extraction tasks such as identifying drug, dose, and relation between them from the EHR [82]. In contrast, health experiences shared on social media platforms contain diverse topics in distinct languages [70]. While health professionals use formal concepts, patients often share their experiences using informal medical terminology, colloquial language, and in the form of a dialogue.

Sentiment analysis was widely studied to find drug/treatment's efficacy from online drug reviews [243]. More recently, aspect based sentiment analysis was investigated [241, 242] on social media to find sentiment towards specific drug/treatment, symptom, and side-effect. Most of the existing studies focused on processing a single sentence [84]. Consider the following post from an online health forum:

- *HI the neurologist started me with* ***pramipexole****. i did* ***not like*** *it. it made me* ***sleepy*** *and* ***didn't*** *seem to* ***help*** *with my* ***movement****.*

In the above example, the negative sentiment towards *pramipexole* is signalled by the phrase *not like* in the second sentence of the post. In addition to this, the patient experiences feeling *sleepy* side-effect, and the drug is not effective for the *movement* symptom, as stated in the third sentence. As can be seen in the above example, *pramipexole* has a negative relation with *movement* a few sentences later. Therefore, a formal approach is required to make sense of the whole post. Additionally, current supervised *aspect-based* sentiment analysis methods are grounded on biomedical NLP tools such as MetaMap [12] in order to recognise sentiment towards a specific concept [242]. However, Gupta et al [77] previously found that MetaMap performed poorly on social media [77].

Supervised machine learning models require a substantial amount of high-quality training data to extract medical concepts from social media content. Producing a large set of annotated data is time-consuming and requires medical expertise. Although MacLean et al [115] found that crowd-based non-expert annotations approximate expert annotations after sufficient training, such annotations are highly dependent on the data source and type of disease. In the past, *Conditional Random Fields (CRF)* [99] and *Bidirectional Long Short-Term Memory Networks (BiLSTM)* [85] were applied to extract medical concepts (e.g. drug/treatment, symptom, side-effect, and ADR) from social media. Most notably, Nikfarjam et al [143] and Cocos et al [40] applied CRF and BilSTM, respectively, for ADR extraction from a Twitter dataset. Specifically, Nikfarjam et al [143] published a Twitter dataset and an ADR lexicon which they used with the CRF model. Lexicons and/or dictionaries are found to be useful with machine learning, and more recently, deep neural networks [130].

Recent advancements in neural network architectures and computational power have allowed NLP methods to take advantage of unsupervised learning [68]. Specifically, neural networks trained on a large corpus can encode distributional semantics [134]. Within the last few years, Transformer [201] based contextual language models such as *Bidirectional Encoder Representations from Transformers (BERT)* [49] have outperformed most of the machine learning models in many NLP tasks [164]. For example, a BERT variant *Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT)* [105], pre-trained on large biomedical corpora outperformed previous state-of-the-art models in various biomedical text mining tasks [105]. Adapting BERT to new datasets and tasks can follow either a fine-tuning or feature-based approach, and only requires a small labelled dataset [49].

1.2 Research hypothesis and questions

The main research hypothesis of this thesis is that NLP techniques can be harnessed to extract actionable information from social media. More specifically, the research questions addressed in this thesis are:

1. How can we build a rule-based concept relationship extraction system to extract a structured representation of drug/treatment's sentiment from social media posts focusing on a chronic disease category (e.g. Parkinsons')?

2. How to develop a machine learning method that uses *minimal supervision* to produce satisfactory results for concept extraction? How can it augment and update labelled datasets and dictionaries?

3. Can we develop an end-to-end NLP pipeline applying a similar rule- and machine learning-based methodologies for an infectious disease category (e.g. COVID-19) for extracting actionable information? Will the concept and relation extraction pipeline be able to triage and diagnose COVID-19 patients from their social media posts?

4. Are concept dictionaries helpful for a deep learning network? Are they transferable?

1.3 Aim and Objectives

The aim of this thesis is to develop a set of methodologies for extracting disease specific actionable health information from social media in a principled manner. To achieve this aim, we outline the following objectives:

1. To develop a methodology for extracting structured representations of sentiment related to drugs and treatments for a chronic disease (i.e. Parkinsons') that are described in patient forum posts.

2. To develop a concept extraction method that can be effective with minimal supervision and adapted to change in data source and disease study.

3. To apply an end-to-end NLP pipeline capable to provide decision makers with actionable information on the symptom severity and prevalence of a respiratory disease (i.e. COVID-19) using social media at the population level.

4. To investigate utility and transferability of manually built dictionaries and pre-trained word embeddings, respectively, by focusing on COVID-19 social media. Specifically, we transfer dictionaries between two types of social media and use them to produce weak labels for training neural networks.

1.4 Contributions

To attain the objectives of this thesis we made the following contributions:

1. We show that structured information can be extracted from posts found in an online Parkinsons' patient forum by forming relationships between a drug/treatment and a symptom or a side-effect, including the polarity/sentiment of the patient's opinion.

2. We develop a novel semi-supervised methodology capable of augmenting and expanding labelled dataset and concept dictionaries, respectively, by utilising a base-line CRF algorithm.

3. We apply an end-to-end NLP pipeline using rule- and machine learning- based methodology for COVID-19 patient-authored social media posts in order to gauge symptom severity and prevalence at the population level by taking a diagnostic approach.

4. We show that performance of a BiLSTM+CRF based deep learning model can be improved by incorporating dictionaries. Moreover, we show transferability of

the model first by producing weak labels using a dictionary from another source and then by performing concept extraction.

1.5 Publications

The following publications by the author are related to this thesis:

1. A. Hasan, M. Levene, D.J. Weston. "Natural language analysis of online health forums". In: *International Symposium on Intelligent Data Analysis*. Springer, 2017, pp. 125–137.

2. A. Hasan, M. Levene, D.J. Weston. "Learning structured medical information from social media". *Journal of Biomedical Informatics*, 2020 Oct;110:103568.

3. A. Hasan, M. Levene, D.J. Weston, R. Fromson, N. Koslover, T. Levene. "Monitoring COVID-19 on Social Media: Development of an End-to-End Natural Language Processing Pipeline Using a Novel Triage and Diagnosis Approach". *Journal of Medical Internet Research*, 550 24(2):e30397.

1.6 Thesis Structure and Overarching Methodology

This section provides both an overview of the thesis structure and a description of the overarching methodology taken in this thesis.

1.6.1 Thesis Structure

The next chapter describes background information and literature reviews of the methodologies followed in this research focusing on medical social media. Specifically, it provides background and literature reviews on rule-, supervised-, semi-supervised-, and deep learning- based information extraction methods. It also discusses different disease monitoring applications developed using social media focusing on infectious diseases. Chapter 3 describes a rule-based concept relationship methodology to extract

structured representations of patients' sentiment regarding drug/treatment focusing on Parkinsons' disease. The method relies on different publicly available and manually built medical dictionaries. In addition, the methodology incorporates an anaphora algorithm that links sentences in a post to extract coherent information. The focus of Chapter 4 is to build a method that is not so reliant on manually curated dictionaries as the rule-based approach described in the previous chapter. It presents a novel semi-supervised methodology which incorporates a dictionary expansion method for automating extraction of the said concepts from an online Parkinsons' forum and from a Twitter dataset. It shows how concept extraction tasks can be improved by augmenting labelled datasets and dictionaries. Chapter 5 describes a case study on a COVID-19 forum dataset where methodologies from the previous two chapters are combined to construct an NLP pipeline. The case study shows how actionable information regarding symptom prevalence and severity can be obtained from social media by taking a triage and diagnostic approach using the pipeline. Chapter 6 extends the work from previous chapters into deep learning, where a COVID-19 symptom extraction task is performed on a large Twitter dataset by producing weak labels using symptom dictionaries. Finally, Chapter 7 revisits the research questions stated above. It also describes the limitations of the proposed methods and potential for future work.

1.6.2 Overarching Methodology

This thesis presents novel methods for extracting actionable health information from social media focusing on two disease categories; one a chronic disease and the other an infectious disease. We selected Parkinsons' disease due to the availability of social media data and the availability of clinical expertise. From the infectious disease category, COVID-19 was chosen because of its prevalence at the population level at the time of this research.

We developed several corpora from online health forums and Twitter related to the diseases mentioned above. For extracting forum datasets we obtained permission

from site administrators and then manually downloaded HTML pages to scrape them automatically offline. For collecting Twitter datasets we used the Twitter Search API. After collection, the datasets were manually annotated with concepts related to drug, symptom, and side-effect using annotation tools developed by the author of this thesis. An example of the annotation tool developed during the course of this research is shown in Figure 1.1.

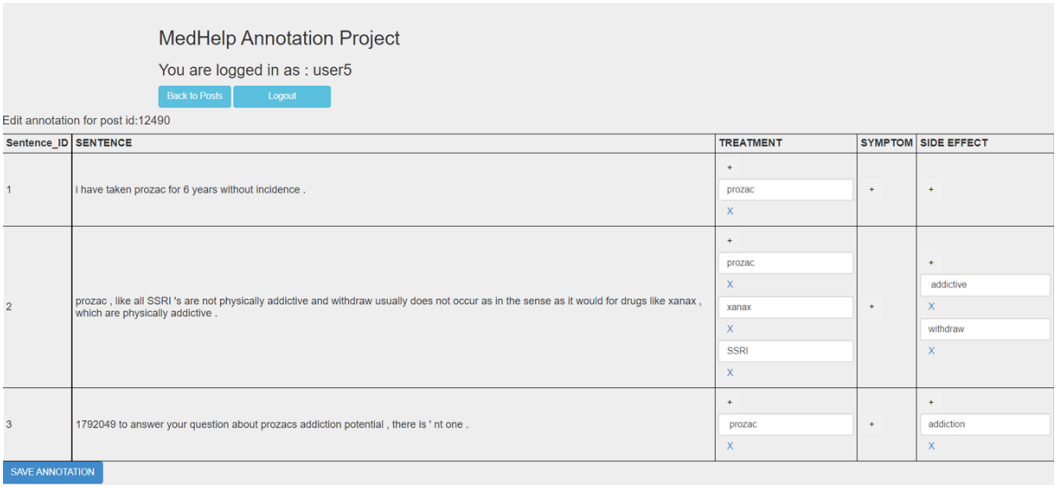


Figure 1.1: An annotation tool developed for annotating social media posts with drug/treatment, symptom, and side-effect. A post is presented sentence by sentence to an annotator. The screenshot is taken from a webpage which was used for validating thesis authors' annotations.

We decided to use words or tokens in a sentence as concept boundaries as opposed to the beginning, inside, and outside (together called BIO) [143] encoding, since number of concepts in our experiments ranged from 5 to 13. This is a relatively large number and adapting to a BIO encoding scheme would double the number of concepts to be recognised. This would have made the annotation procedure highly complex for annotators and potentially degrade the performance of the machine learning algorithms. Moreover, it is possible to convert multi-tokens to the BIO scheme by using a simple mapping program. However, we note that such a program may produce an error

when, for example, two multi-word concepts are mentioned consecutively. We closely followed the annotation guidelines provided in MacLean et al [115] and Nikfarjam et al [143]. Both studies used Cohen's κ for measuring the agreement among annotators. For training and testing of the models we developed repeated cross-validation strategies were used in various experiments. Finally, for evaluating the models, we used F_1 score as a performance measurement and we also performed statistical significance tests where applicable.

Chapter 2

Background and Related Work

The continual growth of social media platforms has enabled health professionals, researchers, and decision-makers to monitor individual and population health both in real time and retrospectively. In particular, health discussion forums and micro-blogging sites, collectively referred to as *medical social media*, are valuable sources for *actionable information* that can aid health professionals to make rapid decisions via the use of text extraction processes. Although much progress is achieved through NLP, structured and coherent information extraction from medical social media for decision-making remains challenging. Previously researchers applied several NLP methodologies such as text classification, sentiment analysis, and information extraction to analyse social media text [70]. Broader ranges of tasks such as detecting flu [10], finding the effectiveness of a treatment/medication [73], and extracting ADR [143] were investigated by applying these methodologies, particularly through the use of NLP methods that utilise supervised, semi-supervised, and unsupervised machine learning algorithms. Sentiment analysis is widely studied on medical social media to analyse patients' disease, vaccine, and drug/treatment experiences [243]. To produce an aggregated representation of patients' experience, analysis at both the sentence and post level is required. Supervised machine learning, specifically deep neural networks, have advanced automatic extraction of medical concepts such as drugs, symptoms and ADR [90, 8]. Producing a large, labelled dataset for the purpose of supervision is a highly intensive and laborious task and the presence of noise in social media makes it chal-

lenging to even find representative samples [70, 186]. Moreover, as new data continuously arrives and novel terms appear, a methodology that works with a small labelled dataset and detects new terms is required.

The remainder of this chapter is organised as follows. Section 2.1 describes social media sources, NLP tasks studied, benchmark datasets and related shared tasks. Section 2.2, describes relevant literature on rule-based concept extraction, relation extraction and anaphora resolution. Section 2.3, discusses supervised concept extraction methods and relevant literature focusing on ADR extraction. Sections 2.4 describes semi-supervised methods and their relevant literature. Section 2.5 describes literature relevant to deep learning based concept extraction methods focusing on COVID-19, incorporating dictionaries with neural architectures, transfer learning, and weak supervision. Finally, in Section 2.6, we describe literature relevant to text classification methods on social media focusing on ADR classification and infectious disease monitoring applications.

2.1 Medical social media sources, NLP tasks and datasets

Gonzalez-Hernandez et al [70] classify medical social media platforms based on user base and length of posts into two categories; (i) generic platforms such as Twitter, Facebook, and Instagram, and (ii) online health discussion forums such as PatientsLikeMe, MedHelp, and Reddit. Generic platforms such as Twitter are targeted for population-level health surveillance tasks such as monitoring flu or influenza [10, 86, 184], detecting ADR [143, 38, 131], and predicting mental health [186]. Although online forums (e.g. Parkinsons' or Breast Cancer) have a smaller user base than Twitter, those are utilised to provide information, not only for population-level surveillance but also at the individual level for a specific disease, symptom, and drug/treatment. Within the two broad categories, there are notable differences in content size and styles [70]. For example, Twitter text has a restriction on word limit and contains shorter descriptions

which sometimes lack context. On the other hand, though the scope of forum discussions is limited to a specific category of condition (e.g. Parkinsons'), they contain individual experiences and opinions and span over multiple sentences.

2.1.1 NLP Tasks

NLP tasks on medical social media can broadly be categorised into sentiment analysis, text classification, and information extraction. Herein, we provide an example of an information extraction task to motivate the discussion.

An example of information extraction

Information extraction (IE) refers to the task of locating relevant entities, for example, drug and/or symptom, in a collection of text documents, thereby extracting coherent and structured information from the text [43]. In general, information extraction involves *named entity recognition (NER)*, which we call *concept extraction* in this thesis, co-reference resolution, relation extraction, and event extraction. As shown in Figure 2.1, we demonstrate each of these tasks by providing an example from our Parkinsons' dataset (described in Chapter 3).

HI the neurologist started me with pramipexole [D]. i did not [NG] like [P] it. it made me sleepy [SD] and didn't [NG] seem to help [P] with my movement [SYM].

Figure 2.1: An example of a social media post. Here, yellow, green, blue, and red colours correspond to drug, symptom, side-effect, and positive polarity concepts, respectively. Similarly, the labels D, SYM, SD, NG, and P also represent drug, symptom, side-effect, negation, and positive polarity labels, respectively.

1. **Concept extraction** is the task of identifying a set of medical concepts and their attributes from the text. In our running example in Figure 2.1 *pramipexole*, *like*,

movement are drug, sentiment, and symptom concepts, respectively.

2. **Co-reference resolution** refers to the identification of multiple mentions of the same concept. For example, in Figure 2.1, the pronoun *it* refers to the drug *pramipexole*.

3. **Relation extraction** [71] is the task of identifying predefined relations between two concept. In this example, several relations exist between *pramipexole* and *sleepy*, *pramipexole* and *movement*.

4. **Event extraction** refers to filling up a predefined template representing structured information. In the running example, we can produce an ordered triple, using sentiment and relations from the post, *(pramipexole, negative, sleepy)*.

2.1.2 Datasets and shared tasks

In order to attain the objectives described in Chapter 1, we manually curate 5 datasets focusing on two diseases; they are (i) Parkinsons', and (ii) COVID-19. Table 2.1 provides a summary of the respective chapters, data sources, and tasks investigated in this thesis, whereby the annotation process is discussed in the corresponding chapters listed in the table.

Research related to medical social media has seen steady growth over the past decade. Though many datasets are available, our focus is on those related to concept and relation extraction tasks. First, we discuss 3 benchmark datasets and then the *Social Media Mining for Health Applications (SMM4H)* [177] shared tasks. The datasets, constructed from Twitter and online medical forums, contain annotations from experts. Publicly available Twitter datasets contain a Twitter identification number and corresponding annotations. However, they exclude Twitter text as the publication of the text is prohibited by the Twitter application program interface license agreement. It is possible that some tweets were deleted by their producers. As a result, those tweets may no longer be available in a benchmark dataset.

Table 2.1: Datasets curated and NLP tasks investigated in this thesis.

Disease	Medical social media	Chapter	Task
Parkinsons'	PatientsLikeMe [149]	Chapter 3	Concept extraction, relation extraction, sentiment analysis
Parkinsons'	MedHelp [126]	Chapter 4	Concept extraction
Parkinsons'	Twitter	Chapter 4	Concept extraction
COVID-19	Patient [148]	Chapter 5	Concept extraction, relation extraction, text classification
COVID-19	Twitter	Chapter 6	Concept extraction

ADRMine

Nikfarjam et al [143] collected Twitter data about 81 drugs and published expert annotations for Drug, ADR, Beneficial Effect, and Indication (which refers to symptom in our terminology). In addition to this, their corpus included corresponding UMLS concept ID and UMLS semantic type. The dataset was released in 2015 as a part of the ADRMine [143] system, hence, why we have decided to call it *ADRMine* in this thesis. It was also shared in 2016 SMM4H for the concept extraction task and subsequently called *PSB2016 Task 2*; see [20]. As stated in [177], the PSB2016 Task 2 dataset contains ADR and indication annotations. Cocos et al [40] compiled a supplementary Twitter dataset for their experiment with ADRMine. The additional dataset contains annotations related to *Attention Deficit Hyperactivity Disorder (ADHD)* drugs which is, henceforth, collectively called *ADRMine and ADHD* dataset in this thesis.

CADEC

The *CSIRO Adverse Drug Event Corpus (CADEC)* [90] corpus, published in 2015, contains social media posts related to 12 drugs from the *AskAPatient* [14] website and

annotation for Drug (1800 concepts), ADR (6,318 concepts), Disease (283 concepts), Symptom (275 concepts), and Findings (435 concepts). The annotations contain a total of 1250 posts, 7,632 sentences, and 101,486 words.

Twimed

This corpus, published in 2017, consists of 1000 tweets and PubMed sentences annotated with three primary concepts; they are Drug, Disease, and Symptom [8]. Additionally, the corpus has 3 relational annotations denoted *Reason-to-use*, *Outcome-positive*, and *Outcome-negative*.

SMM4H shared tasks

The SMM4H shared tasks workshops have been running since 2016 and have been publishing datasets related to text classification, concept extraction, and concept normalisation for their competition. The ADR classification tasks ran for several years, where ADR was defined as accidental injuries resulting from correct medical drug use [177]. Though the overarching goal was to detect ADR concepts from tweets, Sarker et al. [177] reported that only a small proportion of ADR tweets truly contained patients' experiences. Manually separating a small portion of tweets from a large collection for annotation was time-consuming and labour intensive [177]. To overcome this challenge, the ADR extraction task was divided into three different tasks in the first four years; they were (i) ADR classification, (ii) ADR extraction, and (iii) ADR normalisation. The goal of ADR classification was to separate tweets that mentioned ADR from those that did not mention it. ADR extraction and normalisation concerned concept span detection and mapping spans to *Medical Dictionary for Regulatory Activities Terminology (MedDRA)* [1] terms, respectively. In the latest iteration of the SMM4H shared tasks, participants were asked to perform three subtasks of ADR that built cumulatively over the tasks. The final task was termed the *Adverse Drug Event (ADE)*

[1] https://www.meddra.org/

detection task [116]. The highest F_1 scores for three tasks in 2021 were 75.2%, 51.4%, and 34.2%. The scores, in particular, the score of the ADE represented the underlying complexity and challenges of the social media dataset. We discuss the shared tasks in further detail in Section 2.6.1.

2.2 Rule-based information extraction

Rule-based systems, one of the earliest information extraction methods, are constructed using syntactic and semantic rules or patterns in the text. Syntactic and semantic rules usually utilise grammatical properties of words such as *Parts-of-Speech (POS)* tags and dictionaries of semantic classes (i.e. drug, symptom, and side-effect), respectively [165, 62]. The following subsections discuss past research related to rule-based concept extraction, relation extraction, anaphora resolution, and sentiment analysis in medical social media.

2.2.1 Rule-based concept extraction

MetaMap [12, 11] is one of the earliest rule-based automated systems in the biomedical domain which was constructed for processing scientific documents from MEDLINE/PubMed citations. It has an NLP pipeline for processing a piece of text and mapping it into formal UMLS concepts. Several rule-based algorithms are built into MetaMap, to generate candidate concepts for words and phrases in the text and to provide rankings for generated concepts. Moreover, it can be configured to recognise targeted semantic types such as Disease or Syndrome (DSYN), Sign or Symptom (SOSY), and Body Part, Organ, or Organ Component (BPOC). For example, Wu et al [220] target MetaMap to recognise ADR for the sentiment analysis. In this thesis, we utilise MetaMap to recognise drug/treatment, symptom, and body parts in medical social media. The recognition of semantic type is an important stepping stone for information extraction. Although MetaMap is very effective in retrieving concepts from

documents written by experts, it is known to perform poorly on textual sources of social media [77].

Previously, researchers relied on various rule-based methodologies for the ADR recognition task. Leaman et al. [103] were among the early researchers to extract ADR from an online forum using an ADR dictionary. Their rule-based algorithm used a sliding window approach to find concepts from the text. Nikfarjam et al [142] also created an ADR dictionary to apply association rule mining for ADR recognition from social media. Yang et al [225] utilised a similar methodology for extracting ADR from MedHelp [126] website. Wu et al [219] proposed an early warning system to discover the side-effects of a drug using co-occurrence statistics from web pages. They applied *Side Effect Resource (SIDER)* [98] dictionary to locate drug-related web pages. To identify ADR, Yates et al [226] developed a rule-based system using concept dictionaries and a synonym set that included term variants. Researchers also devised methodologies for linking ADRs to respective drugs. For example, Katragadda et al [91] extracted ADRs and connected them to corresponding drugs using a graph algorithm. Moreover, Ru et al [166] created an outcome dictionary to find the effectiveness of drugs from social media.

In the past, researchers also applied unsupervised rule-based approaches to create medical dictionaries automatically from formal and informal documents. Xu et al [223] developed an unsupervised and iterative pattern learning approach for constructing a medical disease dictionary from randomized clinical trial abstracts. On the other hand, creating a dictionary that reflected different ways consumers expressed and thought about health topics called *Consumer Health Vocabularies (CHV)* was the objective of several studies such as in [227] and [203]. A recent relevant example of a colloquial term was *loss of smell* -formally known as *anosmia* among health practitioners. Vydiswaran et al [203] created a CHV by discovering pairs of medical terms and their common alternate names from Wikipedia [203]. More recently, Sarker et al [176] constructed a COVID-19 symptom dictionary of informal terms from Twitter using regular expres-

sions.

Gupta et al [77] deployed a semi-supervised algorithm where they extracted symptoms and conditions (SCs), as well as, drugs and treatments (DTs) from online health forums using lexico-syntactic patterns. At first, they labelled the concepts using dictionaries constructed from publicly available sources. They then created flexible patterns by looking at two to four words before and after the labelled tokens. Patterns were scored by a frequency measure, i.e. the top-*k* most occurring patterns were chosen. They applied these patterns to all the sentences and extracted the matched phrases. These learned phrases were added to the dictionary, and the process was repeated until convergence occurred. This method resulted in an improvement of the F_1-score by approximately 5% over the base-line lexicon-based approach. The authors also reported on the discovery of new DT- and SC- terms that were not present in the seed dictionaries. Moreover, they found that their system outperformed MetaMap for processing social media posts.

2.2.2 Relation extraction

Relation extraction aims to identify pre-defined relations between concepts from unstructured text [15, 81]. For example, the *Outcome-positive* relation from the Twimed corpus represents a positive outcome of a drug used for treating a symptom [8], where both the drug and the symptom co-exist in a sentence. Research related to relation extraction from social media are rare, only a handful of studies related to our work exist. Doan et al [52] developed a rule-based system using a dependency parser, which represented binary grammatical relations between words [120], in order to extract cause-effect relation from Twitter messages for three topics; i.e. *stress*, *insomnia*, and *headache*. For instance, given a tweet *Excessive over thinking leads to insomnia*- their goal was to extract *excessive over thinking* as a cause for *insomnia* [52]. Recently, Ahne et al [1] applied CRF for extracting causal relations from tweets related to Diabetes.

Closest to our work, are the n2c2 shared tasks [82] which uses an EHR dataset, an-

notated with medications (drug/treatment in our terminology) and related concepts. The tasks are: (i) extracting medications and their related concepts from EHR, (ii) establishing a link between a concept and the medication preceding it, and (iii) building an end-to-end pipeline for tasks (i) and (ii). Some medication-related concept examples are strength and dosages, duration and frequency of administration, and reason for administration. Although the shared task organisers reported that deep learning-based models performed well for the relation extraction task, in a later study Alfattni et al [4] found that rule-based systems outperformed a BiLSTM-based deep learning model. Guan et al [75] found that Transformer [201] based models such as BERT [49] improved ADE and Reason (indication) relations extraction task on the n2c2 corpus. In a recent study by Mahendran et al [118], biomedical and clinical contextual language models, BioBERT [105] and ClinicalBERT [7], respectively, outperformed other models on the n2c2 dataset [24]. Research on the extraction of biomedical relations, e.g. protein-protein interaction (PPI) and drug-drug interaction (DDI), from the biomedical literature has seen explosive growth over the past decade thanks to advances in models based on neural networks [232]. Several benchmark corpora exist such as ChemProt [97] and DDI 2013 [182]. Neural approaches, such as hierarchical *Recurrent Neural Network (RNN)* [233] and LSTM [235], achieved state-of-the-art results for the multi-class biomedical relation extraction tasks.

2.2.3 Anaphora resolution

We note that extracting concepts from individual sentences in a post is an important step towards analysing natural language in medical forums. However, we also came to realise that linking the sentences in a post would allow a more comprehensive analysis of the text. Disambiguation of semantic relations between two expressions (sentences in our case) is known as *anaphora resolution*, where a later expression (the *anaphora*) has some semantic relation to an earlier expression (the *antecedent*). The rule-based system described in [71] is a knowledge-centric and pattern-based approach for disambiguat-

ing anaphoric references in clinical records.

2.2.4 Sentiment analysis in medical social media

Sentiment analysis [87] has been widely used to mine opinions from various online customer reviews. According to Deneck et al [48], the application of sentiment analysis in health domains concerns patient's health status, symptoms, and drug/treatments. In addition to this, Bobicev et al [22] assert that sentiment analysis also enables to recognize personal attitude in discussion of one's health. Chee et al. [31] performed sentiment analysis of online health forums using an ensemble of classifiers. Their motivation was to find drugs that could be included under the regulatory watch list. The study in [124] investigated sentiment analysis of health forums in order to find the effectiveness of alternative treatments for cancer (also known as *Complementary and Alternative Medicine (CAM)*). More recently, Ng et at [140] investigated the effectiveness of CAM treatments for COVID-19 from Twitter by applying a general purpose sentiment analysis tool. Furthermore, to quantify the sentiment of vaccines available for the sexually-transmitted Human Papillomavirus (HPV), Massey et al [123] developed a multi-class classification system for Twitter. The Twitter data was analysed to investigate the experience and outcome of chemotherapy in respective patients [230].
Na et al [138] performed a clause-level sentiment analysis by adopting a rule-based linguistic approach. They first divided a sentence into clauses, then, using dependency parsing, they established the relationship between aspects and sentiment. On the other hand, the study in [3] analysed the performance of features for machine learning models on social media. Here we briefly discuss some of the commonly applied features in supervised sentiment analysis tasks:

1. **Word embeddings**: Pre-trained word embeddings, e.g. *Word2Vec* [134] or GloVe [152], are utilised to extract sentence and token level features.

2. **Bag-of-Words (BOW)**: A sentence is represented as a TF*IDF vector, where TF

and IDF respectively represent *term frequency* and *inverse document frequency*, collected from the dataset.

3. **Frequency of sentiment words:** Number of positive and negative words in a sentence recognised by a sentiment lexicon (e.g. MPQA Subjectivity Lexicon [214]), is used as a feature for sentiment classification.

4. **POS tags**: Frequency of selected POS tags, e.g. nouns, adverbs, and adjectives, can also be added as features for sentiment classification.

5. **Negation**: Inclusion of this feature indicates presence of one or more negations, e.g not, nor, neither, in a sentence [3].

In addition to the above features, the UMLS semantic type of a token is often incorporated as a feature in medical settings (refer to [3]). Incorporating emotionally-related features was found to be effective in identifying Twitter users with self-reported mental health conditions (i.e. Bipolar, Depression, PTSD, and SAD) [34, 187]. The study in [163] developed a sentiment analysis tool known as *Tweep* to detect depression among authors of tweets using rule-based, machine learning (Naive Bayes) and *Convolutional Neural Network (CNN)*. Yadav et al [224] applied CNN, to investigate sentiment on different aspects. Such aspects included the severity of a medical condition, the effectiveness of a drug, and ADR. Their model was applied to several disease categories such as depression, anxiety, and asthma.

Moreover, the study in [73] investigated transferability of the supervised sentiment analysis techniques. The researchers specifically created multiple annotations for a drug review with regards to the overall drug rating, side-effects, as well as benefits. Following this, three separate supervised *Logistic Regression (LR)* models were trained for the sentiment analysis on the three said drug review tasks. They performed the transferability experiment on two different data sources and multiple conditions (e.g. pain, anxiety, diabetes). For all prediction tasks they applied a *n-gram* approach to extract features from the post. In the n-gram approach, single tokens or unigrams, and

two or more adjacent tokens (bigrams, trigrams) are used to produce representation of a post or comment.

In recent studies, *Graph Convolution Network (GCN)* was applied by taking account of dependency structures of the sentences in a post [241]. Specifically in [241], each sentence using a dependency parser was parsed to find its graphical representation. Representations of vertices which were recognised as Sign or Symptom concepts by UMLS were classified as sentiments. In a recent study, Žunić et al [242] applied BERT to encode contextual representation of words, as well as to find sentiments for aspects grounded by UMLS. According Žunić et al [242], one of the limitations of their approach was that its aspects (i.e. Sign or Symptom) were a priori known to the deep learning models. In another study, Gupta et al [77] found that UMLS grounded biomedical NLP tools such as MetaMap and that cTAKES performed poorly on social media data. Our approach to sentiment analysis was rule-based and it linked clauses and sentences using anaphoric relations; see Section 2.2.3 and 3.3 for relevant literature and our detailed algorithm. A systematic review of sentiment analysis in health domain can be found in [243].

2.3 Supervised concept extraction

Although rules using dictionaries are effective for concept extraction tasks in the absence of labelled data, they suffer from scalability and portability issues. Nikfarjam et al [143] report that less frequent concepts may be missed by a rule-based system. It is possible that rules generated from a training set may not generalise to an unseen test set constructed from a different type of disease and data source. For various sequence labelling tasks in NLP, e.g. NER or POS tagging, *Hidden Markov Models (HMM)*, CRF, and Neural Network-based models are widely used. In recent times, variants of RNN and Transformers such as BiLSTM and BERT, respectively, have outperformed other models in sequence labelling tasks. These models do not require explicit features

due to the fact that such features are learnt in the training process. Moreover, recent advances in pre-trained word embeddings enabled the researchers to extract features from large unlabelled corpora. In the following, we first provide a discussion about CRF graphical models showing examples for feature generation, and then we discuss joint BiLSTM and CRF models, subsequently called *BiLSTM+CRF*.

2.3.1 Conditional random fields (CRF)

CRFs, are a family of undirected graphical models representing a conditional distribution and can be applied to calculate the conditional probability of a label sequence given a sentence represented as a sequence of tokens. In Figure 2.2 we have shown a linear chain CRF. In the following, we provide a brief description of how CRF is applied for the concept extraction task, for details see [99, 125], and [193].

Let $X = x_0, \ldots, x_t, \ldots, x_T$ be a sequence of T tokens (in our case a token represents

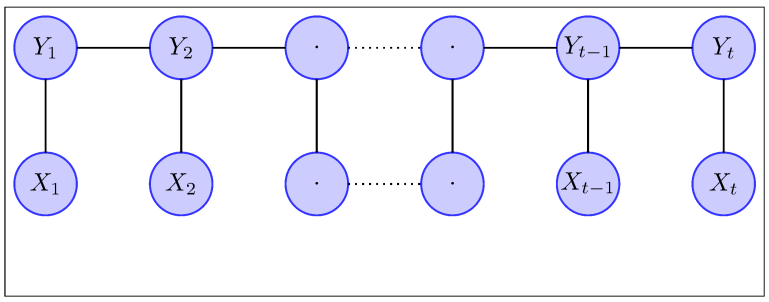


Figure 2.2: Linear chain CRF

a word, we use token and word interchangeably), and $Y = y_0, \ldots, y_t, \ldots, y_T$ be their corresponding labels (semantic type of the token). Let $g_k(y_t, y_{t-1}, \mathbf{x}_t)$ be $k = 1, \cdots, K$ feature functions at position t, and $\mathbf{x}_t$ be a vector of extracted features for the token at location t. The conditional probability of the label sequence Y given the token sequence X is calculated as follows [193]:

$$P(Y|X) = \frac{1}{Z(X)} \prod_{t=1}^{T} \exp \sum_{k=1}^{K} \lambda_k g_k(y_t, y_{t-1}, \mathbf{x}_t). \tag{2.1}$$

Here, $Z(X)$ is a normalisation factor, and λ_k is the weight of the kth feature function. The goal of training is to estimate the weights of feature functions from the labelled instances.

The inference algorithm for the CRF estimates the most likely labels for a token sequence by utilising a dynamic programming algorithm called the Viterbi algorithm. The algorithm produces *Viterbi probabilities* [159] for each possible label sequence by enumerating all labels. Viterbi probabilities are treated as the confidence of the prediction for a label sequence. Finally, to estimate λ parameters in Equation 2.1, the limited-memory BFGS [144] optimisation procedure is used; see [125, 193] for a detailed description of the inference and training procedures for the CRF.

Example of feature processing:

There are two kinds of feature sets; *label-label* and *label-word* features; generating two

> Roprineral[D] tablets from the doc for restless[SYM] leg[SYM], they are helping[P] me.

Figure 2.3: An example of annotated post where D, SYM, and P in brackets denote drug, symptom, and positive polarity labels, respectively.

kinds of probabilities for each sentence; *transition* and *emission probabilities*. A template for a *label-label* feature function, $g_{ij}^{LL}(y_t, y_{t-1}, \mathbf{x}_t)$ is as follows:

$$g_{ij}^{LL}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{y_t=i}\mathbf{1}_{y_{t-1}=j} \forall i, j \in L. \tag{2.2}$$

Where, $\mathbf{1}_{y_t=i}$ and $\mathbf{1}_{y_{t-1}=j}$ denote indicator functions. g_{ij}^{LL} corresponds to unique ordering for g_k in Equation 2.1. L is the set of possible labels. In the example of Table 2.2, the number of labels, L=$\{D, SYM, P, O\}$, is 4, therefore number of label-label features are 16. The second type of feature function, *label-word*, is defined similarly as follows:

$$g_{iv}^{LW}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{y_t=i}\mathbf{1}_{x_t=v} \forall i \in L, v \in V \tag{2.3}$$

Table 2.2: The sentence in Figure 2.3 is tagged with labels accordingly where D, SYM, P, and O represent drug, symptom, positive polarity, and other concepts, respectively.

t	x_t	y_t
1	Roprineral	D
2	tablets	O
3	from	O
4	the	O
5	doc	O
6	for	O
7	restless	SYM
8	leg	SYM
9	,	O
10	they	O
11	are	O
12	helping	P
13	me	O
14	.	O

Here, V is the unique vocabulary of words; i.e. unique words in the total dataset. However, one can add any number of feature templates of *label-word* features. Table 2.3provides a clear example that illustrates how contextual features can be collected from a sentence by considering previous and next tokens for x_t. Similarly, we can process an arbitrary number of *label-word* features by considering many syntactic and lexical patterns. These *label-word* features contribute to the *emission probabilities* of the token.

Table 2.3: Contextual feature extraction for the sentence in Table 2.2 using a window size of 1.

t	x_t	y_t
1	(START, Roprineral, tablelts)	D
2	(Roprineral, tablets, from)	O
3	(tablets, from, the)	O
4	(from, the, doc)	O
5	(the, doc, for)	O
6	(doc, for, restless)	O
7	(for, restless, leg)	SYM
8	(restless, leg, ,)	SYM
9	(leg, ,, they)	O
10	(,, they, are)	O
11	(they, are, helping)	O
12	(are, helping, me)	P
13	(helping, me, .)	O)
14	(me, ., END)	O

2.3.2 Neural architectures

According to Collobert et al [41], feature selection is an empirical and task oriented process implying additional research for each new NLP task. They propose a multilayer neural network architecture whereby feeding it with word indices from a vocabulary. Then the first layer of the network is initialised by mapping these word indices into a feature vector via a lookup table operation. Despite the fact that the network is initialised with features, the internal representation of the subsequent layers is able to generalise on multiple similar NLP tasks [41].

In 2013, Mikolov et al [133] proposed the *Word2Vec* algorithm to learn distributed

representations of words by processing large unlabelled corpora using a simple feed-forward neural network to predict a word given its context (or vice-versa). The result is a dense vector representation for each word where words that have similar meaning are likely to be close in this vector space. The reason derives from the assumption that words that have similar context have similar meaning. Inspired by the success of the word embeddings, Cho et al [37] proposed an RNN Encoder-Decoder neural network architecture for statistical machine translation. It consisted of two RNNs that acted as an encoder and a decoder pair. The encoder mapped a sequence of the source language to corresponding word embeddings, and the decoder mapped it to a sequence of the target language. The networks were trained together to calculate the conditional probability of the target sequence [37].

To replace hand engineered features, in 2016 Lample et al [100] proposed BiLSTM based encoder network for NER; see Figure 2.4. The network first mapped a token sequence to word embeddings and then calculated emission probabilities using a Softmax layer. Herein, we follow the descriptions from [100] to calculate the probability of a label sequence.

For a given sequence of words, $X = x_0, \ldots, x_t, \ldots, x_T$, each word is represented as a d dimensional vector. The forward LSTM takes this sequence and produces a hidden representation for each x_t, denoted $\overrightarrow{h_t}$, and the backward LSTM produces a hidden representation by reversing the sequence denoted as $\overleftarrow{h_t}$. The representation of a word using this model is obtained by concatenating its left and right context representations, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. The hidden representation h_t contains the summaries of both the preceding words and the following words. Let the number distinct labels and the length of the sequence be L and T, respectively. Let $\mathbf{E}$ be the matrix of scores output by the BilLSTM network. Therefore, $\mathbf{E}$ is size of $T \times L$, and $\mathbf{E}_{i,j}$ corresponds to the emission score of jth label for the ith token. For a sequence of predictions $Y = y_0, \ldots, y_t, \ldots, y_T$, the score s is calculated as follows:

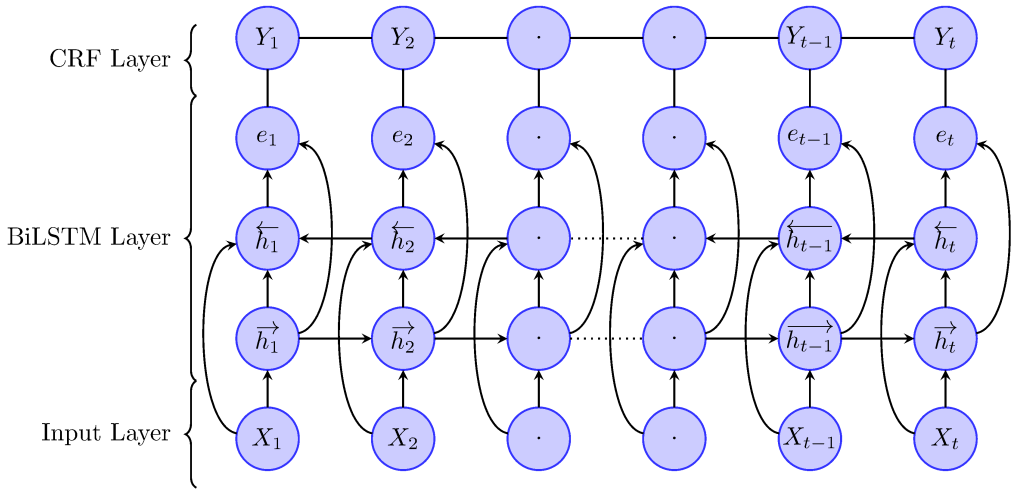


Figure 2.4: BiLSTM+CRF architecture.

$$s(X, Y) = \sum_{i=0}^{T} \mathbf{A}_{y_i, y_{i+1}} + \sum_{i=1}^{T} \mathbf{E}_{i, y_i} \tag{2.4}$$

$\mathbf{A}$ is a matrix of transition scores, where $\mathbf{A}_{i,j}$ represents transition scores from label i to label j. So, the probability of a label sequence is calculated as follows:

$$P(Y|X) = \frac{exp(s(X, Y))}{\sum_{\hat{y}} exp(s(X, \hat{Y}))} \tag{2.5}$$

Finally, the joint network is trained using back propagation; see [100].

2.3.3 Contextual language models

Neural networks initialised with pre-trained word vectors encode prior knowledge of word context which is useful in many NLP tasks including transfer learning. Static embeddings such as Word2Vec and GloVe [152] have been influential in solving NLP tasks. One key issue with them is that words can have more than one meaning, for which static embeddings cannot easily identify. In contrast, deep neural networks, such as a multilayer BiLSTM network trained on a large corpus for a language modelling objective can capture different senses of a word at the sentence level [155]. For

example, *Embedding from Language Model (ELMo)* was able to capture two different senses of the word *play* in two different sentences; i.e. in the context of *drama* and *game* [155]. Following the success of ELMo in encoding polysemy and homonym of words, Devlin et al [49] developed BERT by pre-training a deep Transformer [201] network using a cloze style learning objective. The core idea was called *Masked Language Model (MLM)* which randomly masked words in the text. Later on the network was trained to predict these masked words. *Fine-tuning*, which is training the pre-trained model further for a specific task, has been found to be very computationally efficient. The success of BERT inspired others to pre-train the BERT model with various domain specific corpora. For example, BioBERT [105] was trained on a vast biomedical literature. This model significantly outperformed state-of-the-art models on various biomedical information extraction tasks [105]. Müller et al [137] published a pre-trained *COVID-Twitter-BERT (CT-BERT)* model trained on a corpus of Twitter messages related to COVID-19. The authors reported that their model outperformed BERT models on several sentiment analysis tasks on Twitter datasets [137]. The BERT model known as ClinicalBERT was also trained using clinical notes from hospital admissions [7]. This model outperformed both BERT and BioBERT on a hospital readmission prediction task.

2.3.4 Supervised concept extraction in medical social media

In the past, sequence labelling was performed for ADR extraction from ADRMine, CADEC and Twimed corpora; see Section 2.1.2 for a discussion related to these corpora. A review of past research related to ADR extraction from social media can be found in [174]. MacLean et al [115] in their 2013 study applied CRF successfully for extracting medical concepts from MedHelp [126] forum data. They asked two groups of annotators, experts, as well as non-experts to annotate the text with words and phrases relevant to medical concepts at the sentence level. Specifically, they asked them to extract concepts describing body parts, conditions, symptoms, and treatment. Relevant

Table 2.4: Past ADR extraction methods and results using three benchmark datasets.

Study	Dataset	Model	P	R	F_1
Nikfarjam et al [143]	ADRMine	CRF	0.76	0.68	0.72
Chowdhury et al [38]	ADRMine	BiLSTM	0.72	0.87	0.79
Cocos et al [40]	ADRMine and ADHD	RNN	0.70	0.82	0.75
Ding et al [50]	ADRMine and ADHD	BiGRU	0.78	0.91	0.84
Wang et al [206]	ADRMine and ADHD	BERT+CRF	0.85	0.92	0.88
Miftahutdinov et al [132]	CADEC	CRF	0.85	0.79	0.79
Tutubalina et al[196]	CADEC	BiLSTM+CRF	0.82	0.84	0.81
Scepanovic et al [179]	CADEC	BiLSTM+CRF	0.81	0.82	0.82

words and phrases were labelled either as *medical* or *non-medical*. Aside from creating a gold standard dataset for training with a CRF, their study also included a comparison between expert and non-expert annotations. They found that non-expert annotations were overall an acceptable approximation for expert judgments in the case of social media [115] which motivates our approach to the annotation process. In 2014, Ginn et al [66] published a Twitter corpus annotated with ADR mentions. Note that the corpus was released prior to ADRMine by the same group of researchers. Lin et al [108] investigated the corpus for the ADR extraction task utilising a CRF with Word2Vec [133] and GloVe word vectors[152]. They demonstrated that including such word representations improved performance, moreover, Word2Vec had better clustering properties than GloVe vectors.

Nikfarjam et al [143] achieved state-of-the-art performance for ADR extraction on ADRMine and DailyStrength [47] corpora. Features such as contextual, lexical and semantic

POS tags were added to the CRF classifier. They also added word embedding features, created from Word2Vec [134], trained on unlabelled Twitter and the DailyStrength corpora. In another study, Korkontzelos et al. [96] analysed the effect of sentiment analysis features in ADR classification, which made use of rules such as negation to improve the performance of their system. Cocos et al [40] applied LSTM models on the ADRMine and ADHD dataset. Chowdhury et al [38] applied a BiLSTM based multitask framework and attention mechanism on the ADRMine dataset. Ding et al [50], developed *Bidirectional Gated Recurrent Unit (BiGRU)* with attention mechanism on the ADRMine and ADHD dataset and reported improvements on Cocos et al [40]. Wang et al [206] applied BERT combined with a CRF, henceforth called *BERT+CRF*, which they used to address class imbalance on the same two datasets. Finally, a detailed description of these deep learning models with a recent literature review on ADR extraction in pharmaceutical setting is provided in [104]. Due to the fact that datasets vary in size, we cannot directly compare these models. In Table 2.4 we have provided the Precision (P), Recall (R), and F_1 scores obtained via such methods. In case of the ADRMine dataset, BiLSTM outperformed CRF; see first two rows of Table 2.4. However, precision (P) of CRF is substantially better than that of BiLSTM. For the combined dataset, i.e. ADRMine and ADHD, recent BERT+CRF model outperformed all others.

Miftahutdinov et al [132] also applied CRFs to the CADEC corpus. Their system created word embeddings, similar to those of Nikfarjam et al [143], from unlabelled data by making use of the Word2Vec algorithm, and the resultant word vectors were grouped in predefined clusters that were utilised as features. Tutubalina et al [196] applied BiLSTM+CRF on the CADEC corpus. Scepanovic et al [179] also applied BiLSTM+CRF initialised with *Robustly Optimized BERT Pretraining Approach (RoBERTa)* [111] embeddings. BERT based models have also dominated in recent editions of the SMM4H shared tasks [116]. In the latest three editions of the SMM4H, i.e. 2019 to 2021, BiLSTM+CRF models initialised with BioBERT word embeddings concatenated with dictionary features obtained the best performance [130, 131, 170] on the ADR extrac-

tion tasks.

An HMM was implemented by Sampathkumar et al [171] for extracting ADR from a social media corpus. The HMM provided statistical structure for the forum messages, where drug and side-effects keywords representing the causal relation between the drug, side-effects and other words, were encoded as hidden states. Concepts were extracted from the messages using existing medical lexicons. The model was trained with the positive samples of ADRs, and learnt the association between drugs and side-effects through the presence of keywords. The most likely hidden state sequence is used to provide the predicted labels. The authors conducted various experiments by varying different components of the system. One of their findings was that the F_1-score of the supervised classification model is significantly lowered as the size of the dictionaries is reduced.

2.4 Semi-supervised concept extraction

In a domain such as medical social media where labelled data is scarce, we require a methodology that performs well on a small labelled dataset in conjunction with an unlabelled dataset. Additionally, when data is continuously streamed, concept drift occurs- where the underlying data stream distribution changes gradually over time [72, 236]. The semi-supervised methodology can take advantage of both labelled and unlabelled datasets to tackle the aforementioned challenges. Several categories of semi-supervised methods exist in the literature; see [199] for descriptions of these different categories. We focus on so-called wrapper methods. The simplest wrapper method is called self-training and in this approach, a supervised model is first trained with a labelled dataset. The model is then utilised to collect predictions on an unlabelled dataset. In general, the most confident predicted data points are added to the labelled dataset and the model is re-trained [199]. *Co-training* is an extension of self-training to multiple supervised classifiers. In the case of co-training, two or more

supervised models called *base-learners* are first trained on different views of the dataset and then jointly used for predictions from the unlabelled dataset. In the case of *boosting*, an ensemble of base learners is used for predicting from the unlabelled dataset. Boosting operates by setting weights for each base learner depending on its performance on an earlier iteration. In the following subsection, some of the relevant studies are discussed.

2.4.1 Semi-supervised concept extraction in social media

An iterative semi-supervised active learning based method was proposed to recognise drugs and their side-effects from Twitter data by [27] which includes human annotators in the training loop to augment representative and diversified labelled data. Edo-Osagie et al. [56] used self-training and co-training semi-supervised methods to train different binary classifiers for recognising tweets related to asthma. Lee et al. [106] used a semi-supervised CNN to identify ADE from a publicly available Twitter corpus. They gathered unlabelled corpora from various biomedical sources to learn phrase embeddings using dictionaries. They also expanded a health condition dictionary by selecting similar word vectors from an unlabelled corpus. However, this dictionary expansion did not consider an incremental retraining framework for updating the dictionary at each iteration of the semi-supervised methodology. A semi-supervised methodology was also applied by combining Word2Vec with *brown clustering* [26] features extracted from unlabelled Spanish and Swedish EHRs. These features boosted the performance of different base-line models, including a CRF [153]. More recently, a Chinese drug event report corpus was utilised to compare the effectiveness between the CRF and the BiLSTM+CRF models in recognising ADR concepts when deployed within a co-training style tri-training [238] methodology [35]. Both CRF and BiLSTM+CRF reported achieving comparable performances by leveraging the unlabelled dataset.

2.5 Deep learning for medical concept extraction

Deep learning models have achieved performance improvements on various NLP tasks over traditional machine learning algorithms. Moreover, their utility in *transfer learning* settings, where models trained in one data source are applied to a different data source [72], are well known. In Chapter 6, we investigate the utility of employing manually built COVID-19 dictionaries in deep learning and transfer learning settings. This includes both using dictionaries within the model and also to label the data for weak supervision. In the following subsections, we discuss related literature focusing on studies which performed COVID-19 medical concept extraction utilising social media, dictionary incorporation into deep learning architectures, transfer learning utilising social media, and weak supervision.

2.5.1 COVID-19 medical concept extraction

Since the start of the COVID-19 pandemic, medical social media have been extensively used to track and monitor COVID-19 novel symptoms [94, 76, 176]. Hernandez et al. [83] explored a large scale Twitter dataset [17] and automatically labelled tweets for drugs, conditions/symptoms, and measurements using biomedical taggers such as ScispaCy [139]. They found that existing biomedical/scientific text processing systems for concept extraction do not generalize well when used with non-clinical data sources like Twitter [83]. Batbaatar et al [19] developed a BiLSTM+CRF model to extract disease or syndrome, sign or symptom, and pharmacologic substance concepts from Twitter messages. However, their corpus was annotated automatically using the UMLS ontology which was known to miss concepts written in conversational language [77]. Wang et al [207] developed a tool called *COVID-19 SignSym* that extracted COVID-19 symptoms and their attributes such as body location, severity, and negation from the EHR records. The tool, which was built using the CLAMP [189] software, employed a hybrid approach of combining deep learning-based models, dictionaries and rules.

2.5.2 Combining gazetteers with deep learning

The use of sentiment dictionary features was widespread in high-performing deep learning based systems submitted to various SMM4H shared tasks. For example, Wu et al. [218] utilised a sentiment dictionary with character and word embeddings and combined them with various neural network architectures; this was the best performing system in the 2018 SMM4H ADR classification task. In 2019, the ADR classification was improved significantly by concatenating contextual embeddings with various lexical and syntactical features [211].

Several studies have demonstrated the utility of dictionaries for NER recognition tasks. Chiu et al [36] proposed an LSTM-CNN architecture that incorporated boolean dictionary features each of which indicated a word was part of an element included in a dictionary. In general, there are two types of methods for dictionary matching in the literature: (i) Partial match, and (ii) Full match. For example, consider the dictionary entry *heavy chest pain*, partial matches could be *pain* and *chest pain* whereas the full match is the total dictionary entry itself. Song et al [188] utilised both methods in their BiLSTM+CRF -based architecture for the NER tasks in multiple languages. Magnolini et al [117] demonstrated the use of dictionaries by concatenating such features with the input embeddings of a BiLSTM+CRF architecture. Recent work on incorporating gazetteers/lexicons into neural models focused on creating gazetteer embeddings and gazetteer models [154]. Usually, the gazetteer embeddings are represented as trainable parameters created from dictionary matches. Peshterliev et al [154] used a self-attention mechanism to enhance gazetteer embeddings and concatenated them with ELMo, character CNN and GloVe embeddings. Finally, Sun et al. [192] created gazetteer embeddings using labels and related gazetteers and fused them to a BERT-based encoder.

2.5.3 Transfer learning

Transfer learning has been successfully used in a variety of applications both within NLP and outside [210]. Here we focus on social media applications of transfer learning specifically involving Twitter datasets.

Transfer learning involving Twitter datasets poses many challenges due to the nature of the short sentences, frequent use of informal grammar, and irregular vocabulary (e.g. abbreviations) [141]. Alhuzali et al [5] pre-trained a classifier on a corpus annotated with sentiment information and then successfully applied BiLSTM to extract ADR on the ADRMine dataset. Han et al [80] proposed a *domain-adapted fine-tuning* approach using contextual language models (i.e. BERT and ELMo) for NER and POS tagging tasks in the domain of Early Modern English and Twitter, respectively. The domain-adapted fine-tuning referred to the approach where the language model was further pre-trained on large unlabelled corpora of the target domain. After this process, the language model was fine-tuned on a task in the *source domain* and transferred to the *target domain*. In [80], the base BERT model was further pre-trained on a large Twitter corpus, which was the target domain. The pre-trained model was then fine-tuned utilising the standard CoNLL 2003 NER task [172] which was the source domain. This model was then transferred to predict a NER task on the Twitter dataset published by the 2016 *Workshop on Noisy User Text (WNUT)* [191]. Luo et al [113] first trained a BERT+CRF model on an EHR dataset and transferred the model to extract symptoms from a Twitter dataset that contained mentions of COVID-19. The performance against the Twitter evaluation set achieved an F_1 score of 0.86. Zhang et al [231] trained a BiLSTM+CRF on the PubMed corpus and transferred the model to extract drug, disease and symptom from the Twimed dataset. Nguyen et al [141] recently created a model called BERTweet that was trained on a Twitter corpus of 850 million tweets using the RoBERTa [111] pre-training approach. They achieved state-of-the-art performances on several NLP tasks such as POS tagging, NER, and text classification on the WNUT

datasets.

2.5.4 Weak supervision

Fine-tuning allows pre-trained models, e.g. Word2Vec and BERT, to adapt to a target domain where a small amount of labelled data exists. In contrast, weak supervision, which subsumes another approach called *distant supervision*, is studied widely in cases where a labelled dataset is absent. Usually, the weak supervision approach utilises multiple sources such as dictionaries, ontology, and rules that are collectively called *labelling functions*, to annotate an unlabelled dataset automatically [162, 109]. Lison et al. [109] collated annotations from multiple sources using an HMM, which captured the varying accuracies and inconsistencies of the labelling functions.

While automatically annotating a dataset using labelling functions reduced human labour and cost, the models suffered from noise introduced by imperfect rules and dictionaries. Shang et al [183] proposed AutoNER that trained a BilSTM+CRF model, where a modified CRF optimization procedure was used, by automatically producing labels using a domain dictionary. Additionally, the model made use of a tagging scheme called *tie-or-break* which helped in de-noising label inconsistencies induced by the dictionary. They performed experiments on several biomedical corpora and reported to have achieved comparable results with supervised models. Another challenging problem with weak supervision is designing linguistic patterns or rules which requires a considerable amount of manual effort and domain expertise for automatic annotation. Zhao et al [234] approached the problem by constructing a GCN capable of learning new rules from unlabelled corpora using a set of seed rules within a biomedical weak supervision task. Furthermore, in the biomedical domain, SwellShark [64] used lexicons from different sources and rules to generate a labelled dataset for training with the BiLSTM+CRF network. Recently, the WRENCH [229] benchmark came up with a set of generalised functions to programmatically produce labels for a diverse set of datasets.

2.6 Text classification in social media

8 weeks and 2 days now and still having the hot sweats every night . wake up feeling like im on fire . sweating a horrid smell . some mornings the sweating lasts till midday . Chest pain still left and right side . Trying to figure out if its my oesophagus and heart causing the pain or my lungs (chest xray clear) oh and high heart rate and breathlessness .

Figure 2.5: A COVID-19 patient's social media post. Yellow, green, cyan, and red colours denote duration, symptom, body parts, and severity, respectively.

Medical social media was successfully used to facilitate the detection of influenza epidemics [10, 86] and more recently COVID-19 [184]. In Figure 2.5, we show a social media post from a COVID-19 patient where they share their experience regarding COVID-19 symptoms, severity, duration, and affected body parts. Our objective is to build a NLP pipeline for the classification of patients' posts that incorporates a triage and diagnostic approach in order to extract symptom severity and prevalence of COVID-19 in the population. In the following subsections we first review the text classification tasks in the SMM4H and then NLP applications related to disease detection and tracking focusing on infectuous diseases.

2.6.1 Twitter text classification tasks in SMM4H

In Table 2.5, we show year wise best F_1 scores for ADR classification tasks achieved in the SMM4H shared tasks. In the earlier iterations machine learning based classifiers such as *Support Vector Machines (SVM)* [54, 122], LR, and *Random Forest (RF)* were successful in creating state-of-the-art performance. However, in recent editions BERT based models have dominated. The performance reduction in 2021 as can be seen from Table 2.5, prompted organisers to have a closer look. They found that in addition to the datasets being different, participants in SMM4H 2020 used additional corpora

Table 2.5: Best performing ADR classification results in the SMM4H shared tasks for each year.

Year	Study	Classification method	P	R	F_1
2016	Rastegar-Mojarad et al [161]	Random Forest	0.36	0.50	0.41
2017	Kiritchenko et al [92]	SVM	0.39	0.48	0.43
2018	Wu et al [218]	CNN	0.44	0.63	0.52
2019	Chen [33]	BERT	0.60	0.68	0.64
2020	Wang [205]	RoBERTa	0.62	0.65	0.64
2021	Ramesh [160]	RoBERTa	0.51	0.75	0.61

to train their systems [116]. In Table 2.6, we show various tweet classification tasks from the SMM4H shared tasks starting from 2016 till 2020. In the most recent 2021 SMM4H shared tasks [116], BERT based models dominated in all categories. Relevant tasks that ran in 2021 were: (i) classification of tweets self-reporting potential cases of COVID-19, (ii) classification of COVID-19 tweets containing symptoms, and (iii) classification of self-reported breast cancer posts on Twitter [116]. For task (i), the study in [2] achieved the highest F_1 score of 0.79 among other competitors. They built an ensemble model based on several domain-specific BERT models including BERTweet and CT-BERT. Valdes et al [198] outperformed others in task (ii) by fine-tuning CT-BERT [137]. Finally, Zhou et al [237] achieved the highest F_1 score in task (iii) by utilising BERTweet.

2.6.2 Infectious disease monitoring applications

Public health surveillance tools, which provide population-level mortality and incidence data, are routinely managed by health institutions such as *Centers for Disease Control and Prevention (CDC)*, European Influenza Surveillance Scheme (EISS), and *National Health Services (NHS)*. For example, during the COVID-19 pandemic these institutions put a considerable effort into collecting and analysing COVID-19 clinical data

Table 2.6: English tweet classification tasks ran in SMM4H shared tasks competitions from 2016 to 2020. Table is constructed from [177, 175, 212, 211, 93]

No.	**Task**	**Class description**
1	Automatic classification of tweets mentioning an ADR (2016, 2017, 2018, 2019,2020)	(i) Presence of ADR, (ii) Absence of ADR
2	Medication intake classification (2017, 2018)	(i) Definite intake, (ii) Possible intake, (iii) No intake
3	Vaccine behaviour classification (2018)	(i) Positive, (ii) Negative
4	Automatic classification of personal mentions of health (2019)	(i) Personal health status, (ii) Opinion
5	Automatic classification of tweets that mention medications (2020)	(i) Mention a medication, (ii) Do not mention a medication
6	Automatic characterization of prescription medication abuse chatter in tweets (2020)	(i) Potential abuse/misuse, (ii) Non-abuse/misuse consumption, (iii) Medication mention only without any indication of consumption, (iv) Unrelated
7	Automatic classification of tweets reporting a birth defect pregnancy outcome (2020)	(i) Defect, (ii) Possible defect, (iii) Non-defect

to publish the daily number of new cases, as well as predicting the number of future COVID-19 cases. Social media and internet data can play a crucial role in supplementing such public health data by predicting the hidden tendency of an imminent outbreak [184]. Shen et al [184] retrospectively performed a large-scale study to identify self-reported daily sick posts from a Chinese social media platform using NLP and

machine learning algorithms. The daily count of COVID-19 cases from social media was found to correlate with the data published by the Chinese government. Similarly, Golder et al [69] in the United Kingdom analysed location-based COVID-19 Twitter messages by employing an NLP and machine learning based classification approach to categorise tweets into *Probable*, *Possible*, and *Other* classes. The weekly counts for the first three months in 2020 yielded from the Twitter classification model was shown to correlate with the weekly mortality and new case data published by the UK government [69].

Previously Ginsberg et al [67] showed a correlation between the occurrence of search queries containing flu-related words and *Influenza Like Illness (ILI)* rates published by the CDC. Aramaki et al [10] trained an SVM model to label tweets as flu-related or flu-unrelated and then evaluated the correlation of ILI rates from the Infection Disease Surveillance Center (IDSC) of Japan. Byrd et al [28] demonstrated an approach that employed first a sentiment analysis method to identify authors of tweets affected by influenza and then a visualisation tool to display the result of the analysis. Lin et al [107] proposed a rule-based algorithm which utilised a dependency parser and a NER tagger to filter self-reported tweets mentioning flu. They further deployed several multi-class classification models using n-gram features for the same task. Their experiments achieved a higher classification accuracy when they combined the rule-based algorithm with the machine learning models.

Rudra et al [167] proposed to classify tweets related to Ebola and MERS in several categories related to disease detection using SVM, LR, and Naive-Bayes classifiers. They built two types of models for each classifier using (i) MetaMap, where words of a tweet were mapped to UMLS semantic types, or (ii) BOW, where uni-grams from tweets were extracted. They reported that SVM based models outperformed other models. In addition to this, they found that models built using MetaMap performed better in the case of transfer learning. Serban et al [60] proposed SENTINEL, an end-to-end NLP software system which combined various open source tools for processing pub-

licly available social media data. The NLP system integrated data from heterogeneous sources that are automatically processed for detecting disease outbreaks in real-time. Infectious diseases aside, Esperanca et al [61] demonstrated an NLP and machine learning based framework for tracking chronic diseases such as asthma, cancer, and diabetes from social media. The framework consisted of an automated workflow designed to collect data from social media platforms, filter the data based on geographical criteria (state and national level in the US), and extract tweets relevant to a target disease by deploying an SVM classification model. They collected incidence and mortality statistics from their classification results, despite low correlations with the CDC data at a state level, their findings revealed that it was possible to track diseases using social media at the national level. Finally, [30], [55], and [18] provided a thorough review of the use of Twitter in public health surveillance for the purpose of monitoring, detecting and forecasting ILI and other diseases.

2.7 Conclusion

In this chapter, we provided a broad overview of the information extraction methodologies and relevant literature reviews concerning medical social media. In Section 2.1, we discussed different social media sources, broad categories of NLP tasks on medical social media, benchmark datasets, and shared tasks related to social media. We also provided a brief list of datasets curated for our tasks. Section 2.2 discussed relevant literature on rule-based concept and relation extraction on social media, and sentiment analysis methods on medical social media. One key observation is that anaphora was not considered to analyse a full post comprising multiple sentences. In Section 2.3, we described CRFs and BiLSTM networks, and discussed contextual language models. In addition to this, we provided a literature review concerning supervised concept extraction methodologies on social media. Section 2.4 described several semi-supervised methodologies and relevant literature. We found that most supervised concept extrac-

tion methods employed dictionary features and word embeddings. However, automatically obtaining new concepts from social media posts has received little attention despite the fact that social media is dynamic and concepts change over time. Section 2.5 discussed relevant literature of deep learning models employed for COVID-19 medical concept extraction, gazetteer inclusion with deep learning, transfer learning, and weak supervision. The literature review revealed that dictionaries were useful in deep learning and transfer learning settings. However, specific dictionaries curated for emergent diseases such as COVID-19 have not been investigated. Finally, Section 2.6 discussed various text classification techniques utilised in social media focusing on the SMM4H shared tasks. We identified that at the time of this research no other study took a triage and diagnostic approach to extract actionable information regarding COVID-19 from social media.

Chapter 3

Rule-based Health Information Extraction

The content of this chapter is adapted from our article published in:
A. Hasan, M. Levene, D.J. Weston. "Natural language analysis of online health forums". In: *International Symposium on Intelligent Data Analysis*. Springer, 2017, pp. 125–137.

3.1 Overview

Health related posts in medical forums often contain factual information regarding drug usage, the patients' sentiment of a drug when used to treat a symptom, and the experience of any adverse effects from the drug. In order to extract this type information from free text in medical forums, natural language analysis of the text is required [48]. As has been pointed out by Wang et al. [209], detecting, from unstructured text, the disease, treatment and symptom entities, their attributes and existing relationships, is a major research issue in the NLP domain. Although progress has been made in extracting entities, there is still the challenge of extracting specific relationships between these entities [77]. In particular, here we are interested in extracting relationships of the form (treatment, polarity/sentiment, symptoms/side-effects), which represent relationships that provide us with information on patients' sentiment of various treat-

ments, especially medication. When aggregated over many forum posts such triples could inform practitioners and/or patients on the effectiveness of treatments beyond the information gathered from studies published in medical journals and by the pharmaceutical companies.

In this chapter, we report on a base-line rule-based approach for extracting such triples from about 1000 posts related to Parkinson's disease from the PatientsLikeMe website [149], providing details of the algorithm we deployed and the results from a comprehensive evaluation of the algorithm. It is important to note that patients' comments in a forum, such as the one we are analysing, will contain slang and verbose, informal, descriptions of treatments and side-effects (for example, using "body shaking" instead of the more formal "tremor"). Such informal terms are not normally present in standard ontologies such as the (UMLS) [23]. As a result of this difficulty, much of the previous research in this area has focused on extracting formal medical terms from the free text. We now give a brief summary of our NLP relationship extraction system, whose aim is to build triples, which can be aggregated to provide useful statistical medical information relating to the patients' sentiment of various treatments. Our model makes use of the following *concepts*:

1. Drugs (X), or more generally, treatments.
2. Symptoms (Y), which the drug is meant to treat.
3. Side-effects (Z), which are caused by the drugs.
4. Polarity (P), which indicates how positive a treatment is or how negative a side-effect is from the patient's point of view.

The system first extracts different health related concepts from the forum posts, and then creates structured information by forming a relationship between a drug and a symptom or side-effect, through polarity analysis of the text. We termed such a relationship formally as a *disease triple* henceforth simply a *triple*.

3.2 Contribution

Despite machine learning techniques such as CRF [99, 193, 125] being very effective in NLP information extraction, we have chosen to build a *rule-based system* [120]. We show that dictionaries and rules built here can be used to build a CRF model in the next chapter. Our contributions are as follows:

1. To the best our knowledge, this is the first attempt to extract, from social media, relationships in the form of disease triples, which include patient sentiment. There is no base-line system for such work and in order to attain deep knowledge of the use of natural language, specialised rules are often needed.

2. Dictionaries, lexicons and ontologies in the medical domain are built for extraction tasks from documents written by experts [77]. However, patients are, in most cases, not familiar with this terminology, so they tend to use commonly understood terms. As a result matching to such pre-built dictionaries often results in poor performance. In order to build common domain knowledge, it is first necessary to manually analyse a significant number of posts, and extend publicly available lexicons and gazetteers using a specifically designed set of generalised rules for extracting structured information from the free text.

3. Once the base-line is established it is possible to export the set of designed rules and resulting extended gazetteers to be used in a more sophisticated machine learning such as CRF, to attain transferability and scalability when extracting triples from other forums.

Our overall contribution is the summarisation of health related forum posts by identifying relationship between concepts in the form of disease triples to provide a coherent structure, which can be used to extract meaningful medical statistical information.

Table 3.1: Example of disease triples after processing a post. +, -, symp, side, drug , list, con and intens, denote positive polarity, negative polarity, symptom, side-effects, drug/treatment, list of nouns, conjunction and intensifier, respectively

Sentence or Sentence Segment	**Disease triple**
I take 600mg of gabapentin$_{drug}$ at bedtime, helps$_{+}$ me shake$_{symp}$ and$_{list}$ kick$_{symp}$ less$_{-}$;	(gabapentin,+,shake) (gabapentin,+,kick)
and$_{con}$ a donepezil$_{drug}$ 10mg, settles$_{+}$ me down allowing sleep$_{symp}$.	(donepezil,+,sleep)
Clonazapam$_{drug}$ works$_{+}$ great$_{intens}$	(Clonazapam,+,?)
but$_{con}$ I can't take$_{-}$ the groggy$_{side}$, foggy head$_{side}$ the next day.	(Clonazapam$_{drug\ anaphora}$,-,groggy) (Clonazapam$_{drug\ anaphora}$,-,foggy head)

A motivating example

Let us examine the following example post to motivate the research:
"I take 600mg of gabapentin at bedtime, helps me shake and kick less; and a donepezil 10mg, settles me down allowing sleep. Clonazapam works great but I can't take the groggy, foggy head the next day."

After various pre-processing steps and the application of linguistic rules, we create the disease triples as shown in Table 3.1 . Most of the existing works concentrated on processing a single sentence. However, our method is a formal approach to make sense from the whole post which may contain several sentences by making use of *anaphoric relations* [71] present in the sentences.

3.3 Materials and Methods

The overall methodology is schematically shown in Figure 3.1 and the corresponding pseudo-code of the algorithm is presented in Algorithm 1 below. In the following subsections we describe the dataset, the procedure followed for verifying annotations, and the methodology in more detail.

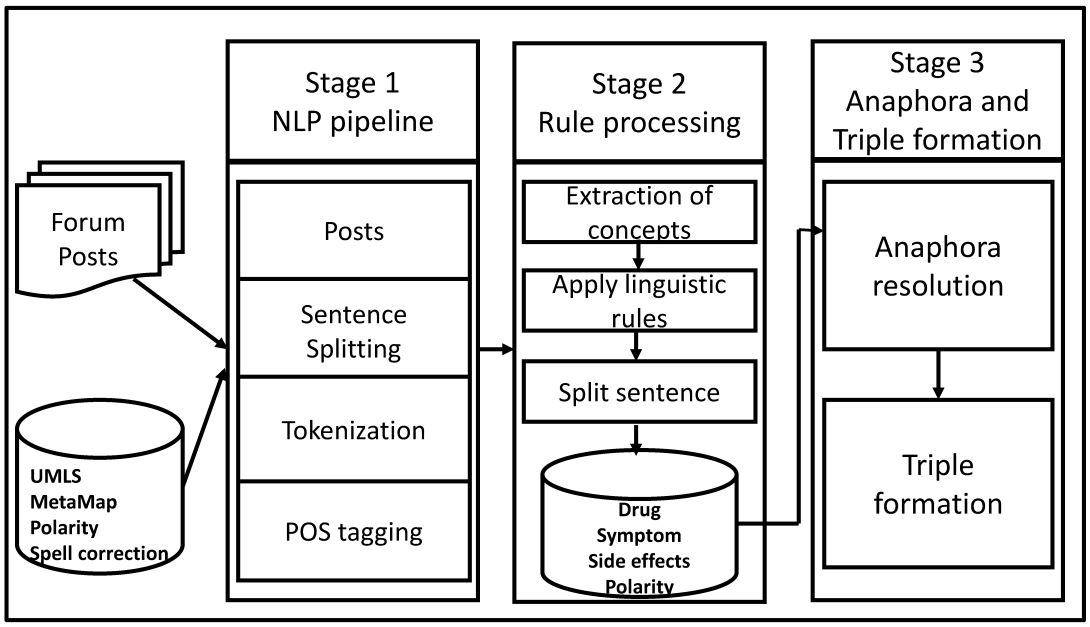


Figure 3.1: Text processing architecture for our rule-based approach.

3.3.1 Dataset

PatientsLikeMe [149] is an online health discussion forum, where patients with chronic health conditions can share their experiences living with disease. The forum is organised in groups of patients experiencing similar health conditions. For our study, we extracted user comments from the Parkinson's disease group. After registering with this website, web pages related to Parkinson's were downloaded manually to conform with terms and conditions of the website before being scraped offline. Finally, the posts were anonymised by removing user IDs. A total of 1058 posts were collected from the

```
Data: P is a list of Posts
foreach p in P do
    Split p into a list of sentences, S;
    foreach s in S do
        Tokenise s into a list of tokens, T;
        foreach t in T do
            Append the POS information;
            Identify the concept class, C matching with the gazetteers;
            Let, C = {X, Y, Z, P} where X, Y, Z, P are drug, symptom,
             side-effect, and polarity, respectively;
            Disambiguate concepts t;
            Calculate the polarity, p by applying linguistic rules for t in P;
        end
        Split s into a list of segments, G using conjunctions, and, but, until;
        foreach g in G do
            Compute polarity score, SC;
            Create a list of triples, L;
            A triple is either (X, Y, SC) or (X, Z, SC);
            Where, "?" is the placeholder for a missing concept;
            foreach l in L do
                Perform anaphora resolution for missing X;
            end
        end
    end
end
```

Algorithm 1: Text processing algorithm

period of April, 2016 to June, 2016.

500 posts were used for training and 400 for testing the system. The remaining posts were used for the annotation validation, described in the next section, where 58 posts were used to train the annotators and the remaining 100 posts for cross validation of the annotations.

3.3.2 Annotation validation

The annotation for the dataset was carried out by the author of this thesis. In order to verify the fidelity of these annotations an experiment was conducted using a small subset of the data, where the level of agreement between the annotator and other annotators was measured.

Ten researchers from Birkbeck's department of Computer Science and Information Systems volunteered for the validation experiment. Annotators were trained by showing annotated posts (20 posts were chosen from the annotator training set of 58) and explaining each concept and triple types. The remaining 100 posts were divided randomly into five sets of 20. Each of the ten annotators were randomly assigned two sets such that each set would get two annotations from different annotators. We followed Nikfarjam et al [143] who used Cohen κ statistic [215] to measure the agreement between pairs of annotators, in addition we used accuracy. A very high level of agreement in recognising drug, symptom, positive and negative strings was achieved. However, agreement and accuracy in recognising triples and side-effects were somewhat lower (72.09% and 88.79% respectively). It was subsequently determined that two of the annotators had not fully understood the task. Table 3.2 shows the results with and without these two 'outlier' annotators. It can be seen that there is a better agreement in identifying triples, however agreement in identifying side-effects has been reduced. This is due to the small number of side-effect concepts in the validation set (7 in total).

Table 3.2: Annotation validation result. κ-O and Accuracy-O are the results after discounting the 2 'outlier' annotators.

Concept	κ	**Accuracy**	κ**-O**	**Accuracy-O**	**Support**
Drug and Treatment	87.37%	94.40%	86.84%	94.16%	446
Symptom	91.49%	96.90%	92.73%	97.32%	162
Side-effects	76.18%	99.40%	71.05%	99.23%	7
Positive polarity	89.71%	96.26%	89.06%	95.97%	112
Negative polarity	90.72%	97.15%	89.77%	97.02%	146
Triples	72.09%	88.79%	76.18%	90.72%	1479

3.3.3 NLP pipeline

Text processing tools, e.g. *General Architecture for Text Engineering (GATE)* [45, 46], Stanford Core NLP [121], and spacy [190], help in constructing an NLP pipeline for the language processing tasks. Our text processing pipeline, constructed using GATE, splits the posts into sentences, tokenises the text and labels the tokens with their POS tags. At this stage, our system recognises drugs, symptoms, side-effects and opinions present in the text by using different lexicons and plug-ins which were RxNorm terminology [169], MetaMap [12], COSTART (Common Standard Thesaurus of Adverse Reaction Terms) [42], and MPQA Subjectivity Lexicon [214], respectively. Apart from using some publicly available resources, we constructed gazetteers using domain knowledge and extended these by augmenting the terms during the training phase of the system. To recognise a symptom from a span of text using dictionaries we produced several *java annotation patterns engine (JAPE)* rules; an example of such a rule is shown in Figure 3.2. The *left-hand side (LHS)* is the part preceding the '–>' and the *right hand side (RHS)* is the part following it. The LHS is the pattern and RHS specifies the annotation on that pattern. In Figure 3.2, *Lookup.majorType=symptom* is the pattern, and *symptom.Symptom* is the annotation.

```
Phase: Symptom
Input: Lookup
Options: control = appelt debug = true
Rule: Symptom1
(
Lookup.majorType == symptom
)
:symptom
 –>
:symptom.Symptom = {rule = ”Symptom1”}
```

Figure 3.2: An example of rule processing using GATE text processing tool. In this rule, a text is annotated as *Symptom* if it is found by a look up operation on the text span.

3.3.4 Rule processing

We disambiguate multiple concepts recognised in the previous stage by applying different linguistic rules, then split the sentences and compute the polarity score of each resulting segment.

Disambiguation and extraction of concepts

We extracted four types of concepts from each sentence, i.e.: drugs, symptoms, side-effects and polarities. We now briefly discuss the concept extraction and disambiguation process.

- **Drug extraction.** We used a subset of the RxNorm terminology [169] as our drug gazetteer, where the vocabulary consisted of drugs and treatments used for Parkinson's disease; RxNorm is a normalised naming system for generic and branded drugs. For our purpose, it is very important to recognise the synonymy of generic and branded drug names to avoid extracting the same type of drug mention multiple times. As such we have constructed two dictionaries for generic and branded drugs. A feature is added to the drug token according to the drug

gazetteer it is extracted from. If a sentence contains both prescription and generic drug mentions, the generic drug mention is subsumed. Spelling mistakes in drug names are corrected using a normalised edit distance of two based on *Levenshtein's edit distance* [119].

- **Symptom extraction.** We used the MetaMap [12] annotation plug-in to annotate symptoms (sign or symptom semantic types) in sentences. We also constructed a separate symptom gazetteer using domain knowledge, and terms such as "voice", "smell" and "restless" were also added to the gazetteer. By default the polarity feature of symptoms recognised by the MetaMap program are set to negative. The symptoms gazetteer contains explicit polarity for each symptom. Not all the symptoms present in the gazetteer are negative, for example concepts such as "sleep" and "energy" are marked as positive. The polarity feature is extracted from the gazetteer for each symptom.

- **Side-effect extraction.** For side-effects, a gazetteer using COSTART [42] is constructed. We extended the dictionary by adding new terms during the training phase.

- **Polarity extraction.** We have collected the polarity terms from the MPQA Subjectivity Lexicon [214], which contains more than 8000 words annotated manually by the authors as positive, negative or neutral. The lexicon also includes the POS information for the terms. Symptom and side-effect terms present in the dictionary are not labelled as polarity terms. The prior polarity score for positive and negative terms are set to +1 and -1 respectively, and neutral words are given as score 0. Our system matches a token with the lexicon if the POS information present in the lexicon is the same as the POS category of the token in the sentence.

Linguistic rules.

Matching polarity words is not enough in order to extract opinion from a sentence, and it often produces wrong result. A description of how valence of a lexical item is

modified by the presence of different lexical items such as "negation", "modifiers" and "presuppositionals" can be found in [157]. In similarity to [157] we applied following heuristic linguistic rules:

- **Negation.** If a negation word such as "not"and "don't" precedes in 0 to 3 tokens of a polarity, symptom or side-effect concept, then the polarity of each concept is reversed and will generate the feature `negation_concept` (for example, "doesn't_work").

- **Modifiers.** Modifiers such as intensifying adverbs (for example, "very", "strongly"), diminishing adverbs (for example, "little", "kind of") increases or decreases the sentiment value of a concept and generate a feature `modifier_concept` (for example, "very_tired"). This rule modifies polarity of the first matched concept at a distance of 0 to 3 tokens.

- **Presuppositional.** The polarity of presuppositional items such as "barely" is multiplied with that of the concepts and polarity is flipped as a result. A feature such as `pre_concept` (for example, "barely_tremor") is generated after applying this rule. As for modifiers, this rule also changes the polarity of the first matched concept at a distance of 0 to 3 tokens.

Split sentence

In this step, we first split a sentence and then calculate the final polarity score. These two steps are described below:

- **Sentence segmentation.** We used common conjunctions (*"and"*, *"but"*) and *"until"* to segment a sentence . Our analysis revealed that *"and"* is sometimes used to connect two or more words to form a list rather than connecting two different parts of a sentence. We constructed a rule to find such conjunctions. For example if POS tags of the two tokens in either end of *"and"* are same, then it is denoted as a list of tokens and we do not segment sentence in such cases.

- **Final polarity score.** We add the polarity scores of all the opinion concepts in a sentence or segment of a sentence. The polarity of a triple is positive if total score is more than 0, negative if it is less than 0. If the score is 0, then polarity of the triple will be that of the symptom or side-effect.

3.3.5 Triple formation

At this stage, we first perform anaphora resolution and then form triples. These two steps are described in the following two paragraphs.

Anaphora resolution.

Messages in a forum contain sentences referring to the concepts mentioned earlier in the text. Our rule for finding anaphoric references for drug is briefly: if the current sentence has a drug mention, then the drug is carried forward to the next sentence in the text. Using this rule, if a triple has a drug mention and a subsequent triple contains the default drug concept ("?"), then we replace the default with the drug found, and repeat the same process for all the sentences in text until we find a new drug mention. If we find multiple mentions of drugs then multiple triples are created containing each drug mention. However, we note there are limitations to this approach; for example, when a sentence has multiple drug mentions with a single symptom mention such triples will not capture the fact that the two drugs jointly lead to the symptom. We also look for the patterns such as "from X to Y", which indicates that the person actually stopped using drug "X" and moved on to using "Y". In such case, we add a feature to the drug ("X") token indicating that the algorithm should stop creating triple for this drug in subsequent sentences.

Disease triples.

We create a list of concepts by ordering them according to their token offset from the beginning of a sentence. Triples are formed using the following format:

Table 3.3: Test set summary

Posts	Sentences	zero triples	one triple	more than one triple
400	2564	544	1447	449

1. *Triple 1*: Drug, Polarity, Symptom

2. *Triple 2*: Drug, Polarity, Side-effect.

The algorithm iteratively finds drug, polarity and symptom or side-effect concepts using the order shown in the formation of triple. A triple is formed by taking three consecutive concepts of a different kind. In our algorithm, consecutive concepts (for example, drugs) of same kind, signals the starting point of a new triple. The algorithm places a default concept which is "?" in case of missing concepts.

3.4 Results

The following subsections describe experiment and its results.

3.4.1 Evaluation

To evaluate our proposed approach the standard measures of accuracy (Acc), precision (P), recall (R) and F_1 [215] were used. Each post was split into segmented sentences as described in Section 3.3.4. Triples formed from a segmented sentence are then merged with those from other segments and subsumed in case of repetition. A sentence can contain zero of more concepts and consequently zero or more disease-triples, as shown in Table 3.3. To evaluate concepts and triples predicted by the system with those present in the actual annotated set. If a concept is correctly predicted by the system, then it is a true positive case (TP). If a concept is predicted by the system, but is missing in the annotation set, then it is a false positive case (FP). If the system failed to recognise a concept, then it is a false negative (FN) case. Lastly, if there is no

concept predicted by the system and same in the actual annotation, then it is a true negative (TN) case. In case of triples, a sentence is segmented (see Section 3.3.4). Each segmented sentence can have zero or more triples. The evaluation metric for triples follows the same definition as concepts.

3.4.2 Training and Test Results

Training of the system was conducted incrementally over 5 iterations. The training data was split into 5 sets of 100 posts each. At the beginning of the first iteration, we analysed the posts from the first set. This meant we annotated them with the concepts as discussed in Section 3.3.2. We then manually generated rules which are shown in Section 3.3.4. The annotation produced new concepts which we included in our dictionaries which we discussed in Section 3.3.4. The system was evaluated on the same set of posts and more rules were manually added until we achieved a satisfactory performance, which meant a precision of approximately 80% or above. After which we then moved on to the next iteration and followed the same procedure. The evaluation was carried out at each iteration by cumulatively adding a new set of posts to the posts from previous iterations. It is interesting to see the overall performance from the training data, Table 3.4. This was achieved in a principled manner, as described in Section 3.3.4, involving the development of as few rules as possible. It should be noted that without anaphora resolution very few triples would have been successfully identified. In addition, dictionaries were extended when necessary.

At the end of training phase, we ran all 500 posts on the system built and evaluated the performance which is shown in Table 3.4. We achieved high precision in recognising drug, symptom and side-effects concepts. This is because, as described earlier (see Section 3.3.4), we scaled the drug dictionary according to the drug and treatments used for Parkinson's, restricted MetaMap[12] to recognise sign and symptom semantic types, extended COSTART[42] dictionary, and used set of generalised rules to disambiguate concepts. For testing the system was run over the remaining 400 post test

Table 3.4: Training results for 500 posts

Concept	**Acc**	**P**	**R**	**F_1**
Drug & Treatment	95.06%	99.63%	99.12%	97.29%
Symptom	95.71%	85.64%	99.06%	91.86%
Side-effects	99.06%	87.95%	98.50%	92.92%
Positive polarity	90.19%	80.98%	96.15%	87.92%
Negative polarity	90.61%	79.60%	94.04%	86.22%
Triple 1	83.06%	81.01%	95.23%	87.54 %
Triple 2	84.76%	82.28 %	94.96%	88.16%

dataset, without any modification to the system. The results are shown in Table 3.5. We can see that the system has generalised well. Naturally the disease triple identification has had the greatest fall in performance, since recognising these relationships is dependent on accurately identifying the concepts from which they are comprised.

Table 3.5: Test results for 400 posts. Acc, P, R, F_1 denote Accuracy, Precision, Recall, and F_1-score, respectively.

Concept	**Acc**	**P**	**R**	**F_1**
Drug & Treatment	90.71%	88.29%	95.14%	91.59%
Symptom	94.26%	84.08%	87.36%	85.69%
Side-effects	98.42%	80.25%	93.53%	86.38%
Positive polarity	86.44%	72.68%	94.42%	82.13%
Negative polarity	87.08%	73.57%	88.52%	80.35%
Triple 1	73.93%	71.11%	96.02%	81.71%
Triple 2	74.47%	71.31%	96.81%	82.13%

Table 3.6: Error analysis of triples created from the sentence in Example 1. Bold denotes wrong triple. For details of subscripts see Table 3.1

Sentence or Sentence segment	**Actual Disease triple**	**Predicted Disease triple**
My wife was on $\text{Rytary}_{\text{drug}}$ 36.25/145 mg for 5 days and_{con} returned to C/L_{drug} today because it was not working. and_{con} she was getting side effects.	(Rytary,-,?) (C/L,?,?)	(Rytary,-,?) **(C/L,-,?)**

3.5 Discussion

Though we are very successful in recognising concepts, the system makes a few mistakes in disambiguating polarity terms. As a result, the performances at triple level are lower, which resembles that of recognising positive and negative polarity terms. This result is in line with our hypothesis (see Section 3.1) that we can establish a relation between drug and symptom and drug and side-effects through the polarity of a sentence. Although the polarity dictionary [214] has been extended by incorporating common phrases and is also supported by set of generalised rules, there is still room for improvement.

3.5.1 Error analysis

In Table 3.6, the system makes a mistake as the reference to "it" in the second segment refers to the previous drug mentioned in the sentence.

Example 1. *"My wife was on Rytary 36.25/145 mg for 5 days and returned to C/L today because it was not working and she was getting side effects"*

Table 3.7: Error analysis of triples created from the sentence in Example 2. Bold denotes wrong triple. For details of subscripts see Table 3.1

Sentence or Sentence segment	**Actual Disease triple**	**Predicted Disease triple**
My update- -I have settled$_{+}$ out on Rytary$_{\text{drug}}$ 23.75/95 (4 capsules) 4x a day–MUCH$_{\text{intense}}$ better$_{+}$ than on Stalevo$_{\text{drug}}$—no side effects$_{+}$–except perhaps a bit more$_{\text{intense}}$ energetic$_{\text{side}}$ than I should be-lol	(Rytary,+,?) (Rytary,-,energetic) (Stalevo,-,?)	(Rytary,+,?) **(Stalevo,+,energetic)**

Example 2. *" My update- -I have settled out on Rytary 23.75/95 (4 capsules) 4x a day–MUCH better than on Stalevo—no side effects–except perhaps a bit more energetic than I should be-lol"*

In Table 3.7 the system makes two mistakes by putting the side-effects concept with the wrong drug/treatment mention.

3.6 Conclusion

In this chapter, we proposed representing potential useful medical information in free-form unstructured text, with disease-triples. To attain our first objective outlined in Section 1.3, we have developed a strong base-line system. We achieved an F_1 score of over 80%, in identifying these disease-triples using traditional NLP methods, and have demonstrated that this approach can generalise successfully. In the next chapter, our objective is to build a supervised and semi-supervised medical concept extraction methods utilising labelled and unlabelled datasets.

Chapter 4

Supervised and semi-supervised concept extraction

The content of this chapter is adapted from our article published in:
A. Hasan, M. Levene, D.J. Weston. "Learning structured medical information from social media". *Journal of Biomedical Informatics*, 2020 Oct;110:103568.

4.1 Overview

In the previous chapter, we built a framework for concept relation extraction using a NLP methodology by utilising health related concepts augmented with sentiment as expressed in the text. Medical social media such as MedHelp [126] and Twitter, often contain experiential information from patients who share symptoms and side-effects of the prescribed treatments. These shared experiences from a group of patients have been proven to be useful for public health monitoring [150, 103]. To further advance such findings, a group of researchers from the University of Pennsylvania has been organising the SMM4H shared tasks to detect ADR from tweets [211]. However, as shown in the examples of Figure 4.1, social media posts contain not just ADR mentions, they can also include other useful information such as a patients' sentiment regarding their medical condition. In this chapter, we wish to identify not just ADRs but also effects of a drug that may not be the intended therapeutic outcome and indeed

might be considered beneficial, hence we use the term side-effect, [57].
The methodology, developed in Chapter 3, was rule-based together with lexicon matching. It is possible that terms contained in social media text may not exist in the publicly available dictionaries and ontologies such as UMLS [23]. Consequently, recognising concepts from such colloquial text using lexicon matching algorithms often produce poor results [143]. To address this challenge, researchers have applied supervised machine learning methods, which requires manually annotated training data.

In this chapter, we present a semi-supervised methodology based on CRFs [99, 193, 125], which classifies tokens in a sentence belonging to one of the categories shown in Table 4.1. We believe these classes cover the semantics pertaining to the objective of the research. The rules in Chapter 3 Section 3.3.4 were, heavily dependent on the lexicon matching, and inferred from the training dataset by manually analysing the text. Some lexicons were publicly available, and others were curated manually from the training set. Whereas in this chapter, the aim is to automate, with minimal supervision, the dependency on the labelled data and the manually created lexicon.

First, a small number of posts and tweets are sampled and annotated. The CRF model was trained on a proportion of the sample, and then this model was applied to the unlabelled data iteratively in order to tag and collect highly confident labelled sentences, symptom and side-effect terms. In an online setting, where data becomes available continuously, as the language changes, the semi-supervised methodology would allow us to automate the incorporation of new terms into dictionaries and be able to adapt to the domain changes with minimal human effort. Here we show that within a single disease category, i.e. Parkinson's, such a continuous training process will either improve or maintain the F_1 score. We thus believe that our method has the additional potential to be used across disease categories with minimal effort, and can be scaled to the practical use needed in medical applications.

In contrast to studies in [143, 132], which focused on medical social media, we deal with more classes and extend the self-training technique [239, 59] to enlarge the

Example

T I envy you I can take 75[CD] mg[DOSE] of melatonin[D] and never[NG] fall[SYM] asleep[SYM].

T I take those mirapex[D] now for Percocet[D] withdrawals[SD].

T Roprineral[D] tablets from the doc for restless[SYM] leg[SYM] , they are helping[P] me .

M Sinemet[D] increases CNS[SYM] dopamine[SYM] which can lead to psychosis[SD].

M My sinemet[D] had been worn[N] off[N] for 5[CD] hours[TMCO] I take it every two[CD] to three[CD] hours[TMCO].

M My hands[BPOC] feel really[INT] weak[SYM] , but I am able[P] to still[INT] function[P].

Figure 4.1: T1, T2 and T3 are examples from Twitter and M1, M2 and M3 are those from the MedHelp dataset. The class label of a token is given inside a square bracket. The description of labels is listed in Table 4.1.

training dataset within a semi-supervised framework. Moreover, our methodology involves in minimal human supervision as opposed to the study in [27]. Our system architecture is relatively simple, each class label is coupled with a dictionary feature, and in addition MetaMap [23] is used to determine a small number of useful UMLS

semantic types from the text.

Accumulating structured information in the form of a dictionary, which is another point of difference with previous research in this area, has direct impact on the prediction of concepts in a supervised classification task. For example, in 2019 SMM4H shared task, the KFU-NLP team, combined contextual word embeddings and BERT [49] with dictionary features, and achieved the top result in the identifying ADR extraction task [211]. Other supervised methodologies such as [143] also rely heavily on lexicons for the improvement in the classification task. Moreover, in Chapter 6, we show that a COVID-19 symptom lexicon is able to produce *weak labels*. Thus, we believe that automatic expansion of dictionaries will allow us to perform incremental learning which is a different task from ontology population [223] and expansion of consumer health vocabulary [74].

4.2 Contribution

We make several contributions, as follows:

1. We show that with a small amount of manually labelled training data we obtain very good performance, and this can be achieved using a semi-supervised methodology which add labelled sentences to the training data in an iterative fashion.

2. Our methodology incrementally augments symptom and side-effect dictionaries by collecting the most confident terms classified by the model. To the best of our knowledge no other previous work attempted to collect learnt health related concepts from the unlabelled data and reuse them in the dictionaries.

3. We combine the above contributions to extend the traditional self-training method [239], by sharing the knowledge in the training data and dictionaries so that sentences, which were rejected at an earlier iteration can still be added when terms

Table 4.1: Class label, description of the class, and the number of words in the class with the percentage inside a bracket are shown separately for MedHelp and Twitter dataset.

Class	**Description**	**Medhelp**	**Twitter**
D	Drug/Treatment	1233(0.93%)	1822(5.59%)
P	Positive polarity	3453(2.6%)	989(3.04%)
N	Negative polarity	4101(3.0%)	1088(3.34%)
NG	Negation	2265(1.7%)	851(2.61%)
PRE	Pre-suppositionals	325(0.24%)	82(0.25%)
INT	Intensifiers	2514(1.89%)	620(1.9%)
SYM	Symptom	5505(4.14%)	1105(3.39%)
SD	Side-effect	756(0.57%)	580(1.78%)
BPOC	Body parts	2475(1.86%)	233(0.72%)
TMCO	Temporal functions	3572(2.69%)	871(2.67%)
CD	Numbers	3347(2.52%)	564(1.73%)
DOSE	Dosage information	206(0.15%)	211(0.65%)
O	Other	103194(77.62%)	23549(72.31%)
Total		132946	32565

are correctly classified at a later iteration.

We tested our methodology on two datasets: the first with posts on Parkinson's from MedHelp, and the second with tweets from Twitter. We then evaluated the performance of the semi-supervised methodology on both data sources by using 100 runs with repeated cross validation; see Section 4.4.1. To compare the models, we have devised a methodology that can detect potential improvement of the semi-supervised algorithm over the base-line.

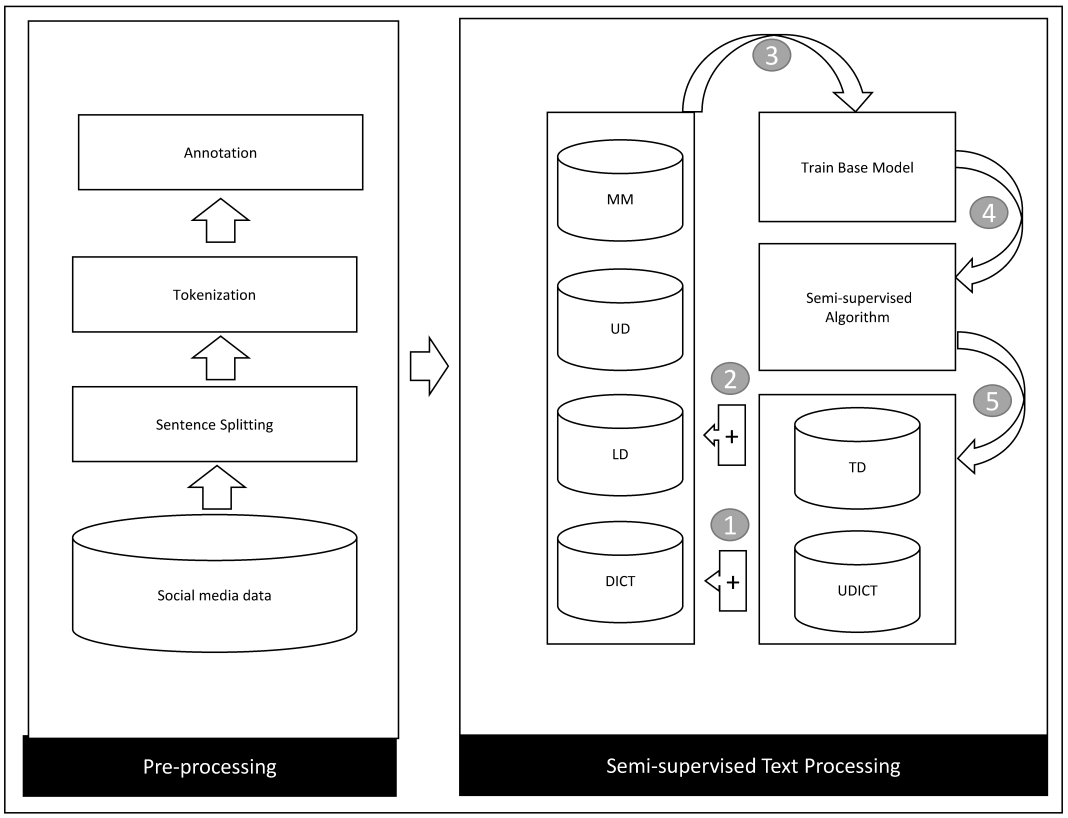


Figure 4.2: The semi-supervised text processing framework. *MM* denotes the MetaMap plug-in for UMLS, *LD* denotes the labelled dataset, *UD* denotes the unlabelled dataset, *DICT* denotes the different publicly available dictionaries used, *TD* denotes the tagged data using the base-line model trained on *UD*, and *UDICT* denotes the dictionaries learnt with the semi-supervised algorithm from the unlabelled dataset. Arrows labelled 1 and 2 denote *UDICT* and *TD* are augmented to *DICT* and *LD*. All other arrows denote sequence order.

4.3 Materials and Methods

Our overall methodology is shown schematically in Figure 4.2 and the corresponding pseudo-code of the semi-supervised algorithm is presented in Algorithm 2. In the following subsections, we first describe the data collection and annotation procedures of the text and then describe the methodology in more detail.

4.3.1 Data collection and annotation

We collected 1000 user posts discussing Parkinson's disease from the MedHelp forum [126] and the same number of tweets from Twitter. MedHelp posts were manually collected from January 2018 to March 2018 using the search system provided by the website [1]. To collect tweets, we used the Twitter search API providing it with a list of known Parkinson's drugs; tweets containing links and photos were excluded from the dataset. The posts and tweets were then anonymised and annotated using the labels shown in Table 4.1. For the semi-supervised learning algorithm, we collected an additional 4,000 tweets and 15,000 MedHelp posts. The sample size for labelled data was calculated at a 95% confidence interval with 4%-6% error margin, which gave us the size of 600-400 posts/tweets, respectively. Here, the labelled and unlabelled datasets are denoted as LD and UD, respectively; see Figure 4.2.

Annotation Validation

In order to verify the fidelity of the annotations carried out by the author of this thesis, an experiment was conducted using a small subset of the data, where the level of agreement between the annotator and other annotators was measured. Eight annotators were trained by showing them annotated posts explaining drug, symptom and side-effect concepts. Each annotator received an average of 18 posts and tweets from a total of 150. The agreement between the first author and the annotators was calculated using Cohen's κ statistic [215]. The overall agreement reached was 75% after discounting an outlier. Though we achieved a very high level of agreement (81%) for the drug concept, the agreement for the symptom and side-effect concepts were lower at 69% and 74% respectively. However, when we combined the symptom and side-effect into a single class, Cohen's κ reached 75%.

[1] `https://www.medhelp.org/search?&query=parkinsons` Accessed: March 2018

Data: LD: Labelled data divided in $train_0$, $test$, and $valid$ sets
$DICT$: Existing dictionaries
UD: Unlabelled data
α: Confidence interval threshold
n: Number of sentences
i_{max}: Maximum number of iterations

$UDICT_0 \leftarrow$ Empty dictionary to store symptom and side-effect predicted from UD;
f_0: *base-line model* trained on $train_0$ and $DICT$;
$i \leftarrow 0$;
repeat
  $TD \leftarrow$ Tag UD by f_i;
  $viterbi_i \leftarrow$ The set of n highest viterbi sentences from TD;
  $train_i \leftarrow train_{i-1} \cup viterbi_i$;
  $UD_i \leftarrow UD_{i-1} - viterbi_i$;
  $UDICT_{i+1} \leftarrow UDICT_i \cup$ Symptom and side-effect terms predicted;
  from TD by f_i according to α;
  $TD_{mark} \leftarrow$ Mark sentences from TD using $UDICT_i$;
  if $i > 0$ **then**
    $mark_i \leftarrow$ The set of n highest Viterbi sentences from corrected TD_{mark};
    $train_i \leftarrow train_i \cup mark_i$;
    $UD_i \leftarrow UD_i - mark_i$;
  $f_{i+1} \leftarrow$ Re train base-line model using $train_i$;
  Extract features from UD using $UDICT_i$;
  Test f_{i+1} on $valid$ and store F_1 score;
until $i < i_{max}$;

Algorithm 2: Semi-supervised training assuming separate symptom and side-effect classes, where $train_0$ is the base line model trained using the labelled dataset.

4.3.2 Training the base-line model

We pre-processed the labelled and unlabelled data, LD and UD respectively, through a built-in feature extraction program in GATE [45] using an NLP pipeline. The NLP pipeline splits the text into sentences and tokens, performs POS tagging, applies lexicons and gazetteers (denoted as $DICT$ in Figure 4.2) to find the membership of a token and integrate it with MetaMap (denoted as MM in Figure 4.2) to infer the UMLS semantic class. The labelled dataset, LD, is divided into training, test and validation sets denoted *train*, *test*, and *valid* respectively. We built a model denoted as the *base-line model* by applying a linear-chain CRF [193]; see Chapter 2 Section 2.3 for a detail description. Let $X = x_0, \ldots, x_t, \ldots, x_T$ be a sequence of tokens. Then our linear chain CRF relies on the following boolean features for each token t:

1. Word based features: There are three word based features; the first is the token itself, t, and two context tokens which are its predecessor and successor namely $t-1$ and $t+1$. This differs from [132] who used a wider context and also differs from Nikfarjam et al [143] who built seven features based on six nearest neighbours to the token, t.

2. Lexicon features: These features represents whether t is a member of one of the following publicly available lexicons. A token can be a member of multiple dictionaries. We make use of the following lexicons:

   (a) The MPQA Subjectivity Lexicon [214] for polarity detection.

   (b) The RXNORM [169] drug lexicon.

   (c) Prepositional, negation and intensifier lexicons built from Chapter 3 Section 3.3.4. These dictionaries were built using common language usage. Prepositionals flip the polarity of a symptom (e.g., *hardly, barely*), intensifiers are used to intensify the polarity of an expression (e.g., *more*), and negations change the positive polarity to negative and vice versa. A token is matched

with all these dictionaries to set these features on/off.

(d) Symptom dictionary with 180 terms commonly used with Parkinson's disease.

(e) The SIDER [98] dictionary extended with the terms that occurred frequently in the training sequences.

3. MetaMap mapping: The feature extraction program integrates MetaMap to map tokens to their corresponding semantic classes. We set three features depending on the semantic class the token is mapped to:

   (a) Organic Chemical, *ORCH* and Pharmacologic Substance, *PHSU*,

   (b) Sign or Symptom, *SOSY*, and Disease or Syndrome, *DSYN*, and

   (c) Body Part, Organ, or Organ Component, *BPOC*.

4. Rule-based: Our feature extraction program identifies whether:

   (a) The token, t is a member of the built-in temporal gazetteer in GATE, and

   (b) The POS tag of t is *CD* type.

Once features are extracted from the text, the base-line model is trained and tested using the *train* and *test* datasets, respectively by using a Python wrapper [44] for CRFsuite, see [147]. For training , we used the limited-memory BFGS [144] gradient descent technique, which is in-built. The training procedure is set with the default regularisation parameters and a maximum of 100 iterations. See Section 4.4.1 for a discussion on the distribution of the training and test datasets and procedure for cross-validation followed in this study.

The pre-trained base-line model is applied to the unlabelled data, UD, to obtain the tagged sentences, TD, and new symptom and side-effect terms in $UDICT$, as shown in Figure 4.2. TD and $UDICT$ are then selected by the semi-supervised algorithm to augment the original training data and the existing dictionaries, *train* and $DICT$, respectively.

4.3.3 The semi-supervised algorithm

Semi-supervised learning involves using both labelled and unlabelled data to train a model [239]. In our approach we first build a CRF based purely on the labelled training data (the base-line model). This model is then used to predict the labels for the unlabelled training data. We then analyse these predicted labels to identify new words to be included in the dictionaries, which are described in the previous section. We also identify sentences to be included in the labelled training data and removed from the unlabelled set. The CRF is then rebuilt using these updates and the process is repeated until a stopping criterion has been met. We first summarise our semi-supervised method as follows:

1. Train the base-line model using the labelled data which we call $train_0$ in Algorithm 2.
2. Repeat the following steps until the stopping criteria is satisfied:
   (a) Tag the unlabelled data using the base-line model.
   (b) Identify most confident sentences from the tagged data, add them to the labelled set.
   (c) Identify new symptom and side-effect terms, add them to their relevant dictionary.
   (d) Flag sentences where newly identified symptom and side-effect terms are misclassified.
      - Identify any flagged sentences that subsequently have had their misclassified terms correctly classified. Add these sentences to the labelled set.
   (e) Rebuild the base-line CRF model with the above updates and record the performance on the validation set.

The method is shown in Algorithm 2 and is described in more detail as follows.

Identifying new dictionary terms

We are interested in identifying new symptoms and side-effects, consequently we restrict our search for new terms to these two labels. For each unique word, that does not already exist in a dictionary and is not a known stop word, we collate all the predicted labels. A word will be added to the dictionary corresponding to its most frequent label provided we are confident of that predicted label. Our confidence is measured by estimating the standardised *Wald confidence interval*, CI, [204] at at the 95% level i.e.,

$$CI = \hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \tag{4.1}$$

where $\hat{p}$ is estimated probability that the word is assigned its most frequent label and n is the number of instances of the word. If the lower bound of the confidence interval, $\hat{p} - CI$, is greater than a threshold (set to 0.5), denoted by α in the Algorithm 2, we proceed to augment the dictionary with this word.

Identifying sentences

The linear-chain CRF produces the best tagged sequence for a sentence with a score similar to the inference probability produced by a HMM known as the *Viterbi* probability [159, 193]. We use this probability to rank sentences and select at most the top five that have a probability above a threshold of 0.9. Ideally we wish to include only one sentence per iteration, however due to computational constraints we increase this figure to five. There are notable shortcomings of using the Viterbi probability for ranking sentences. First sentences with short length will tend to have a high Viterbi probability and second class label imbalance (the *Other* label dominates) in our data generally results in higher probabilities for sequences labelled with *Other*. To mitigate this bias, we only consider sentences of length greater than 3 that also contain at least one of the drug, symptom or side-effect labels.

The set of highest Viterbi sentences are likely to be similar to sequences that are in the labelled training set [39]. As a result, the augmented training data may lack

in diversity. To overcome this, we introduce an additional approach for identifying sentences to include in the labelled training set. As described above, a word will be added to a dictionary provided there is sufficient consistency in the predicted labels. The sentences that contain the word but which have been mislabelled (i.e. not predicted the most frequent label) are flagged. At any subsequent iteration, all flagged sentences are checked to see if any new dictionary word has now been relabelled correctly. These corrected sentences are ranked by their Viterbi probability and the top five are transferred to the labelled training set.

Stopping Criterion and Model Selection

The model parameters of the CRF are recorded after each iteration and the model that provides the highest F_1 score on the validation set is selected as the final model. The maximum number of iterations is fixed at 100. In our experiment, we found that typically the model converges in 30 iterations for the MedHelp data and 15 iterations for the Twitter data. We believe this difference is due to the difference in size of these datasets.

4.4 Results

We evaluate the performance of the basline and semi-supervised algorithms using precision (P), recall (R) and F-score (F_1). True positives (TP), false positives (FP) and false negatives (FN) are calculated by comparing the model's extracted concept with the actual annotation via exact matching at the individual token level. Here we report both macro and micro averaged F_1 scores. Macro scores are computed by considering the score independently for each class and then taking the average, while micro scores are computed by considering all the classes together. The F_1 scores are calculated by averaging over 100 runs of repeated cross validation [200] described next.

4.4.1 Repeated cross-validation

We permuted our dataset to produce 100 different runs to estimate the average F_1 score, see Table 4.2. The F_1 score is calculated for each run using a 5-fold cross validation strategy of 20% training and 80% test. This is different from the traditional 80%-20% train-test split because we wish to emulate the situation where there is minimal labelled data. From the test set, which is 80% of the total dataset, one-third is reserved as a validation set to be used with the semi-supervised model.

We make use of fractional training sets which are a subset of the full 20% training set. These smaller subsets have sizes 10%, 25%, 50%, and 75% are used as separate training sets and the F_1 score is averaged over them. For example, 10% of the training data yielded 10 disjoint sets from the full training set, which were ran independently to get the F_1 score and then averaged. This evaluation procedure is repeated starting with empty symptom and side-effect dictionaries, and then incrementing the size of the dictionaries by 25%, 50%, 75% and 100%.

Repeated cross-validation is a time and space constrained procedure, which required runs over a network of multiple machines using the Condor [194] distributed batch computing system. More specifically, we employed a network of 135 machines simultaneously, where each machine had 8 to 12 CPU cores, and the algorithm ran over several days.

The base-line CRF produces high macro F_1 scores of 88.90% and 84.3% for MedHelp and Twitter dataset respectively at larger(>=50%) training and dictionary sizes. The score is further improved to 90.90% and 87.2% for MedHelp and Twitter respectively when we combine symptom and side-effect classes to one single class; see the bottom part of Table 4.2. For micro F_1 scores see Table 4.3. It is also evident that the improvement of the macro and micro F_1 score by the semi-supervised model is about 1% when we do not use symptom and side-effect dictionaries and the training size is less than 50%. Although, this improvement is not significant at larger dictionary and training

Table 4.2: Macro-average F_1 scores are calculated omitting the *Other* class. Averages of 100 runs of repeated cross-validation across different dictionary sizes are shown. Here, *Base* and *Semi* denote the results from the base-line and semi-supervised models, respectively.

		Training Size					
		10%		**25%**		**>=50%**	
Dataset	**Dictionary size**	**Base**	**Semi**	**Base**	**Semi**	**Base**	**Semi**
MedHelp	0%	0.773	0.786	0.839	0.844	0.873	0.875
	25%	0.797	0.807	0.859	0.863	0.887	0.888
	>=50%	0.800	0.811	0.863	0.867	0.889	0.891
Twitter	0%	0.597	0.606	0.732	0.738	0.809	0.813
	25%	0.629	0.640	0.773	0.780	0.841	0.844
	>=50%	0.631	0.643	0.775	0.782	0.843	0.846
All classes after combining symptom and side-effect to one single class							
MedHelp	0%	0.822	0.833	0.880	0.885	0.906	0.908
	25%	0.832	0.841	0.888	0.891	0.909	0.910
	>=50%	0.833	0.842	0.889	0.892	0.909	0.911
Twitter	0%	0.645	0.656	0.786	0.794	0.859	0.862
	25%	0.671	0.682	0.807	0.814	0.871	0.874
	>=50%	0.673	0.684	0.808	0.815	0.872	0.875

sizes, it shows that the performance of the semi-supervised model dominates that of the base-line model. Next, by running an accuracy test on both models, we quantify more precisely how much more accurate is the semi-supervised model in comparison to the base-line model, and whether the difference is significant. See Table 4.5 and 4.6 for macro and micro accuracy scores, respectively. We discuss this comparison in Section 4.4.3. In the next section, we compare our results with some of the previous studies.

Table 4.3: Micro-average F_1 scores are calculated omitting the *Other* class. Averages of 100 runs of repeated cross-validation across different dictionary sizes are shown. Here, *Base* and *Semi* denote the results from the base-line and semi-supervised models, respectively.

		Training Size					
		10%		**25%**		**>=50%**	
Dataset	**Dictionary size**	**Base**	**Semi**	**Base**	**Semi**	**Base**	**Semi**
MedHelp	0%	0.815	0.823	0.846	0.852	0.875	0.878
	25%	0.831	0.836	0.86	0.863	0.882	0.883
	>=50%	0.834	0.838	0.863	0.866	0.884	0.886
Twitter	0%	0.723	0.728	0.790	0.793	0.838	0.840
	25%	0.742	0.749	0.813	0.817	0.856	0.858
	>=50%	0.745	0.751	0.815	0.819	0.858	0.860
All classes after combining symptom and side-effect to one single class							
MedHelp	0%	0.827	0.833	0.858	0.863	0.886	0.888
	25%	0.845	0.848	0.871	0.872	0.890	0.892
	>=50%	0.848	0.850	0.873	0.875	0.892	0.894
Twitter	0%	0.734	0.74	0.802	0.807	0.854	0.856
	25%	0.773	0.779	0.834	0.837	0.873	0.875
	>=50%	0.777	0.783	0.837	0.840	0.876	0.877

4.4.2 Comparison with related work

The proposed semi-supervised model builds on our work from Chapter 3. Our method is not directly comparable with other methods in the literature due to it having different objectives (see Section 4.1). The ADRmine system [143] achieved an F_1 score of 82.1% and 72.1%, on DailyStrength [2] and on Twitter, respectively, for an ADR detec-

[2] http://www.dailystrength.org/

tion task. Miftahutdinov et al. [132] attained 79.9% for a multi-label classification task using the CADEC corpus. For a discussion, related to the objectives and the methodologies utilised by these two related works, see Chapter 2 Section 2.3.4. In the 2019 SMM4H shared task [211] for ADR detection from Twitter, the KFU NLP team [130] reached the best F_1 score of 65.8% in the competition. The team reported to have used the readily available BioBERT-CRF implementation from [105] with standard parameters deployed for BERT-based models. The results from these studies suggest that our proposed semi-supervised model's performance is competitive, see Table 4.2 and 4.3.

4.4.3 Comparing the base-line and semi-supervised models

The difference in performance between the base-line and the semi-supervised models is small, to investigate this difference further we constructed a 2×2 contingency table, as shown in Table 4.4. Here, X_{11} denotes the total count when base-line and semi-supervised models both predicted a concept correctly, whereas, X_{12} represents the number of times the base-line model predicted a concept correctly but the semi-supervised did not. On the other hand, X_{21} is the total count of correct prediction by the semi-supervised model when the base-line model was incorrect. Finally, the cell containing X_{22} represents number of times both models' predictions were incorrect. If N is the number of tokens in the test set, then the accuracy percentage for the semi-supervised model over the base-line model is $100 \times X_{21}/N$, and similarly the percentage of accuracy for the base-line over the semi-supervised model is $100 \times X_{12}/N$. To assess the significance of improvement, we computed the χ^2 value for 1 degree of freedom by making use of X_{12} and X_{21}, which is known as McNemar's non-parametric test [65].

We ran the accuracy test along with the repeated cross validation strategy described above, and the calculated average macro and micro percentages are shown in Table 4.5 and 4.6, respectively. In the case of micro average accuracy, we considered all the tokens in the test set by ignoring their class labels. To calculate the macro average accu-

Table 4.4: Contingency table template for comparing accuracy between the semi-supervised and the base-line model.

	Semi-supervised model		
Base-line model	Correct	Incorrect	Total
Correct	X_{11}	X_{12}	$X_{1,}$
Incorrect	X_{21}	X_{22}	$X_{2,}$
Total	$X_{,1}$	$X_{,2}$	N

racy, the score is considered separately for all the classes and then averaged. The result, shown in Table 4.5 and 4.6, suggests that the semi-supervised model is generally 1-2% more accurate than the base-line model at every division of dictionary and train sizes over 100 runs. This implies that the semi-supervised model always improves the prediction of base-line model. We now discuss the significance of this improvement by the semi-supervised model.

4.5 Discussion

In order to compute the statistical significance of a possible improvement of the semi-supervised model over the base-line model we made use of the McNemar's test, as described above, with respect to the symptom and side-effect classes and the combined symptom and side-effect class. The results show that the difference is significant for the symptom class for all dictionary and training sizes. Regarding the side-effect class for Twitter, although the semi-supervised model performed better than the base-line model, it is not generally significant, probably due to the imbalance between the side-effect and symptom classes; in particular the side-effect class is much smaller in size than the symptom class, which gives a priori preference to the symptom class over the side-effect during the classification process. Moreover, for MedHelp, the symptom class is also larger than the side-effect class, even greater than in Twitter. In this case

Table 4.5: Macro-average Accuracy Test: Scores are calculated omitting the *Other* class. Averages of 100 runs of repeated cross-validation across different dictionary sizes are shown. Here, *Base* and *Semi* denote the result from the base-line and semi-supervised models, respectively.

		Training Size					
		10%		**25%**		**>=50%**	
Dataset	**Dictionary size**	**Base**	**Semi**	**Base**	**Semi**	**Base**	**Semi**
MedHelp	0%	0.968	3.457	0.818	2.031	0.661	1.317
	25%	0.815	2.337	0.697	1.503	0.658	1.151
	>=50%	0.799	2.304	0.711	1.534	0.666	1.177
Twitter	0%	1.164	2.362	1.242	2.237	0.955	1.557
	25%	1.360	3.003	1.160	2.519	0.872	1.515
	>=50%	1.260	2.885	1.455	2.635	0.937	1.442
All classes after combining symptom and side-effect to one single class							
MedHelp	0%	0.806	2.665	0.564	1.582	0.429	0.942
	25%	0.621	1.915	0.440	1.162	0.417	0.817
	>=50%	0.614	1.910	0.426	1.162	0.414	0.811
Twitter	0%	1.365	3.151	1.336	2.681	0.881	1.457
	25%	1.127	2.995	1.091	1.952	0.659	1.156
	>=50%	1.031	2.781	0.939	2.018	0.616	1.091

it seems that the misclassification of side-effects as symptoms by the semi-supervised model is accentuated further due to this large class imbalance.

In Figures 4.3, 4.4 and 4.6, we have shown the comparison between the models in predicting symptom, side-effect and the combined symptom and side-effect classes at different dictionary sizes. As described earlier, the averages of X_{12} and X_{21} from 100 repeated cross validated runs are plotted on y-axis against different training sizes on x-axis. These averages are used as input for McNemar's test. In addition, we considered

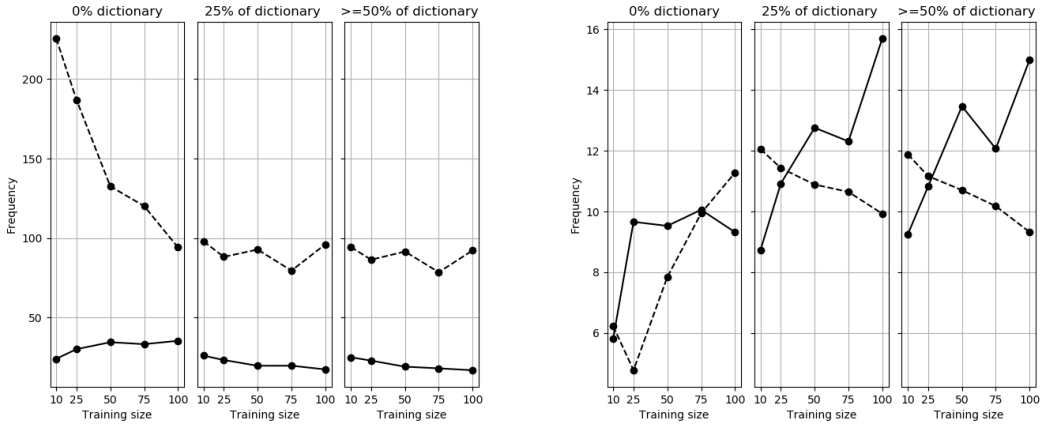
Table 4.6: Micro-average Accuracy Test: Scores are calculated omitting the *Other* class. Averages of 100 runs of repeated cross-validation across different dictionary sizes are shown. Here, *Base* and *Semi* denote the result from the base-line and semi-supervised models, respectively.

		Training Size					
		10%		**25%**		**>=50%**	
Dataset	**Dictionary size**	**Base**	**Semi**	**Base**	**Semi**	**Base**	**Semi**
MedHelp	0%	0.078	0.287	0.067	0.169	0.054	0.110
	25%	0.066	0.190	0.057	0.124	0.053	0.096
	>=50%	0.065	0.187	0.058	0.126	0.054	0.099
Twitter	0%	0.348	0.688	0.375	0.675	0.29	0.463
	25%	0.408	0.887	0.347	0.743	0.263	0.453
	>=50%	0.376	0.846	0.444	0.791	0.286	0.430
All classes after combining symptom and side-effect to one single class							
MedHelp	0%	0.060	0.202	0.042	0.12	0.032	0.072
	25%	0.046	0.142	0.033	0.088	0.031	0.062
	>=50%	0.046	0.141	0.032	0.087	0.031	0.062
Twitter	0%	0.289	0.754	0.247	0.552	0.174	0.306
	25%	0.385	0.88	0.384	0.783	0.25	0.419
	>=50%	0.299	0.811	0.282	0.557	0.187	0.314

the minimum of both X_{12} and X_{21}, which calculates a *conservative* estimate for the said significance test.

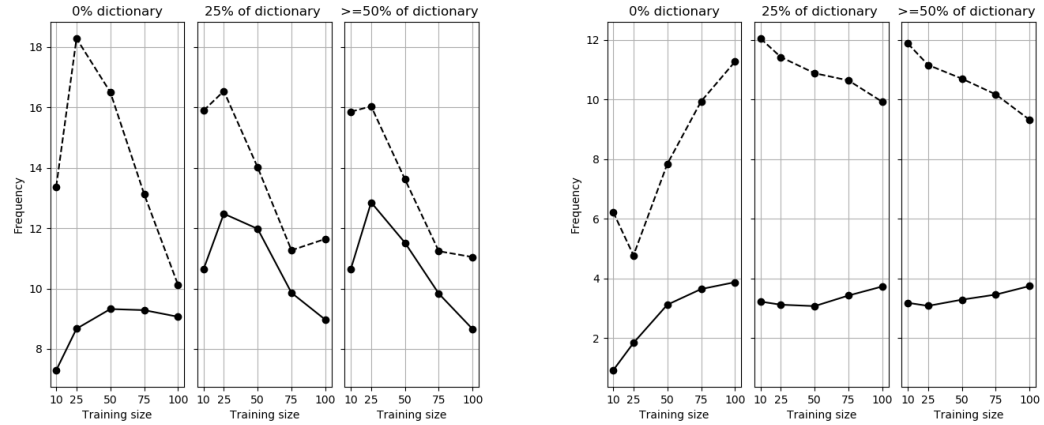
4.5.1 Symptom prediction

For the MedHelp dataset the semi-supervised model correctly predicted, on average, 100 symptom terms more than the base-line model; see Figure 4.3a. This difference is significant for the symptom class at the 95% confidence level using Mcnemar's test;



(a) Symptom class (b) Side-effect class

Figure 4.3: MedHelp: Comparison of base-line and semi-supervised models in predicting (a) symptom and (b) side-effect classes by using MedHelp dataset. Lines and dots represent base-line and semi-supervised model, respectively.



(a) Symptom class (b) Side-effect class

Figure 4.4: Twitter: Comparison of base-line and semi-supervised models in predicting (a) symptom and (b) side-effect classes using Twitter dataset. Lines and dots represent base-line and semi-supervised model, respectively.

the test makes use of the conservative estimate as described above. Although for the symptom class, the margin of difference for Twitter dataset is smaller, as seen in Figure

4.4a, this difference is also significant. In Figure 4.5 we see an example of improvement over the base-line model. In this case, the semi-supervised model correctly recognises *shakes* as symptom while the base-line model classifies *shakes* as *Other*.

4.5.2 Side-effect prediction

In the case of the MedHelp dataset, we found that the accuracy of predicting side-effect degrades slightly. In general many symptom and side-effect terms are common in both respective dictionaries creating ambiguity and possible misclassification. The cause of the ambiguity is most likely due to symptom and side-effect often appearing in common contexts. We found that in such cases, even a human annotator may find it difficult to distinguish between these classes. The large class imbalance for MedHelp, as shown in Table 4.1, causes the transition probabilities of symptom terms to be higher than those of side-effect terms. Thus during test phase, the semi-supervised model gives priority to symptom over side-effect. As a consequence, the semi-supervised model collects more symptom terms than side-effects and the misclassification of side-effect as symptom occurs occasionally. In Figure 4.5, we can see this in action; the semi-supervised model misclassified the term *pain* as symptom, denoted by *SYM*, where as the underlying base-line model classified it correctly as side-effect, denoted as *SD*. The term *pain*, exists simultaneously in the symptom and side-effect dictionaries. Moreover, as the transition probability is higher for symptom classes, the model marginally predicts an incorrect label. However, this problem is not present in case of Twitter dataset as the symptom classes are only about twice more in size than side-effect. Though, in the case of Twitter, the improvement over the base-line model is not significant in the conservative estimate; in the average case, it is significant except at larger dictionary and training sizes. Next, we combined the symptom and side-effect terms into a single term and reran the whole procedure again. The result of this process is described next.

Example of an improvement

1. The worst$_{N}$ part is that it has affected$_{N}$ my left$_{BPOC}$ hand$_{BPOC}$ moreINT and my pinky$_{BPOC}$ and ring$_{BPOC}$ finger$_{BPOC}$ have really$_{INT}$ fast$_{INT}$ **shakes$_{SYM}$**.
2. Base : [‘O’, ‘N’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘BPOC’, ‘BPOC’, ‘INT’, ‘O’, ‘O’, ‘BPOC’, ‘O’, ‘BPOC’, ‘BPOC’, ‘O’, ‘INT’, ‘INT’, **‘O’**]
3. Semi : [‘O’, ‘N’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘BPOC’, ‘BPOC’, ‘INT’, ‘O’, ‘O’, ‘BPOC’, ‘O’, ‘BPOC’, ‘BPOC’, ‘O’, ‘INT’, ‘INT’, **‘SYM’**]

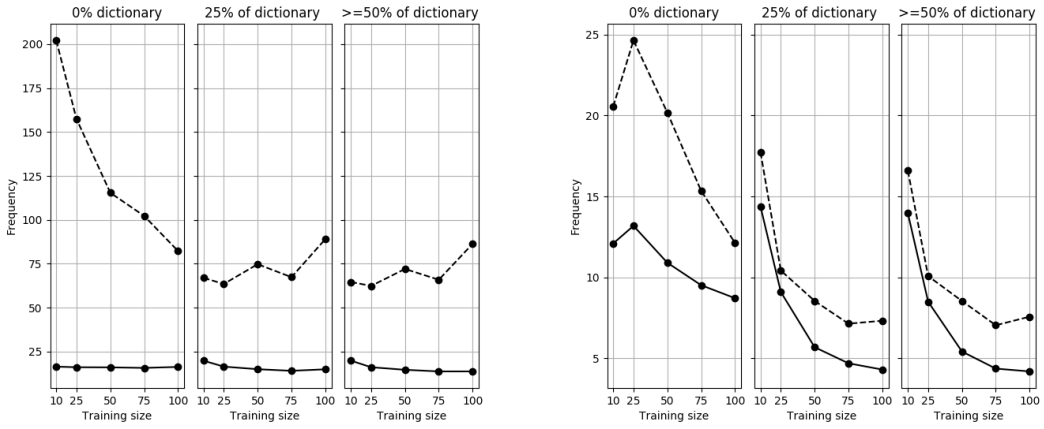
(a) Example of an improvement, where *shakes* was correctly classified by *Semi* as *SYM*.

Example of a misclassification

1. The side$_{N}$ effects$_{N}$ were noted to be mild$_{P}$ and included diarrhea$_{SD}$ neck$_{SD}$ pain$_{SD}$ and dry$_{SD}$ mouth$_{SD}$.
2. Base : [‘O’, ‘N’, ‘N’, ‘O’, ‘O’, ‘O’,‘O’, ‘P’, ‘O’, ‘O’, ‘SD’, ‘SD’,**‘SD’**, ‘O’, ‘SYM’, ‘SYM’, ‘O’]
3. Semi : [‘O’, ‘N’, ‘N’, ‘O’, ‘O’, ‘O’, ‘O’, ‘P’, ‘O’, ‘O’, ‘SD’, ‘SD’, **‘SYM’**, ‘O’,‘SYM’, ‘SYM’, ‘O’]

(b) Example of a misclassification by *Semi*, where *pain* was incorrectly classified as *SYM* instead of *SD*.

Figure 4.5: Examples of (a) an improvement and (b) a misclassification made by the semi-supervised model. Here, at 1, we have a sentence with annotated labels in the subscript, at 2 and 3 the predicted labels by the base-line and semi-supervised models are, respectively, given. The boldface letters signal either an improvement or a misclassification.



(a) MedHelp (b) Twitter

Figure 4.6: Comparison of base-line and semi-supervised models in predicting after combining symptom and side-effect classes in (a) MedHelp and (b) Twitter dataset. Lines and dots represent base-line and semi-supervised model, respectively.

4.5.3 Combining symptom and side-effect

When we combine the symptom and side-effect classes into a single class, the F_1 score for the base-line model improved significantly for both datasets, even more for the semi-supervised model; see the bottom part of Tables 4.2 and 4.3. McNemar's test shows a significant difference between the models, and the semi-supervised model is generally more accurate than the base-line, see Table 4.5. For MedHelp the prediction of the combined symptom and side-effect class by the semi-supervised model is significantly better than that of the base-line model; see Figure 4.6a. Although the experiment with Twitter dataset shows slightly less improvement, it is also significant in most cases; see Figure 4.6a, except at the 50% of training and dictionary sizes in conservative estimate. In the average case, the semi-supervised improves over the base-line model significantly for all cases; see Figure 4.6b.

4.6 Conclusion

We have proposed a semi-supervised algorithm, designed to enhance an underlying pre-trained base-line model, for extracting health related concepts from social media. This algorithm improves on the base-line model when a small amount of labelled data is available, this means that manual annotation can be kept to a minimum. Central to our approach is a procedure for automatically expanding dictionaries of medical concepts, in particular, symptoms and side-effects. These additional words/phrases are also used to identify a diversified set of sentences with which to augment the training data. Although the performance of our method does not drastically improve on that of the base-line model, this process has the potential to be applied in practical usage where the language changes continuously. In such a setting the proposed model will be able to adapt to the changes, as is shown in our experiments. In the next chapter, we build on this methodology to develop an end-to-end NLP pipeline for analysing COVID-19 social media posts.

Chapter 5

Case Study

The content of this chapter is adapted from our article published in:
A. Hasan, M. Levene, D.J. Weston, R. Fromson, N. Koslover, T. Levene. "Monitoring COVID-19 on Social Media: Development of an End-to-End Natural Language Processing Pipeline Using a Novel Triage and Diagnosis Approach". *Journal of Medical Internet Research*, 550 24(2):e30397.

5.1 Overview

In Chapter 3, we developed a concept relational extraction method in order to extract a structured representation of patients sentiment regarding Parkinsons' drug/treatment. Moreover, in Chapter 4, we proposed a semi-supervised method using a base-line CRF algorithm that is capable to work with a small labelled dataset. In this chapter, we make use of the these methodologies, to attain the third objective set out in the Chapter 1. Specifically, our goal is to build an actionable information extraction method using medical social media related to COVID-19.

During the coronavirus pandemic, hospitals were 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 the NHS helpline if they needed to [195]. Consequently, there is an urgent need to develop novel practical approaches to assist medical staff. A variety of meth-

ods have been recently developed that involve NLP; the concerns of these methods range from the level of the individual, see for example [146, 181], up to the population level [176, 158].

Herein, we take a diagnostic approach and propose an end-to-end 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 diagnosis could help in gauging the prevalence of infections in the population. Attempting a clinical diagnosis of influenza, or in our case a diagnosis of COVID-19, purely on the information provided in a social media post is unlikely to be sufficiently accurate to be actionable on an individual level, since the quality of this information will be typically noisy and incomplete. However, it is not necessary to have actionable diagnoses at the individual level in order to identify interesting patterns at the population level, which may be useful within public health surveillance systems. One of our key concerns is in the production of a high-quality human labelled dataset on which to build our pipeline. In Section 5.2, we address this issue by providing an overview of our pipeline and the development procedure of our dataset. Section 5.3 discusses the contribution made in this chapter. In Section 5.4 we discuss prior research related to COVID-19 symptom tracking tools, prediction models from clinical features, and diagnosis using textual sources. Section 5.5 describes the dataset and its annotation and our methodology. Section 5.6 discusses evaluation procedures, experimental setup, and the outcome of the evaluation of our NLP pipeline. Section 5.7 provides discussions related to our findings and comparison with a prior work. Finally, Section 5.8 provides concluding remarks.

5.2 The NLP pipeline

The first step in the NLP pipeline is attained by developing an annotation application that detects and highlights COVID-19 related symptoms with their severity and dura-

tion in a social media post. 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*.

The author of this thesis manually annotated the data with concepts and relations. Annotation allowed us to present posts highlighted with identified concepts and relations to three experts along with several questions, as shown in Figure 5.1. 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 [145].

The three experts are junior 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 5.1. The posts labelled by the doctors were then used to construct two types of predictive machine learning model using SVM [54, 122]; see Step 4 Subsection 5.5.3. The *triage models* employ multi-stage binary classifiers, which consider the risk averseness or tolerance of the doctors when making the diagnosis [13]. 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

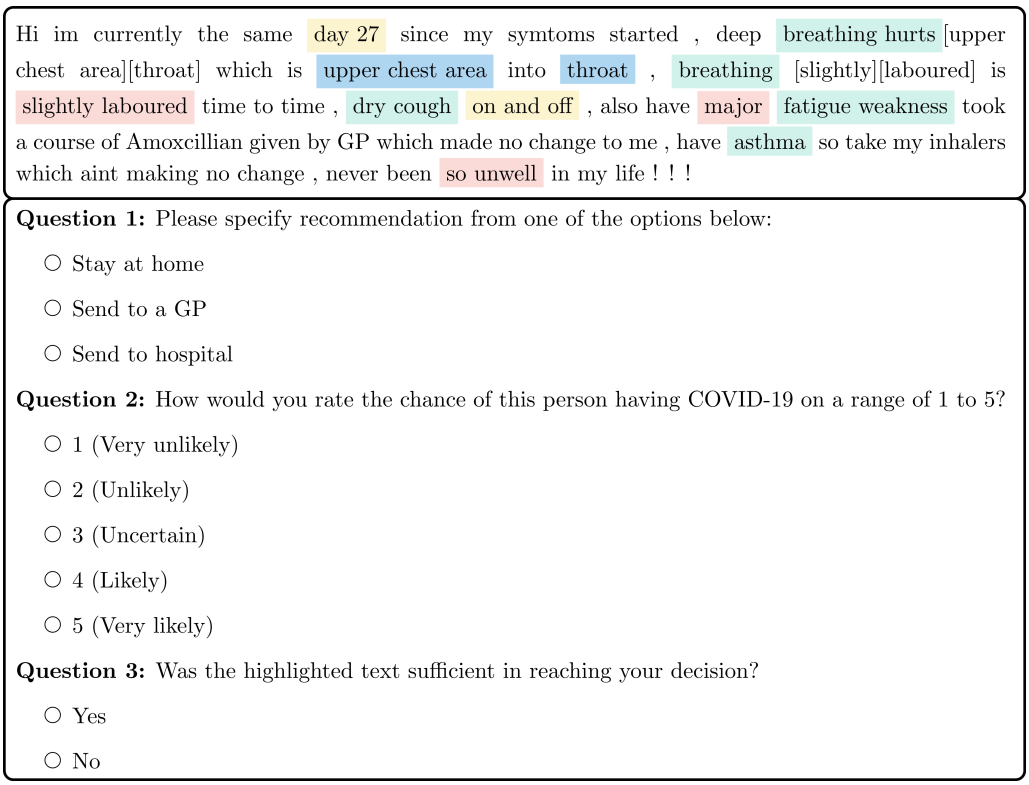


Figure 5.1: A patient-authored social media post is annotated with symptoms (light green), affected body parts (pale blue), duration (light yellow) and severities (pink). The phrases in the square brackets show relations between a symptom and a body part/duration/severity, when the distance was greater 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.

classifying *COVID* and *NO_COVID* classes; these are detailed in the Problem setting section in Material and Methods (Subsection 5.5.2).

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 CRF [99]; see Step 1 of the Methodology Subsection in Materials and Methods (Subsection 5.5.3) for implementation details. Relations are obtained from these predicted concepts using an unsupervised *Rule-Based (RB)* classifier; see Step 2 in the Materials

and Methods section (Subsection 5.5.3). We also discuss the feature importance obtained from the constructed COVID-19 diagnostic models, and compare them with the most frequent symptoms from [176] 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 not rank in the top 5 most frequent symptoms; see Discussion.

5.3 Contribution

Overall, we make several contributions as follows:

1. We show that it is possible to take an approach which aims at disease detection to augment public health surveillance systems, by constructing 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.

5.4 Related work

In this chapter, we focus on features extracted from a textual source to triage and diagnose COVID-19 for the purpose of providing population level statistics in the context of public health surveillance. We discussed several NLP applications used for tracking infectious diseases in Chapter 2. In the following subsections we provide related COVID-19 symptom tracking tools, machine learning based prediction models using clinical features, and diagnostic models using features extracted from textual sources.

5.4.1 COVID-19 symptom tracking tools

Since the start of the COVID-19 pandemic, a number of mobile app-based self-reported symptom tools have emerged, to track novel symptoms [228, 6, 53, 95]. The mobile application in [127] applied LR to predict the percentage of probable infected cases among the total app users in the US and UK combined. A recent review on mobile symptom trackers used for COVID-19 can be found in [180]. The authors in [135] performed a statistical analysis on primary care EHR records to find longitudinal dynamics of symptoms prior to and throughout the infection. Sarabadani et al [173] extracted COVID-19 symptoms from Reddit discussion forums utilising an active learning methodology to understand the longitudinal impact of COVID-19 symptoms before and after recovery. In addition, researchers analysed Twitter messages to conduct studies on self reported long-term post-COVID symptoms [16] known as *Long COVID*. Recently, Miao et al [129] published a study that analysed Long COVID symptoms from Twitter messages utilising a combination of rule- and machine learning- based methods. Specifically, the hybrid method extracted symptoms, gender, symptom duration, and locations from a bulk collection of Twitter messages.

5.4.2 COVID-19 prediction models from clinical features

In general, COVID-19 clinical prediction models can broadly be categorised into risk, diagnosis and prognosis models [222]. In Judson et al [89], a portal-based COVID-19 self-triage and self-scheduling tool was employed to segment patients into four risk categories: emergent, urgent, no-urgent and self-care. Whereas, the online telemedicine system in [110] used LR to predict low, moderate and high risk patients, by utilising demographic information, clinical symptoms, blood tests and *Computed Tomography (CT)* scan results. Moreover, machine learning algorithms, such as decision trees, have shown promising results in detecting COVID-19 from blood test analyses [25]. In Schwab et al [181], various machine learning models were developed to predict patient

outcome from clinical, laboratory and demographic features found in EHR [58]. They reported that *Gradient Boosting (XGB)*, RF 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 [208], where they developed predictive models for the inpatient mortality in Wuhan, using an ensemble of XGB models. Similarly, in Vaid et al [197], mortality and critical events for patients using XGB classifiers were predicted. Zoabi et al [240] used machine learning models to retrospectively predict COVID-19 test results using eight binary features (i.e. sex, age 60 years or above, known contact with an infected individual, and five initial clinical symptoms) collected from patients during their PCR testing. A critical review on various diagnostic and prognostic models of COVID-19 used in clinical settings, can be found in [222]. Finally, Alyasseri et al [9] provided a thorough review on machine learning and deep learning based COVID-19 diagnostic models.

5.4.3 COVID-19 diagnosis using textual sources

EHR records contain valuable patient information that can be harnessed to build strong disease prognosis models. Izquierdo et al [88] applied deep learning to extract various medical concepts, which were used as features for a decision tree model, to predict ICU admissions for COVID-19 in-patients.

Meystre et al [128] analysed patients' unstructured notes from a telemedicine system to predict COVID-19 test results using an NLP system. The NLP system first extracted several features such as demographics and social history, medical risk factors, laboratory tests, medications, and environmental risk factors; for a description of these features see [128]. Then, the system utilised these features with two machine learning models to predict COVID-19 test results. Both machine learning models, i.e. LR and SVM, achieved comparable performance. Notably, the NLP system employed both rule-based and deep learning methods for feature extraction from textual sources.

López et al [112] utilised radiological text reports from lung CT scans to diagnose COVID-19. Similar to our approach, they first extracted concepts using a popular medical ontology [23] and then constructed a document representation using word embeddings [134] and concept vectors [112]. 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.

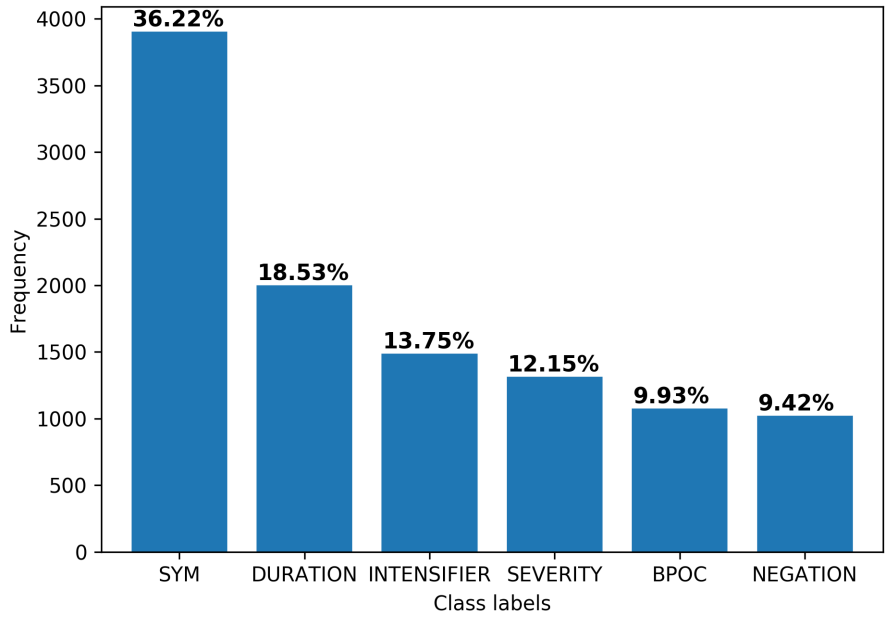


Figure 5.2: Frequency distribution of annotated classes/concepts from the text are shown. We also show the percentage of each class after discounting the *OTHER* labels. The average number of tokens per post is 130.17(SD = 97.83). Here, *SYM, DURATION, INTENSIFIER, SEVERITY, BPOC and NEGATION* denote symptoms, duration, intensifiers, severity, body parts and negations, respectively.

5.5 Materials and Methods

5.5.1 Data

We collected social media posts discussing COVID-19 medical conditions from a forum called *Patient* [148]. This is a public forum that was created at the onset of the coronavirus outbreak in the United Kingdom. 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 anonymized, and duplicates were removed, we randomly selected 500 distinct posts. The thesis author annotated these posts with the classes shown in Figure 5.2. The class labels represent symptoms and the related concepts: (1) duration; (2) intensifier, which increases the level of symptom severity; (3) severity; (4) negation, which denotes the presence or absence of the symptom or severity; and (5) affected body parts. We also annotated relations between a symptom and other concepts that exist at the 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 5.1. Each marked post was presented to the doctors using a web application, and they were asked 3 questions independently; see Figure 5.1. We called the doctors' answers to questions 1 and 2 as the COVID-19 symptom triage and diagnosis, respectively. Thus, for each post, we had 3 independent answers from 3 doctors, which we denoted as A, B, and C, respectively; these corresponded to the last 3 authors of the paper and were assigned randomly.

Measurement of agreement

To measure the agreement between the answers (recommendations and ratings) of the 3 doctors to questions 1 and 2 of Figure 5.1, we first calculated the proportion of observed agreement (p_o), as suggested by de Vet et al. [202], who stipulated that Cohen

Table 5.1: Pair-wise agreement between pairs of doctors answers for Question 1 and 2; see Figure 5.1 for an example.

Question 1				**Question 2**		
Pair	p_o	κ	**AC1**	p_o	κ	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
Average	0.68	0.22	0.60	0.66	0.56	0.58

κ is actually a measure of reliability rather than than agreement, and observe that p_o is high in all cases as can be seen in Table 5.1. We noted that the paradoxical behavior of Cohen κ can arise when the absolute agreement (p_o) is high [63]. This may occur when there is a substantial imbalance in the marginal totals of the answers, which we observed in the answers to question 1. Consequently, in addition to Cohen κ, we deployed a common solution to this problem, called the AC1 statistic devised by Gwet and coworkers [78, 217].

We found that for question 1 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 [101] for benchmark scale for the strength of agreement. For question 2 it turns out that the said paradox did not occur, resulting in similar values for Kappa and AC1. The agreement between A and B (κ=0.64, AC1=0.67) and between B and C (κ=0.64, AC1=0.67) are substantial, while the agreement between A and C (κ=0.40, AC1=0.40) is on the boundary of fair and moderate; see Table 5.1.

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 doc-

tors 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.

5.5.2 Problem setting

Triage classification for Question 1

We map the doctors' recommendation from question 1 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. To combine recommendations from 2 or more doctors, we first took 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 [13], we categorise the ceiling of the average to be *risk averse*, denoted by, for example, AB(R-a), and the floor to be *risk tolerant*, denoted by, for example, AB(R-t). Thus for each patient's post, we have in total eleven recommendations from three doctors for question 1. We construct a multi-stage classification model for each of these recommendations, where the goal is to classify a post into 1 of the 3 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 termed as the *ground truth probability* (abbreviated *GTP*), was simply:

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

Given our ground truth probability estimates are discrete we investigated 3 decision boundaries, denoted by LE, LT, and NEQ, 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 that NEQ ignores cases on the 0.5 boundary. Each of the three decision functions shown above is a stand-alone algorithm which is applied to construct a dataset for the classification task. This implies that there are three datasets, and each post may exist in multiple datasets.

5.5.3 Methodology

A schematic of our methodology to triage and diagnose patients from their social posts is shown in Figure 5.3. We now detail each of the steps in COVID-19 triage and diagnosis pipeline shown in the figure.

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 [45] built-in 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: (1) Symptom, (2) Severity, (3) Duration, (4) Intensifier, and (5) Negation. The dictionaries were built by analysing the posts while annotating them. We also utilise the MetaMap system [12], 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 5.2, by applying a CRF [193]. A detailed description of our

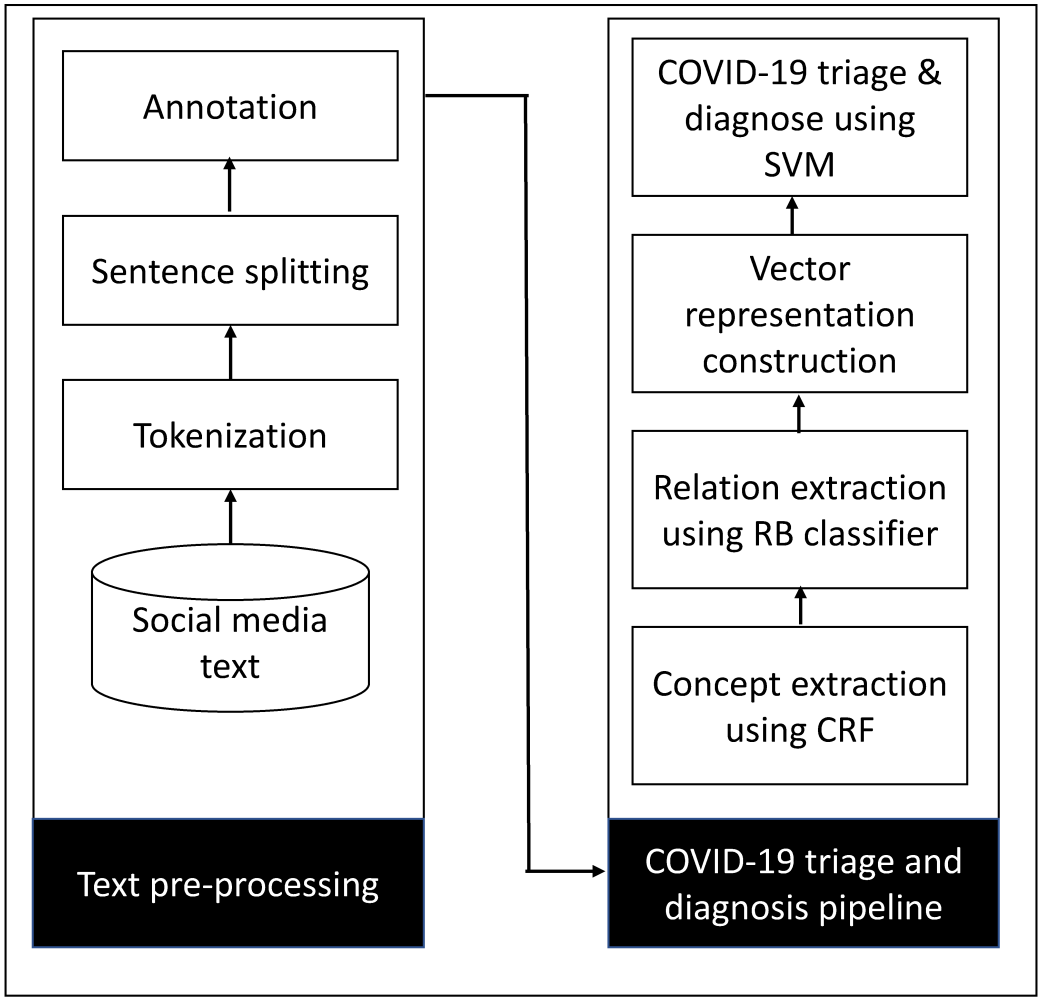


Figure 5.3: A block diagram of COVID-19 triage and diagnosis text processing pipeline. Here, CRF, RB classifier and SVM are acronyms for Conditional Random Fields, Rule-Based classifier and Support Vector Machine, respectively.

CRF training methodology can be found in Chapter 2 Section 2.3 and Chapter 4 Section 4.3.2. 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. A relation can be defined as an ordered pair of concepts (c_1, c_2) that exist in a sentence, where c_1 is a symptom, formally the *principal entity*, and c_2 could be a body parts, severity, or a duration [15]; more formally, the c_2 concepts are called *modifiers* [51]. To retrieve such relations, we first find all pairs of (c_1, c_2) from a sentence, where c_1 and c_2 are within a distance D of each other; we set the allowable distance to be between $D = 2$ and $D = 7$ based on observation. Then we apply the following rules:

[Rule 1] Filter (c_1, c_2) pairs from a sentence by choosing the closest symptom (c_1) to a modifier (c_2) within the distance D.

[Rule 2] Once a candidate modifier, c_2 is chosen for the relation with a symptom, c_1, it cannot be used with other symptoms in the same sentence.

The relations extracted in the form of (c_1, c_2) allow a formal semantic representation [51] of a post. We also consider the relation $(c_1, ?)$, in cases where a suitable modifier cannot be found for a symptom in the sentence. To retrieve such relations, we first find all pairs of (c_1, c_2) from a sentence, where c_1 and c_2 are within a distance D of each other; we set the allowable distance to be between $D = 2$ and $D = 7$. We do not consider $D = 1$ here; it misses a lot of modifiers and performs poorly in compare to other distances.'

In total, we extracted 5 kinds of relations as follows: *(SYM, SEVERITY), (SYM, DURATION), (SYM, BPOC), (SYM, NEGATION)* and *(SYM, ?)*—here, *SYM* and *BPOC* refer to symptoms, and body part, organ, or organ component, respectively.

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 3 appended vectors containing: (1) the absence or presence of 110 unique body parts, (2) the absence or value of a symptom duration, and (3) the absence, negation or value or a symptom severity. The construction of the representation is described as follows:

First, the symptom-only vector, $\langle s_0, s_1 \ldots, s_n \rangle$, is transformed by setting its default values to -2; this initially represents the non-existence of a symptom in the post. For each symptom, s_i, if there exists a severity modifier, or in other words, if a *(SYM, SEVERITY)* relation is found for the said concept, then its value is set to that of the modifier. More specifically, the value in this case is in between 1 and 5. We also consider negations here; if there exist a *(SYM, NEGATION)* relation in the post for a symptom, then the default value is subsumed by -1. In addition, it is 0, when the symptom is mentioned but no severity or negation modifier relation exist; in other word, here, we consider the *(SYM, ?)* relation. Let, the modified vector be $\langle ss_0, ss_1 \ldots, ss_n \rangle$.

Similarly, we construct another vector from $\langle s_0, s_1 \ldots, s_n \rangle$ by transforming it to reflect *(SYM, DURATION)* relations in a post. Here, if there exists a duration modifier of a symptom, s_i, then its value is set to that of the modifier, otherwise it is set to 0. Thus we modify the symptom-only vector to $\langle sd_0, sd_1 \ldots, sd_n \rangle$.

Next, we take the body parts vocabulary and construct a discrete vector of 110 dimensions; the number of terms in the body parts vocabulary. Let this vector be $\langle bp_0, bp_1, \ldots, bp_m \rangle$, where 1 and 0 represent, respectively, the presence and absence of a body part mentioned in the post.

Finally, we concatenate the vectors, $\langle ss_0, ss_1 \ldots, ss_n \rangle$, $\langle sd_0, sd_1 \ldots, sd_n \rangle$, and $\langle bp_0, bp_n, \ldots, bp_m \rangle$, which forms *Symptom-modifier relation vector* representation for the post.

Step 4: Triage and diagnosis

We utilised SVM classification and regression models to triage and diagnose patients' posts, respectively, from the vector representations described earlier. For question 1, the recommendation from a doctor or combination of doctors is the class label of the post; see section Problem setting in Materials and Methods (see Subsection 5.5.2) 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 multi-stage classifier for COVID-19 triage.

For diagnosing COVID-19 cases, we deploy a variant of SVM, called *Support Vector Regression (SVR)* [54], to estimate the probability of COVID-19. We use the GTP that is derived from answers to question 2 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 5.2: The concept extraction using CRF on 3-fold cross validation.

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	

Table 5.3: Relation extraction using RB classifier on 3-fold cross validation.

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

Table 5.4: Question 1: Multi-stage classification results for RBF kernel using the symptom-modifier relation vector trained on the ground truth.

	SVM Classifier 1			**SVM Classifier 2**		
Model	**P**	**R**	F_1	**P**	**R**	F_1
A	0.82	0.91	0.86	0.73	0.95	0.83
B	0.73	0.77	0.75	0.81	0.99	0.89
C	0.85	0.98	0.91	—	—	—
AB(R-a)	0.70	0.75	0.72	0.80	0.96	0.88
AB(R-t)	0.84	0.96	0.89	0.85	1.00	0.92
BC(R-a)	0.72	0.75	0.73	0.92	1.00	0.96
BC(R-t)	0.86	0.99	0.92	—	—	—
AC(R-a)	0.79	0.87	0.83	0.89	1.00	0.94
AC(R-t)	0.88	0.98	0.93	—	—	—
ABC(R-a)	0.70	0.76	0.73	0.89	0.99	0.93
ABC(R-t)	0.88	0.99	0.93	—	—	—

5.6 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 [120]. 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 was not balanced with *COVID* and *NO_COVID* classes, as can be seen in Figure 5.4, and we wished to give equal weight to all instances, we reported micro-averaged scores for the SVR classification. In contrast, in the case of concept extraction, the Other class dominated. So, in this

Table 5.5: Question 1: Multi-stage classification results of two classifiers for RBF kernel using the symptom-modifier relation vector trained on the CRF prediction.

	SVM Classifier 1			**SVM Classifier 2**		
Model	**P**	**R**	F_1	**P**	**R**	F_1
A	0.81	0.89	0.85	0.72	0.91	0.80
B	0.74	0.74	0.74	0.81	0.99	0.89
C	0.85	0.96	0.90	—	—	—
AB(R-a)	0.73	0.71	0.71	0.81	0.96	0.88
AB(R-t)	0.84	0.94	0.88	0.84	1.00	0.92
BC(R-a)	0.74	0.71	0.72	0.92	1.00	0.96
BC(R-t)	0.88	0.98	0.93	—	—	—
AC(R-a)	0.81	0.85	0.83	0.89	1.00	0.94
AC(R-t)	0.88	0.98	0.93	—	—	—
ABC(R-a)	0.72	0.72	0.72	0.89	1.00	0.94
ABC(R-t)	0.89	0.98	0.93	—	—	—

case, we reported the macro-averaged scores for the CRF classification results.

Experimental setup

For the CRF we report 3-fold cross validated macro-averaged results. Specifically, we trained each fold by a Python wrapper [44] for CRFsuite, see [147]. 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 [29] implemented in Sklearn [151] with both Linear and Gaussian *Radial Basis Function* (RBF) kernels [122]. Similarly, the SVR [114], imple-

Table 5.6: Question 1: Multi-stage classification results of two classifiers for RBF kernel using the symptom-only relation vector trained on the ground truth.

	SVM Classifier 1			**SVM Classifier 2**		
Model	**P**	**R**	F_1	**P**	**R**	F_1
A	0.83	0.91	0.87	0.74	0.85	0.79
B	0.71	0.81	0.76	0.81	0.98	0.89
C	0.87	0.97	0.92	—	—	—
AB(R-a)	0.69	0.75	0.72	0.83	0.96	0.89
AB(R-t)	0.85	0.94	0.89	0.85	1.00	0.92
BC(R-a)	0.71	0.79	0.75	0.92	0.99	0.95
BC(R-t)	0.88	0.98	0.93	—	—	—
AC(R-a)	0.80	0.86	0.83	0.89	1.00	0.94
AC(R-t)	0.90	0.98	0.94	—	—	—
ABC(R-a)	0.68	0.74	0.71	0.90	1.00	0.95
ABC(R-t)	0.90	0.98	0.94	—	—	—

mented 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 [151].

We simulated two cases for COVID-19 triage and diagnosis. First SVM and SVR models were 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 resembled an end-to-end NLP application. To obtain 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.

Table 5.7: Question 1: Multi-stage classification results for RBF kernel using the symptom-only relation vector trained on the CRF prediction.

	SVM Classifier 1			**SVM Classifier 2**		
Model	**P**	**R**	F_1	**P**	**R**	F_1
A	0.84	0.89	0.87	0.74	0.82	0.78
B	0.74	0.79	0.77	0.82	0.98	0.89
C	0.86	0.95	0.90	—	—	—
AB(R-a)	0.72	0.76	0.73	0.83	0.92	0.87
AB(R-t)	0.87	0.93	0.90	0.84	0.98	0.90
BC(R-a)	0.72	0.78	0.75	0.92	0.99	0.95
BC(R-t)	0.87	0.97	0.92	—	—	—
AC(R-a)	0.80	0.86	0.83	0.89	1.00	0.94
AC(R-t)	0.89	0.95	0.92	—	—	—
ABC(R-a)0.71	0.76	0.73	0.89	0.99	0.93	
ABC(R-t)	0.90	0.95	0.92	—	—	—

Evaluation outcomes

The concept and relation extraction phases produce excellent and very good predictive performances, respectively; see Table 5.2 and 5.3. The triage classification results from Q1 are shown in Table 5.4, 5.5, 5.6 and 5.7; the full enumeration can be seen in the first column. 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 (see Table 5.4 and 5.5). The Symptom-only vector representations produces results in the range of 71-94% and 79-95%; see Table 5.6 and 5.7. 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,

Table 5.8: Question 2: Micro-averaged F_1 results for different models and decision functions trained on ground truth. Here A, B, C are three medical doctors (abbreviated as Dr) who took part in the experiment.

	Symptom-modifier			**Symptom-only**		
Model	**LE**	**LT**	**NEQ**	**LE**	**LT**	**NEQ**
A	0.72	0.61	0.78	0.70	0.59	0.74
B	0.78	0.61	0.76	0.78	0.62	0.77
C	0.87	0.75	0.87	0.88	0.75	0.87
AB	0.72	0.66	0.74	0.74	0.65	0.75
BC	0.84	0.76	0.84	0.85	0.79	0.86
AC	0.81	0.73	0.81	0.83	0.74	0.83
ABC	0.74	0.67	0.76	0.75	0.67	0.77

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 5.5) we cannot report them. We report macro-averaged F_1 score results since question 1 was framed as a decision problem, where weights for the classes are a priori equal. The results obtained after training with CRF predictions were 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 were in the range of 72-87%, 61-76%, and 74-87% for the LE, LT, and NEQ decision functions, respectively; see Table 5.8. 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

Table 5.9: Question 2: Micro-averaged F_1 results for different models and decision functions trained on the CRF predictions. Here A, B, C are three medical doctors (abbreviated as Dr) who took part in the experiment.

	Symptom-modifier			**Symptom-only**		
Model	**LE**	**LT**	**NEQ**	**LE**	**LT**	**NEQ**
A	0.68	0.64	0.76	0.50	0.79	0.74
B	0.76	0.64	0.77	0.78	0.57	0.74
C	0.86	0.75	0.87	0.87	0.74	0.86
AB	0.70	0.65	0.73	0.71	0.66	0.74
BC	0.83	0.76	0.83	0.85	0.78	0.86
AC	0.80	0.74	0.82	0.80	0.73	0.81
ABC	0.72	0.69	0.76	0.74	0.69	0.77

each model for different decision functions, is shown in Figure 5.4. When we trained the models with the symptom-modifier vector representation from the CRF predictions, the results were in the range of 68-86%, 64-76%, and 73-87% for the *LE*, *LT*, and *NEQ* decision functions, respectively; see Table 5.9. This indicated that, for diagnosis as well as triage, an end-to-end NLP application is likely to perform similarly to stand-alone applications. Here, we report micro-averaged F_1 scores since, in our dataset, *NO_COVID* cases dominated; this largely resembled 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 [213].

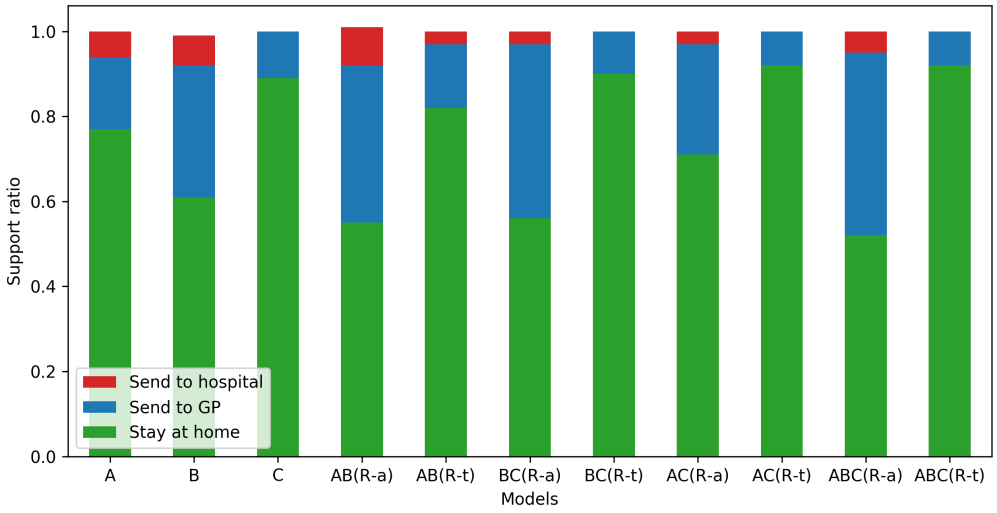


Figure 5.4: Support ratio of triage classes across models for Question 1 classification tasks. Absolute numbers for the *Send to hospital* class in test sets are as follows: A=10, B=12, AB(R-a)=14, AB(R-t)=5, BC(R-a)=6, CA(R-a)=5, ABC(R-a)=9; the value for the remaining models is zero.

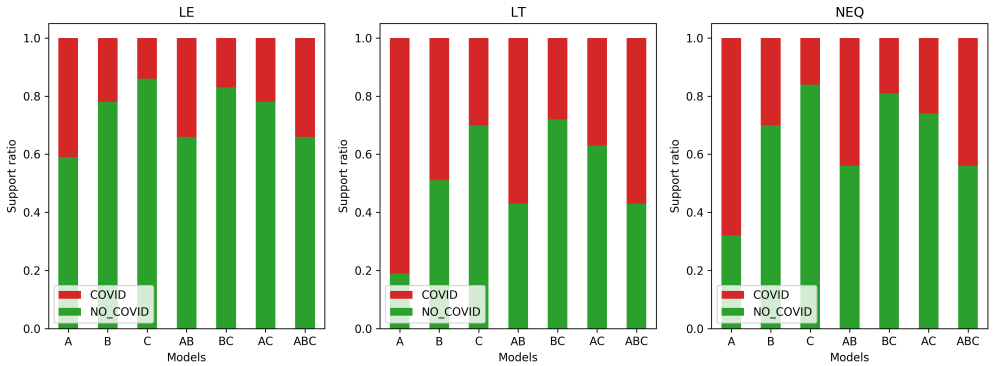


Figure 5.5: Support ratio of diagnosis classes across models and three decision functions for Question 2 classification tasks.

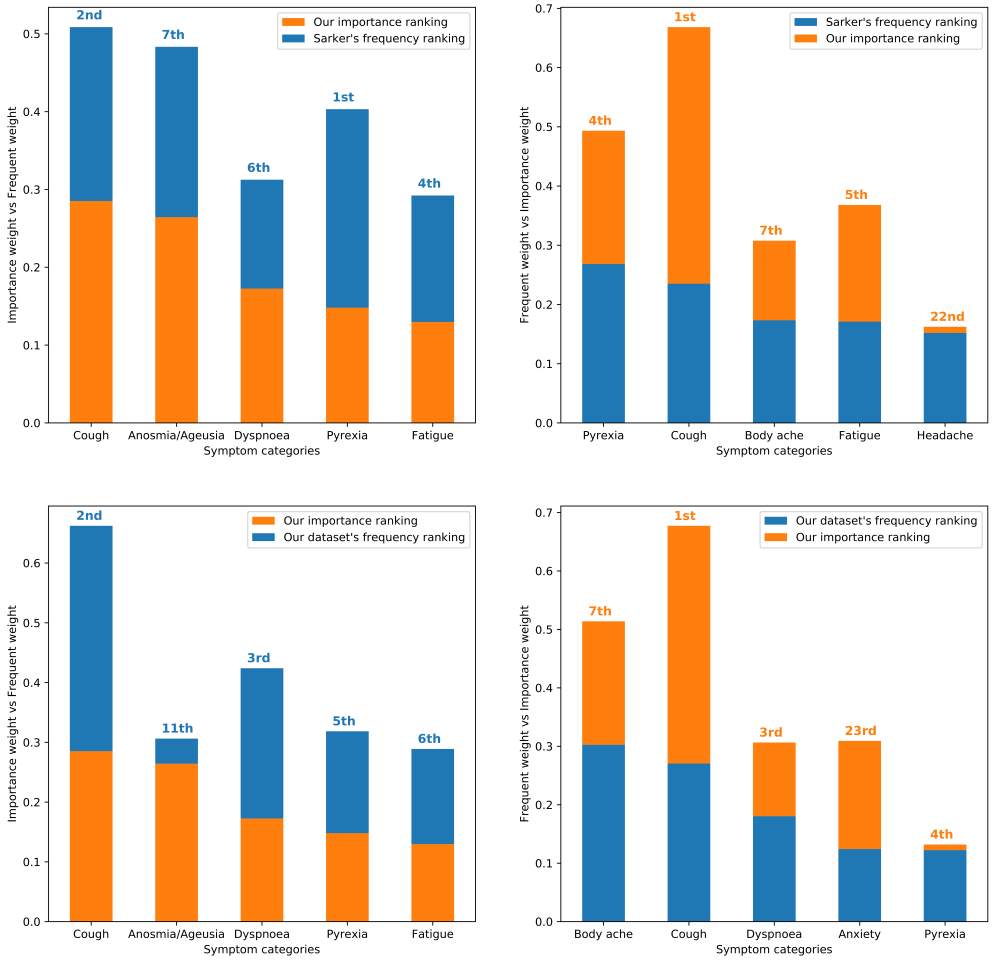


Figure 5.6: Feature comparison between our most important features and Sarker's most frequent symptoms (top row), and between our most important features and our most frequent symptoms (bottom row). The feature importance rankings are obtained from an SVM linear kernel using the symptom-only vector representation.

5.7 Discussion

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 were trained in two ways; using (i) ground truth labels, and (ii) predictions obtained from the NLP pipeline. Trained models were 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.

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 7 models and 3 decision function; see Problem Settings in Materials and Methods (see Subsection 5.5.2). The features are then mapped to the categories found in the Twitter COVID-19 lexicon complied by Sarker et al. [176]. We rank the mapped features, select the top 5 highest weighted features (which we refer to as the important features), and the weights are then normalised. The top 5 important features in our dataset are: *cough, anosmia/agusia, dyspnea, pyrexia, and fatigue*. Mizrahi et al. [135] 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 [176] frequent categories, we compiled the corresponding frequencies of our 5 most important symptoms. Normalised weights and frequencies are then plotted in Figure 5.6. The top-left stacked bar chart compares our 5 most important features with Sarker et al's [176] frequencies. *Cough* was the most important symptom from our dataset, where it was the second-most frequent. *Anosmia/ageusia* ranked second in our importance list, while it

was seventh in the most frequent list. *Pyrexia* came first and fourth in both the frequent and importance lists, respectively.

The top-right chart in Figure 5.6 shows a comparison between Sarker et al's [176] frequent ranking and our importance ranking. Here, we selected top 5 most frequent symptoms from Sarker et al's [176] frequency list and normalise them. These are: *pyrexia, cough, body ache, fatigue*, and *Headache*. We took the corresponding importance weights of these symptoms and plotted them in a stacked bar chart. Here, *headache* ranked 22^{nd} in our importance ranking, while it was 5^{th} in the frequent ranking. We found a large difference between the 2 rankings, implying that the top most frequent symptoms are not necessarily the most important ones.

Next we compared 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 5.6, we observed that anosmia/ageusia are a relatively low in order in the frequency ranking (ie, 11^{th}). As in Sarker et al's [176] ranking, *cough* came 2^{nd} in our dataset's frequency ranking.

Finally, the bottom-right chart in Figure 5.6 refers to the comparison between our dataset's frequency and importance rankings of the corresponding symptoms. We observed that *anxiety* ranked 4^{th} in the most frequent list, where it was low, (ie, 23^{rd}) in the most importance ranking.

5.8 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 take an approach that aims at the detection of COVID-19 using an automated triage and diagnosis system in order to augment public health surveillance systems, despite the

heterogeneous nature of typical social media posts. The output of such an approach is actionable information for decision makers as it will enable them to estimate severity and prevalence of the disease in the population. In the next chapter we investigate the transferability of the concept extraction model and the dictionaries built for this case study.

Chapter 6

Deep learning for concept extraction

6.1 Overview

Extracting COVID-19 symptoms both from social media and from medical documents has been found to be useful for tracking this disease [94] and for building prognosis models to predict mortality in hospitals [185].

In Chapter 4, we developed a CRF that included manually created dictionary features to extract medical concepts from social media. In contrast, a BiLSTM network initialised with word and character embeddings [100] demonstrated improvements in concept extraction over a feature based CRF from a variety of formal document datasets [79]. More recently, BERT [49] achieved state-of-the-art performance in a variety of NLP tasks including biomedical text mining [105]. However, one of the bottlenecks for deploying such architectures is the need for labelled data. Recent work in medical concept extraction such as [192], utilised dictionaries/gazetteers with deep learning models to leverage external knowledge when labelled datasets are scarce.

Herein, we investigate two variants of a BiLSTM based deep learning architecture that essentially differ in the type of input provided. In Figure 6.1. the input sequence is replaced with vectors from a pre-trained static word embedding. In Figure 6.2 the input static word embedding vector is concatenated with a vector from a contextual word embedding. We are interested in the effect of including dictionary features to these models. In this chapter, we investigate the effect of appending a dictionary vec-

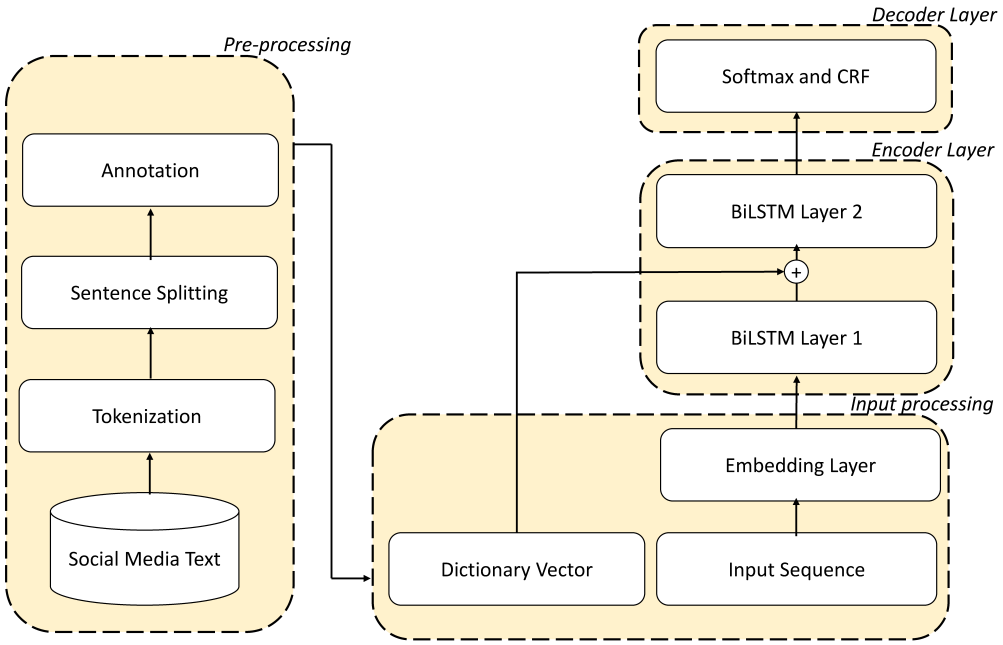


Figure 6.1: BiLSTM+CRF architecture for extracting COVID-19 medical concepts from social media.

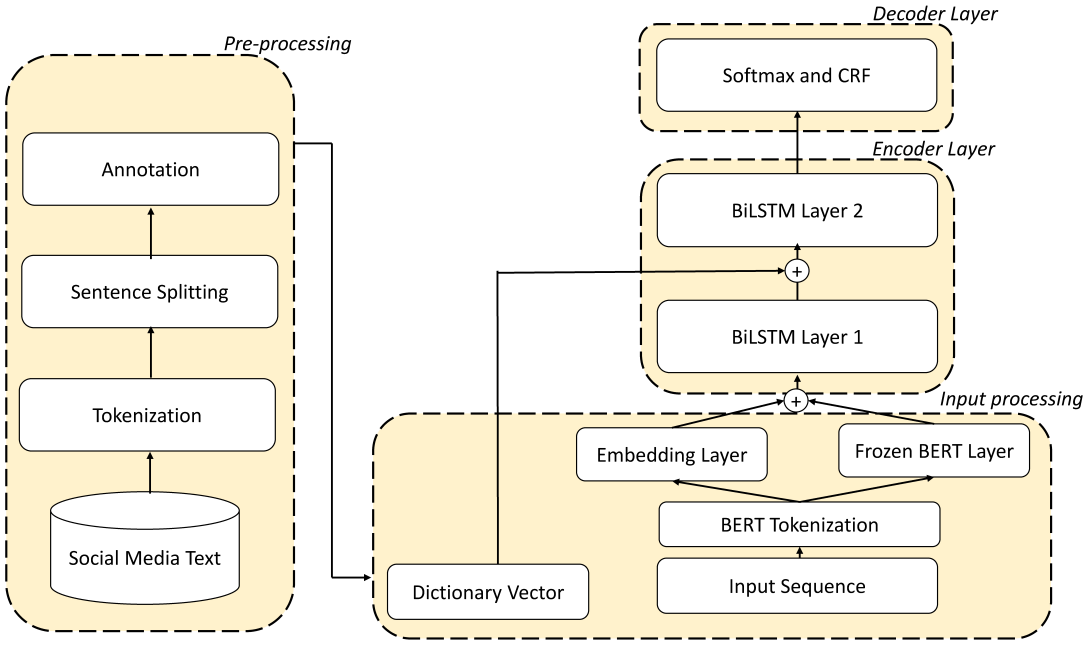


Figure 6.2: BERT+BiLSTM+CRF architecture for extracting COVID-19 medical concepts from social media.

tor (which we describe in Section 6.3) to either the input of the first or second BiLSTM layer. The figures show the case where the dictionary vector is injected into the input of Layer 2. The investigation is performed using a COVID-19 discussion forum [148] dataset which was annotated with several medical concepts (see Dataset description in Section 6.3.1). The results show that models built incorporating dictionaries perform better than those without them. Furthermore, in order to check the transferability of the models and dictionaries to a publicly available COVID-19 Twitter dataset [32], weak supervision methodology is developed. Specifically, we utilise the manually built symptom dictionary from the previous chapter and a publicly available COVID-19 symptom dictionary from [176]. These are termed as (i) *Our dictionary*, (ii) *Sarker dictionary*, and (iii) *Combined dictionary*. The combined dictionary is the merger of the former two dictionaries. First, two base-line models are trained using labelled dataset by Our and Sarker dictionaries separately. Then the models are retrained by incrementally adding terms from either dictionaries; i.e. the coverage of Our dictionary is increased by Sarker's and vice-versa. These models are tested with the dataset tagged by Our, Sarker, and Combined dictionaries, respectively. Furthermore, they are tested with a manually annotated ground truth data. For the Twitter dataset, we use COVID-19 version of BERTweet [141], and for the forum dataset we use BERT base model [49].

6.2 Contributions

Our contributions are as follows:

1. With combination of static and contextual word embeddings and by leveraging dictionary features, we obtain a very good performance in extracting COVID-19 medical concepts from social media text.

2. We show that dictionaries are useful as weak learners and the neural model achieve a very good performance when we transfer it to extract COVID-19 symp-

Tokens	d_1	d_2	d_3	d_4	d_5	d_6	d_7
...	0	0	0	0	0	0	0
headaches	1	0	0	0	0	0	1
a	0	0	1	0	0	0	0
week	0	0	1	0	0	0	0
...	0	0	0	0	0	0	0
absolutely	0	1	0	0	0	0	0
exhausted	1	0	0	0	0	0	1
...	0	0	0	0	0	0	0
high	0	1	0	0	0	0	0
temp	1	0	0	0	0	0	0

Hi I started with headaches a week ago, then on Sunday morning I woke up feeling absolutely exhausted and with a high temp. I had a dry cough for a day but now its subsided. But I have had a complete loss of taste and smell, and I mean completely! Has this happened to anyone else? Im a nurse and able to get tested tomorrow. When I have the results I will post.

Figure 6.3: An example post and its feature matrix for a selected sequence. Here, green, yellow, and red denote symptom, duration, and severity concepts. Moreover, d_1 to d_6 denote symptom, severity, duration, intensifier, negation, and body parts dictionaries, respectively, and d_7 represents MetaMap.

toms from a larger Twitter dataset.

6.3 Materials and Methods

Schematics of our architectures are shown in Figures 6.1 and 6.2. In the following subsections we demonstrate our data collection procedure and the architectures of the models.

6.3.1 Data

We extracted 3000 posts related to COVID-19 from a patient social media forum called Patient [148], from this we randomly selected 500 social media posts to manually annotate. These posts were annotated with the class labels representing symptoms and

the related concepts: (1) duration; (2) intensifier, which increases the level of symptom severity; (3) severity; (4) negation, which denotes the presence or absence of the symptom or severity; and (5) affected body parts. The details of data collection procedure can be found in Chapter 5 Section 5.5.1.

We collected tweets from the first 3 months of 2020 that contained at least one symptom, amounting to 36204 tweets, from a multilingual COVID-19 dataset published through the Github [1] repository by the authors of [32], we manually annotated 1000 randomly selected tweets.

6.3.2 Neural Network Architecture

In this chapter, we develop two variants of the BiLSTM+CRF architecture. Each architecture comprises Input Processing, Encoder Layer, and Decoder Layer. They differ in Input Processing unit, where we add a BERT layer, hence the architecture in Figure 6.1 is denoted as *BiLSTM+CRF* and the one in Figure 6.2 is denoted as *BERT+BiLSTM+CRF*. We now give details of these units as follows.

Input Processing

The input processing unit consists of (i) input sequence, (ii) wordpiece tokenization, (iii) dictionary vector, (iv) embedding layer, and (v) frozen BERT layer. The units (i), (iii), and (iv) are common in both architectures whereas (ii) and (v) are used with BERT+BiLSTM+CRF architecture.

Input Sequence: An input sequence is either a sentence (from the forum dataset) or a complete tweet (from the Twitter dataset). The sequences are tokenized using the GATE [45] software package. For a given sequence, S of length l, from this tokenization procedure we obtain $w_1, w_2, \ldots, w_l$ *tokens*. The vocabularies are constructed from the unique tokens of these datasets.

Dictionary Vector: The sequences of a post or a tweet is processed to construct a dic-

[1] `https://github.com/echen102/COVID-19-TweetIDs`

tionary vector. The dictionary vector for token w_i consists of 7 bits of information, where each bit is denoted as d_i, and represent either a dictionary or UMLS semantic types. The example of a dictionary matrix for a selected sequence is shown in Figure 6.3. Here, after tokenization, we processed the sequence using a NLP pipeline constructed using the GATE software. For dictionary/gazetteer matches we configure the pipeline for full match. We have five dictionaries in our pipeline. They are as follows:(d_1) Symptom, (d_2) Severity, (d_3) Duration, (d_4) Intensifier, and (d_5) Negation. The dictionaries were built by analysing the posts while annotating them. We also utilized the MetaMap, to map tokens to *Body Part, Organ, or Organ Component*, and *Sign or Symptom*, and Disease or Syndrome semantic concepts to represent d_6 and d_7 bits in our dictionary vector. Thus, for a given sequence, S, we collect $\vec{d_1}, \vec{d_2}, \ldots, \vec{d_l}$ vectors.

BERT Tokenization: BERT uses wordpiece tokenization. In this procedure known subword tokens are iteratively merged based on maximum likelihood. Here, we denote the tokens of S as $wp_1, wp_2, \ldots, wp_m$ after the wordpiece tokenization. Note that l and m may not match and $m >= l$. For example the word "COVID" splits into two sub words ["CO", "#VID"] in the case of wordpiece tokenization.

Embedding Layer: If a token is wordpieced at position i, then the token and dictionary vector, wi and $\vec{d_i}$, respectively, are repeated for the same number of times it is pieced. As a result, for the architecture in Figure 6.2, the length of S is extended to m. We collected pre-trained word embeddings $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_m}$ for the sequence S from the google news corpus [134]. If a word is not present in the vocabulary, the embedding is initialized randomly. Similarly the dictionary vectors are mapped to $\vec{d_1}, \vec{d_2}, \ldots, \vec{d_m}$ for S.

Frozen BERT Layer: For BERT+BiLSTM+CRF architecture in Figure 6.2, we utilize BERT models by freezing their parameters and fed wordpieced tokens into the layer before inputting them to the encoder layer. This produces contextual BERT vectors for a sequence and denoted as $\vec{b_1}, \vec{b_2}, \ldots, \vec{b_m}$. Specifically, we took representations from the last BERT layer.

Encoder Layer

The encoder in our architecture is a 2 layered BiLSTM network which is similar to the architecture presented in [100]. The *BiLSTM Layer 1* in the BiLSTM+CRF architecture is fed with pre-trained word vectors $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_l}$. For BERT+BiLSTM+CRF architecture, we feed concatenation of $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_m}$ and $\vec{b_1}, \vec{b_2}, \ldots, \vec{b_m}$. After processing, the first BiLSTM layer produces hidden representations $\vec{h_i^1}$. This representation is concatenated with $\vec{d_i}$ and fed into the second layer. Let the output of the second layer is $\vec{h_i^2}$. We also incorporated an attention layer on top of the last hidden layer for BiLSTM+CRF architecture. We examined two types of attention: (i) self attention, and (ii) cross attention. In case of self attention, the query, key and value vectors come from the same hidden representations for each token. Whereas, in case of cross attention, the query is the last hidden representation of the BiLSTM which is deemed as the sentence representation, and the key and value vectors are each token's hidden representations. We found that cross attention works well the BiLSTM+CRF architecture, however, the BERT+BiLSTM+CRF architecture are found to perform worse when attention is added. So we removed them from BERT+BiLSTM+CRF architecture.

Decoder Layer

Our decoder layer comprises a Softmax layer and a neural CRF. The hidden representations, $\vec{h_i^2}$, of a token from the final encoder layer is fed into a Softmax layer to produce the emission probability of a tag j, $\mathbf{E}_{i,j}$. The transition scores are calculated using the neural CRF [100]. Finally, the probability of a label sequence is calculated using Equation 2.4 from Chapter 2 Section 2.3.2.

6.3.3 Models

From BiLSTM+CRF architecture we build the following models:

1. BiLSTM+CRF: The *BiLSTM Layer 1* is initialised with pre-trained word vectors

$\vec{v}_i$.

2. +DICT(1): The *BiLSTM Layer 1* is initialised with the concatenation of dictionary and word vectors $\vec{d}_i$, $\vec{v}_i$, respectively.

3. +DICT(2): The *BiLSTM Layer 1* is initialised with pre-trained word vectors $\vec{v}_i$ to produce $\vec{h}_i^1$. The dictionary vector $\vec{d}_i$ is concatenated with $\vec{h}_i^1$ and fed into *BiLSTM Layer 2*.

From BERT+BiLSTM+CRF architecture we build the following models:

1. BERT+BiLSTM+CRF: Concatenation of $\vec{v}_i$ and $\vec{b}_i$ are fed in to *BiLSTM Layer 1*.

2. +DICT(1): Concatenation of $\vec{v}_i$, $\vec{b}_i$, and $\vec{d}_i$ are fed in to *BiLSTM Layer 1*.

3. +DICT(2): Concatenation of $\vec{v}_i$ and $\vec{b}_i$ are fed in to *BiLSTM Layer 1* to produce $\vec{h}_i^1$. The dictionary vector $\vec{d}_i$ is concatenated with $\vec{h}_i^1$ and fed into *BiLSTM Layer 2*.

	BiLSTM+CRF			**+DICT(1)**			**+DICT(2)**		
Label	P	R	F_1	P	R	F_1	P	R	F_1
SYM	0.84	0.77	0.80	0.92	0.94	0.93	0.93	0.95	0.94
SEVERITY	0.67	0.51	0.58	0.74	0.77	0.75	0.75	0.80	0.77
BPOC	0.82	0.89	0.85	0.91	0.88	0.89	0.90	0.90	0.90
INTENSIFIER	0.82	0.90	0.86	0.87	0.94	0.91	0.88	0.94	0.91
DURATION	0.79	0.79	0.79	0.81	0.91	0.86	0.85	0.89	0.87
NEGATION	0.81	0.88	0.84	0.83	0.83	0.83	0.81	0.87	0.84
O	0.96	0.97	0.97	0.98	0.98	0.98	0.99	0.98	0.98
MACRO	0.82	0.82	0.81	0.87	0.89	0.88	0.87	0.90	**0.89**

Table 6.1: Results of concept extraction from forum dataset using **BiLSTM+CRF** architecture. For the descriptions of **BiLSTM+CRF**, **+DICT(1)**, and **+DICT(2)** models see Subsection 6.3.3

	BERT+BiLSTM+CRF			**+DICT(1)**			**+DICT(2)**		
Label	P	R	F_1	P	R	F_1	P	R	F_1
SYM	0.79	0.86	0.82	0.92	0.92	0.92	0.93	0.97	0.95
SEVERITY	0.70	0.38	0.49	0.75	0.66	0.69	0.75	0.85	0.80
BPOC	0.91	0.77	0.83	0.87	0.92	0.89	0.93	0.90	0.91
INTENSIFIER	0.82	0.80	0.81	0.84	0.95	0.89	0.87	0.94	0.90
DURATION	0.78	0.82	0.80	0.83	0.87	0.84	0.84	0.91	0.87
NEGATION	0.83	0.90	0.86	0.84	0.89	0.86	0.83	0.93	0.88
O	0.96	0.96	0.96	0.98	0.97	0.98	0.99	0.97	0.98
MACRO	0.83	0.78	0.80	0.86	0.88	0.87	0.88	0.92	**0.90**

Table 6.2: Results of concept extraction from forum dataset using **BERT+BiLSTM+CRF** architecture. For the descriptions of **BERT+BiLSTM+CRF**, **+DICT(1)**, and **+DICT(2)** models see Subsection 6.3.3. In all cases BERT parameters are frozen.

6.3.4 Transfer learning/Weak supervision

For investigating how well our dictionaries based on the forum data can transfer to another dataset, we focus on extracting symptoms only. We introduce one further dictionary that has been developed by analysing the same Twitter dataset we use in our forthcoming experiments which was published by Sarker et al. [176]. We prune this dictionary by removing terms related to anxiety, stress & general mental health symptoms and some phrases related to pyrexia or fever such as *102 fever*, *103+ fevers*, and *fever spiked to 107*. The number surrounding the term *fever* is annotated as *Severity* in our forum dataset.

To distinguish between the two dictionaries we call the forum built dictionary, *Our dictionary* and the Twitter based dictionary the *Sarker dictionary*.

From the Twitter dataset we extracted tweets which had at least one symptom. We

did this by simply using Sarkar's dictionary to identify them. We removed 1000 tweets and annotated them which we used for our ground truth experiments. We used weak learning to train models in two ways:

1. *Our base-line*: The train dataset is tagged using Our dictionary.

2. *Sarker base-line*: The train dataset is tagged by the Sarker dictionary.

We also looked at the effect of incrementally combining the dictionaries. Starting with Our dictionary we include 20% of Sarker's dictionary, which we then use to tag the training data. We repeatedly add a further 20% of Sarker's dictionary and tag the data again until we have the union of both dictionaries. We repeat the process of tagging the training data starting with the Sarker dictionary and incrementally including 20% of Our dictionary. For test set we not only evaluate on the ground truth but also on a *weakly learnt test set* that is generated by tagging using the union of both dictionaries, i.e. Combined dictionaries. This latter test set is useful to see how well each individual dictionary can represent a dictionary that is generated from the combined datasets. For completeness we also look at the performance when we tag the test set using only the individual dictionaries separately.

6.4 Results

6.4.1 Experimental setup

For both datasets we reported a 3-fold cross-validated F_1 scores. Training is done with the batch size of 16. The maximum sequence length for forum posts and tweets are 512 and 130, respectively. The number of features in the hidden state for each BiLSTM layer is 100. We use the Adam optimizer with a learning rate of 0.01 and a weight decay of 1e-5. The experiments were performed using the `transformers` library [216] and all models were trained on an NVIDIA Tesla P100.

	Test datasets								
Dictionary	**Combined**			**Our**			**Sarker**		
Size %	P	R	F_1	P	R	F_1	P	R	F_1
0%	1.00	0.83	0.90	1.00	1.00	1.00	0.63	0.72	0.67
20%	1.00	0.93	0.96	0.94	1.00	0.97	0.67	0.88	0.76
40%	1.00	0.96	0.98	0.92	1.00	0.96	0.68	0.94	0.79
60%	1.00	0.98	0.99	0.91	1.00	0.95	0.69	0.96	0.80
80%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82
100%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82

Table 6.3: Results of weakly supervised symptom extraction task using Our base-line **BiLSTM+CRF+DICT(2)** model and for incremental additions from the Sarker dictionary.

	Test datasets								
Dictionary	**Combined**			**Our**			**Sarker**		
Size %	P	R	F_1	P	R	F_1	P	R	F_1
0%	1.00	0.59	0.74	0.81	0.51	0.63	1.00	1.00	1.00
20%	1.00	0.88	0.94	0.88	0.86	0.87	0.77	1.00	0.87
40%	1.00	0.92	0.96	0.88	0.90	0.89	0.74	1.00	0.85
60%	1.00	0.92	0.96	0.88	0.89	0.89	0.73	0.97	0.84
80%	1.00	0.98	0.99	0.89	0.98	0.93	0.71	1.00	0.83
100%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82

Table 6.4: Results of weakly supervised symptom extraction task using Sarker base-line **BiLSTM+CRF+DICT(2)** model and for incremental additions from Our dictionary.

	Test datasets								
Dictionary	**Combined**			**Our**			**Sarker**		
Size %	P	R	F_1	P	R	F_1	P	R	F_1
0%	1.00	0.82	0.90	1.00	1.00	1.00	0.63	0.72	0.67
20%	1.00	0.92	0.96	0.94	1.00	0.97	0.67	0.87	0.76
40%	1.00	0.96	0.98	0.91	1.00	0.96	0.69	0.94	0.79
60%	1.00	0.98	0.99	0.91	1.00	0.95	0.69	0.96	0.81
80%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82
100%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82

Table 6.5: Results of weakly supervised symptom extraction task using Our base-line **BERT+BiLSTM+CRF+DICT(2)** model and for incremental additions of the Sarker dictionary. All experiments are performed using COVID-19 version of BERTweet.

	Test datasets								
Dictionary	**Combined**			**Our**			**Sarker**		
Size %	P	R	F_1	P	R	F_1	P	R	F_1
0%	1.00	0.60	0.75	0.81	0.52	0.63	1.00	1.00	1.00
20%	1.00	0.88	0.94	0.87	0.86	0.86	0.77	1.00	0.87
40%	1.00	0.92	0.96	0.88	0.90	0.89	0.75	1.00	0.85
60%	1.00	0.93	0.96	0.88	0.92	0.90	0.74	1.00	0.85
80%	1.00	0.98	0.99	0.89	0.98	0.93	0.71	1.00	0.83
100%	1.00	1.00	1.00	0.89	1.00	0.94	0.70	1.00	0.82

Table 6.6: Results of weakly supervised symptom extraction task using Sarker baseline **BERT+BiLSTM+CRF+DICT(2)** model and for incremental additions of Our dictionary. All experiments are performed using COVID-19 version of BERTweet.

6.4.2 Evaluation

Supervised Concept Extraction

Results for the supervised experiments of the forum dataset are shown in Table 6.1 and 6.2 for BiLSTM+CRF and BERT+BiLSTM+CRF models, respectively. In both cases, incorporating the dictionary information into the input of the second BiLSTM layer, **+DICT(2)**, performs better than incorporating it into the first layer,**+DICT(1)**. Including BERT performs marginally better than not having it.

Weak Supervision

For weak supervision results we focus only on the **+DICT(2)** models since that was the best performing location for a dictionary in the supervised extraction experiments. Tables 6.3 and 6.4 show the result from the BiLSTM+CRF model when Our and Sarker dictionary is used as the base-line, respectively. We note the final row from each table is the same, since the training labels are identical. We note also for the Combined test set, the labels for both the training and test are generated using the same dictionaries, hence we get an F_1 of 1.

Overall we see from Tables 6.3 and 6.4 that combining the dictionaries, even incrementally, improves performance. Focussing on the first row and final column of Table 6.3 we see that our dictionary performs favourably (F_1 of 0.67) when the test data has been weakly labelled using a dictionary derived from that data compared to when we swap the roles of the dictionaries, which can be seen in Table 4 first row middle column (F_1 of 0.63).

Tables 6.5 and 6.6 show results for when we include BERT features. We see that we achieve similar results which suggests that the information that BERT provides does not contribute much more than we already have through the static embedding and our dictionary features.

Ground truth

For the experiments involving the ground truth we found that replacing BERT with a COVID-19 version of BERTweet produced better results, which we report here in Tables 6.7 and 6.8. As one would expect the results are lower than the experiments performed with weakly labelled test sets, nevertheless the performance is still good. Notably the finding that including a language model does not appreciably improve performance is observed here too.

	BiLSTM			**BERTweet**		
%	P	R	F_1	P	R	F_1
0%	0.80	0.56	0.66	0.82	0.57	0.68
20%	0.82	0.65	0.72	0.83	0.65	0.73
40%	0.83	0.72	0.77	0.84	0.70	0.77
60%	0.82	0.74	0.78	0.84	0.73	0.78
80%	0.83	0.78	0.80	0.84	0.76	0.80
100%	0.83	0.78	0.81	0.84	0.76	0.80

Table 6.7: Results of symptom extraction from the ground truth test set using Our base-line with incremental additions from Sarker dictionary. BiLSTM, and BERTweet correspond models with and without the language model, see main text.

6.5 Discussion

We show some example tweets with tagging results from our models in Table 6.9. All the examples are taken from the BiLSTM model when Sarker dictionary is used as a base line and and for incremental additions from our dictionary. The test data is labelled with the ground truth. The Example 1 shows that the tweet contain COVID-19 symptoms such as *headache, fatigue, soar throat, and cough* which are common in both

	BiLSTM			BERTweet		
%	P	R	F_1	P	R	F_1
0%	0.92	0.60	0.72	0.92	0.57	0.70
20%	0.87	0.74	0.80	0.88	0.72	0.79
40%	0.86	0.75	0.80	0.87	0.73	0.80
60%	0.85	0.75	0.80	0.87	0.74	0.80
80%	0.84	0.78	0.81	0.85	0.76	0.80
100%	0.83	0.78	0.81	0.84	0.76	0.80

Table 6.8: Results of symptom extraction from the ground truth test set using Sarker base-line with incremental additions from Our dictionary. BiLSTM, and BERTweet correspond models with and without the language model, see main text.

dictionaries. In the Example 2, the concept *blood oxygen levels* does not have a presence in the Sarker dictionary. However, when it reaches 100% with our dictionary, the model finds the symptom. Similarly, in Example 3, with the addition of our dictionary the model finds out *digestive symptoms*. In Example 4, we show that though *loss of taste and smell* exist in the Sarker dictionary, due to the longest match operation, the dictionary does tag the single word *smell* as symptom. However, since our dictionary has *smell* in it, the model is able to correctly find it when its coverage is increased.

Table 6.9: Examples of mistakes made by the models. Green and red background colours denote correct and incorrect predictions, respectively.

Example 1
Ground truth: headache , fatigue , sore throat , cough and chest pressure since sunday night. no fever though! but if its not covid, i dont know what it is.
Sarker base-line: headache , fatigue , sore throat , cough and chest pressure since sunday night. no fever though! but if its not covid, i dont know what it is.

Sarker base-line + Our dictionary: headache , fatigue , sore throat , cough and chest pressure since sunday night. no fever though! but if its not covid, i dont know what it is.

Example 2

Ground truth: he left the house, got off sick bed, hardly able to stand, muscle spasms , not cognisant enough to remember speaking to pm, extremely low blood oxygen levels enough to be in hospital, broke lock down and self isolating rules and probable road traffic laws drove to hospital.

Sarker base-line: he left the house, got off sick bed, hardly able to stand, muscle spasms , not cognisant enough to remember speaking to pm, extremely low blood oxygen levels enough to be in hospital, broke lock down and self isolating rules and probable road traffic laws drove to hospital.

Sarker base-line + Our dictionary: he left the house, got off sick bed, hardly able to stand, muscle spasms , not cognisant enough to remember speaking to pm, extremely low blood oxygen levels enough to be in hospital, broke lock down and self isolating rules and probable road traffic laws drove to hospital.

Example 3

Ground truth: covid - 19 patients experience loss of appetite , diarrhoea and other digestive symptoms .

Sarker base-line:covid - 19 patients experience loss of appetite , diarrhoea and other digestive symptoms.

Sarker base-line + Our dictionary:covid - 19 patients experience loss of appetite , diarrhoea and other digestive symptoms .

Example 4

Ground truth: smell that? if not, you should probably call your doctor study finds loss of taste and smell can indicate covid-19.

Sarker base-line: smell that? if not, you should probably call your doctor study finds loss of taste and smell can indicate covid-19.

Sarker base-line + Our dictionary: smell that? if not, you should probably call your doctor study finds loss of taste and smell can indicate covid-19.

6.6 Conclusion

Our experiments have shown that building a small domain specific set of dictionaries can be beneficial for COVID-19 medical concept extraction. These dictionaries have the advantage that they are easy to produce and are interpretable. Moreover, models built using these dictionaries can generalize well and it is possible to transfer them to different datasets on a similar task. The results are encouraging in that a small domain specific set of dictionaries based on forum data can perform commensurately with BERTweet on Twitter data when they are included as features in a model.

Chapter 7

Conclusion

People use everyday conversational language when they discuss health conditions on social media platforms. To extract actionable health information from these platforms, researchers have been using NLP techniques employing rule- and machine learning-based methodologies. Despite advances in these areas, extracting meaningful, coherent, and structured medical information from social media still remains challenging. We approached this problem in a principled manner, first by building a rule-based NLP pipeline utilising dictionaries and linguistic rules and then by automating the concept extraction process using machine learning. We also applied a hybrid NLP pipeline by combining rule-based and machine learning approaches in a case study to extract actionable health information regarding COVID-19. Moreover, we demonstrated the utility of dictionaries by incorporating them within deep neural architectures. Finally, we showed the transferability of our dictionaries within a weak supervision approach. In this chapter, we revisit the research questions from Chapter 1 and provide concluding remarks.

7.1 Revisiting research questions

In Chapter 1, we described our research questions which led us to propose a set of rule-based, supervised, semi-supervised, and weakly supervised methodologies to extract actionable health information from social media. Those questions also had a one-to-

one correspondence to the objectives set in Chapter 1 Section 1.3. In this section, we link our research questions back to the objectives, summarise them, and discuss their limitations.

Q1 How can we build a rule-based concept relationship extraction system to extract a structured representation of drug/treatment's sentiment from social media posts focusing on a chronic disease category (e.g. Parkinsons')?

The above question links to our first objective in Chapter 1 Section 1.3 which states the following:

- To develop a methodology for extracting structured representation of sentiment related to drug/treatment from a chronic disease (i.e. Parkinsons') patient forum posts.

In Chapter 3, we developed a text processing pipeline for extracting structured information regarding sentiment of Parkinsons' drug/treatment from a forum dataset. We demonstrated how actionable information can be extracted from posts by forming relationships between a drug/treatment and a symptom or side-effect, including the polarity/sentiment of the patient. In particular, we made use of several publicly available and manually built dictionaries to recognise drug/treatment, symptom, and side-effect concepts. Using various linguistic rules we calculated polarity at the sentence level. In order to create relations we segmented a sentence using common conjunctions(*and, but, and until*) and selected the closest symptom and side-effect concept to a drug. Thus we formed two types of *disease triples*; they are (i) (Drug, Polarity, Symptom), and (ii) (Drug, Polarity, Side-effect). Finally, we linked the sentences using anaphora resolution [71]. Our methodology is detailed in Algorithm 1.

We annotated our dataset with concepts and relations and validated the annotation by several researchers from Birkbeck. Our NLP relationship extraction system achieved an F_1 score of 81.71% and 82.13%, respectively, on an unseen

test dataset in discovering the said relationships.

A key limitation of our system was that we did not record useful temporal/quantitative data such as dosages or frequency of recurrence of side-effects. Consider the following example:

- *My wife was on Rytary **36.25/145 mg** for **5 days** and returned to C/L today because it was not working and she was getting side effects.*

Here, information regarding dosage, i.e. *36.25/145 mg*, of the drug *Rytary* together with its duration, i.e. *5 days*, will complement the structured medication information that can be obtained from such social media posts, allowing a more thorough assessment of side-effects.

Another key limitation is the limited size and number of datasets used, this applies to all our methodologies. However, this is somewhat mitigated by using formal statistical testing.

Q2 How to develop a machine learning method that uses *minimal supervision* to produce satisfactory results for concept extraction? How can it augment and update labelled datasets and dictionaries?

The above question links to our second objective in Chapter 1 Section 1.3 which states the following:

- To develop a concept extraction method that can be effective with minimal supervision and adapted to change in data source and disease study.

In Chapter 4, we first developed a feature-based CRF algorithm, which we called the *base-line* model, utilising dictionaries and rules from the relationship extraction system of Chapter 3. The base-line model made use of a small number of targeted UMLS semantic types as features; see Section 4.3.2 for a detailed description of features. Next, a semi-supervised algorithm, capable of learning

new concepts from a large unlabelled corpus, was developed using the base-line model. It iteratively augmented highly confident labelled sentences to the training set. Our methodology differed from previous studies in two ways, it expanded the concept dictionaries (i.e. symptom and side-effects) and then reused them in the training procedure; see Section 4.3.3. In addition to this, new terms identified were utilised further to select diversified labelled sentences to augment the training data. Thus, it allowed us to devise a methodology that will adapt to continuous training and changes of concept over time.

We performed extensive experiments using two data sources; (i) MedHelp [126] medical forum, and (ii) Twitter. Our repeated cross-validation strategy, containing 100 different runs, followed a 5-fold cross-validation for each run. The semi-supervised model outperformed the base-line models by 1% when the base-line did not use any dictionary. The base-line model produced high macro F_1 scores of 88.90% and 84.3% for MedHelp and Twitter dataset; see Table 4.2. Though the improvement of the semi-supervised model over the base-line was not significant over larger training and dictionary size, it showed that the semi-supervised model always dominated. In order to measure the performance of the dictionary expansion procedure, we used McNamer's test [65]; see Section 4.5. This test showed that for the MedHelp dataset the semi-supervised model correctly predicted, on average, 100 symptom terms more than the base-line model; see Figure 4.3a. The results validated that the semi-supervised methodology was successful in augmenting labelled datasets and expansion of dictionaries.

This work did not investigate the utility of using word embeddings to improve the dictionary expansion procedure. Distributional vectors or word embeddings represent lexical items such as words according to the context in which they occur in a corpus of documents. Such a methodology can recognise words with similar meanings from an unlabelled corpus by deploying a similarity measure [136]. In addition, word embeddings could further be investigated for improving the

augmentation procedure of diversified training data within the semi-supervised methodology. Furthermore, an active learning methodology, where a human being feeds low confident sentences back into the supervised model, has potential to improve the proposed semi-supervised model.

Q3 Can we develop an end-to-end NLP pipeline applying a similar rule- and machine learning-based methodologies for an infectious disease category (e.g. COVID-19) for extracting actionable information? Will the concept and relation extraction pipeline be able to triage and diagnose COVID-19 patients from their social media posts?

The above question links to our third objective in Chapter 1 Section 1.3 which states the following:

- To apply an end-to-end NLP pipeline capable to provide decision makers with actionable information on the symptom severity and prevalence of a respiratory disease (i.e. COVID-19) using social media at the population level.

In Chapter 5, we developed an NLP pipeline which applied concept and relationship extraction methods to provide decision-makers information regarding COVID-19 severity and prevalence at the population level. The extraction of such information required us to design a study using an expert labelled dataset. We asked three doctors two questions by showing posts from COVID-19 patients to rate the likelihood of COVID-19 on a Likert Scale of 1 to 5, and segment them into three risk categories; see Section 5.1. The expert annotations allowed us to build text classification models by taking a triage and diagnostic approach. The inputs to the models were represented by the concepts and relations extracted using our the CRF and RB classifiers. We tested the models using the NLP pipeline in two ways, first by providing human labelled data, and then by predictions from CRF and RB classifiers. The tests were always performed on the ground truth

datasets. Our NLP pipeline achieved 71%-96% and 61%-87%, respectively, for the triage and diagnosis of COVID-19 when the models were trained on human-labelled data. It achieved similar results when tested on the joint predictions from CRF and RB classifier. Moreover, we discussed important features of the diagnostic machine learning models and compared them with the most frequent symptoms revealed in the study from Sarker et al [176]. We found that the most important COVID-19 symptoms are cough, anosmia/ageusia, dyspnea, pyrexia, and fatigue, whereas the most frequent ones are pyrexia, cough, body ache, fatigue, and headache.

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. The aim of this study was to extract useful information at a population level, rather than to provide an actionable decision for an individual via social media posts. Our manually annotated dataset has 2 main limitations. First, having only 3 experts limited 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.

Q4 Are concept dictionaries helpful for a deep learning network? Are they transferable?

The above question links to our fourth objective in Chapter 1 Section 1.3 which states the following:

- To investigate utility and transferability of manually built dictionaries and pre-trained word embeddings, respectively, by focusing on COVID-19 social media. Specifically, we transfer dictionaries between two types of social media and use them to produce weak labels for training neural networks.

To investigate the utility of dictionaries in neural network settings, we constructed two BiLSTM+CRF based architectures in Chapter 6. They differed in their inputs where the first was initialised with a pre-trained Word2Vec embedding and the second was initialised with a concatenation of Word2Vec and a pre-trained BERT embedding. Each token of the sequence was represented by a fixed size boolean vector where each element represented dictionary membership. The dictionary vector was appended to the architecture, and several models were constructed. We found that models built with dictionaries performed better than those without them. For a supervised concept extraction task on the COVID-19 forum data, our best model achieved a macro F_1 score of 90%. Furthermore, to check the transferability of our models and dictionaries, we investigated Twitter data. Specifically, we produced a large labelled set automatically by utilising our symptom dictionary and then tested the models with a ground truth set built for COVID-19 symptom extraction from Twitter. Additionally, another publicly available symptom dictionary [176] was investigated together with a domain-dependent variant of BERTweet. We found that models built using our dictionary produced similar results to those built using publicly available domain dictionaries and contextual language models. Our models produced an F_1 score of 80% on these transfer learning settings. Thus, we successfully built deep learning models using dictionaries without a labelled dataset and showed their transferability to a different data source.

The experiments focused on one disease and we extracted useful information arising from two different unstructured datasets. Although this was necessary in order to develop the experiments, we note that a stronger case could be made

if we analysed other diseases and datasets. In addition, the weak supervision methodology did not consider other possible sources and rules for generating automatic labels.

7.2 Future work

There are several ways in which the work presented in this thesis can be extended in addition to addressing the limitations presented in the previous section. Employing a multi-sense approach would be useful since people use the same words to express different meanings. For example, people may use *temp* to express *fever/temperature* or a *temporary job* in the context of COVID-19. Our investigation shows that a deep contextual language model such as BERTweet fails to encode such homonyms. Therefore, encoding different senses of a term into a language model requires further research.

Our models do not consider abbreviations, spelling mistakes, and different stemming variants. In future research, we want to investigate models that are robust against such variants.

For our deep learning models, we apply a simple weak learning methodology using a single source. This methodology could be extended further by incorporating multiple dictionary sources. Similarly, linguistic and pattern rules are other good candidate sources for automatic annotation. In addition, learning good word representations for out of vocabulary terms requires further investigation.

7.3 Conclusion

This thesis investigated information extraction tasks from health related social media focusing on Parkinson's and COVID-19. While there has been a great deal of previous research into extracting information from social media narratives, to the best of our knowledge, no previous study applied a triage and diagnostic perspective for detecting COVID-19 in order to provide researchers and decision-makers with the symptom

severity and prevalence in the population. Specifically, we proposed a set of supervised and semi-supervised machine learning methods as well as deep learning models in a principled manner for medical concept extraction tasks. We believe such concept extraction methodologies utilising social media will generalise and can be applied to other diseases.

Bibliography

[1] Ahne, A., Khetan, V., Tannier, X., Rizvi, M.I.H., Czernichow, T., Orchard, F., Bour, C., Fano, A., Fagherazzi, G.: Identifying causal associations in tweets using deep learning: Use case on diabetes-related tweets from 2017-2021. arXiv preprint arXiv:2111.01225 (2021)

[2] Aji, A.F., Nityasya, M.N., Wibowo, H.A., Prasojo, R.E., Fatyanosa, T.: BERT goes brrr: A venture towards the lesser error in classifying medical self-reporters on Twitter. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 58–64. Association for Computational Linguistics, Mexico City, Mexico (2021)

[3] Carrillo-de Albornoz, J., Rodriguez Vidal, J., Plaza, L.: Feature engineering for sentiment analysis in e-health forums. PloS One 13(11), e0207996 (2018)

[4] Alfattni, G., Belousov, M., Peek, N., Nenadic, G., et al.: Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study. JMIR medical informatics 9(5), e24678 (2021)

[5] Alhuzali, H., Ananiadou, S.: Improving classification of Adverse Drug Reactions through Using Sentiment Analysis and Transfer Learning. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 339–347. Association for Computational Linguistics, Florence, Italy (2019)

[6] Allen, W.E., Altae-Tran, H., Briggs, J., Jin, X., McGee, G., Shi, A., Raghavan, R., Kamariza, M., Nova, N., Pereta, A., et al.: Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. Nature Human Behaviour 4(9), 972–982 (2020)

[7] Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)

[8] Alvaro, N., Miyao, Y., Collier, N., et al.: Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. JMIR Public Health and Surveillance 3(2), e6396 (2017)

[9] Alyasseri, Z.A.A., Al-Betar, M.A., Doush, I.A., Awadallah, M.A., Abasi, A.K., Makhadmeh, S.N., Alomari, O.A., Abdulkareem, K.H., Adam, A., Damasevicius, R., et al.: Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. Expert systems 39(3), e12759 (2022)

[10] Aramaki, E., Maskawa, S., Morita, M.: Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1568–1576. EMNLP '11 (2011)

[11] Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17(3), 229–236 (2010)

[12] Aronson, A.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)

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

[14] AskAPatient: `https://www.askapatient.com/`, accessed: 2019-07-11

[15] Bach, N., Badaskar, S.: A Review of Relation Extraction. Literature review for Language and Statistics II 2, 1–15 (2007)

[16] Banda, J.M., Singh, G.V., Alser, O.H., Prieto-Alhambra, D.: Long-term patient-reported symptoms of COVID-19: an analysis of social media data. medRxiv (2020)

[17] Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., Chowell, G.: A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. Epidemiologia 2(3), 315–324 (2021)

[18] Barros, J.M., Duggan, J., Rebholz-Schuhmann, D.: The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review. Journal of medical internet research 22(3), e13680 (2020)

[19] Batbaatar, E., Ryu, K.H.: Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach. International journal of environmental research and public health 16(19), 3628 (2019)

[20] Belousov, M.: Learning explainable representations of concepts in specialised languages: experiments in healthcare social media. Ph.D. thesis, University of Manchester (2020)

[21] Bengio, Y., Ducharme, R., Vincent, P.: A Neural Probabilistic Language Model. Journal of Machine Learning Research 3, 1137–1155 (2003)

[22] Bobicev, V., Sokolova, M.: Thumbs Up *and* Down: Sentiment Analysis of Medical Online Forums . In: Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task. pp. 22–26 (2018)

[23] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(suppl_1), D267–D270 (2004)

[24] Bose, P., Srinivasan, S., Sleeman IV, W.C., Palta, J., Kapoor, R., Ghosh, P.: A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. Applied Sciences 11(18), 8319 (2021)

[25] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., Cabitza, F.: Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. Journal of Medical Systems 44(8), 1–12 (2020)

[26] Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-Based *n*-gram Models of Natural Language. Computational Linguistics 18(4), 467–479 (1992)

[27] Burkhardt, S., Siekiera, J., Glodde, J., Andrade-Navarro, M.A., Kramer, S.: Towards identifying drug side effects from social media using active learning and crowd sourcing. In: Pacific Symposium on Biocomputing 2020. vol. 2020. World Scientific (2020)

[28] Byrd, K., Mansurov, A., Baysal, O.: Mining Twitter Data For Influenza Detection and Surveillance. In: Proceedings of the International Workshop on Software Engineering in Healthcare Systems. pp. 43–49 (2016)

[29] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011)

[30] Charles-Smith, L.E., Reynolds, T.L., Cameron, M.A., Conway, M., Lau, E.H., Olsen, J.M., Pavlin, J.A., Shigematsu, M., Streichert, L.C., Suda, K.J., et al.: Using

Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PloS One 10(10), e0139701 (2015)

[31] Chee, B.W., Berlin, R., Schatz, B.: Predicting Adverse Drug Events from Personal Health Messages. In: AMIA Annual Symposium Proceedings. vol. 2011, p. 217. American Medical Informatics Association (2011)

[32] Chen, E., Lerman, K., Ferrara, E., et al.: Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health and Surveillance 6(2), e19273 (2020)

[33] Chen, S., Huang, Y., Huang, X., Qin, H., Yan, J., Tang, B.: HITSZ-ICRC: A report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets. In: Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. pp. 47–51. Association for Computational Linguistics, Florence, Italy (2019)

[34] Chen, X., Sykora, M., Jackson, T., Elayan, S., Munir, F.: Tweeting Your Mental Health: Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions. In: Proceedings of the 51st Hawaii International Conference on System Sciences (2018)

[35] Chen, Y., Zhou, C., Li, T., Wu, H., Zhao, X., Ye, K., Liao, J.: Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. Journal of biomedical informatics 96, 103252 (2019)

[36] Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics 4, 357–370 (2016)

[37] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on

Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)

[38] Chowdhury, S., Zhang, C., Yu, P.S.: Multi-Task Pharmacovigilance Mining from Social Media Posts. In: WWW'18: Proceedings of the 2018 World Wide Web Conference. pp. 117–126 (2018)

[39] Clark, K., Luong, M.T., Manning, C.D., Le, Q.: Semi-Supervised Sequence Modeling with Cross-View Training. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1914–1925. Association for Computational Linguistics, Brussels, Belgium (2018)

[40] Cocos, A., Fiks, A.G., Masino, A.J.: Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. Journal of the American Medical Informatics Association 24(4), 813–821 (2017)

[41] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12, 2493–2537 (2011)

[42] COSTART: `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/`, accessed: 2016-06-21

[43] Cowie, J., Lehnert, W.: Information Extraction. Communications of the ACM 39(1), 80–91 (1996)

[44] python crfsuite: `https://python-crfsuite.readthedocs.io/en/latest/`, accessed: 2018-03-14

[45] Cunningham, H., Maynard, D., Bontcheva, K.: Text processing with gate. CA: Gateway Press (2011)

[46] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: an Architecture for Development of Robust HLT Applications. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 168–175 (2002)

[47] DailyStrength: `https://www.dailystrength.org/`, accessed: 2017-05-04

[48] Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine 64(1), 17–27 (2015)

[49] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)

[50] Ding, P., Zhou, X., Zhang, X., Wang, J., Lei, Z.: An Attentive Neural Sequence Labeling Model for Adverse Drug Reactions Mentions Extraction. IEEE Access 6, 73305–73315 (2018)

[51] Dligach, D., Bethard, S., Becker, L., Miller, T., Savova, G.K.: Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association 21(3), 448–454 (2014)

[52] Doan, S., Yang, E.W., Tilak, S.S., Li, P.W., Zisook, D.S., Torii, M.: Extracting health-related causality from twitter messages using natural language processing. BMC medical informatics and decision making 19(3), 71–77 (2019)

[53] Drew, D.A., Nguyen, L.H., Steves, C.J., Menni, C., Freydin, M., Varsavsky, T., Sudre, C.H., Cardoso, M.J., Ourselin, S., Wolf, J., et al.: Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science 368(6497), 1362–1367 (2020)

[54] Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support Vector Regression Machines. Proceedings of the 9th International Conference on Neural Information Processing Systems 9, 155–161 (1996)

[55] Edo-Osagie, O., De La Iglesia, B., Lake, I., Edeghere, O.: A scoping review of the use of Twitter for public health research. Computers in Biology and Medicine 122, 103770 (2020)

[56] Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., De La Iglesia, B.: Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. PloS One 14(7) (2019)

[57] Edwards, I.R., Aronson, J.K.: Adverse drug reactions: definitions, diagnosis, and management. The Lancet 356(9237), 1255–1259 (2000)

[58] Einstein, D.: Diagnosis of COVID-19 and its clinical spectrum AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr). Kaggle . `https://www.kaggle.com/einsteindata4u/covid19`, accessed:2021-02-24

[59] van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning 109(2), 1–68 (2019)

[60] Șerban, O., Thapen, N., Maginnis, B., Hankin, C., Foot, V.: Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. Information Processing & Management 56(3), 1166–1184 (2019)

[61] Esperanca, A., Miled, Z.B., Mahoui, M.: Social Media Sensing Framework for Population Health. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). pp. 0298–0304. IEEE (2019)

[62] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-Scale Information Extraction in KnowItAll

(Preliminary Results). In: Proceedings of the 13th International Conference on World Wide Web, WWW2004. pp. 100–110 (2004)

[63] Feinstein, A.R., Cicchetti, D.V.: High agreement but low Kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology 43(6), 543–549 (1990)

[64] Fries, J., Wu, S., Ratner, A., Ré, C.: SWELLSHARK: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. arXiv preprint arXiv:1704.06360 (2017)

[65] Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference. Berlin Heidelberg: Springer (2011)

[66] Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K., Gonzalez, G.: Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and classification Benchmark. In: Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. pp. 1–8. Citeseer (2014)

[67] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012–1014 (2009)

[68] Goldberg, Y.: A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research 57, 345–420 (2016)

[69] Golder, S., Klein, A.Z., Magge, A., O'Connor, K., Cai, H., Weissenbacher, D., Gonzalez-Hernandez, G.: Extending A Chronological and Geographical Analysis of Personal Reports of COVID-19 on Twitter to England, UK. medRxiv (2020)

[70] Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., Savova, G.: Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. Yearbook of Medical Informatics 26(01), 214–227 (2017)

[71] Gooch, P., Roudsari, A.: Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. Journal of Biomedical Informatics 45(5), 901–912 (2012)

[72] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press (2016)

[73] Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S.: Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In: Proceedings of the 2018 International Conference on Digital Health. pp. 121–125 (2018)

[74] Gu, G., Zhang, X., Zhu, X., Jian, Z., Chen, K., Wen, D., Gao, L., Zhang, S., Wang, F., Ma, H., Lei, J.: Development of a Consumer Health Vocabulary by Mining Health Forum Texts Based on Word Embedding: Semiautomatic Approach. JMIR Medical Informatics 7(2), e12704 (2019)

[75] Guan, H., Devarakonda, M.: Leveraging contextual information in extracting long distance relations from clinical notes. In: AMIA Annual Symposium Proceedings. vol. 2019, p. 1051. American Medical Informatics Association (2019)

[76] Guo, J.W., Radloff, C.L., Wawrzynski, S.E., Cloyes, K.G.: Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nursing 37(6), 934–940 (2020)

[77] Gupta, S., MacLean, D.L., Heer, J., Manning, C.D.: Induced lexico-syntactic patterns improve information extraction from online medical forums. Journal of the American Medical Informatics Association 21(5), 902–909 (2014)

[78] Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61(1), 29–48 (2008)

[79] Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33(14), i37–i48 (2017)

[80] Han, X., Eisenstein, J.: Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4238–4248. Association for Computational Linguistics, Hong Kong, China (2019)

[81] Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., Liu, Z., Li, P., Zhou, J., Sun, M.: More data, more relations, more context and more openness: A review and outlook for relation extraction. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. pp. 745–758. Association for Computational Linguistics, Suzhou, China (2020)

[82] Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, O.: 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association 27(1), 3–12 (2020)

[83] Hernandez, L.A.R., Callahan, T.J., Banda, J.M.: A biomedically oriented automatically annotated Twitter COVID-19 dataset. Genomics & Informatics 19(3) (2021)

[84] Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science 349(6245), 261–266 (2015)

[85] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (1997)

[86] Hu, H., Wang, H., Wang, F., Langley, D., Avram, A., Liu, M.: Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. Scientific Reports volume 8(1), 1–8 (2018)

[87] Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177 (2004)

[88] Izquierdo, J.L., Ancochea, J., Soriano, J.B., Group, S.C..R., et al.: Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing. Journal of medical Internet research 22(10), e21801 (2020)

[89] Judson, T.J., Odisho, A.Y., Neinstein, A.B., Chao, J., Williams, A., Miller, C., Moriarty, T., Gleason, N., Intinarelli, G., Gonzales, R.: Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. Journal of the American Medical Informatics Association 27(6), 860–866 (2020)

[90] Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C.: CADEC: A corpus of adverse drug event annotations. Journal of Biomedical Informatics 55, 73–81 (2015)

[91] Katragadda, S., Karnati, H., Pusala, M., Raghavan, V., Benton, R.: Detecting Adverse Drug Effects Using Link Classification on Twitter Data. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 675–679. IEEE (2015)

[92] Kiritchenko, S., Mohammad, S.M., Morin, J., de Bruijn, B.: NRC-Canada at SMM4H Shared Task: Classifying Tweets Mentioning Adverse Drug Reactions and Medication Intake. arXiv preprint arXiv:1805.04558 (2018)

[93] Klein, A., Alimova, I., Flores, I., Magge, A., Miftahutdinov, Z., Minard, A.L., O'Connor, K., Sarker, A., Tutubalina, E., Weissenbacher, D., Gonzalez-

Hernandez, G.: Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. pp. 27–36. Association for Computational Linguistics, Barcelona, Spain (Online) (2020)

[94] Klein, A.Z., Magge, A., O'Connor, K., Amaro, J.I.F., Weissenbacher, D., Hernandez, G.G.: Toward using Twitter for tracking COVID-19: a natural language processing pipeline and exploratory data set. Journal of Medical Internet Research 23(1), e25314 (2021)

[95] Kondylakis, H., Katehakis, D.G., Kouroubali, A., Logothetidis, F., Triantafyllidis, A., Kalamaras, I., Votis, K., Tzovaras, D., et al.: COVID-19 Mobile Apps: A Systematic Review of the Literature. Journal of medical Internet research 22(12), e23170 (2020)

[96] Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., Gonzalez, G.H.: Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of Biomedical Informatics 62, 148–158 (2016)

[97] Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., López, J.A., Nandal, U., et al.: Overview of the BioCreative VI chemical-protein interaction Track. In: Proceedings of the sixth BioCreative challenge evaluation workshop. vol. 1, pp. 141–146 (2017)

[98] Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Research 44(1), 1075–1079 (2015)

[99] Lafferty, J., McCallum, A., Pereira, F.C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). pp. 282–289 (2001)

[100] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (2016)

[101] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)

[102] Lavertu, A., Vora, B., Giacomini, K.M., Altman, R., Rensi, S.: A New Era in Pharmacovigilance: Toward Real-World Data and Digital Monitoring. Clinical Pharmacology & Therapeutics 109(5), 1197–1202 (2021)

[103] Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., Gonzalez, G.: Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. pp. 117–125. Association for Computational Linguistics, Uppsala, Sweden (2010)

[104] Lee, C.Y., Chen, Y.P.P.: Prediction of drug adverse events using deep learning in pharmaceutical discovery. Briefings in Bioinformatics 22(2), 1884–1901 (2021)

[105] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)

[106] Lee, K., Qadir, A., Hasan, S.A., Datla, V., Prakash, A., Liu, J., Farri, O.: Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In: Proceedings of the 26th International Conference on World Wide Web. pp. 705–714 (2017)

[107] Lin, G., Zaeem, R.N., Sun, H., Barber, K.S.: Trust Filter for Disease Surveillance: Identity. In: 2017 Intelligent Systems Conference (IntelliSys). pp. 1059–1066. IEEE (2017)

[108] Lin, W.S., Dai, H.J., Jonnagaddala, J., Chang, N.W., Jue, T.R., Iqbal, U., Shao, J.Y.H., Chiang, I.J., Li, Y.C.: Utilizing Different Word Representation Methods for Twitter Data in Adverse Drug Reactions Extraction. In: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI). pp. 260–265. IEEE (2015)

[109] Lison, P., Barnes, J., Hubin, A., Touileb, S.: Named entity recognition without labelled data: A weak supervision approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1518–1533. Association for Computational Linguistics, Online (2020)

[110] Liu, Y., Wang, Z., Ren, J., Tian, Y., Zhou, M., Zhou, T., Ye, K., Zhao, Y., Qiu, Y., Li, J., et al.: A COVID-19 Risk Assessment Decision Support System for General Practitioners: Design and Development Study. Journal of Medical Internet Research 22(6), e19786 (2020)

[111] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)

[112] López-Úbeda, P., Díaz-Galiano, M.C., Martín-Noguerol, T., Luna, A., Ureña-López, L.A., Martín-Valdivia, M.T.: COVID-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine 127, 104066 (2020)

[113] Luo, X., Gandhi, P., Storey, S., Huang, K.: A deep language model for symptom extraction from clinical text and its application to extract covid-19 symptoms

from social media. IEEE Journal of Biomedical and Health Informatics 26(4), 1737–1748 (2021)

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

[115] MacLean, D.L., Heer, J.: Identifying medical terms in patient-authored text: a crowdsourcing-based approach. Journal of the American Medical Informatics Association 20(6), 1120–1127 (2013)

[116] Magge, A., Klein, A., Miranda-Escalada, A., Ali Al-Garadi, M., Alimova, I., Miftahutdinov, Z., Farre, E., Lima López, S., Flores, I., O'Connor, K., Weissenbacher, D., Tutubalina, E., Sarker, A., Banda, J., Krallinger, M., Gonzalez-Hernandez, G.: Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 21–32. Association for Computational Linguistics, Mexico City, Mexico (2021)

[117] Magnolini, S., Piccioni, V., Balaraman, V., Guerini, M., Magnini, B.: How to Use Gazetteers for Entity Recognition with Neural Models. In: Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5). pp. 40–49. Association for Computational Linguistics, Macau, China (2019)

[118] Mahendran, D., McInnes, B.T.: Extracting Adverse Drug Events from Clinical Notes. AMIA Summits on Translational Science Proceedings 2021, 420 (2021)

[119] Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to Information Retrieval, vol. 1. Cambridge University Press (2008)

[120] Manning, C.D., Schütze, H., et al.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press (1999)

[121] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)

[122] Marsland, S.: Machine Learning: An Algorithmic Perspective. Chapman and Hall/CRC, Boca Raton, FLA, 2 edn. (2014)

[123] Massey, P.M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., Klassen, A.C.: Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter. Journal of Medical Internet Research 18(12), e6670 (2016)

[124] Mazzocut, M., Truccolo, I., Antonini, M., Rinaldi, F., Omero, P., Ferrarin, E., De Paoli, P., Tasso, C., et al.: Web Conversations About Complementary and Alternative Medicines and Cancer: Content and Sentiment Analysis. Journal of Medical Internet Research 18(6), e5521 (2016)

[125] McCallum, A., Li, W.: Early results for Named Entity Recognition with Conditional Random Fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 188–191. Association for Computational Linguistics (2003)

[126] MedHelp: `https://www.medhelp.org/`, accessed: 2018-06-18

[127] Menni, C., Valdes, A.M., Freidin, M.B., Sudre, C.H., Nguyen, L.H., Drew, D.A., Ganesh, S., Varsavsky, T., Cardoso, M.J., El-Sayed Moustafa, J.S., et al.: Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature Medicine 26(7), 1037–1040 (2020)

[128] Meystre, S.M., Heider, P.M., Kim, Y., Davis, M., Obeid, J., Madory, J., Alekseyenko, A.V.: Natural language processing enabling COVID-19 predictive ana-

lytics to support data-driven patient advising and pooled testing. Journal of the American Medical Informatics Association 29(1), 12–21 (2022)

[129] Miao, L., Last, M., Litvak, M.: An interactive analysis of user-reported long COVID symptoms using Twitter data. In: Proceedings of the 2nd Workshop on Deriving Insights from User-Generated Text. pp. 10–19. Association for Computational Linguistics, (Hybrid) Dublin, Ireland, and Virtual (2022)

[130] Miftahutdinov, Z., Alimova, I., Tutubalina, E.: KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue. In: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task. pp. 52–57 (2019)

[131] Miftahutdinov, Z., Sakhovskiy, A., Tutubalina, E.: KFU NLP team at SMM4H 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. pp. 51–56. Association for Computational Linguistics, Barcelona, Spain (Online) (2020)

[132] Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying Disease-related Expressions in Reviews Using Conditional Random Fields. In: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog. pp. 155–166 (2017)

[133] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: ICLR Workshop Papers (2013)

[134] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)

[135] Mizrahi, B., Shilo, S., Rossman, H., Kalkstein, N., Marcus, K., Barer, Y., Keshet, A., Shamir-Stein, N., Shalev, V., Zohar, A.E., et al.: Longitudinal symptom dynamics of COVID-19 infection. Nature Communications 11(1), 1–10 (2020)

[136] Mudinas, A., Zhang, D., Levene, M.: Bootstrap domain-specific sentiment classifiers from unlabeled corpora. Transactions of the Association for Computational Linguistics 6, 269–285 (2018)

[137] Müller, M., Salathé, M., Kummervold, P.E.: Covid-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv preprint arXiv:2005.07503 (2020)

[138] Na, J.C., Kyaing, W.Y.M.: Sentiment Analysis of User-Generated Content on Drug Review Websites. Journal of Information Science Theory and Practice 3(1), 6–23 (2015)

[139] Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 319–327. Association for Computational Linguistics, Florence, Italy (2019)

[140] Ng, J.Y., Abdelkader, W., Lokker, C.: Tracking discussions of complementary, alternative, and integrative medicine in the context of the COVID-19 pandemic: a month-by-month sentiment analysis of Twitter data. BMC Complementary Medicine and Therapies 22(1), 1–15 (2022)

[141] Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14. Association for Computational Linguistics, Online (2020)

[142] Nikfarjam, A., Gonzalez, G.H.: Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments. In: AMIA Annual Symposium Proceedings. vol. 2011, p. 1019. American Medical Informatics Association (2011)

[143] Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., Gonzalez, G.: Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association 22(3), 671–681 (2015)

[144] Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of Computation 35(151), 773–782 (1980)

[145] Norman, G.: Likert scales, levels of measurement and the "laws" of statistics. Advances in Health Sciences Education 15(5), 625–632 (2010)

[146] Obeid, J.S., Davis, M., Turner, M., Meystre, S.M., Heider, P.M., O'Bryan, E.C., Lenert, L.A.: An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. Journal of the American Medical Informatics Association 27(8), 1321–1325 (07 2020)

[147] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), `http://www.chokkan.org/software/crfsuite/`

[148] Patient: `https://patient.info/forums/discuss/browse/coronavirus-covid-19--4541` (2021), accessed: 2021-01-18

[149] PatientsLikeMe: `https://www.patientslikeme.com/`, accessed: 2017-04-21

[150] Paul, M., Dredze, M.: You Are What You Tweet: Analyzing Twitter for Public Health. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 5, pp. 265–272 (2011)

[151] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[152] Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[153] Perez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M., Dalianis, H.: Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. Journal of Biomedical Informatics 71, 16–30 (2017)

[154] Peshterliev, S., Dupuy, C., Kiss, I.: Self-Attention Gazetteer Embeddings for Named-Entity Recognition. arXiv preprint arXiv:2004.04060 (2020)

[155] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018)

[156] Pilipiec, P., Liwicki, M., Bota, A.: Using Machine Learning for Pharmacovigilance: A Systematic Review. Pharmaceutics 14(2), 266 (2022)

[157] Polanyi, L., Zaenen, A.: Contextual Valence Shifters. In: Computing Attitude and Affect in Text: Theory and Applications, pp. 1–10. Springer (2006)

[158] Qin, L., Sun, Q., Wang, Y., Wu, K.F., Chen, M., Shia, B.C., Wu, S.Y.: Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. International Journal of Environmental Research and Public Health 17(7) (2020)

[159] Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE. pp. 257–286 (1989)

[160] Ramesh, S., Tiwari, A., Choubey, P., Kashyap, S., Khose, S., Lakara, K., Singh, N., Verma, U.: BERT based transformers lead the way in extraction of health information from social media. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 33–38. Association for Computational Linguistics, Mexico City, Mexico (2021)

[161] Rastegar-Mojarad, M., Elayavilli, R.K., Yu, Y., Liu, H.: Detecting signals in noisy data - can ensemble classifiers help identify adverse drug reaction in tweets? In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing (2016)

[162] Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid Training Data Creation with Weak Supervision. In: Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases. p. 269. NIH Public Access (2017)

[163] Razak, C.S.A., Zulkarnain, M.A., Hamid, S.H.A., Anuar, N.B., Jali, M.Z., Meon, H.: Tweep: A System Development to Detect Depression in Twitter Posts. In: Computational Science and Technology, pp. 543–552. Springer (2020)

[164] Reis, E.S.D., Costa, C.A.D., Silveira, D.E.D., Bavaresco, R.S., Righi, R.D.R., Barbosa, J.L.V., Antunes, R.S., Gomes, M.M., Federizzi, G.: Transformers aftermath: Current research and rising trends. Communications of the ACM 64(4), 154–163 (2021)

[165] Riloff, E., Jones, R., et al.: Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In: AAAI-99 Proceedigns. pp. 474–479 (1999)

[166] Ru, B., Harris, K., Yao, L.: A Content Analysis of Patient-Reported Medication Outcomes on Social Media. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). pp. 472–479. IEEE (2015)

[167] Rudra, K., Sharma, A., Ganguly, N., Imran, M.: Classifying Information from Microblogs during Epidemics. In: Proceedings of the 2017 International Conference on Digital Health. pp. 104–108 (2017)

[168] Rumelhart, D., Hinton, G., Williams, R.: Learning internal representations by backpropagating errors. Nature 323(533-536), 807 (1986)

[169] RXNORM: `https://www.nlm.nih.gov/research/umls/rxnorm/`, accessed: 2016-06-21

[170] Sakhovskiy, A., Miftahutdinov, Z., Tutubalina, E.: KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 39–43. Association for Computational Linguistics, Mexico City, Mexico (2021)

[171] Sampathkumar, H., Chen, X.w., Luo, B.: Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. BMC Medical Informatics and Decision Making 14(1), 1–18 (2014)

[172] Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Conference on Natural Language Learning (CoNLL). pp. 142–147 (2003)

[173] Sarabadani, S., Baruah, G., Fossat, Y., Jeon, J., et al.: Longitudinal Changes of COVID-19 Symptoms in Social Media: Observational Study. Journal of medical Internet research 24(2), e33959 (2022)

[174] Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: a review. Journal of Biomedical Informatics 54, 202–212 (2015)

[175] Sarker, A., Gonzalez-Hernandez, G.: Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. Training 1(10,822), 1239 (2017)

[176] Sarker, A., Lakamana, S., Hogg-Bremer, W., Xie, A., Al-Garadi, M.A., Yang, Y.C.: Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Journal of the American Medical Informatics Association 27(8), 1310–1315 (07 2020)

[177] Sarker, A., Nikfarjam, A., Gonzalez, G.: Social Media Minning Shared Task Workshop. In: Biocomputing 2016: Proceedings of the Pacific Symposium. pp. 581–592. World Scientific (2016)

[178] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)

[179] Scepanovic, S., Martin-Lopez, E., Quercia, D., Baykaner, K.: Extracting Medical Entities from Social Media. In: CHIL'20: Proceedings of the ACM Conference on Health, Inference, and Learning. p. 170–181. CHIL '20, Association for Computing Machinery, New York, NY, USA (2020)

[180] Schmeelk, S., Davis, A., Li, Q., Shippey, C., Utah, M., Myers, A., Turchioe, M.R., Creber, R.M., et al.: Monitoring Symptoms of COVID-19: Review of Mobile Apps. JMIR mHealth and uHealth 10(6), e36065 (2022)

[181] Schwab, P., DuMont Schütte, A., Dietz, B., Bauer, S.: Clinical Predictive Models for COVID-19: Systematic Study. Journal of Medical Internet Research 22(10), e21439 (Oct 2020)

[182] Segura-Bedmar, I., Martínez, P., Herrero-Zazo, M.: SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 341–350 (2013)

[183] Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., Han, J.: Learning Named Entity Tagger using Domain-Specific Dictionary. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2018)

[184] Shen, C., Chen, A., Luo, C., Zhang, J., Feng, B., Liao, W., et al.: Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. Journal of Medical Internet Research 22(5), e19421 (2020)

[185] Silverman, G.M., Sahoo, H.S., Ingraham, N.E., Lupei, M., Puskarich, M.A., Usher, M., Dries, J., Finzel, R.L., Murray, E., Sartori, J., et al.: NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models. Journal of Artificial Intelligence Research 72, 429–474 (2021)

[186] Skaik, R., Inkpen, D.: Using Social Media for Mental Health Surveillance: A Review. ACM Computing Surveys 53(6) (dec 2020)

[187] Skaik, R., Inkpen, D.: Using Twitter Social Media for Depression Detection in the Canadian Population. In: 2020 3rd Artificial Intelligence and Cloud Computing Conference. pp. 109–114 (2020)

[188] Song, C.H., Lawrie, D., Finin, T., Mayfield, J.: Improving Neural Named Entity Recognition with Gazetteers. arXiv preprint arXiv:2003.03072 (2020)

[189] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., Xu, H.: CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association 25(3), 331–336 (2018)

[190] spaCy: `https://spacy.io/` (2022), accessed: 2022-07-29

[191] Strauss, B., Toma, B., Ritter, A., De Marneffe, M.C., Xu, W.: Results of the WNUT16 Named Entity Recognition Shared Task. In: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). pp. 138–144 (2016)

[192] Sun, Q., Bhatia, P.: Neural Entity Recognition with Gazetteer based Fusion. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3291–3295. Association for Computational Linguistics, Online (2021)

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

[194] Thain, D., Tannenbaum, T., Livny, M.: Distributed computing in practice: The Condor experience. Concurrency and Computation: Practice and Experience 17(2-4), 323–356 (2005)

[195] The NHS website: `https://web.archive.org/web/20200316223405/https://www.nhs.uk/conditions/coronavirus-covid-19/` (2020), accessed: 2021-07-07

[196] Tutubalina, E., Nikolenko, S.: Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. Journal of Healthcare Engineering 2017 (2017)

[197] Vaid, A., Somani, S., Russak, A.J., De Freitas, J.K., Chaudhry, F.F., Paranjpe, I., Johnson, K.W., Lee, S.J., Miotto, R., Richter, F., et al.: Machine Learning to Pre-

dict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. Journal of Medical Internet Research 22(11), e24018 (2020)

[198] Valdes, A., Lopez, J., Montes, M.: UACH-INAOE at SMM4H: a BERT based approach for classification of COVID-19 Twitter posts. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 65–68. Association for Computational Linguistics, Mexico City, Mexico (2021)

[199] Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning 109(2), 373–440 (2020)

[200] Vanwinckelen, G., Blockeel, H.: On Estimating Model Accuracy with Repeated Cross-Validation. In: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning. pp. 39–44 (2012)

[201] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. Advances in Neural Information Processing Systems 30 (2017)

[202] de Vet, H.C., Mokkink, L.B., Terwee, C.B., Hoekstra, O.S., Knol, D.L.: Clinicians are right not to like Cohen's κ. BMJ 346 (2013)

[203] Vydiswaran, V.V., Mei, Q., Hanauer, D.A., Zheng, K.: Mining Consumer Health Vocabulary from Community-Generated Text. In: AMIA Annual Symposium Proceedings. vol. 2014, p. 1150. American Medical Informatics Association (2014)

[204] Wallis, S.: Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. Journal of Quantitative Linguistics 20(3), 178–208 (Aug 2013)

[205] Wang, C.K., Dai, H.J., Zhang, Y.C., Xu, B.C., Wang, B.H., Xu, Y.N., Chen, P.H., Lee, C.H.: ISLab system for SMM4H shared task 2020. In: Proceedings of the

Fifth Social Media Mining for Health Applications Workshop & Shared Task. pp. 42–45. Association for Computational Linguistics, Barcelona, Spain (Online) (2020)

[206] Wang, J., Yu, L.C., Zhang, X.: Explainable detection of adverse drug reaction with imbalanced data distribution. PLoS Computational Biology 18(6), e1010144 (2022)

[207] Wang, J., Abu-el Rub, N., Gray, J., Pham, H.A., Zhou, Y., Manion, F.J., Liu, M., Song, X., Xu, H., Rouhizadeh, M., et al.: COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model. Journal of the American Medical Informatics Association 28(6), 1275–1283 (2021)

[208] Wang, K., Zuo, P., Liu, Y., Zhang, M., Zhao, X., Xie, S., Zhang, H., Chen, X., Liu, C.: 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)

[209] Wang, S., Li, Y., Ferguson, D., Zhai, C.: SideEffectPTM: An Unsupervised Topic Model to Mine Adverse Drug Reactions from Health Forums. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. p. 321–330. BCB '14, Association for Computing Machinery, New York, NY, USA (2014)

[210] Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data 3(1), 1–40 (2016)

[211] Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M.J., Gonzalez-Hernandez, G.: Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In: Proceedings of the Fourth Social

Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. pp. 21–30. Association for Computational Linguistics, Florence, Italy (Aug 2019)

[212] Weissenbacher, D., Sarker, A., Paul, M.J., Gonzalez-Hernandez, G.: Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. pp. 13–16. Association for Computational Linguistics, Brussels, Belgium (2018)

[213] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature Selection for SVMs. In: Advances in Neural Information Processing Systems. vol. 13 (2000)

[214] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). pp. 347–354 (2005)

[215] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2016)

[216] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)

[217] Wongpakaran, N., Wongpakaran, T., Wedding, D., Gwet, K.L.: A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Medical Research Methodology 13(1), 1–7 (2013)

[218] Wu, C., Wu, F., Liu, J., Wu, S., Huang, Y., Xie, X.: Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representa-

tion and multi-head self-attention. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. pp. 34–37. Association for Computational Linguistics, Brussels, Belgium (2018)

[219] Wu, H., Fang, H., Stanhope, S.J.: An Early Warning System for Unrecognized Drug Side Effects Discovery. In: WWW '12 Companion: Proceedings of the 21st International Conference on World Wide Web. pp. 437–440 (2012)

[220] Wu, L., Moh, T.S., Khuri, N.: Twitter opinion mining for adverse drug reactions. In: 2015 IEEE International Conference on Big Data (Big Data). pp. 1570–1574. IEEE (2015)

[221] Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al.: Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association 27(3), 457–470 (2020)

[222] Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Bonten, M.M., Dahly, D.L., Damen, J.A., Debray, T.P., et al.: Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369 (2020)

[223] Xu, R., Supekar, K., Morgan, A., Das, A., Garber, A.: Unsupervised Method for Automatic Construction of a Disease Dictionary from a Large Free Text Collection. In: AMIA Annual Symposium Proceedings. vol. 2008, p. 820. American Medical Informatics Association (2008)

[224] Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P.: Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)

[225] Yang, C.C., Yang, H., Jiang, L., Zhang, M.: Social media mining for drug safety signal detection. In: SHB'12: Proceedings of the 2012 international workshop on Smart health and wellbeing. pp. 33–40 (2012)

[226] Yates, A., Goharian, N.: ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In: European Conference on Information Retrieval. pp. 816–819. Springer (2013)

[227] Zeng, Q.T., Tse, T.: Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association 13(1), 24–29 (2006)

[228] Zens, M., Brammertz, A., Herpich, J., Südkamp, N., Hinterseer, M., et al.: App-Based Tracking of Self-Reported COVID-19 Symptoms: Analysis of Questionnaire Data. Journal of Medical Internet Research 22(9), e21956 (2020)

[229] Zhang, J., Yu, Y., Li, Y., Wang, Y., Yang, Y., Yang, M., Ratner, A.: WRENCH: A Comprehensive Benchmark for Weak Supervision. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[230] Zhang, L., Hall, M., Bastola, D.: Utilizing Twitter data for analysis of chemotherapy. International Journal of Medical Informatics 120, 92–100 (2018)

[231] Zhang, T., Lin, H., Ren, Y., Yang, Z., Wang, J., Duan, X., Xu, B.: Identifying adverse drug reaction entities from social media with adversarial transfer learning model. Neurocomputing 453, 254–262 (2021)

[232] Zhang, Y., Lin, H., Yang, Z., Wang, J., Sun, Y., Xu, B., Zhao, Z.: Neural network-based approaches for biomedical relation classification: A review. Journal of Biomedical Informatics 99, 103294 (2019)

[233] Zhang, Y., Zheng, W., Lin, H., Wang, J., Yang, Z., Dumontier, M.: Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 34(5), 828–835 (2018)

[234] Zhao, X., Ding, H., Feng, Z.: GLaRA: Graph-based labeling rule augmentation for weakly supervised named entity recognition. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 3636–3649. Association for Computational Linguistics, Online (2021)

[235] Zheng, W., Lin, H., Liu, X., Xu, B.: A document level neural model integrated domain knowledge for chemical-induced disease relations. BMC Bioinformatics 19(1), 1–12 (2018)

[236] Zheng, X., Li, P., Hu, X., Yu, K.: Semi-supervised classification on data streams with recurring concept drift and concept evolution. Knowledge-Based Systems 215, 106749 (2021)

[237] Zhou, T., Li, Z., Gan, Z., Zhang, B., Chen, Y., Niu, K., Wan, J., Liu, K., Zhao, J., Shi, Y., Chong, W., Liu, S.: Classification, extraction, and normalization : CASIA_Unisound team at the social media mining for health 2021 shared tasks. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. pp. 77–82. Association for Computational Linguistics, Mexico City, Mexico (2021)

[238] Zhou, Z.H., Li, M.: Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. IEEE Transactions on Knowledge & Data Engineering (11), 1529–1541 (2005)

[239] Zhu, X., Goldberg, A.B.: Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 3(1), 1–130 (Jun 2009)

[240] Zoabi, Y., Deri-Rozov, S., Shomron, N.: Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine 4(1), 1–5 (2021)

[241] Žunić, A., Corcoran, P., Spasić, I.: Aspect-based sentiment analysis with graph convolution over syntactic dependencies. Artificial Intelligence in Medicine 119, 102138 (2021)

[242] Žunić, A., Corcoran, P., Spasić, I.: The case of aspect in sentiment analysis: Seeking attention or co-dependency? Machine Learning and Knowledge Extraction 4(2), 474–487 (2022)

[243] Žunić, A., Corcoran, P., Spasic, I., et al.: Sentiment Analysis in Health and Well-Being: Systematic Review. JMIR Medical Informatics 8(1), e16023 (2020)