



Natural Language Processing for Health System Messages: Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content

Citation

Sun, Mary Daotung. 2021. Natural Language Processing for Health System Messages: Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content. Master's thesis, Harvard University Division of Continuing Education.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37370747>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

Natural Language Processing for Health System Messages:
Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content

Mary Sun

A Thesis in the Field of Software Engineering
for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2022

Abstract

Recent efforts in clinical natural language processing research have focused on mining social media data and other user-generated data relevant to COVID-19. An increasing number of machine learning tools have been developed for annotation, topic modeling, and various classification tasks, including the now widely used Bidirectional Encoder Representations from Transformer (BERT) language model. In this study, we improve the domain specificity of COVID-Twitter-BERT for secure health system messages through supervised fine-tuning on a large corpus of patient-originated messages. Following sentiment annotation with a validated rule-based heuristic tool, an auxiliary sentence approach was implemented to conduct aspect-based sentiment analysis. We find that existing machine learning tools optimized on social media tools demonstrate moderate-to-good generalizability to secure patient messages. Furthermore, significant performance improvements to a weighted average F1 score of 0.881 were realized upon implementation of fine-tuning and auxiliary sentence approaches. We also observe temporal trends in positive, negative, and compound (overall) sentiment measured on over 3.2 million patient messages related to COVID-19. We believe that secure health system messages are an increasingly valuable source of clinical text data and can be effectively mined using emerging deep transfer learning techniques.

Dedication

To my incredible partner, David, for your loving advice and steadfast support during this roller coaster of an educational journey. You are my safe harbor.

To my inspiring parents, Michelle and John, who have made every sacrifice to provide their children with unimaginable opportunity. Thank you a million times over.

Acknowledgments

I am incredibly fortunate to have met and worked with my thesis director, Dr. Junko Takeshita. In addition to being a fount of knowledge, she has been a constant source of inspiration and learning during the thesis process. Her patient guidance and infectious passion for clinical epidemiology have informed all aspects of this work.

I extend my heartfelt gratitude to my research advisor, Dr. Hongming Wang. Her valuable advice, never-ending patience, and encouragement at every turn has made this project possible. I am also grateful to Dr. Sylvain Jaume for his thoughtful advice and guidance during the thesis process and my overall course of study.

Table of Contents

Abstract.....	3
Dedication.....	iv
Acknowledgments.....	v
List of Tables	viii
List of Figures	ix
Chapter I Introduction.....	1
Chapter II Background and Technical Context	2
2.1 Introduction.....	2
2.2 Clinical NLP	4
2.2.1 Acuity Triage	4
2.2.2 Event Detection.....	6
2.2.3 Case (Diagnostic) Detection	9
2.2.4 Patient Sentiment	10
2.3 Technical Approaches in Clinical NLP	12
2.3.1 Systematized Nomenclature.....	12
2.3.2 Public Annotated Corpora.....	14
2.3.3 Named Entity Recognition.....	22
2.3.4 Medical Concept Normalization	25
Chapter III Intersections of Clinical NLP and COVID-19	33
3.1 User-Generated Corpora	33

3.2 Deep Learning for Clinical Knowledge	37
Chapter IV Aspect-Based Sentiment Analysis of Health System Messages	39
4.1 Approach	39
4.2 Text Preprocessing	40
4.2.1 Empirical Techniques	41
4.2.2 Rule-Based Heuristics	43
4.3 Parsimonious Lexicon and Rule-Based Annotation	45
4.4 Fine-Tuning of CT-BERT	46
4.5 Auxiliary Sentence Approach	49
Chapter V Development Tools and Resources	51
5.1 Electronic Patient Portal	51
5.2 Deep Transfer Learning in R	51
5.3 Hyperparameters	52
Chapter VI Results	53
6.1 VADER Output	53
6.2 Fine-Tuned CT-BERT with Auxiliary Sentence Approach	55
Chapter VII Discussion and Conclusions	57
7.1 Experimental Outcomes	57
7.2 Limitations	59
7.3 Future Directions	60
References	63

List of Tables

Table 1. Normalized Averages of VADER Output Scores for All Tokenized Message Components Relevant to COVID-19, Segmented by Time Period.	54
Table 2. Comparative Evaluation of Aspect-Based Sentiment Analysis Models Evaluated on a Novel Corpus of Patient-Originated Secure Health System Messages Containing COVID-19 Content.	56

List of Figures

Figure 1. Proposed Fine-Tuned Deep Transfer Learning Approach to Aspect-Based Sentiment Classification.	40
Figure 2. Architecture of Proposed CT-BERT Fine-Tuning Process.	48
Figure 3. Overview of Proposed CT-BERT Fine-Tuning Process.	49

Chapter I

Introduction

This thesis presents the design, architecture, and implementation of a rule- and deep transfer learning-based algorithm for improved sentiment analysis of secure health system messages. Despite the increasing volume of health system messages and telehealth-related communications in the era of COVID-19, most natural language processing (NLP) studies have focused on more freely minable social media data. By collating and utilizing a large, richly featured corpus of health system messages, this work provides a unique comparison of extant approaches to COVID-related sentiment analysis. We also investigate the generalizability of popular transformation-based models to the analysis of patient-originated secure messages through the evaluation of recently described rule-based and machine fine-tuning strategies. Our proposed implementation aims to demonstrate a feasible method for domain-specific, aspect-based sentiment analysis of COVID-19 content that broadens the existing scope of communication corpora in clinical NLP.

Chapter II

Background and Technical Context

2.1 Introduction

Natural language processing (NLP) is a computational subfield at the intersection of linguistics, computer science, and artificial intelligence. Though the term is used broadly, nearly all applications of NLP seek to functionally understand human language through computational analysis. To this end, machine NLP generally treats a language as a set of symbols arranged according to some discrete, discoverable set of rules (Khurana et al., 2017). By utilizing a variety of frameworks to modularize, manipulate, and otherwise transform these representations of language, such algorithms attempt to accurately receive or transmit information from large volumes of text data. Given the extraordinary rate at which unstructured text data is generated in today's digital age, ongoing NLP efforts primarily focus on the extraction of context and meaning from unstructured free text.

Challenges in NLP can be categorized into issues of speech recognition, natural language understanding (deconstruction), and natural language generation (construction). Speech recognition, often referred to as "speech-to-text," is the machine processing of spoken language into a meaningful textual format. Natural language understanding refers to machine "comprehension" of human-originated text that informs machine-based action. Finally, natural language generation is the programmatic transformation of various types of input data into human-readable text. The level of task complexity varies

based on a diverse array of application goals, input data, and analysis layers, with many real-world applications involving aspects across all three domains.

In recent years, healthcare has become an important focus of applied NLP research. Indeed, NLP is the most widely used big data analytical technique in healthcare and frequently applied to internal health system data sources as well external data originating from state agencies, federally-funded laboratories, pharmaceutical companies, and medical journals (Mehta & Pandit, 2018). The availability of rich internal data is largely due to the widespread use of electronic medical record (EMR) systems, which were federally mandated for all United States healthcare providers starting in 2014 (Congress, 2009). Given the limitations of coding structured EMR fields and ever-increasing volumes of regulatory documentation, many clinicians opt to capture the nuances of diagnostic reasoning in free text notes. This data is often the only source of patient status indicators, provider recommendations, and interprofessional consults, among other critical pieces of clinical information (Lovis et al., 2000; Tange et al., 1997; Walsh, 2004). Additionally, many health systems have implemented patient portals and telemedicine apps that give patients and providers the option to exchange messages. These messaging systems tend to mimic email services and are almost always designed using a free text format. As a result, an estimated 80-90% of all medical record and patient-provider communications data is stored as unstructured free text (Vashisth et al., 2020).

Despite the vast array of text-based data available in EMRs and patient-provider messaging portals, this information is widely underutilized due to a lack of accessible end-user interface tools (Broseta, 2021). Accessibility problems are further exacerbated

by the fragmentation of software development companies and the lack of interoperability across many EMR programs (Jensen et al., 2012; Technology, 2017). By design, NLP methods are well-positioned to turn the existing landscape of feature-rich, text-based clinical data repositories into usable drivers of downstream healthcare operations. Examples of common development focuses include the extraction of actionable insights from clinical notes, differentiation of low vs. high-risk patients, and real-time augmentation of diagnostic decisions (Broseta, 2021). NLP-enabled text mining applications are uniquely suited to address data challenges in modern medicine and could potentially drive transformative advances across all aspects of healthcare delivery.

2.2 Clinical NLP

Over the past decade, the growing adoption of EMR systems has digitized clinical information and driven an unprecedented increase in the volume of unorganized data. Given the increasing regulatory complexity of modern medicine, the lack of standardization across health systems, and the growing ubiquity of data-rich environments in healthcare, there is an urgent need for tools that can extract actionable value from unstructured text. Recent estimates predict that the value of the global healthcare NLP market will rise by 20% over the next few years to nearly \$7.5 billion USD in 2027 (Research, 2020). Investment in large-scale clinical research projects has risen accordingly, with NLP development concentrated in several key areas.

2.2.1 Acuity Triage

Patients who present acutely are often triaged by medical staff as part of the clinical intake process. NLP algorithms that can identify patients at higher risk of

imminent adverse events are being actively developed to deploy resources more efficiently across healthcare. For instance, a study of 139,150 adult emergency department visits found that NLP analyses of free-text reason for visit data collected during triage improved models predicting whether advanced diagnostic imaging would be needed (Zhang, Kim, et al., 2019). Similar models have also been developed for pediatric emergency visits (Zhang, Bellolio, et al., 2019). Other neural network-based algorithms have been used to ingest clinical notes at the point of triage to predict features ranging from emergency department resource requirements, patient acuity levels, and even future need for hospitalization with moderate to excellent accuracy (Choi et al., 2019; Jiang et al., 2021; Patel et al., 2018; Sterling et al., 2020). Recently, a deep learning tool showed excellent performance in triaging radiation oncology incident reports based on the degree of dosage irregularity (Syed et al., 2020). These and other recent studies suggest predictive NLP tools can be used to improve outcomes in emergent care.

Even before the point of care, clinical NLP can also be used to triage incoming messages from patients. In 2017, rule-based and machine learning approaches were used to automate the classification of patient messages sent through secure patient portals. Messages were categorized under all applicable communication types (i.e., informational, medical, prescriptions, and so on), with several models demonstrating excellent performance (Cronin et al., 2017; Elmessiry et al., 2017). These results have since been replicated with other neural networks across health systems and content types; newer studies have begun to also account for socioemotional tone (De et al., 2021; Hogan et al., 2018; Sulieman et al., 2017). Interestingly, researchers have also applied NLP analysis to messages coordinating patient care between medical staff and administrators.

Assessments of breast cancer treatment teams have identified areas where clinicians' clerical burdens can be reduced to combat provider burnout (Steitz et al., 2021). This emerging body of evidence demonstrates the utility of NLP "pre-processing" in the management of increasing patient-provider messages volumes.

2.2.2 Event Detection

Electronic health records contain significant amounts of phenotypic information that are not captured in structured (coded) EMR fields. While structured fields tend to be used for billing processes and demographic surveys, unstructured data contains clinically pertinent information authored and primarily utilized by providers. Patient status updates, progress notes, and other indicators of medical status are included to help determine medical status and appropriate next steps for patients. These records are frequently used in clinical research to determine the occurrence of adverse events, investigate symptoms that may contribute to or characterize diseases, and retrospectively investigate clinical outcomes. NLP algorithms capable of identifying parameterized medical events from the continuous, unorganized stream of EMR updates could increase real-time adherence to clinical protocols and enhance clinical research.

Detecting and preventing adverse events is an important public health priority. Adverse drug reactions (ADRs) drive up to 28% of visits to emergency department visits and 6.4% of hospital fatalities (Patel & Zed, 2002; Wester et al., 2008). However, voluntary reporting by most institutions remains largely unsuccessful (Hazell & Shakir, 2006). Over the last decade, text mining algorithms have become increasingly valuable in identifying ADRs in narrative EMR documents. As algorithm complexity increases from

simple free text searches to NLP-enabled analyses, identification of ADRs generally improves (Warrer et al., 2012; Wong et al., 2018; Wunnava et al., 2019; Young et al., 2019). Recently, a public server was developed to automatically ingest, aggregate, and report gene expression changes associated with drug resistance by disease category. By combining NLP and rule-based methods, researchers were able to construct a hybrid set of textual features that identify different drug response relationships with excellent precision and accuracy (Khalid & Sezerman, 2017). Future applications of these and similar tools show great promise for evaluating ADRs in complex disease.

In addition to the discovery of ADRs, narrative EMR documentation plays a critical role in the assessment of medical status. NLP methods have advanced the algorithmic detection of symptoms across multiple disease categories, most notably in cardiac and other emergency medical events, chronic diseases, surgery, and more recently mental health (Hernandez-Boussard et al., 2019; Koleck et al., 2019; Le Glaz et al., 2021; S. Lee et al., 2019; Mellia et al., 2021; Sheikhalishahi et al., 2019). For example, a recent United Kingdom (UK) study applied a rule-based natural language engine to the medical records of 2000 randomized patients with a recorded myocardial infarction (MI). Researchers extracted 10,927 diagnoses and thousands of related test results, negation statements, and suspected secondary diagnoses. Compared to information in structured fields, their results showed that the proportion of patients with preceding chest pain was significantly higher, that subtypes of MI could be specifically defined, and that recorded causes of death were often incomplete (Shah et al., 2019). Many studies find that NLP-based processing similarly outperforms billing codes in evaluating symptom burden across hemodialysis, allergic drug reactions, non-alcoholic fatty liver disease, epilepsy,

and other medical conditions (Banerji et al., 2020; Chan et al., 2020; Decker et al., 2021; Van Vleck et al., 2019). Given the algorithmic nature of clinical protocols and the importance of staging/grading in disease classification, NLP systems provide clear value in patient management scenarios.

Finally, free-text documentation is a rich source of information for clinical researchers seeking to improve our understanding of disease course and current treatment regimes. NLP mitigates several key challenges in outcomes research, including the difficulty of modeling outcomes at an individual or group level, prohibitive costs of large-sample studies, and inaccessibility of other non-EMR text data at meaningful scale (Velupillai et al., 2018). NLP opportunities are particularly exciting in mental health research, where most medical documentation is in free-text format and external data is increasingly populated. For example, deep learning has been used to characterize outcomes in patients with psychiatric conditions using social media posts, transcripts of online and text-based psychotherapy sessions, and notes from interprofessional consults (Althoff et al., 2016; Gkotsis et al., 2017; Goldberg et al., 2020; Obeid et al., 2019). In a recent ophthalmologic surgery study, a rule-based laterality classifier built using the records of 6347 patients achieved complete accuracy in categorizing surgery types when compared to human annotators (S. Y. Wang et al., 2020). NLP algorithms have also been used in oncology to quantify factors associated with cancer recurrence and overall survival (Carrell et al., 2014; Kehl et al., 2020; Ping et al., 2013; Strauss et al., 2013; Yim et al., 2016). These and other text-processing pipelines are important methods for capturing clinical events and advancing clinical outcomes research.

2.2.3 Case (Diagnostic) Detection

Achieving diagnostic accuracy, usually as early as possible in a patient's clinical course, is integral to providing effective care. Machine learning applications may be able to synthesize large volumes of dynamic, unstructured text data to augment clinical diagnoses in real-time. In the medical literature, rapid advances in the automated understanding of radiology reports are at the forefront of NLP-based case detection. A systematic review published in 2021 identified 164 publications describing applications of NLP in radiology. Of models that extracted information from free text, nearly 40% surveilled the progression of known disease while 20% attempted to classify disease types and/or stages (Casey et al., 2021). Recently, an increasing number of studies have applied spatial role labeling annotation schemes to radiologic reports in the development of deep learning-based entity extraction methods (Datta & Roberts, 2020). Algorithms have demonstrated accuracy in the investigation of tumor characteristics such as size, progression status, change over time, and treatment response (Bozkurt et al., 2019; P. H. Chen et al., 2018; Cheng et al., 2010; Hassanpour et al., 2017). Classification algorithms have also been developed to identify and further classify a range of conditions including ischemic stroke, pulmonary embolism, bone fractures, ophthalmologic lesions, and late-life depression, among many others (Casey et al., 2021; M. C. Chen et al., 2018; DeSouza et al., 2021; Kim et al., 2019; C. Lee et al., 2019; Yang et al., 2021).

Patient-reported outcomes (PROs) are another important aspect of patient phenotyping. Due to their subjective nature, PROs are rarely systematically recorded and are most often found within unstructured text in EMRs. Programmatic NLP can be an efficient way to convert unstructured PRO narratives to structured labels that augment

diagnostics (Gonzalez-Hernandez et al., 2017). In a recent study, a hybrid NLP pipeline was developed to utilize PROs in the classification of urinary incontinence. Within a cohort of 3642 prostate cancer patients, this model achieved excellent performance (Bozkurt et al., 2020). Another study utilized over 11 million patient-generated posts in online support group networks run by a large health system. NLP analysis found statistically significant correlations between statin use and discussions of memory impairment, aphasia, cognitive impairment, and emotional change (Timimi et al., 2019). In a study of acute coronary syndromes, patient-reported post-discharge pain was associated with shorter periods before rehospitalization (J. Chen et al., 2021). Furthermore, early studies suggest that NLP analyses may also improve cohort identification in the case of relatively common diseases such as diabetes and rare diseases such as X-linked hypophosphatemia (Broseta, 2021; Turchin & Florez Builes, 2021). In these scenarios and likely many others, the extraction of targeted features from patient-reported information can improve diagnostic accuracy.

2.2.4 Patient Sentiment

Improving the patient experience has become an important priority for many healthcare institutions. Since patient feedback is often text-based and solicited via electronic surveys or captured within patient portal messages, NLP algorithms are well positioned to address this need. Specific techniques designed to perform sentiment analysis, also referred to as “opinion mining” or “emotion AI,” allow investigators to measure affective states expressed in free text. A recent narrative review identified 19 studies where language analysis techniques were applied to both unsolicited patient feedback posted to social media sites and solicited feedback provided through structured

surveys. Researchers found that a relatively small vocabulary set accounted for 92% of content, suggesting that the perspectives and language domains used in patient feedback are distinctly fixed. The results of topic modelling analyses were also similar to those of topic classification analyses (Khanbhai et al., 2021). These data characteristics improve the chances that NLP classifiers can reliably and reproducibly evaluate sentiment in patient feedback (Smith, 2017). Interestingly, several studies found no relationship between conventional metrics of patient sentiment and sentiment in matched Twitter posts (Greaves et al., 2014; Hawkins et al., 2016). Social media analysis may instead be better used to gain insight into how patients and their families experience disease (Gabarron et al., 2019; Zunic et al., 2020).

Patient perception also plays a key role in the adoption of new health system technologies. Automated conversational agents, which often employ underlying NLP algorithms to interpret and respond to input text, are increasingly utilized in healthcare. Agents have been developed for various uses including health screenings, patient triage, counseling, and at-home health management; they demonstrate effectiveness in efficiently allocating resources and advancing patient-centered care (Chang et al., 2008; Fitzpatrick et al., 2017; Håvik et al., 2018; Isaza-Restrepo et al., 2018). In addition to supporting their technical architecture, NLP can also be applied to understand how patients perceive and utilize conversational agents. Many studies have identified specific features that tend to be rated highly by patients and appear to improve individual experiences of the healthcare system. The overwhelming majority of studies find that patients feel positively towards automated agents due to perceptions of usefulness, ease of use, appropriateness, and overall satisfaction (Milne-Ives et al., 2020). As with

feedback provided directly by patient sources, sentiment analysis regarding conversational agents can be applied to improve healthcare delivery.

2.3 Technical Approaches in Clinical NLP

The field of computational NLP has progressed significantly over the past decade. With the advent of commercially available artificial intelligence applications, adoption of NLP across healthcare, law, finance, and other data-reliant enterprise industries has increased at an unprecedented rate (Vashisth et al., 2020). While many shared analytical approaches can be utilized in NLP analyses regardless of data type, certain methodologies are more common in healthcare due to the unique characteristics of narrative clinical data. For example, vocabulary and linguistic structures used in clinical text corpora tend to be highly stable and can be captured within a consistent set of domains (Khanbhai et al., 2021). However, the underlying meaning of these vocabularies is differentiated from that of other industries due to the complex coding system implemented by health insurance payers. In addition, the use of patient/customer/user-generated data is relatively more common in clinical NLP initiatives than in other industry applications. Given the extensive regulations regarding the use of private health information (PHI), researchers have developed differentiated methods of data annotation, extraction, and analysis specific to clinical NLP that have important implications for technical approach.

2.3.1 Systematized Nomenclature

The complexities of disease pathologies and clinical practice necessitate a method for producing unambiguous medical information. First released in 2002, the Systematized

Nomenclature of Medicine-Clinical Terms (SNOMED CT) was developed by a multinational group of researchers to ensure consistent reuse of data and support semantic interoperability (Lee et al., 2014). The SNOMED system is widely considered to be the most comprehensive catalogue of global standards for health terms and is maintained by SNOMED International, a nonprofit organization with 39 member countries (Agrawal et al., 2013; Elkin et al., 2006; Khorrami et al., 2018). In the United States, SNOMED was recently implemented as a standard for meaningful use in electronic health record programs and is near-ubiquitously used to encode diagnoses and problem lists (Agrawal et al., 2013).

SNOMED CT consists of three components: concepts, descriptions, and relationships. The system is continuously expanding, with the current version containing over 350,000 concepts and accompanying descriptions and mapping over one million semantic relationships. From an ontological perspective, SNOMED is structured as a complex graph containing labeled edges and vertices, allowing for features such as compositional grammar, expression constraint queries, and the creation of postcoordinated concepts (Gaudet-Blavignac et al., 2021). Given its usefulness across clinical practice, SNOMED is the basis of many structured fields in EMRs. However, manual encoding of clinical free text into SNOMED is prohibitively time intensive and lacks scalability. To circumvent these costs, NLP techniques are used to encode clinical corpora into SNOMED vocabulary structures. An evolving body of work evaluates the utility of current concept mapping strategies across rule-based, manual, hybrid, and machine learning approaches; extant methodologies have been found to be primarily rule-based (Gaudet-Blavignac et al., 2021).

2.3.2 Public Annotated Corpora

Clinical text analyses frequently utilize forms of supervised machine learning and deep learning. To perform at or above acceptable predictive standards, these and other artificial intelligence methods require large volumes of annotated data for algorithmic training and evaluation. Additionally, the ubiquitous SNOMED CT ontology requires a significant investment in data annotation for mapping free text to over 350,000 classes. In clinical machine learning applications, the SNOMED ontology is often used as a proxy for semantic content to simplify classification between free text and structured data (Gaudet-Blavignac et al., 2021). Automated mapping of clinical free text is therefore a non-trivial task and subject to many constraints; as described in a recent systematic review, this may lead to the reuse of non-evaluated, non-open-source tools such as *Medtex* (Gaudet-Blavignac et al., 2021). Furthermore, clinical NLP efforts can be limited by the expense and/or lack of accessibility of relevant, high-quality PHI, as such data is subject to extensive patient privacy and confidentiality restrictions. Therefore, the public availability of several annotated clinical text corpora has played a key role in supporting clinical NLP studies. However, results generated from these shared datasets may be limited in their generalizability (D'Avolio et al., 2011; Gobbel et al., 2014; Stanfill et al., 2010).

The i2b2-shared task dataset was developed for the 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records and is one of the most popular datasets used in clinical NLP studies. Created as a partnership between the UCLA Informatics for Integrating Biology and the Bedside (i2b2), an NIH-funded National Center for Biomedical Computing, and the Veterans Affairs Salt Lake City

Health Care system, the dataset consists of 394 training reports, 477 test reports, and 877 unannotated de-identified reports contributed by Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center (Uzuner et al., 2011). Training and test reports were manually annotated by providers at the three institutions and shared in the public domain under data use agreements. To date, this dataset is unique in its composition of multiple report types sourced from multiple independent institutions. Multi-institutional participation is thought to support the ongoing diversification of clinical NLP datasets and mitigate concerns regarding generalizability. Additionally, the construction of the 2010 dataset extends that of previous i2b2 datasets by adding new types of data, concepts, assertions, and relationships (Hirschman et al., 2005; Jones, 1995; Uzuner, Solti, & Cadag, 2010; Uzuner, Solti, Xia, et al., 2010; Uzuner et al., 2011).

The applications of i2b2 data in clinical NLP are diverse and ongoing. Though originally created as reference standard corpora for the three challenge tasks of concept extraction, assertion classification, and relation classification, the 2010 dataset has since been used to develop algorithms for rapid medical phrase mapping, medication information extraction and processing, clinical knowledge discovery, and events and temporal expression extraction from clinician notes across specialties (Gobbel et al., 2014; Jindal & Roth, 2013; Patrick & Li, 2010; Patrick et al., 2011; Zhou et al., 2011). The corpus has also been used to develop guidelines for and validate state-of-art parsing technologies in clinical text (Fan et al., 2013; Jiang et al., 2015). More recently, researchers have also used i2b2 data to adapt word vectors to better encapsulate lexical and clinical synonymy concepts and to build an NLP pipeline that accurately extracts

physician-reported pain from clinical reports (Naseri et al., 2021; Nath et al., 2021). The dataset has also been used to examine two deep neural network architectures, finding that a recurrent neural network (RNN) model achieved state-of-the-art performance in named entity recognition compared to a convolutional neural network (CNN) model (Wu et al., 2017). Datasets from other i2b2 challenge tasks, such as the 2008 Obesity Challenge task and the 2009 Medical Extraction Challenge, have also been used to phenotype patients with certain conditions and improve clinical entity recognition in EMR systems (Gligic et al., 2020; Hong et al., 2019). An estimated 200 individuals use data from past i2b2 challenge tasks in global academic NLP research (Uzuner et al., 2011).

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II and III are comprehensive intensive care unit (ICU) datasets that are publicly accessible upon formal request. The MIMIC databases were first released in 2011 and have been developed by the Beth Israel Deaconess Medical Center since 2003 with funding from the National Institutes of Health. MIMIC II consists of 40,426 ICU admissions across 32,356 patients admitted to medical, surgical, ICU, recovery, and critical care units (*The MIMIC II Clinical Database*, 2011), while MIMIC III includes additional procedure, medication, and progress note information. In addition to narrative data comprised of clinical notes and structure fields, a subset of MIMIC II records contain high temporal resolution data including physiological waveforms, laboratory results, and minute-by-minute data recordings from bedside monitor devices (Saeed et al., 2011). The availability of detailed records in MIMIC is unmatched across publicly available annotated corpora and especially in the subfield of critical care informatics. Additionally, many MIMIC records include International Classification of Disease 9th edition (ICD-9)

codes directly used in administrative processes. MIMIC records have also been deidentified in accordance with legislative standards. While these data are available for a large and diverse patient population, they are representative of only a single tertiary care hospital in the United States.

Given their convenience, depth, and large sample size, the MIMIC datasets have been utilized across clinical NLP. Researchers have mined MIMIC text corpora for a wide variety of medical knowledge, including the alcohol and substance abuse status of patients discharged from the ICU, the effect of hypotension exposure on the risk of acute kidney injury, and differences in circadian variation between septic patients who survive and those who succumb in critical care situations (Lehman et al., 2016; Topaz et al., 2019; Wang et al., 2014). Many MIMIC-based studies describe the processes of building, testing, and refining machine learning models intended to predict clinical outcomes. For example, a recent study compared the performance of six commonly used machine learning programs in predicting acute hypotensive episodes. Only non-invasive physiological parameters were used in an ensemble prediction model to achieve good-to-excellent accuracy (G. Zhang et al., 2021). Many other studies have extracted thousands of MIMIC cases to predict ICU readmission risk, hospital mortality and other morbidity correlates, and even personalized illness scores (Awad et al., 2017; Data & Pirracchio, 2016; Ghosh et al., 2017; Han et al., 2021; Lee & Maslove, 2017; Venugopalan et al., 2017; Xue et al., 2019; X. Zhang et al., 2021). MIMIC data has also been used to identify synonymy in SNOMED and train clinical language models used in clinical information extraction pipelines (Henriksson et al., 2013; Kim et al., 2021).

The MIMIC databases have also contributed significantly to NLP-driven innovation. A particularly novel study employed MIMIC II data to extract features from electrocardiogram and photoplethysmography monitoring data. By running these features through a hybrid feature selection algorithm and multi-penalty regularized regression technique, researchers were able to implement a successful cuff-less blood pressure management system (Khan Mamun, 2021). Another study used MIMIC III to determine that positive sentiment in nursing notes written within 12 hours prior to death was inversely correlated with in-hospital mortality rates. Certain features generated from sentiment analysis, including sentiment polarity and sentiment subjectivity, may be promising indicators of in-hospital mortality that can be assessed in real-time (Gao et al., 2021). Furthermore, MIMIC has even been used to advance bioinformatics education, where a subset of data from MIMIC II is integrated as part of a case study demonstrating the concepts of propensity score analysis in EMR research (Chen & Moskowitz, 2016). In 2013, the Shared Annotated Resources (ShARe) project and the CLEF Initiative (Conference and Labs of the Evaluation Forum, previously known as Cross-Language Evaluation Forum) created a “silver standard” concept annotations corpora derived from the MIMIC-III database (Goldberger et al.; Mowery, 2013; Suominen et al., 2013). Notably, the ShARe/CLEF corpus has been used to test several concept recognition clinical NLP tools including *cTAKES*, the National Center for Biomedical Ontology (NCBO) Annotator, the Biomedical Concept Annotation System (*BeCAS*) and MetaMap/*MMTx* (Collier et al., 2015; Oellrich et al., 2015; Tang et al., 2015).

Other public clinical corpora are notable for their robust annotation and specific focuses. The Clinical E-Science Framework (CLEF) corpus, developed by the Royal

Marsden Hospital, consists of clinical narratives, radiology reports, and histopathology reports associated with over 20,000 deceased cancer patients (Roberts et al., 2007). CLEF was constructed using two semantic annotation schemes and achieves good performance with gold standard inputs (Hong et al., 2020; Angus Roberts et al., 2008). A UK-based research group successfully used CLEF data to extract important clinical relationships from text (A. Roberts et al., 2008). Another corpus, PhenoCHF, is derived from an obesity identification-related subset of i2b2 semantically annotated with phenotypic information concerning congestive heart and renal failure. PhenoCHF has been used to develop a series of algorithms that can extract phenotype information across diseases. Consistent with previous literature, investigators found that solutions based on conditional random fields and feature-rich datasets were optimal (Alnazzawi et al., 2015).

Conversely, the Temporal Histories of Your Medical Events (THYME) corpus is also based on clinical notes and pathology reports but focuses on temporal semantic relationships. Developed using 1,254 deidentified notes from the Mayo Clinic pertaining to oncology, brain cancer, and colon cancer, the THYME guidelines extend the ISO-TimeML tool to differentiate between linguistically- and inferentially-derived temporal orderings of clinical data (Bethard et al., 2015; Bethard et al., 2016; Styler et al., 2014). A systematic review found that 28 studies have used the THYME corpus to identify clustered events and other significant temporal groupings (Alfattni et al., 2020; Savova et al., 2017). Researchers found that associative attention networks can be used to reduce the number of entity pairings and improve performance in models of health signals (Zhao et al., 2019). THYME has also been used to validate the first open-source temporal relation discovery program, able to programmatically infer temporal relationships

between clinical events and time expressions using a multilayered modeling strategy (Lin et al., 2016).

Importantly, few clinical corpora have been made available in non-English languages. The MedNLPDoc corpus, comprised of fewer than 1000 medical examination notes and annotation only at the document level, is published in Japanese (Aramaki et al., 2016). The Turku Clinical TreeBank and PropBank are available in Finnish and were developed using physician and nursing notes consisting of 2800 sentences across only eight cardiology ICU patients (Ginter et al., 2010; Haverinen et al., 2013; Laippala et al., 2014). Similarly, the Stockholm EPR PHI Pseudo Corpus consists of 100 pseudonymized patients records and is published in Swedish (Alfalahi et al., 2012). There is also an ongoing effort to develop clinical corpora in Chinese, where researchers have applied a novel iterative annotation method to annotate nearly 1000 clinical documents (He et al., 2017). These are other multilanguage NLP efforts have identified syntactical challenges specific to individual language structures (He et al., 2017; Xu et al., 2014; Zhang et al., 2016).

Several ancillary resources can help support multilanguage clinical NLP efforts. Most well-known is the Unified Medical Language System (UMLS), an aggregate of over 100 biomedical terminologies and ontologies that serves as an interoperability standard across biomedical vocabularies. The current UMLS Metathesaurus release includes over 9.1 million terms in English, 1.3 million terms in Spanish, and less than 5% total terms in all other languages including Japanese, Dutch, and French (Bodenreider, 2004; Névél et al., 2018). Additionally, the Clinical Trials for Evidence-Based Medicine in Spanish (CT-EBM-SP) corpus is a biomedical resource that spans 1200 clinical trial

texts annotated with various entities from UMLS groups (Campillos-Llanos et al., 2021). While not directly applicable to clinical narrative text, the CT-EBM-SP employs related machine learning methods to develop a pipeline for processing evidence-based medicine articles. Similarly, Chia is a newly developed annotated corpus of eligibility criteria extracted from 1,000 Phase IV interventional trials registered on ClinicalTrials.gov, comprising 41,487 entities with 25,017 relationship types (Kury et al., 2020). Finally, the multilingual clinical entity recognition tool *Cimind* was developed and validated using translated versions of the 2016 and 2017 editions of the CLEF database (Cabot et al., 2019).

Accessible annotated corpora serve as a valuable tool in the advancement of clinical NLP. However, their relative scarcity, expensive, and lack of multilingual support can pose significant barriers to research (Jiang et al., 2015). Recently, a Norwegian research team has attempted to mitigate this burden through the systematic generation and development of annotation guidelines for synthetic clinical corpora. Double loops of corpus production and guideline development were implemented to generate a corpus of 477 sentences associated with over 6000 entity labels. A linear classifier trained to extract family history information regarding genetic cardiac disease achieved good precision and recall on both synthetic data and a genuine corpus of 2,276 outpatient cardiologist notes (Brekke et al., 2021). Another study which implemented recurrent neural network, generative adversarial network, generate pre-training, and conditional transformer language models to generate 500 “history of present illness” clinical notes. Similar to Brekke et al., named entity recognition models achieved good performance on synthetic data compared to a subset of i2b2 data (Li et al., 2021). Recent

evidence also suggests that machine-generated medical phrases can effectively substitute clinical features in free text (Abdollahi et al., 2021). Synthetic data generation should therefore be considered as a promising data augmentation strategy in clinical NLP.

2.3.3 Named Entity Recognition

In NLP, the mapping of language expressions to a systematized nomenclature is referred to as clinical named entity recognition. Entity recognition is critical to critical NLP and often the first step in more complex analyses (Sun et al., 2018). To meet the needs of clinical researchers, an uncoordinated set of NLP mapping technologies has been developed with significant variation in methods, measurement, and output terminologies. Indeed, a recent literature review identified close to a dozen software-based tools used in clinical text mining studies (Gaudet-Blavignac et al., 2021). However, many of these algorithms have not been validated externally or at all (Kersloot et al., 2020). The most popular tool is *Medtex*, which was developed by the Australian eHealth research center as a pipeline layer built on other tools (Nguyen et al., 2009). *Medtex* utilizes the General Architecture for Text Engineering (GATE) platform to map clinical terms, abbreviations, acronyms, and clinical short-hand onto SNOMED concepts (Cunningham, 2002; Nguyen et al., 2009). Following term retrieval, underlying part-of-speech parsing algorithms and variant generation techniques from the open-sourced MetaMap (MMTx) and NegEx programs are applied to annotate text with SNOMED concepts and negation symbols (Aronson, 2001; Chapman et al., 2001). Furthermore, *Medtex* implements a messaging architecture of parallel message consumers that can receive messages from one queue simultaneously. This high-throughput capability makes *Medtex* especially popular in the analysis of electronic pathology feeds (Gaudet-

Blavignac et al., 2021; Monson-Haefel & Chappell, 2002; Nguyen et al., 2015). No evaluation of *Medtex* mapping has been published, though excellent sensitivity, specificity, and positive predictive value has been observed in cancer identification from pathology reports and extraction of tumor prognostic markers (Khor et al., 2019; Nguyen et al., 2016).

Another commonly used entity recognition tool is the *Mayo Clinic Vocabulary Server (MCVS)*, also referred to as the Multi-threaded Clinical Vocabulary Server. The *MCVS* uses a combination of platform- and language-neutral server tools, which create an extensive set of indices to underlying SNOMED terminology, and user-facing modular GUI tools, which implement automated compositional expression and automated term dissection methods to navigate the terminology and build conditional coded expressions (Elkin et al., 2006). Considering only exact conceptual matches, an early study of the *MCVS* demonstrated a sensitivity of 99.7% and specificity of 97.9% across the 4,996 most common problem statements in the Mayo Master Sheet Index (Elkin et al., 2006). Though not as widely utilized as the *MCVS*, *LingoEngine* has been cited as its commercially available iteration (Brown et al., 2008).

A third system called the *Medical Text Extraction Reasoning and Mapping System (MTERMS)* uses shallow and deep parsers to directly mine free text for SNOMED terminology matches and to propose relationships between terminologies (Zhou et al., 2011). Each implementation of *MTERMS* is customized and requires the addition of specific ad hoc dictionaries, as well as other syntactical parameters. *MTERMS* has been used to mine clinical information regarding medications, allergens, allergic reactions, and familial relationships, with moderate to excellent measures of success. Specifically, two

studies of allergic reactions and familial relationships demonstrated precision of 84.4% and 100.0% and recall of 91.0-97.4%, respectively (Goss et al., 2014; Plasek et al., 2016; Zhou et al., 2014). Generally, *MTERMS* is recommended for more restricted, single subject use.

Other clinical ontology tools used to flag relevant text in clinical NLP studies include the *Open Biomedical Annotator (OBA)*, a web service that annotates free text within a variety of ontologies, the *Medical Language Extraction and Encoding System*, which transforms clinical data into controlled vocabularies, the *clinical Text Analysis and Knowledge Extraction System (cTAKES)*, which includes a dictionary lookup component for mapping free text and a named entity recognition component, *Snapper*, which allows users to input free text data to map from external terminologies to SNOMED CT, and the *Rapid Text Annotation Tool*, which is a token order-specific naive Bayes-based machine learning system that predicts associations between phrases and concepts (Gaudet-Blavignac et al., 2021). Of these, the *Medical Language Extraction and Encoding System* and the *Rapid Text Annotation Tool* have demonstrated precision of 89% and 95% and recall of 83% and 85%, respectively (Friedman et al., 2004; Uzuner et al., 2011). Additionally, *cTAKES* was recently evaluated in a phenotyping efficacy study along with the *MetaMapLite*, *MedLEE*, and *ClinPhen* NLP systems, all of which were found to demonstrate improved performance when ensembled with at least one other system (C. Liu et al., 2019). These and other recent studies suggest that available NLP tools can be ensembled to achieve better recognition of generic as well as patient-specific phenotype concepts (Kreimeyer et al., 2017; Kuo et al., 2016).

Though *Medtex*, *MCVS*, and *MTERMS* are the most frequently used programs for automated mapping of free text to SNOMED CT concepts, they are not publicly available. Conversely, *MMTx* – one of the foundational packages used in *Medtex* – is open-source, as is *cTAKES*. *OBA* is not open-source, but its functionality is publicly accessible through a web-based interface. Importantly, *MMTx*, *cTAKES*, and *OBA* were not specifically designed to map the SNOMED ontology, though they can be used for that purpose. They also lack support for features such as multiple languages and postcoordination, which can be critical in the context of clinical data mining. Given this landscape, there is currently not a publicly available “out of the box” solution for mapping free text to SNOMED CT concepts. Of those that are widely used, few are accessible at the implementation level. This landscape of mapping tools and the ubiquity of SNOMED CT should be considered when designing clinical NLP studies.

2.3.4 Medical Concept Normalization

Mapping clinical free text to systematized nomenclature for entity recognition, knowledge extraction, and disease phenotyping – as described previously – is generally the first order analysis in clinical NLP. However, the use of these models can be limited by the high variability of linguistic expressions that convey the same concept. For example, “heart attack” is more commonly referred to by medical professionals as “myocardial infarction” and abbreviated in clinician notes as “MI,” though all three terms map to the same identifier in the UMLS. Furthermore, while certain medical abbreviations are commonly used, they are often significantly different between geographies, specialties, clinical roles, and even institutions. These differences can be difficult even for external clinicians to decipher. Additionally, the same abbreviations can

have different meanings depending on clinical context. In clinical trial and other experimentation-focused settings, words like “transport” can refer to concepts as divergent as cellular function and human movement (Henry et al., 2020). Therefore, even “data-hungry” deep learning models may achieve high accuracy on training datasets but fail to generalize to other medical records (Lee et al., 2020; Nath et al., 2021). This can delay research efforts and prevent or limit real-world implementation into clinical practice.

Medical entity normalization is an emerging subfield of clinical NLP that seeks to address these issues. Often performed as a second-order analysis task, the goal of normalization is to link all flagged medical entities to a standardized concept vocabulary or schema. Entity normalization algorithms consist of an (1) ontology processor, which widens or narrows the scope of the reference ontology, (2) string matcher, which compares flagged expressions (entities) against the reference ontology, and (3) filter that removes false positive flags using rules or other models (Wang & Matthews, 2008). Generally, the ontology processor relies on manually constructed sources of in-domain knowledge to develop the linking processes used in NLP model training. The string matcher, arguably the most important component in these systems, can be created using a variety of approaches including exact or partial matching, rule-based heuristics, or machine learning models. Emerging studies incorporate all three approaches types with relative success (Suryawati & Widiantoro, 2017). Finally, filter models can be embedded in the string matcher layer or integrated from extant language filtering libraries using across NLP. In practice, it follows that multi-step clinical NLP analyses operate by first running named entity recognition tasks to flag expressions of interest, followed by

running named entity normalization tasks to retrieve a normalized set of medical concepts for the flagged expressions (Henry et al., 2020). In conjunction, these two processes are thought to improve the exchange of information across distinct sets of EMRs and could be instrumental to effective communication between medical contexts.

Traditional approaches to entity normalization include the application of string matching techniques, which attempt to find exact or partial (also referred to as “soft” or “fuzzy”) language expression matches (Kalyan & Sangeetha, 2021). Though more simplistic than partial matching, exact string matching has been successful in a variety of clinical cases. For example, a recent study applied exact matching techniques to a manually annotated corpus of 1345 clinical trials from ClinicalTrials.gov and 9422 abstracts from PubMed in the detection of clinical screening and diagnostic instruments for posttraumatic stress disorders. Resulting trends analyses of instrument prevalence were consistent with clinical experience and have been made freely available (Zong et al., 2021). In contrast, partial matching approaches require the implementation of similarity measures often considered under the umbrella of string-similarity matching. Such measures can be hardcoded manually by domain experts, or developed based on the characteristics of flagged entities or that of the corpus itself (Markić et al., 2020). Early partial matching experiments on large-scale gene-protein dictionaries found that logistic regression-based similarity measures outperformed others in dictionary look-up programs (Tsuruoka et al., 2007). Subsequent experiments have used various metrics including edit-distance, fast heuristic comparators, and token-based distance to develop effective similarity measures (Cohen et al., 2003). Since then, many tools and algorithms have been released that demonstrate high efficiency in matching genomic and cellular-level

concepts (Alser et al., 2020; Karcioglu & Bulut, 2021a, 2021b; Nicolas, 2012). Though few clinical notes reference basic science concepts such as these, matching algorithms based on these methods are used in the analysis of drug names, laboratory results, and histopathology and radiology report expressions in clinical text corpora (Bonert et al., 2021; Grundmeier et al., 2016; Mughli et al., 2021).

String matching approaches are limited by high computational costs. Though various methods exist for reducing computing power, matching techniques usually require the incremental parsing of enormously large corpora as well as the iterative execution of lookup commands. As a result, the use of rule-based heuristic models has become widespread within medical entity normalization due to their smaller computational footprint and domain flexibility. In contrast to matching techniques, rule-based algorithms alter the lexical forms of entities of interest according to sequential rules designed to convert all possible expressions of a particular concept to the universal lexicon (Wang & Matthews, 2008). The modified first entry can now be matched to the underlying medical concept dictionary and return a standardized concept. Importantly, the heuristic rules-based design allows dictionary lookup to be performed in constant time, regardless of corpus size (Tsuruoka et al., 2008). A recent scoping review of NLP in clinical pathology found that rule-based methods were the most commonly normalization mechanism before 2003, ostensibly due to constraints on computing power relative to current studies (Burger et al., 2016). This approach achieves success in the extensive utilization of expert knowledge, as demonstrated by early diagnostic histopathology and cytopathology systems, genotypic drug resistance interpretation systems, and high-performing teams within the BioCreAtIvE 2 Gene Normalization Task

(Bartels & Hiessl, 1989; Hakenberg et al., 2007; Schmidt et al., 2002). Interestingly, heuristic approaches were recently used to develop a novel algorithmic plugin that identifies clinical events matching a user-specified temporal order (Shrestha et al., 2021).

Both string matching and heuristic rule-based techniques are subject to significant limitations. Matching techniques can also fail when there is no overlap between a certain language expression and its corresponding medical concept (e.g., “pins and needles” mapping to paresthesia). Rule-based techniques are limited by bottlenecks in acquiring or refining the domain knowledge required to construct useful rules limits, as well as to scale data normalization approaches. As a result, the use of probabilistic machine learning techniques has become widespread in clinical NLP entity normalization (Burger et al., 2016; Mehta & Pandit, 2018). The first machine learning-based approach was described in 2013 and applied pairwise-learning-to-rank, which ranks entities based on differences in relevance, to disease name normalization (Leaman et al., 2013). These and subsequent machine learning methods such as low-rank approximation and machine-phrase translation were found to outperform baseline methods when applied to clinical text and medically-relevant social media posts (Leaman & Lu, 2014; Limsopatham & Collier, 2015). Simultaneously, efforts to scale-up NLP pipelines to accommodate increasingly larger and more complex EMR corpora were developed. In some cases, the application of foundational computing techniques such as replication via multithreading, intra-annotator threading for further decomposed tasks, and the creation of remote annotator services achieved improvements on the order of twelve-fold processing speeds (Divita et al., 2015).

Recently, deep learning-based systems have risen to the forefront of medical entity normalization for their powerful semantic analysis capabilities (Wu et al., 2019). By taking a machine learning approach and conceptualizing normalization tasks as multi-class text classification problems, investigators have developed algorithms that can be trained to automatically process complex linguistic features such as grammatical structure, inter-item lexical relationships, and collocations. The inclusion of a varied and complex set of features allows for accurate machine predictions of textual meaning and conceptual relationships. In practice, these systems have improved on existing architectures by implementing ensembles of CNN, RNN, and linear model layers with input word embeddings trained on clinical domain-specific sources (Vashishth et al., 2021). Investigators have achieved promising results on benchmark clinical text datasets, data from online health communities and social media platforms, and drug reviews (Belousov et al., 2017; Lee et al., 2017; Tutubalina et al., 2018). Importantly, variation exists within approaches to word embeddings. These can be traditionally learned using shallow neural network models such as Word2Vec, though lower-quality word vectors are produced due to a more limited capacity to code information as vector representations. Conversely, embeddings can also be trained downstream from scratch. Downstream-trained models require more training examples to improve performance and are infeasible with small datasets (Kalyan & Sangeetha, 2021).

Currently, there are several pre-trained deep learning-based language representation models popularly described in the literature. These include Embeddings from Language Models (ELMo), Universal Language Model Fine-tuning (ULMFiT), General Pre-Training (GPT), and Bidirectional Encoder Representations from

Transformers (BERT), the last of which was developed by Google AI Research (Devlin et al., 2018; Howard & Ruder, 2018; Peters et al., 2018; Radford et al., 2018). Of these, the BERT algorithm is most frequently implemented in medical concept mapping studies. BERT and is pre-trained on large volumes of unlabeled text data, iterating through stacked transformer encoders that allow for large-scale learning and the creation of bidirectional language representations (Kalyan & Sangeetha, 2021). BERT serves as the foundation for additional task-specific layers, which can then be further refined using task-specific annotated datasets of all sizes. By expanding the scope of inputs that can be used in the model refinement process, BERT has achieved superior performance compared to CNN or RNN-based models trained from scratch in the clinical narrative, biomedical, and consumer health NLP domains (Alsentzer et al., 2019; Ji et al., 2020; Kalyan & Sangeetha, 2021; Kilicoglu et al., 2018; Miftahutdinov et al., 2021; Omoregbe et al., 2020; Rasmy et al., 2021; Wei et al., 2019). A recent study of 16,526 symptoms within 1108 discharge summaries found that BERT's performance was comparable to those of human annotators and slightly improved compared to RNN-based models (Steinkamp et al., 2020). Encouragingly, BERT-based models were able to extract and high-level symptoms, acronyms, typographical errors, and grouping statements in ways that generalized effectively to other clinical note corpora (Steinkamp et al., 2020). To date, BERT has been cited in nearly thirty thousand published studies and its success has led to the development of other open-source, transformer-based models that demonstrate good-to-excellent performance in clinical NLP tasks (Alsentzer et al., 2019; X. Yang et al., 2020).

Medical entity normalization is an important clinical NLP tool that is rapidly advancing across a variety of technologies. Development is now concentrated in the refinement of machine learning models applied to large EMR text corpora and gold standard annotated corpora as described previously. Many studies find that CNN, RNN, and stacked encoder models perform superiorly to other models such as support vector machine and random forest, though there is ongoing debate. More commonly, machine learning approaches are ensembled with rule-based and heuristic models to achieve state-of-the-art performance. For example, one of several recent studies found that a rule-based clinical text classification paradigm using weak supervision and deep language representation models achieved near-perfect classification of smoking status and hip fractures at Mayo Clinic (Wang et al., 2019). Similar to those described previously for named entity recognition, publicly available annotated corpora are being developed for medical concept normalization (Luo et al., 2019). Recent evidence also suggests that synthetic datasets can successfully be used to train deep language representations of medical concepts (Pattisapu et al., 2020). Emerging challenges in medical concept normalization include addressing string ambiguity in clinical text, developing robust model evaluation measures, and developing coverage for multilingual terminologies that can generalize across datasets (Newman-Griffis et al., 2021; Wajsbürt et al., 2021).

Chapter III

Intersections of Clinical NLP and COVID-19

The emergence of the SARS-CoV-2 virus and the ongoing global pandemic has had a profound influence on clinical artificial intelligence research (Adamidi et al., 2021; Grabar & Grouin, 2021). Shifts from in-person visits to telehealth sessions have had lasting impacts, many of which include increased utilization of remote healthcare services. These changes are particularly relevant to clinical NLP, as they have catalyzed the rapid implementation of infrastructure supporting electronic patient-provider communication and video visits. Recent studies report a growing uptick in the number of electronic health messages sent across SMS, email, and EMR platforms in a variety of specialties (Boyajian et al., 2021; Musaoğlu et al., 2021). Over the past two years, many individuals have also turned to social media and online communities to discuss COVID-19 developments. NLP-enabled analysis of electronic health messages, social media posts, and other text-based media has and continues to provide valuable insights into public sentiment. Given the topic of the current work, the following section presents a brief discussion of the relevant impacts that COVID-19 has had on clinical NLP (Robles Hernandez et al., 2021).

3.1 User-Generated Corpora

Given the heightened importance of digital communication during the pandemic lockdowns, as well as the ubiquity of social media platforms, there has been significant interest in utilizing user-generated data in NLP analyses of COVID-19. In May 2021, a high-profile digital lab released a variant of the BERT deep learning model called

COVID-Twitter-BERT (CT-BERT). This transformer-based model was pretrained on over 160 million tweets about the coronavirus collected from January to April 2020, often considered to be the “peak” period of infections in the United States. CT-BERT showed an improvement of 10-30% over the BERT-Large model across five major classification datasets and is further optimized for use on COVID-19 social media content (Müller et al., 2020). To date, CT-BERT has been used as the base algorithm for several models that demonstrate efficacy in measuring the fact-checking potential of tweets published by the CLEF 2020 CheckThat! challenge. Augmentations of CT-BERT have improved baseline performance and demonstrate evidence of reliable generalizing to the classification and semantic analysis of other types of user-generated posts (Alkhalifa et al., 2020; B. Chen et al., 2021; Glazkova et al., 2021; Hettiarachchi & Ranasinghe, 2020; Møller et al., 2020; Tran et al., 2020). Additionally, machine learning methods were used to automatically annotate a biomedically oriented corpus of over 120 million tweets regarding the characterization, symptomology, and public experience of COVID-19

Potentially biased or misleading information regarding the coronavirus was rapidly disseminated through networked media sites and algorithmized social media platforms, a phenomenon also referred to as “the global COVID-19 infodemic” (Islam et al., 2020). In response, user-generated datasets have also been mined to help identify, track, and contain the spread of misinformation. Due to novel language expressions and the rapid exchange of digital information, extant training data were not as effective in the identification of COVID-specific misinformation. In early 2021, a Twitter corpus called COVIDLies was released earlier this year to help address this issue. Comprising 6591 expert-annotated tweets across 62 unique entities of COVID-19-based misinformation,

this corpus was intended for the detection of misinformation regarding COVID-19. Alongside the corpus itself, investigators also described an evaluation of existing NLP systems on COVIDLies, including behavior tests of the best-performing models (Hossain, 2021). Applications of augmented algorithms to this dataset are ongoing. The 2021 Workshop on NLP for Internet Freedom, supported by the National Science Foundation, recently launched two multilingual shared tasks for fighting the COVID-19 infodemic by predicting worthiness of COVID-related tweets and predicting which tweets would be censored, respectively (NLPIF, 2021). BERT- and transformer-based systems have achieved good performance, (Henia et al., 2021; Tziafas et al., 2021; Uyangodage et al., 2021).

Given its success in other predictive NLP applications, many incrementally augmented systems have ensembled variants of the foundational BERT deep learning model to add COVID and Twitter domain-targeted layers. Advanced word embedding techniques, the creation of long short-term memory networks, and the addition of other classification models allow for the improved identification of contextual relationships between rare and unseen words, as often appear in text corpora relevant to COVID-19 (Alsayat, 2021). Such adaptations appear to significantly improve the performance of CT-BERT text classification tasks in sentiment analysis of COVID-related tweets (Lin & Moh, 2021). Importantly, expressed sentiment is thought to be an important precursor of human behavior and could potentially inform public health responses (Ballinari & Behrendt, 2021; Mangono et al., 2021; Mehta et al., 2021). A recent linguistic study of health platform rating data associated with 5521 physicians found that negative sentiment was a significant factor in patient decision-making (Shah et al., 2021). Furthermore, a

multilabel version of CT-BERT achieved incremental improvements in predicting COVID-related events from Twitter data, including positive and negative COVID test results, COVID-related deaths, and developments in COVID-related medical treatment and prevention (Xiangyu Yang et al., 2020). NLP-based analysis of text messages from a tele-mental health platform may also provide insight into pandemic-related psychological responses and psychiatric symptoms (Hull et al., 2021).

Other NLP deep learning techniques have been used to conduct annotation analyses on a variety of datasets. A CNN and RNN-based system, trained on over 300,000 posts on Facebook and Twitter in the UK and US, found that public sentiment concerning the safety, development, and effectiveness of COVID-19 vaccine correlated broadly with nationwide surveys in both countries (Hussain et al., 2021). Additionally, a named entity recognition algorithm trained on tweets geotagged to the UK was able to extract temporal trends of COVID cases, identify popular locations for the use of face-coverings, and uncover attitudes responses to face-covering requirements, vaccines, and national testing and contact tracing programs (Alamoodi et al., 2021; Cheng et al., 2021). A web-based system named EVIDENCEMINER was recently developed to allow completely automatic annotation based on background COVID-19 corpora given a user-generated query of a NLP statement (X. Wang et al., 2020). Intelligent frameworks that implement sentiment analysis have also been developed to predict the socioeconomic impacts of and corresponding public sentiments around COVID-19 (Nasir et al., 2021). As recent events demonstrate, understanding community-specific responses to aspects of the coronavirus response can be critical to the development of culturally competent public health messaging (Budhwani et al., 2021). Furthermore, the preponderance of

COVID-related literature has catalyzed the development of novel text mining methods intended to identify new concepts related to COVID-19 and reduce repeated effort within clinical NLP (Colic et al., 2020; Wang & Lo, 2020).

3.2 Deep Learning for Clinical Knowledge

From January through December 2020, over 101 artificial intelligence studies developed or validated predictive models for screening, diagnosis, and prognostication of COVID-19 (Adamidi et al., 2021). While many algorithms were based on established deep learning approaches, other studies implemented logistic regression, score, machine learning, and XGBoost models to diagnose COVID-19 in hospital-admitted adults. Notably, the best performing of these achieved excellent an all-area under the ROC curve measure of 0.94 (Locquet et al., 2021). Similarly, linear and non-linear classification methods achieved 82-83% accuracy in predicting the level of care needed for 2566 consecutive COVID-19 patients at five health systems in Massachusetts. Clinical features such as opacities on chest imaging, age, male gender, clinical presentation at admission, and diabetes were some of the most highly associated features that predicted for ICU admission and need for ventilation (Hao et al., 2020). Several of these results were replicated in another study of 10,504 patients conducted in Spain (Izquierdo et al., 2020). A variety of other physiological variables and laboratory results from electronic healthcare records have also been used with generic as well as deep machine learning algorithms to accurately model patient severity (Adamidi et al., 2021; Shorten et al., 2021; Tayarani N, 2021). Interestingly, an active learning NLP system was proposed to efficiently filter relevant from non-relevant scientific articles regarding COVID-19 to aid clinical research (Verspoor, 2021). Neural network algorithms have also been used to

develop semi-automatic screening tools that can be applied to sort through COVID-19 publications (Catelli et al., 2020).

As with existing efforts in general clinical NLP, several public corpora have been made available to advance coronavirus-related NLP research. In May 2021, COVID-19 Annotated Clinical Text (CACT) Corpus was developed and made publicly available. The CACT consists of 1,472 clinical notes from the University of Washington and contains comprehensive annotations detailing COVID-19 diagnoses, testing, and medical presentation. Investigators also released a pretrained neural network developed to extraction diagnoses and symptomology from free clinical text (Lybarger, Ostendorf, Thompson, et al., 2021). A research group at Columbia University released another publicly available annotated corpus in June of the same year. Their database consists of 700 COVID-19 trials and trials' corresponding eligibility criteria (COVIC); developers utilized a hierarchical annotation schema based on the OMOP Common Data Model, which accommodates four levels of granularity for the study cohort, eligibility criteria, named entity, and standard concept. COVIC attempts to centralize unstructured clinical trial eligibility criteria to enable automatic eligibility screening for participation in COVID-19 trials. Within COVIC, over 17,000 entities are mapped to standardized medical concepts and 1009 entities are normalized to computable representations. Future uses of COVIC include semantic indexing for COVID-19 trial analytics and benchmarking of machine learning-based feature extraction (Sun et al., 2021). These and other NLP studies continue employ predictive modeling approaches to better understand the symptomology, clinical course, and medical and social determinants of COVID-19.

Chapter IV

Aspect-Based Sentiment Analysis of Health System Messages

4.1 Approach

This chapter discusses the approach taken to conduct a novel aspect-based sentiment analysis of a large corpus of secure patient portal messages containing COVID-19 content. Prior to model training, a combination of empirical NLP packages and a set of manually developed rule-based heuristics were used for text preprocessing. Automated annotation was performed using a parsimonious lexicon and rule-based sentiment model specifically attuned to sentiments expressed in user-generated social media data (Hutto & Gilbert, 2014). Predictive sentiment analysis models were then trained using a domain-targeted state-of-the-art bidirectional encoder-based deep learning transfer algorithm, CT-BERT. Since CT-BERT is pretrained on Twitter posts containing COVID-19 content, supervised fine-tuning on an original large corpus of patient-originated messages was performed to add a layer of domain specificity for secure health system message (Dang et al., 2020; Lee et al., 2020; Li et al., 2019; Malla & Alphonse, 2021; Müller et al., 2020; Rasmy et al., 2021; W. Yu et al., 2021). Finally, an auxiliary sentence approach was used to expand the input corpora by treating unique sentiments as aspects and assigning polarity to each paired sentence entity (Adi et al., 2016; Jindian et al., 2020). In brief, this allowed for the conversion of single-sentence classification to pair-sentence classification (Lin & Moh, 2021; Sun, Huang, et al., 2019; Yu et al., 2019). This work builds on state-of-the-art deep learning models applied to social media data and offers a targeted, aspect-based approach to patient-originated secure health messages in the era of COVID-19.

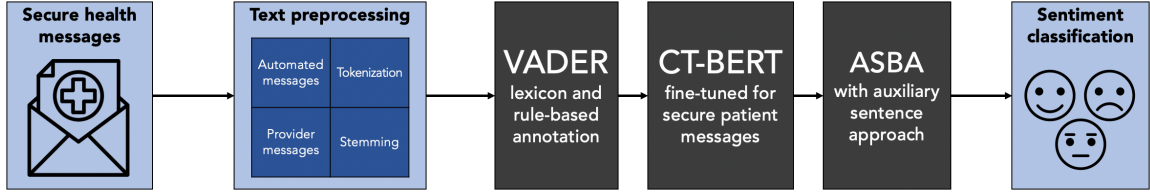


Figure 1. Proposed Fine-Tuned Deep Transfer Learning Approach to Aspect-Based Sentiment Classification.

High level overview of our proposed fine-tuned deep transfer learning approach to aspect-based sentiment classification of COVID-19 content in secure patient-originated health system messages. The above system captures additional domain specificity for secure health system messages and employs an auxiliary sentence approach to improve sentiment classification accuracy. We also apply a parsimonious rule-based heuristics model for processing user-generated data for sentiment annotation of a novel corpus.

4.2 Text Preprocessing

The present work utilized a large corpus of secure health system messages from a major tertiary health system. Raw data consisted of text files containing all patient-originated, provider-originated, and system-originated (i.e., automated appointment reminders, scheduling information, medication notifications) messages exchanged through the associated provider-patient portal from early 2020 to mid-2021. Messages were collected across all available clinical departments and specialties during this period. Prior to machine analysis, a combination of automated preprocessing and annotation packages were used to collate and convert text data to a standardized format. The goal of preprocessing was to establish a high quality of data for use in subsequent machine learning and inference-based models, through the accurate identification of fundamental linguistic units (i.e., characters, words, and sentences) appropriate for passage to further processing stages. The goal of annotation was to add specific metadata appropriate for an aspect-based sentiment analysis task of COVID-19 data. After preprocessing, a set of

rule-based heuristics were manually developed for additional data labeling and to filter messages into subsets of interest. Only patient-originated messages determined to include COVID-19 content were included in analysis.

4.2.1 Empirical Techniques

The corpus used in this work consisted of generally well-written messages in standard American English and did not contain microtext specific to social media posts, emails, or web pages (Xue et al., 2011). We therefore applied a standardized set of empirical text preprocessing methods widely used in NLP studies (Kannan et al., 2014; Oyeboode et al., 2021; Rustam et al., 2021). First, a tokenization pipeline was utilized to normalize text to a more easily machine-readable state. Non-text related components were removed, including special number and date formats, hyperlinks, unescape HTML characters, HTML tags, and repeated characters (e.g., “loooong time” becomes “long time”). Unicode characters, including emoticons, image attachments/GIFs, and other special characters such as at-symbols were removed. Prepositions, articles, pronouns, question words and other common words that are historically of low value in sentiment mining were also removed (Kannan et al., 2014; Smelyakov et al., 2020). Most punctuation marks were removed, though those with semantic value were preserved. These included sentence markers such as periods and exclamation points, as well as commas (Oyeboode et al., 2021). Contractions were expanded, and slang words converted to their English word equivalents using online sentiment analysis dictionaries available for public academic use under the GNU General Public License (Oyeboode et al., 2021; SentiStrength, 2014, 2018). Duplicated text, caused by systematized inclusion of prior message history with each consecutive message, was removed.

Next, the standard tokenizing pipeline was internally modified to preserve several word types that are normally removed in NLP applications. During initial preprocessing, we found that these components had semantic value in the specific context of COVID-19 content. Numeric words unrelated to timekeeping (e.g., “three o’ clock”) were preserved, as they often had value in communicating the extent of concerns related COVID-19. For example, patients used numeric words to describe their number of missed health maintenance appointments, the number of times they had called a provider’s office, the duration of COVID-related symptoms, their number of potential COVID-19 exposures, and their vaccine dose number. We also preserved common domain-specific abbreviations and acronyms, including clinical shorthand, medical jargon, and degree signifiers (e.g., “MD” and “NP”). These components provided useful information correlated to patient message types, logistical timelines, and symptomology. This was also consistent with other clinical NLP studies, where only a limited amount of preprocessing is performed to preserve medical terminology. Of note, clinical normalizers have been applied to handle relevant challenges related to ad-hoc formatting, medical jargon, and ambiguous acronyms (Leaman et al., 2015). Text was not lowercased to preserve characteristics of sentiment emphasis and expression.

Finally, stop words were removed to minimize noise in future analyses. Stop words are common words or phrases such as “hey,” “please,” and “thanks” that do not convey semantic information (Kulkarni & Shivananda, 2021; Rajman & Besançon, 1998). A compiled library of stop words was imported from various open-source software packages and biomedical resources including PubMed (easyPubMed, 2021; Nothman et al., 2018). In combination, these lists captured general English stop words

and domain (COVID)-specific stop words as described in other works (Älgå et al., 2020). Previous evaluations of publicly available stop word lists found that individual libraries can be lacking in completeness, domain specificity, and compatibility with specific tokenizers, which we attempted to mitigate through our multi-source approach (Nothman et al., 2018; Sarica & Luo, 2021). Per previous recommendations, we also dynamically adapted imported stop lists where necessary for patient-originated message and COVID-19 content contexts (Nothman et al., 2018).

Additional simple text preprocessing methods including stemming, lemmatizing, and multiword grouping are also widely used. Generally, their aim is to conflate variant words and/or word forms into a common representation. These procedures are more common in information retrieval studies, with the common assumption that queries including stems of certain words indicate a user’s interest in documents containing the various word forms of that stem (Kannan et al., 2014). Though an intriguing consideration for future studies, none of these techniques were applied to the current analysis given its focus on aspect-based sentiment inference. Furthermore, an extensive benchmarking study of neural text classifier performance in text categorization and sentiment analysis studies, which compared tokenizing, lemmatizing, lowercasing, and multiword grouping preprocessing decisions, found that simple tokenization of input text was adequate (Camacho-Collados & Pilehvar, 2017).

4.2.2 Rule-Based Heuristics

A rule-based heuristic approach was implemented to filter and subset patient-originated messages. Given the specificity of our dataset to hospital-specific

departments/divisions and healthcare operations in the mid-Atlantic and immediate Philadelphia region, as well as the idiosyncrasies of the institution itself, these rules were developed through manual exploration of the dataset. In many cases, filtering logic was developed through the manipulation and recategorization of existing metadata labels. For example, duplicative and payer-specific labels corresponding to different subspecialties were collapsed into clinically relevant categories. Demographic labels were recorded and applied across a variety of domains based on provided census categories, as well as institution- and state-specific information such as payer groups, billing codes, and social services. Individual messages within the same exchange chain were categorized as either sent by the provider, patient, or automated system based on grouped metadata and distinguishing key phrases. Through the utilization and modification of existing metadata, an initial taxonomy of communication types was created (i.e., main categories of medical events, information-seeking, medications, and so on). Additionally, ancillary data files containing problem lists, active medications, and active diagnoses were cleaned based on matching records in preprocessed message files and matched on unique identifiers. These approaches are aligned with previously described manual annotation of patient portal messages (Cronin et al., 2017; Elmessiry et al., 2017).

Notably, a keyword-based approach was used to filter messages containing COVID-19 content. Messages containing variations of “COVID-19,” “COVID,” “coronavirus,” “SARS-CoV-2,” and “pandemic” were labeled as relevant to COVID-19. These keywords were developed based on manual exploration of the dataset and are consistent with previous COVID-19 topic modelling and social media mining studies (Bahja & Safdar, 2020; Boon-Itt & Skunkan, 2020; Bose et al., 2021; Cronin et al., 2017;

Malla & Alphonse, 2021; Oyeboode et al., 2021; Shamrat et al., 2021; Wang et al., 2021).

A similar approach was used to sublabel messages relevant to the COVID-19 vaccine (Kwok et al., 2021). Additionally, message dates were used to segment messages into early, middle, and late-stage COVID with respect to relevant developments in the United States. Only messages in English were considered during analysis.

4.3 Parsimonious Lexicon and Rule-Based Annotation

The Valence Award Dictionary and sEntiment Reasoner (VADER) is an open-source lexicon and rule-based sentiment analysis tool specifically developed for social media data. When applied to input corpora, VADER first constructs and empirically validates a gold-standard list of lexical features attuned to microblog-like contexts. Associated sentiment intensity measures are also recorded, then combined with five rule-based heuristics that encapsulate syntactical conventions for expressing sentiment intensity (Hutto & Gilbert, 2014). These rules are specifically designed to reflect generalizable, human-centered markers of sentiment in user-authored posts including the use of exclamation marks, ALL-CAPS, degree modifiers, and contrastive conjunctions signaling a shift in polarity. Upon its release in 2014, VADER was validated against eleven typical state-of-practice benchmark models including those with foundational machine learning layers utilizing Naïve Bayes, Maximum Entropy, and Support Vector Machine methods. VADER was found to have excellent classification accuracy, generalized more favorably across contexts than any of the benchmark models, and outperformed individual human raters in sentiment analysis (Hutto & Gilbert, 2014). Since its release, VADER has been modified and validated in a variety of user-generated

and user-reported data types (Awajan et al., 2021; Larsen et al., 2021; Luber et al., 2021; Margus et al., 2021; Sasangohar et al., 2021; Shrestha, 2021).

Recently, VADER has been used extensively in analyses of public responses and patient sentiment regarding COVID-19. Performance has been validated by human raters and semi-automated mixed-method models across a diverse set of topics, including time series analyses of mental health during the pandemic, public and private responses to face-covering mandates and other public health measures, perceptions of the COVID-19 vaccine, and general sentiment regarding the coronavirus pandemic (Chandrasekaran et al., 2020; Chiny et al., 2021; Chum et al., 2021; Gbashi et al., 2021; Liew & Lee, 2021; Liu & Liu, 2021; Monselise et al., 2021; Mudassir et al., 2021; Sediqin, 2021; Shrestha, 2021; Valdez et al., 2020; Vargas et al., 2021). An increasing number of these studies utilize user- or patient-authored text data not derived from Twitter, including interview transcripts and private messages. Furthermore, a recent study evaluating various deep learning and lexical-semantic based sentiment analyses methods on telemedicine data found that VADER demonstrated superior performance compared to other deep learning methods (Talpada et al., 2019). Given its demonstrated generalizability across domains, direct accessibility for extension/modification, high interpretability, and computational efficiency, VADER was utilized for scored sentiment annotation of the current dataset.

4.4 Fine-Tuning of CT-BERT

BERT, described previously, is a widely used state-of-the-art large-scale autoencoding transformer-based language model with pretrained word embeddings (Devlin et al., 2018). BERT evaluates both positionality and order, with representations

jointly conditioned on both left and right-side contexts in all model layers (Lin & Moh, 2021). Excellent performance is attained on downstream NLP tasks with little backpropagation and fine-tuning required, reportedly on the order of less than 5 epochs. Importantly, BERT utilizes encoder components within a transformer model which can be split into embedding layers that ingest a sequential sentence (Lin & Moh, 2021; Vaswani et al., 2017). BERT has been adapted for multiple domains; CT-BERT is a recently released version of BERT that has been optimized for use on COVID-19 content (Müller et al., 2020). The fine-tuning process is contextualized within existing CT-BERT architecture as shown in Figure 2.

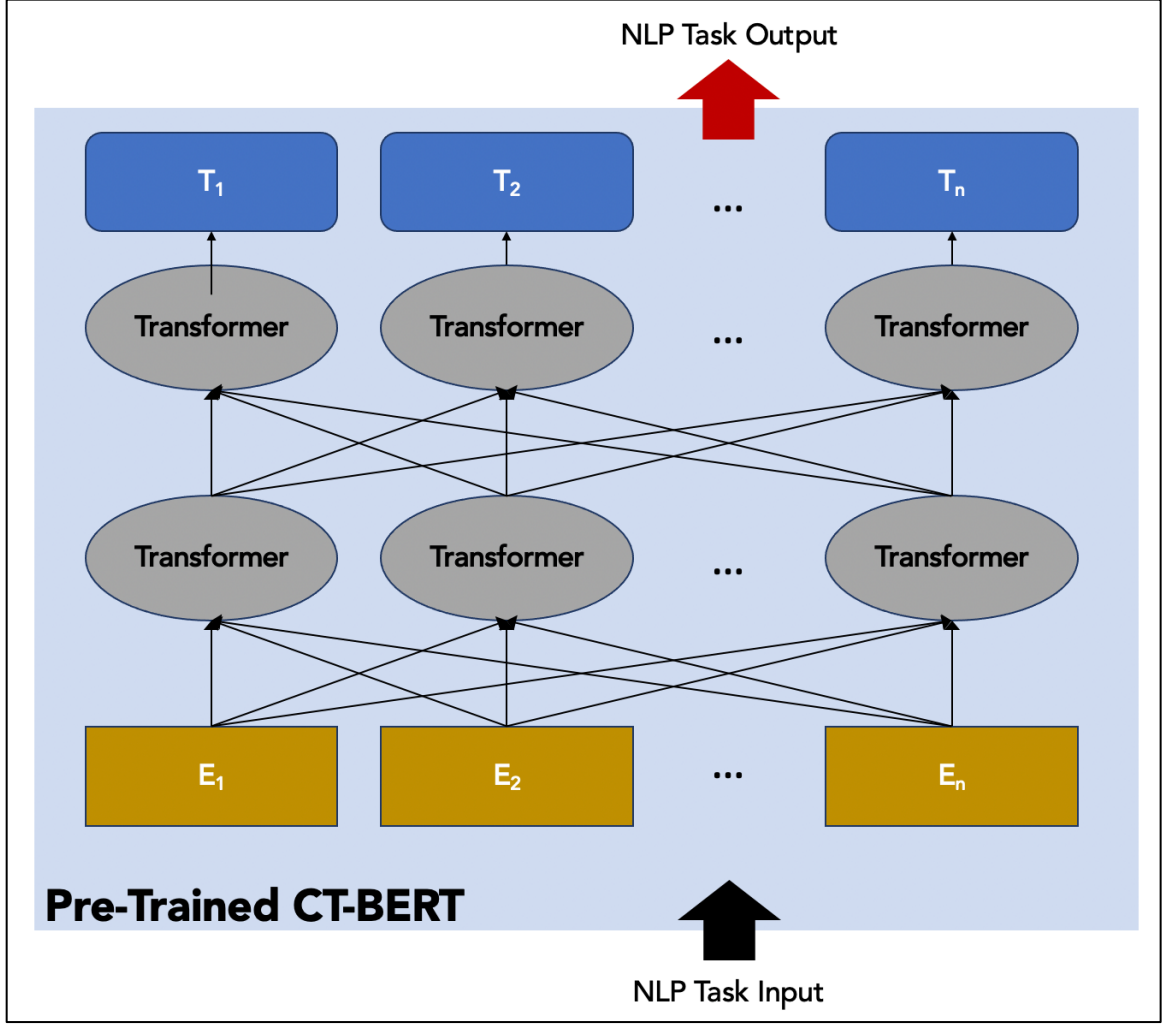


Figure 2. Architecture of Proposed CT-BERT Fine-Tuning Process.

Architecture of CT-BERT fine-tuning process trained on patient-originated secure health system messages containing COVID-19 content. Adapted from Devlin et al., 2019 and Lee et al., 2020.

To improve the domain specificity of CT-BERT for secure health system messages, supervised fine-tuning was performed using the present dataset (Xu et al., 2020). An additional neural output layer was trained on annotated patient-originated COVID-19 messages and evaluated on standard named entity recognition, relation extraction, and question answering tasks (Ji et al., 2021; Lee et al., 2020; Müller et al.,

2020). An appropriate layer-wise decreased learning rate was used to overcome the BERT fine-tuning catastrophic forgetting problem (Merchant et al., 2020; Sun, Qiu, et al., 2019). The present model was then initialized with pretrained weights from CT-BERT and updated with the fine-tuned weights and final output layer. Compared to the original CT-BERT model, the updated version more accurately captures the characteristics and vocabulary of secure health system messages. An overview of the fine-tuning process is provided in Figure 3.

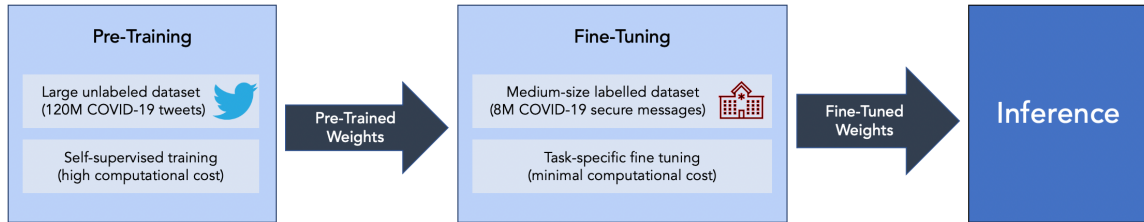


Figure 3. Overview of Proposed CT-BERT Fine-Tuning Process.

High level overview of CT-BERT fine-tuning methodology.

4.5 Auxiliary Sentence Approach

Recent studies of CT-BERT-based language models for sentiment analysis demonstrate good overall performance, but only marginal improvement compared to other predictive models (Li et al., 2019). These limited improvements are thought to be due to a lack of relevant data and low task-specific awareness, as BERT was pretrained on BookCorpus and Wikipedia (Nakov et al., 2019). Studies of the general BERT model find that implementing an auxiliary sentence approach substantially improves performance in sentiment analysis tasks. Indeed, auxiliary-modified BERT outperforms both feature-based and fine-tuning methods to achieve state-of-the-art performance on

multi-class classification data (Yu et al., 2019). Similarly, a recent study of CT-BERT demonstrated superior performance in pair classification tasks requiring text generation and prediction (Lin & Moh, 2021). Therefore, an auxiliary sentence approach was applied to expand the dataset and convert the original binary-sentence classification task to pair-sentence classification. Evaluation of binary sentence pairs allows for aspect-based sentiment analysis, where the aspect and polarity of each sentence is extracted to classify message sentiment more accurately. This can be especially important when evaluating data regarding a potentially polarizing and complex topic like COVID-19 (Jang et al., 2021).

Chapter V

Development Tools and Resources

In this chapter, the tools and development frameworks used to implement the approaches described in Chapter 4 are briefly discussed.

5.1 Electronic Patient Portal

Secure health system messages were collected from a proprietary instance of MyChart® by Epic, an electronic patient portal service licensed from Epic Systems Corporation. Patients of both inpatient and outpatient facilities at the previously described health system have access to the MyChart via web application and smartphone app. MyChart functionality includes sending direct messages to providers on a patient's care team, scheduling, checking in for, and canceling appointments, sharing medical records to other providers, requesting medication refills, paying medical bills, and accessing medical information including lab and test results. Patient-provider messages and all associated metadata are stored in a HIPAA-protected data lake. Data export was performed by the data science division of the health system and raw data was accessed in a machine-readable format.

5.2 Deep Transfer Learning in R

All text preprocessing, annotation, and machine learning analyses were performed using R statistical software, version 4.1.2 (R-Core-Team, 2017). R packages *tm* and *VADER* were used for text preprocessing, automated annotation, and other NLP tasks (Feinerer, 2020; Roehrick, 2021). The R implementation of several Python modules and

their dependencies including *BERT*, *tensorflow*, *COVID-Twitter-BERT*, and *Huggingface* were utilized for fine-tuning and model training (Bratt, 2020; Digital-Epidemiology-Lab, 2021; RStudio, 2021). Internal modifications were made to open-source code as needed.

5.3 Hyperparameters

Various deep neural network model parameters were tested for their impact on the performance of the final model. These included learning rate, batch size, maximum epochs, dropout rate, momentum, and the number of hidden units. As documented in previous reports of fine-tuning BERT, the loss did not converge with a high learning rate $5e-2$ and batch size 9. After several rounds of experimentations, loss stopped improving at 12 epochs. Ultimately, a decaying learning rate of $2e-5$ and batch size 4 were chosen for both single and pair-sentence classification tasks. Dropout rate remained at 0.1 (Gao et al., 2019).

An early stop mechanism was set to prevent overfitting during model training. The AdamW optimizer, accessed through TensorFlow for R, was used to improve model training speed while stabilizing weight decay as previously described in the literature (Lin & Moh, 2021; Loshchilov & Hutter, 2017, 2018; Xie et al., 2020). Due to the higher memory cost of AdamW, 128 was chosen as the maximum sequence length. This sequence length is consistent with other studies of BERT-based models, though other systems not limited by out-of-memory issues have run epochs with maximum sequence length of up to 512 tokens (Dumitrescu et al., 2020; Lin & Moh, 2021; Y. Liu et al., 2019; Novo, 2021; Ulčar & Robnik-Šikonja, 2020).

Chapter VI

Results

After text preprocessing and filtering rules were applied, over 3.2M patient-originated messages sent from 209,000 unique patients were included in machine analysis. Eighteen medical subspecialties were represented, with most messages appearing to be exchanged in the context of outpatient care services. Messages were sent to all levels of clinical care providers, most commonly physicians, physician assistants, and nurse practitioners. After text preprocessing, the corpus of included messages was found to comprise over 62,000 unique entities and nearly 21,500 discrete relationship types.

The following sections briefly summarize the results of (1) sentiment annotation using VADER and (2) aspect-based sentiment analysis with auxiliary sentence approach using CT-BERT, fine-tuned on a large corpus of patient-originated secure health system messages containing COVID-19 content.

6.1 VADER Output

The VADER model of sentiment annotation outputs four scores normalized from 0-1. “Positive,” “Negative,” and “Neutral” scores measures the proportion of text classified into the corresponding sentiment category. These scores are translated to percentages and sum to 1. The “Compound” score is the sum of all lexicon ratings normalized between -1, representing the most extreme negative sentiment, to +1, representing the most extreme positive sentiment (Barros et al., 2021; Hutto & Gilbert,

2014; Roehrick, 2021). Table 1 reports the VADER sentiment score averages across all tokenized language components in the present corpus.

As described previously, messages were segmented by date to reflect major developments regarding COVID-19 in the United States (CDC, 2021). “Early COVID” was defined as March through August 2020, representing the period from which a global pandemic was first declared by the World Health Organization and the U.S. death toll surpassed 100,000. “Middle COVID” was defined as September through December 2020, representing the period from which the U.S. death toll surpassed 300,000 to the vaccination of more than 1 million Americans against COVID-19. “Late COVID” was defined as January through May 2021, representing the period from which the U.S. death toll surpassed 400,000 to the administration of over 200 million vaccinations.

Table 1. Normalized Averages of VADER Output Scores for All Tokenized Message Components Relevant to COVID-19, Segmented by Time Period.

Sentiment Category	Early COVID	Middle COVID	Late COVID
	<i>March-August 2020</i>	<i>September-December 2020</i>	<i>January-May 2021</i>
Positive	0.113	0.104	0.391
Neutral	0.476	0.405	0.287
Negative	0.411	0.491	0.322
Compound	-0.152	-0.033	0.336

6.2 Fine-Tuned CT-BERT with Auxiliary Sentence Approach

Model performance for aspect-based sentiment analysis was benchmarked on the present VADER-annotated dataset. Results were evaluated using 10-fold cross-validation to compute accuracy (Q. Chen et al., 2021; Savoy, 2012). This accounts for classification effectiveness across all sentiment classes by counting the number of correct predictions divided by the number of cases (tokenized language components). Evaluative measures of model performance are based on the following formulas:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Cases}$$

$$F1 = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall\ (Sensitivity) = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Sentiment size was used to weight F1 scores, precision, and recall as reported in Table 2. Our proposed model achieved a weighted average F1 score of 0.881 for aspect extraction and predicted sentiment intensity scores with a mean square error of 0.008.

Table 2. Comparative Evaluation of Aspect-Based Sentiment Analysis Models Evaluated on a Novel Corpus of Patient-Originated Secure Health System Messages Containing COVID-19 Content.

Model	Accuracy	Macro F1	Precision
CT-BERT	76.0	75.0	72.0
CT-BERT, fine-tuned, single	84.1	83.7	82.0
CT-BERT, fine-tuned, auxiliary/pair	88.5	88.1	85.0

Chapter VII

Discussion and Conclusions

7.1 Experimental Outcomes

VADER-generated sentiment annotation of the processed corpus demonstrates several temporal trends. First, there is a significant increase in positive sentiment from both early and middle COVID to late COVID ($p < 0.001$). Patients may have expressed more positive sentiments to their providers as lockdown periods ended, vaccines became widely available, and infection rates declined in the northeastern United States. Negative sentiment followed a similar but inverted trend, where messages in the late COVID category had a significantly larger proportion of negative expressions compared to early and middle COVID ($p < 0.05$). These changes were somewhat associated with changes in positive sentiment and could be attributed to similar causes. No significant changes in neutral sentiment were detected between early, middle, and late COVID. However, the largest proportion of text within all COVID-relevant messages was classified as having neutral sentiment. This may signify that the nature of patient-provider messages is more formal and therefore less likely to contain strong positive or negative sentiment. Therefore, domain-specific methods targeted to secure health system messages may be required for the meaningful application of widely used COVID sentiment analysis models.

Notably, compound measures showed a statistically significant period-over-period increase ($p < 0.05$), with the largest absolute difference observed between early and late COVID. Based on compound scores, the average patient-originated message during early

and middle COVID could be described as weakly negative to neutral compared to weakly or mildly positive during late COVID. These changes could be due to improvements in COVID-related harms and public response developments. Previous analyses further suggest that changes in COVID-related sentiment may be mediated by spatiotemporal interactions related to geographic location, socioeconomic status, and political affiliation (Ali et al., 2021; Surano et al., 2021). Trends may also differ between distinct subsets of COVID-19 message. For example, compound sentiment across Twitter posts related to COVID-19 symptoms, outbreak sources, and political impacts was found to be consistently negative, while sentiment reversed from negative to positive in posts discussing economy recovery, government response, and COVID-19 treatment (Chandrasekaran et al., 2020). Consistent with our findings, a recent study of COVID-19 vaccine-related social media posts found that negative sentiment decreased from November 2020 to February 2021 (Karami et al., 2021; Vargas et al., 2021). Other studies also demonstrate that public health mandates and announcements regarding COVID-19 prevalence have a significant impact on expressions of public sentiment (Chalkiadakis et al., 2021; Ebadi et al., 2020; Thorpe Huerta et al., 2021). For a limited number of topics, these temporal trends in pandemic-related sentiment appear to generalize to non-US countries including China and the UK (Han et al., 2020; S. Yu et al., 2021; Zou et al., 2020).

Compared to both CT-BERT and fine-tuned CT-BERT with single-sentence classification, fine-tuned CT-BERT with pair-sentence classification achieved superior performance. Classification measures derived from a standardized confusion matrix were increased in the latter model. However, the magnitude of improvement seems to be

largest when moving from CT-BERT to either of the fine-tuned, domain-specific CT-BERT models. We also observe that the construction of an auxiliary sentence, and consequent conversion of a single-sentence classification task to a pair-sentence classification task, modestly improves model performance. These findings are consistent with a prior study of CT-BERT in COVID-19 social media data, though our results suggest that fine-tuned CT-BERT has a better result on patient-originated health system messages (Lin & Moh, 2021). Even standalone, our fine-tuned model achieves good to excellent performance in pair-sentence classification tasks.

These findings suggest that (1) domain-specific training improves the performance of CT-BERT on secure health system messages and (2) the implementation of an auxiliary sentence approach improves the predictive ability of sentiment analysis models when applied to COVID-related message data. We believe that our proposed model can be directly applied to future analyses of secure health system message corpora.

7.2 Limitations

The present study has several limitations. First, the foundational CT-BERT model was pretrained on COVID-19 social media data. Additionally, CT-BERT was trained on corpora that included some non-English tokens (Müller et al., 2020). In addition to fine-tuning, an optimized approach to BERT modification would include full pretraining on English-only, patient-originated secure health system messages. Pretraining was not performed due to limited computational power. Second, VADER was used for automated sentiment annotation given data volume. Given that manual human review is considered the gold standard for annotation, the accuracy of our findings may be limited.

Additionally, VADER is optimized for sentiment analysis of social media data and may not capture expressions within more formal communication texts such as patient portal messages. Finally, the dataset used in this study was obtained from a single health system. Therefore, our results may have limited institutional and regional generalizability. Our exclusion of non-English messages also limits cross-language use.

7.3 Future Directions

Manual, semi-automated, and other fully automated approaches to annotation could be applied to the present dataset. Consumer health frameworks such as the Fast Health Care Interoperability Resources (FIHR) data model have previously been used to analyze a large corpus of secure patient messages, though not in the setting of COVID-19 (De et al., 2021). The UMLS has also been used in machine learning studies of patient messages, though on a significantly smaller volume data drawn from only a few medical specialties (Mastorakos et al., 2021). Other open-domain biomedical knowledge tools including DBpedia and MetaMap have been applied to consumer health inquiry data released by the National Library of Medicine (NLM) (Cui et al., 2014; Demner-Fushman et al., 2020; Kilicoglu et al., 2016; Kilicoglu et al., 2018; Mrabet et al., 2016; Roberts et al., 2014). These and other semantic approaches could support the development of automated tools that inform the triage, routing, and management of incoming patient messages, especially in the setting of COVID-19.

Future topic modelling and entity recognition studies can utilize our data for a variety of potentially valuable applications. This feature-rich, clinically diverse corpus of secure health system messages in the setting of the COVID-19 pandemic can be mined

for information regarding patient beliefs, patient-provider interactions, specialty-specific approaches to COVID-19, impacts of various public health measures on outpatient care, and other significant topics. A granular understanding of patient information-seeking trends in response to COVID-related measures could provide valuable guidance in the future development of public health campaigns (Bashar et al., 2020). Similarly, surveilling disease-related “emotional contagion” on social media sites could improve preventive measures in the future (Crocamo et al., 2021; Rajput et al., 2020). Investigators could also explore ensembling and hybrid approaches – instead of or in addition to the pre-training and fine-tuning strategies utilized here – to augment the performance of BERT-based models on patient-originated communications (Kumar et al., 2021; Møller et al., 2020).

From a clinical perspective, few studies have examined the complexity of medical decision-making represented in patient portal messages (Robinson et al., 2017). Understanding characteristics associated with patient messages could be of increasing importance given the rapid spread of telehealth adoption. Symptom extraction models that achieve good performance on social media data could be augmented to identify symptomology on the present data, especially since there may be systematic differences between symptoms reported to clinical providers versus those shared informally on social networks (Luo et al., 2021; Lybarger, Ostendorf, Thompson, et al., 2021). Importantly, studies that incorporate ancillary demographic information could provide high-resolution characterizations of social determinants of health at scale (Lybarger, Ostendorf, & Yetisgen, 2021). Further utilization of our dataset and the proposed fine-tuned deep transfer learning model have significant potential to address these knowledge gaps as

they relate to secure health system communications. Ultimately, we believe that future applications of this work can help advance clinical and technical understanding of the COVID-19 pandemic.

References

- Abdollahi, M., Gao, X., Mei, Y., Ghosh, S., Li, J., & Narag, M. (2021). Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artif Intell Med*, 120, 102167.
- Adamidi, E. S., Mitsis, K., & Nikita, K. S. (2021). Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review. *Comput Struct Biotechnol J*, 19, 2833-2850.
- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.
- Agrawal, A., He, Z., Perl, Y., Wei, D., Halper, M., Elhanan, G., & Chen, Y. (2013). The readiness of SNOMED problem list concepts for meaningful use of electronic health records. *Artif Intell Med*, 58(2), 73-80.
- Alamoodi, A. H., Zaidan, B. B., Al-Masawa, M., Taresh, S. M., Noman, S., Ahmaro, I. Y. Y., Garfan, S., Chen, J., Ahmed, M. A., Zaidan, A. A., Albahri, O. S., Aickelin, U., Thamir, N. N., Fadhil, J. A., & Salahaldin, A. (2021). Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Computers in Biology and Medicine*, 139, 104957.
- Alfalahi, A., Brissman, S., & Dalianis, H. (2012). Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) Held in Conjunction with LREC,
- Alfattni, G., Peek, N., & Nenadic, G. (2020). Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, 108, 103488.
- Älgå, A., Eriksson, O., & Nordberg, M. (2020). Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study. *J Med Internet Res*, 22(11), e21559.

- Ali, G., Rahman, M. M., Hossain, M. A., Rahman, M. S., Paul, K. C., Thill, J. C., & Samuel, J. (2021). Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics. *Healthcare (Basel)*, 9(9).
- Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., & Liakata, M. (2020). QMUL-SDS at CheckThat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. *arXiv preprint arXiv:2008.13160*.
- Alnazzawi, N., Thompson, P., Batista-Navarro, R., & Ananiadou, S. (2015). Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Med Inform Decis Mak*, 15 Suppl 2(Suppl 2), S3.
- Alsayat, A. (2021). Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model. *Arab J Sci Eng*, 1-13.
- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.
- Alser, M., Shahroodi, T., Gómez-Luna, J., Alkan, C., & Mutlu, O. (2020). SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs. *Bioinformatics*.
- Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4, 463-476.
- Aramaki, E., Morita, M., Kano, Y., & Ohkuma, T. (2016). Overview of the NTCIR-12 MedNLPDoc task. 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan.
- Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.
- Awad, A., Bader-El-Den, M., McNicholas, J., & Briggs, J. (2017). Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inform*, 108, 185-195.

- Awajan, I., Mohamad, M., & Al-Quran, A. (2021). Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data From Online Reviews. *IEEE Access*, 9, 47338-47353.
- Bahja, M., & Safdar, G. A. (2020). Unlink the Link Between COVID-19 and 5G Networks: An NLP and SNA Based Approach. *IEEE Access*, 8, 209127-209137.
- Ballinari, D., & Behrendt, S. (2021). How to gauge investor behavior? A comparison of online investor sentiment measures. *Digit Finance*, 3(2), 169-204.
- Banerji, A., Lai, K. H., Li, Y., Saff, R. R., Camargo, C. A., Jr., Blumenthal, K. G., & Zhou, L. (2020). Natural Language Processing Combined with ICD-9-CM Codes as a Novel Method to Study the Epidemiology of Allergic Drug Reactions. *J Allergy Clin Immunol Pract*, 8(3), 1032-1038.e1031.
- Barros, L., Trifan, A., & Oliveira, J. L. (2021). VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining.
- Bartels, P. H., & Hiessl, H. (1989). Expert systems in histopathology. II. Knowledge representation and rule-based systems. *Anal Quant Cytol Histol*, 11(3), 147-153.
- Bashar, M. A., Nayak, R., & Balasubramaniam, T. (2020). Topic, sentiment and impact analysis: Covid19 information seeking on social media. *arXiv preprint arXiv:2008.12435*.
- Belousov, M., Dixon, W. G., & Nenadic, G. (2017). Using an Ensemble of Linear and Deep Learning Models in the SMM4H 2017 Medical Concept Normalisation Task.
- Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., & Verhagen, M. (2015). Semeval-2015 task 6: Clinical tempeval. proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).
- Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).
- Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1), D267-D270.

- Bonert, M., Naqvi, A., Rahman, M., Marshall, J. K., Xenodemetropoulos, T., Arora, P., Slater, J., & Major, P. (2021). Stability of diagnostic rate in a cohort of 38,813 colorectal polyp specimens and implications for histomorphology and statistical process control. *Sci Rep*, 11(1), 16942.
- Boon-Itt, S., & Skunkan, Y. (2020). Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR public health and surveillance*, 6(4), e21978.
- Bose, P., Roy, S., & Ghosh, P. (2021). A Comparative NLP-Based Study on the Current Trends and Future Directions in COVID-19 Research. *IEEE Access*, 9, 78341-78355.
- Boyajian, R., Gordon, W. J., Kowtoniuk, A. M., Boyajian, K. R., Mackin, M., Martin, N. E., Orio, P. F., 3rd, Walsh, G. P., & Nguyen, P. L. (2021). Development & Impact of a Virtual PSA Monitoring Clinic for Follow-up of Prostate Cancer Patients: An Efficient Model With Unique Benefits Relevant to COVID-19. *Int J Radiat Oncol Biol Phys*, 111(3s), S65.
- Bozkurt, S., Alkim, E., Banerjee, I., & Rubin, D. L. (2019). Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. *J Digit Imaging*, 32(4), 544-553.
- Bozkurt, S., Paul, R., Coquet, J., Sun, R., Banerjee, I., Brooks, J. D., & Hernandez-Boussard, T. (2020). Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. *Learn Health Syst*, 4(4), e10237.
- Bratt, J. (2020). *RBERT*. In <https://github.com/jonathanbratt/RBERT>
- Brekke, P. H., Rama, T., Pilán, I., Nytrø, Ø., & Øvrelid, L. (2021). Synthetic data for annotation and extraction of family history information from clinical text. *Journal of Biomedical Semantics*, 12(1), 11.
- Broseta, J. J. (2021). Different approaches to improve cohort identification using electronic health records: X-linked hypophosphatemia as an example. *Intractable Rare Dis Res*, 10(1), 17-22.
- Brown, S. H., Elkin, P. L., Rosenbloom, S. T., Fielstein, E., & Speroff, T. (2008). eQuality for all: Extending automated quality measurement of free text clinical narratives. *AMIA Annu Symp Proc*, 2008, 71-75.

- Budhwani, H., Maycock, T., Murrell, W., & Simpson, T. (2021). COVID-19 Vaccine Sentiments Among African American or Black Adolescents in Rural Alabama. *J Adolesc Health*.
- Burger, G., Abu-Hanna, A., de Keizer, N., & Cornet, R. (2016). Natural language processing in pathology: a scoping review. *J Clin Pathol*.
- Cabot, C., Darmoni, S., & Soualmia, L. F. (2019). Cimind: A phonetic-based tool for multilingual named entity recognition in biomedical texts. *J Biomed Inform*, 94, 103176.
- Camacho-Collados, J., & Pilehvar, M. T. (2017). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.
- Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., & Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Med Inform Decis Mak*, 21(1), 69.
- Carrell, D. S., Halgrim, S., Tran, D. T., Buist, D. S., Chubak, J., Chapman, W. W., & Savova, G. (2014). Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol*, 179(6), 749-758.
- Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua, V., Tobin, R., Whiteley, W., Wu, H., & Alex, B. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak*, 21(1), 179.
- Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., & Esposito, M. (2020). Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Applied Soft Computing*, 97, 106779.
- CDC. (2021). *CDC Museum COVID-19 Timeline*. U.S. Department of Health & Human Services. Retrieved October 20 from <https://www.cdc.gov/museum/timeline/covid19.html>
- Chalkiadakis, I., Yan, H., Peters, G. W., & Shevchenko, P. V. (2021). Infection rate models for COVID-19: Model risk and public health news sentiment exposure adjustments.

- Chan, L., Beers, K., Yau, A. A., Chauhan, K., Duffy, Á., Chaudhary, K., Debnath, N., Saha, A., Pattharanitima, P., Cho, J., Kotanko, P., Federman, A., Coca, S. G., Van Vleck, T., & Nadkarni, G. N. (2020). Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients. *Kidney Int*, 97(2), 383-392.
- Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. (2020). Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. *J Med Internet Res*, 22(10), e22624.
- Chang, P., Sheng, Y. H., Sang, Y. Y., & Wang, D. W. (2008). Developing a wireless speech- and touch-based intelligent comprehensive triage support system. *Comput Inform Nurs*, 26(1), 31-38.
- Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5), 301-310.
- Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Meng, X., Ren, W., & Zhou, Y. (2021). Transformer-based language model fine-tuning methods for covid-19 fake news detection. International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation.
- Chen, J., Kiefe, C. I., Gagnier, M., Lessard, D., McManus, D., Wang, B., & Houston, T. K. (2021). Non-specific pain and 30-day readmission in acute coronary syndromes: findings from the TRACE-CORE prospective cohort. *BMC Cardiovasc Disord*, 21(1), 383.
- Chen, K. P., & Moskowitz, A. (2016). Comparative Effectiveness: Propensity Score Analysis. In *Secondary Analysis of Electronic Health Records* (pp. 339-349).
- Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., Langlotz, C. P., Amrhein, T. J., & Lungren, M. P. (2018). Deep Learning to Classify Radiology Free-Text Reports. *Radiology*, 286(3), 845-852.
- Chen, P. H., Zafar, H., Galperin-Aizenberg, M., & Cook, T. (2018). Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *J Digit Imaging*, 31(2), 178-184.

- Chen, Q., Leaman, R., Allot, A., Luo, L., Wei, C.-H., Yan, S., & Lu, Z. (2021). Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing. *Annual Review of Biomedical Data Science*, 4(1), 313-339.
- Cheng, I. K., Heyl, J., Lad, N., Facini, G., & Grout, Z. (2021). Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Sci Rep*, 11(1), 19009.
- Cheng, L. T., Zheng, J., Savova, G. K., & Erickson, B. J. (2010). Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging*, 23(2), 119-132.
- Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2021). Analysis of sentiments conveyed through Twitter concerning COVID-19. SHS Web of Conferences.
- Choi, S. W., Ko, T., Hong, K. J., & Kim, K. H. (2019). Machine Learning-Based Prediction of Korean Triage and Acuity Scale Level in Emergency Department Patients. *Healthc Inform Res*, 25(4), 305-312.
- Chum, A., Nielsen, A., Bellows, Z., Farrell, E., Durette, P. N., Banda, J. M., & Cupchik, G. (2021). Changes in Public Response Associated With Various COVID-19 Restrictions in Ontario, Canada: Observational Infoveillance Study Using Social Media Time Series Data. *J Med Internet Res*, 23(8), e28716.
- Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks.
- Colic, N., Furrer, L., & Rinaldi, F. (2020). Annotating the pandemic: Named entity recognition and normalisation in COVID-19 literature.
- Collier, N., Oellrich, A., & Groza, T. (2015). Concept selection for phenotypes and diseases using learn to rank. *J Biomed Semantics*, 6, 24.
- American Recovery and Reinvestment Act of 2009, (2009).
<https://www.congress.gov/bill/111th-congress/house-bill/1/text>
- Crocamo, C., Viviani, M., Famiglini, L., Bartoli, F., Pasi, G., & Carrà, G. (2021). Surveilling COVID-19 Emotional Contagion on Twitter by Sentiment Analysis. *European Psychiatry*, 64(1), e17, Article e17.

- Cronin, R. M., Fabbri, D., Denny, J. C., Rosenbloom, S. T., & Jackson, G. P. (2017). A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform*, 105, 110-120.
- Cui, L., Tao, S., & Zhang, G. Q. (2014). A Semantic-based Approach for Exploring Consumer Health Questions Using UMLS. *AMIA Annu Symp Proc*, 2014, 432-441.
- Cunningham, H. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. Proc. 40th annual meeting of the association for computational linguistics (ACL 2002).
- D'Avolio, L. W., Nguyen, T. M., Goryachev, S., & Fiore, L. D. (2011). Automated concept-level information extraction to reduce the need for custom software and rules development. *Journal of the American Medical Informatics Association*, 18(5), 607-613.
- Dang, H., Lee, K., Henry, S., & Uzuner, O. (2020). Ensemble BERT for classifying medication-mentioning tweets. Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task.
- Data, M. I. T. C., & Pirracchio, R. (2016). Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project. In *Secondary Analysis of Electronic Health Records* (pp. 295-313).
- Datta, S., & Roberts, K. (2020). A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations. *Data Brief*, 32, 106056.
- De, A., Huang, M., Feng, T., Yue, X., & Yao, L. (2021). Analyzing Patient Secure Messages Using a Fast Health Care Interoperability Resources (FIHR)-Based Data Model: Development and Topic Modeling Study. *J Med Internet Res*, 23(7), e26770.
- Decker, B. M., Hill, C. E., Baldassano, S. N., & Khankhanian, P. (2021). Can antiepileptic efficacy and epilepsy variables be studied from electronic health records? A review of current approaches. *Seizure*, 85, 138-144.
- Demner-Fushman, D., Mrabet, Y., & Ben Abacha, A. (2020). Consumer health information and question answering: helping consumers find answers to their

- health-related information needs. *Journal of the American Medical Informatics Association*, 27(2), 194-201.
- DeSouza, D. D., Robin, J., Gumus, M., & Yeung, A. (2021). Natural Language Processing as an Emerging Tool to Detect Late-Life Depression. *Front Psychiatry*, 12, 719125.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Digital-Epidemiology-Lab. (2021). *COVID-Twitter-BERT*. In <https://github.com/digitalepidemiologylab/covid-twitter-bert>
- Divita, G., Carter, M., Redd, A., Zeng, Q., Gupta, K., Trautner, B., Samore, M., & Gundlapalli, A. (2015). Scaling-up NLP Pipelines to Process Large Corpora of Clinical Notes. *Methods Inf Med*, 54(06), 548-552.
- Dumitrescu, S. D., Avram, A.-M., & Pyysalo, S. (2020). The birth of Romanian BERT. *arXiv preprint arXiv:2009.08712*.
- easyPubMed. (2021). PubMed Query Stopwords. https://rdrr.io/github/dami82/easyPubMed/man/PubMed_stopwords.html
- Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., & Wong, A. (2020). Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. *Scientometrics*, 1-15.
- Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., & Speroff, T. (2006). Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clin Proc*, 81(6), 741-748.
- Elmessiry, A., Cooper, W. O., Catron, T. F., Karrass, J., Zhang, Z., & Singh, M. P. (2017). Triaging Patient Complaints: Monte Carlo Cross-Validation of Six Machine Learning Classifiers. *JMIR Med Inform*, 5(3), e19.
- Fan, J. W., Yang, E. W., Jiang, M., Prasad, R., Loomis, R. M., Zisook, D. S., Denny, J. C., Xu, H., & Huang, Y. (2013). Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *J Am Med Inform Assoc*, 20(6), 1168-1177.

- Feinerer, I. (2020). *tm: Text Mining Package*. In <https://cran.r-project.org/web/packages/tm/>
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health*, 4(2), e19.
- Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*, 11(5), 392-402.
- Gabarron, E., Dorronzoro, E., Rivera-Romero, O., & Wynn, R. (2019). Diabetes on Twitter: A Sentiment Analysis. *J Diabetes Sci Technol*, 13(3), 439-444.
- Gao, Q., Wang, D., Sun, P., Luan, X., & Wang, W. (2021). Sentiment Analysis Based on the Nursing Notes on In-Hospital 28-Day Mortality of Sepsis Patients Utilizing the MIMIC-III Database. *Comput Math Methods Med*, 2021, 3440778.
- Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7, 154290-154299.
- Gaudet-Blavignac, C., Foufi, V., Bjelogrić, M., & Lovis, C. (2021). Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. *J Med Internet Res*, 23(1), e24594.
- Gbashi, S., Adebo, O. A., Doorsamy, W., & Njobeh, P. B. (2021). Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study. *JMIR Med Inform*, 9(3), e22916.
- Ghosh, S., Li, J., Cao, L., & Ramamohanarao, K. (2017). Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. *J Biomed Inform*, 66, 19-31.
- Ginter, F., Haverinen, K., Laippala, V., & Viljanen, T. (2010). *Turku Clinical TreeBank and PropBank*.

- Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci Rep*, 7, 45141.
- Glazkova, A., Glazkov, M., & Trifonov, T. (2021). g2tmn at constraint@ aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation.
- Gligic, L., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2020). Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. *Neural Netw*, 121, 132-139.
- Gobbel, G. T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S. H., Elkin, P. L., & Matheny, M. E. (2014). Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. *Journal of Biomedical Informatics*, 48, 54-65.
- Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., Narayanan, S. S., & Atkins, D. C. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *J Couns Psychol*, 67(4), 438-448.
- Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., & Stanley, H. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220.
- Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearb Med Inform*, 26(1), 214-227.
- Goss, F. R., Plasek, J. M., Lau, J. J., Seger, D. L., Chang, F. Y., & Zhou, L. (2014). An evaluation of a natural language processing tool for identifying and encoding allergy information in emergency department clinical notes. *AMIA Annu Symp Proc*, 2014, 580-588.
- Grabar, N., & Grouin, C. (2021). Year 2020 (with COVID): Observation of Scientific Literature on Clinical Natural Language Processing. *Yearb Med Inform*, 30(1), 257-263.

- Greaves, F., Lavery, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. (2014). Tweets about hospital quality: a mixed methods study. *BMJ Qual Saf*, 23(10), 838-846.
- Grundmeier, R. W., Masino, A. J., Casper, T. C., Dean, J. M., Bell, J., Enriquez, R., Deakyne, S., Chamberlain, J. M., Alpern, E. R., & The Pediatric Emergency Care Applied Research, N. (2016). Identification of Long Bone Fractures in Radiology Reports Using Natural Language Processing to support Healthcare Quality Improvement. *Appl Clin Inform*, 07(04), 1051-1068.
- Hakenberg, J., Royer, L., Plake, C., Strobel, H., & Schroeder, M. (2007). Me and my friends: gene mention normalization with background knowledge. Proc 2nd BioCreative Challenge Evaluation Workshop.
- Han, D., Xu, F., Li, C., Zhang, L., Yang, R., Zheng, S., Wang, Z., & Lyu, J. (2021). A Novel Nomogram for Predicting Survival in Patients with Severe Acute Pancreatitis: An Analysis Based on the Large MIMIC-III Clinical Database. *Emerg Med Int*, 2021, 9190908.
- Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China. *Int J Environ Res Public Health*, 17(8).
- Hao, B., Sotudian, S., Wang, T., Xu, T., Hu, Y., Gaitanidis, A., Breen, K., Velmahos, G. C., & Paschalidis, I. C. (2020). Early prediction of level-of-care requirements in patients with COVID-19. *eLIFE*, 9.
- Hassanpour, S., Bay, G., & Langlotz, C. P. (2017). Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. *J Digit Imaging*, 30(3), 314-322.
- Haverinen, K., Laippala, V., Kohonen, S., Missilä, A., Nyblom, J., Ojala, S., Viljanen, T., Salakoski, T., & Ginter, F. (2013). Towards a dependency-based PropBank of general Finnish. Proceedings of the 19th Nordic Conference of Computational Linguistics.
- Håvik, R., Wake, J. D., Flobak, E., Lundervold, A., & Guribye, F. (2018). A Conversational Interface for Self-screening for ADHD in Adults. International Conference on Internet Science.

- Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., McIver, D. J., Rozenblum, R., Wright, A., Bourgeois, F. T., & Greaves, F. (2016). Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*, 25(6), 404-413.
- Hazell, L., & Shakir, S. A. (2006). Under-reporting of adverse drug reactions : a systematic review. *Drug Saf*, 29(5), 385-396.
- He, B., Dong, B., Guan, Y., Yang, J., Jiang, Z., Yu, Q., Cheng, J., & Qu, C. (2017). Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *J Biomed Inform*, 69, 203-217.
- Henia, W., Rjab, O., Haddad, H., & Fourati, C. (2021). iCompass at NLP4IF-202Q Fighting the COVID-19 Infodemic.
- Henriksson, A., Conway, M., Duneld, M., & Chapman, W. W. (2013). Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. *AMIA Annu Symp Proc*, 2013, 600-609.
- Henry, S., Wang, Y., Shen, F., & Uzuner, O. (2020). The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. *J Am Med Inform Assoc*, 27(10), 1529-1537.
- Hernandez-Boussard, T., Monda, K. L., Crespo, B. C., & Riskin, D. (2019). Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *Journal of the American Medical Informatics Association*, 26(11), 1189-1194.
- Hettiarachchi, H., & Ranasinghe, T. (2020). Infominer at wnut-2020 task 2: Transformer-based covid-19 informative tweet extraction. *arXiv preprint arXiv:2010.05327*.
- Hogan, T. P., Luger, T. M., Volkman, J. E., Rocheleau, M., Mueller, N., Barker, A. M., Nazi, K. M., Houston, T. K., & Bokhour, B. G. (2018). Patient Centeredness in Electronic Communication: Evaluation of Patient-to-Health Care Team Secure Messaging. *J Med Internet Res*, 20(3), e82.
- Hong, J., Davoudi, A., Yu, S., & Mowery, D. L. (2020). Annotation and extraction of age and temporally-related events from clinical histories. *BMC Med Inform Decis Mak*, 20(Suppl 11), 338.

- Hong, N., Wen, A., Stone, D. J., Tsuji, S., Kingsbury, P. R., Rasmussen, L. V., Pacheco, J. A., Adekkanattu, P., Wang, F., Luo, Y., Pathak, J., Liu, H., & Jiang, G. (2019). Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J Biomed Inform*, 99, 103310.
- Hossain, T. T. (2021). *COVIDLies: Detecting COVID-19 misinformation on social media* UC Irvine.
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Hull, T. D., Levine, J., Bantilan, N., Desai, A. N., & Majumder, M. S. (2021). Analyzing Digital Evidence From a Telemental Health Platform to Assess Complex Psychological Responses to the COVID-19 Pandemic: Content Analysis of Text Messages. *JMIR Form Res*, 5(2), e26190.
- Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A., & Sheikh, A. (2021). Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. *J Med Internet Res*, 23(4), e26627.
- Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media.
- Isaza-Restrepo, A., Gómez, M. T., Cifuentes, G., & Argüello, A. (2018). The virtual patient as a learning tool: a mixed quantitative qualitative study. *BMC Med Educ*, 18(1), 297.
- Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A. H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19-Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *Am J Trop Med Hyg*, 103(4), 1621-1629.
- Izquierdo, J. L., Ancochea, J., & Soriano, J. B. (2020). Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing. *J Med Internet Res*, 22(10), e21801.

- Jang, H., Rempel, E., Roth, D., Carenini, G., & Janjua, N. Z. (2021). Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *J Med Internet Res*, 23(2), e25431.
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*, 13(6), 395-405.
- Ji, S., Hölttä, M., & Marttinen, P. (2021). Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*, 139, 104998.
- Ji, Z., Wei, Q., & Xu, H. (2020). BERT-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020, 269.
- Jiang, H., Mao, H., Lu, H., Lin, P., Garry, W., Lu, H., Yang, G., Rainer, T. H., & Chen, X. (2021). Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *Int J Med Inform*, 145, 104326.
- Jiang, M., Huang, Y., Fan, J. W., Tang, B., Denny, J., & Xu, H. (2015). Parsing clinical text: how good are the state-of-the-art parsers? *BMC Med Inform Decis Mak*, 15 Suppl 1(Suppl 1), S2.
- Jindal, P., & Roth, D. (2013). Extraction of events and temporal expressions from clinical narratives. *J Biomed Inform*, 46 Suppl, S13-s19.
- Jindian, S., Yu, S., & Luo, D. (2020). Enhancing Aspect-based Sentiment Classification with Auxiliary Sentence and Domain Knowledge. *IEEE Access*, PP, 1-1.
- Jones, K. S. (1995). Reflections on TREC. *Information Processing & Management*, 31(3), 291-314.
- Kalyan, K. S., & Sangeetha, S. (2021). BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and highway network. *Artif Intell Med*, 112, 102008.

- Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.
- Karami, A., Zhu, M., Goldschmidt, B., Boyajieff, H. R., & Najafabadi, M. M. (2021). COVID-19 Vaccine and Social Media in the U.S.: Exploring Emotions and Discussions on Twitter. *Vaccines (Basel)*, 9(10).
- Karcioglu, A. A., & Bulut, H. (2021). Improving hash-q exact string matching algorithm with perfect hashing for DNA sequences. *Comput Biol Med*, 131, 104292.
- Karcioglu, A. A., & Bulut, H. (2021). The WM-q multiple exact string matching algorithm for DNA sequences. *Comput Biol Med*, 136, 104656.
- Kehl, K. L., Xu, W., Lepisto, E., Elmarakeby, H., Hassett, M. J., Van Allen, E. M., Johnson, B. E., & Schrag, D. (2020). Natural Language Processing to Ascertain Cancer Outcomes From Medical Oncologist Notes. *JCO Clin Cancer Inform*, 4, 680-690.
- Kersloot, M. G., van Putten, F. J. P., Abu-Hanna, A., Cornet, R., & Arts, D. L. (2020). Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *J Biomed Semantics*, 11(1), 14.
- Khalid, Z., & Sezerman, O. U. (2017). ZK DrugResist 2.0: A TextMiner to extract semantic relations of drug resistance from PubMed. *J Biomed Inform*, 69, 93-98.
- Khan Mamun, M. M. R. (2021). Cuff-less blood pressure measurement based on hybrid feature selection algorithm and multi-penalty regularized regression technique. *Biomed Phys Eng Express*, 7(6).
- Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., & Mayer, E. (2021). Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform*, 28(1).
- Khor, R. C., Nguyen, A., O'Dwyer, J., Kothari, G., Sia, J., Chang, D., Ng, S. P., Duchesne, G. M., & Foroudi, F. (2019). Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. *Int J Med Inform*, 121, 53-57.

- Khorrami, F., Ahmadi, M., & Sheikhtaheri, A. (2018). Evaluation of SNOMED CT Content Coverage: A Systematic Literature Review. *Stud Health Technol Inform*, 248, 212-219.
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:1708.05148*.
- Kilicoglu, H., Abacha, A. B., Mrabet, Y., Roberts, K., Rodriguez, L., Shooshan, S., & Demner-Fushman, D. (2016). Annotating named entities in consumer health questions. Proceedings of the Tenth International Conference on Language Resources and Evaluation.
- Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., & Demner-Fushman, D. (2018). Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1), 34.
- Kim, C., Zhu, V., Obeid, J., & Lenert, L. (2019). Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLOS ONE*, 14(2), e0212778.
- Kim, Y., Heider, P. M., Lally, I. R., & Meystre, S. M. (2021). A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System. *JMIR Med Inform*, 9(4), e22797.
- Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc*, 26(4), 364-379.
- Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73, 14-29.
- Kulkarni, A., & Shivananda, A. (2021). Deep learning for NLP. In *Natural language processing recipes* (pp. 213-262). Springer.
- Kumar, A., Gupta, P., Balan, R., Neti, L. B. M., & Malapati, A. (2021). BERT Based Semi-Supervised Hybrid Approach for Aspect and Sentiment Classification. *Neural Processing Letters*, 1-18.

- Kuo, T. T., Rao, P., Maehara, C., Doan, S., Chaparro, J. D., Day, M. E., Farcas, C., Ohno-Machado, L., & Hsu, C. N. (2016). Ensembles of NLP Tools for Data Element Extraction from Clinical Notes. *AMIA Annu Symp Proc, 2016*, 1880-1889.
- Kury, F., Butler, A., Yuan, C., Fu, L. H., Sun, Y., Liu, H., Sim, I., Carini, S., & Weng, C. (2020). Chia, a large annotated corpus of clinical trial eligibility criteria. *Sci Data*, 7(1), 281.
- Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. *J Med Internet Res*, 23(5), e26953.
- Laippala, V., Viljanen, T., Airola, A., Kanerva, J., Salanterä, S., Salakoski, T., & Ginter, F. (2014). Statistical parsing of varieties of clinical Finnish. *Artificial Intelligence in Medicine*, 61(3), 131-136.
- Larsen, A. S. T., Norheim, K. L., Marandi, R. Z., Hansen, E. A., & Madeleine, P. (2021). A field study investigating sensory manifestations in recreational female cyclists using a novel female-specific cycling pad. *Ergonomics*, 64(5), 571-581.
- Le Glaz, A., Haralambous, Y., Kim-Dufoir, D. H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVyllder, J., Walter, M., Berrouguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J Med Internet Res*, 23(5), e15708.
- Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22), 2909-2917.
- Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57, 28-37.
- Leaman, R., & Lu, Z. (2014). Automated disease normalization with low rank approximations. *Proceedings of BioNLP 2014*.
- Lee, C., Kim, Y., Kim, Y. S., & Jang, J. (2019). Automatic Disease Annotation From Radiology Reports Using Artificial Intelligence Implemented by a Recurrent Neural Network. *AJR Am J Roentgenol*, 212(4), 734-740.

- Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. *J Am Med Inform Assoc*, 21(e1), e11-19.
- Lee, J., & Maslove, D. M. (2017). Customization of a Severity of Illness Score Using Local Electronic Medical Record Data. *J Intensive Care Med*, 32(1), 38-47.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
- Lee, K., Hasan, S. A., Farri, O., Choudhary, A., & Agrawal, A. (2017). Medical Concept Normalization for Online User-Generated Texts. 2017 IEEE International Conference on Healthcare Informatics (ICHI).
- Lee, S., Mohr, N. M., Street, W. N., & Nadkarni, P. (2019). Machine Learning in Relation to Emergency Medicine Clinical and Operational Scenarios: An Overview. *West J Emerg Med*, 20(2), 219-227.
- Lehman, L. W. H., Feng, M., Yang, Y., & Mark, R. G. (2016). Blood Pressure and the Risk of Acute Kidney Injury in the ICU: Case-Control Versus Case-Crossover Designs. In M. C. Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 369-375).
- Li, J., Zhou, Y., Jiang, X., Natarajan, K., Pakhomov, S. V., Liu, H., & Xu, H. (2021). Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *J Am Med Inform Assoc*, 28(10), 2193-2201.
- Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting BERT for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.
- Liew, T. M., & Lee, C. S. (2021). Examining the Utility of Social Media in COVID-19 Vaccination: Unsupervised Learning of 672,133 Twitter Posts. *JMIR public health and surveillance*, 7(11), e29789.
- Limsopatham, N., & Collier, N. (2015). Adapting phrase-based machine translation to normalise medical terms in social media messages. *arXiv preprint arXiv:1508.02285*.
- Lin, C., Dligach, D., Miller, T. A., Bethard, S., & Savova, G. K. (2016). Multilayered temporal modeling for the clinical domain. *J Am Med Inform Assoc*, 23(2), 387-395.

- Lin, H. Y., & Moh, T.-S. (2021). *Sentiment analysis on COVID tweets using COVID-Twitter-BERT with auxiliary sentence approach* Proceedings of the 2021 ACM Southeast Conference, Virtual Event, USA.
- Liu, C., Ta, C. N., Rogers, J. R., Li, Z., Lee, J., Butler, A. M., Shang, N., Kury, F. S. P., Wang, L., Shen, F., Liu, H., Ena, L., Friedman, C., & Weng, C. (2019). Ensembles of natural language processing systems for portable phenotyping solutions. *J Biomed Inform*, 100, 103318.
- Liu, S., & Liu, J. (2021). Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. *Vaccine*, 39(39), 5499-5505.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Locquet, M., Diep, A. N., Beaudart, C., Dardenne, N., Brabant, C., Bruyère, O., & Donneau, A. F. (2021). A systematic review of prediction models to diagnose COVID-19 in adults admitted to healthcare centers. *Arch Public Health*, 79(1), 105.
- Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Loshchilov, I., & Hutter, F. (2018). Fixing weight decay regularization in adam.
- Lovis, C., Baud, R. H., & Planche, P. (2000). Power of expression in the electronic patient record: structured data or narrative text? *Int J Med Inform*, 58-59, 101-110.
- Luber, M., Weisser, C., Säfken, B., Silbersdorff, A., Kneib, T., & Kis-Katos, K. (2021). Identifying Topical Shifts in Twitter Streams: An Integration of Non-negative Matrix Factorisation, Sentiment Analysis and Structural Break Models for Large Scale Data. Multidisciplinary International Symposium on Disinformation in Open Online Media.
- Luo, X., Gandhi, P., Storey, S., & Huang, K. (2021). 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*, 1-1.

- Luo, Y. F., Sun, W., & Rumshisky, A. (2019). MCN: A comprehensive corpus for medical concept normalization. *J Biomed Inform*, 92, 103132.
- Lybarger, K., Ostendorf, M., Thompson, M., & Yetisgen, M. (2021). Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *J Biomed Inform*, 117, 103761.
- Lybarger, K., Ostendorf, M., & Yetisgen, M. (2021). Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform*, 113, 103631.
- Malla, S., & Alphonse, P. (2021). COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. *Applied Soft Computing*, 107, 107495.
- Mangono, T., Smittenaar, P., Caplan, Y., Huang, V. S., Sutermaster, S., Kemp, H., & Sgaier, S. K. (2021). Information-Seeking Patterns During the COVID-19 Pandemic Across the United States: Longitudinal Analysis of Google Trends Data. *J Med Internet Res*, 23(5), e22933.
- Margus, C., Brown, N., Hertelendy, A. J., Safferman, M. R., Hart, A., & Ciottone, G. R. (2021). Emergency Physician Twitter Use in the COVID-19 Pandemic as a Potential Predictor of Impending Surge: Retrospective Observational Study. *J Med Internet Res*, 23(7), e28615.
- Markić, I., Štula, M., Zorić, M., & Stipaničev, D. (2020). Entropy-Based Approach in Selection Exact String-Matching Algorithms. *Entropy (Basel)*, 23(1).
- Mastorakos, G., Khurana, A., Huang, M., Fu, S., Tafti, A. P., Fan, J., & Liu, H. (2021). Probing Patient Messages Enhanced by Natural Language Processing: A Top-Down Message Corpus Analysis. *Health Data Science*, 2021.
- Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *Int J Med Inform*, 114, 57-65.
- Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Comput Sci*, 7, e476.

- Mellia, J. A., Basta, M. N., Toyoda, Y., Othman, S., Elfanagely, O., Morris, M. P., Torre-Healy, L., Ungar, L. H., & Fischer, J. P. (2021). Natural Language Processing in Surgery: A Systematic Review and Meta-analysis. *Ann Surg*, 273(5), 900-908.
- Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What Happens To BERT Embeddings During Fine-tuning? *arXiv preprint arXiv:2004.14448*.
- Miftahutdinov, Z., Kadurin, A., Kudrin, R., & Tutubalina, E. (2021). Medical Concept Normalization in Clinical Trials with Drug and Disease Representation Learning. *Bioinformatics*.
- Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E., & Meinert, E. (2020). The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J Med Internet Res*, 22(10), e20346.
- The MIMIC II Clinical Database*. (2011).
<https://archive.physionet.org/physiobank/database/mimic2cdb/>
- Møller, A. G., Van Der Goot, R., & Plank, B. (2020). NLP North at WNUT-2020 Task 2: Pre-training versus Ensembling for Detection of Informative COVID-19 English Tweets. Proceedings of the Sixth Workshop on Noisy User-generated Text (WNUT 2020).
- Monselise, M., Chang, C.-H., Ferreira, G., Yang, R., & Yang, C. C. (2021). Detecting Topics and Sentiments of Public Concerns on COVID-19 Vaccines with Social Media Trend Analysis. *Journal of medical Internet research*.
- Monson-Haefel, R., & Chappell, D. A. (2002). *Java message service*. O'Reilly Media, Inc.
- Mowery, D. (2013). *ShAReCLEF eHealth 2013: Natural Language Processing and Information Retrieval for Clinical Care Version 1.0*.
- Mrabet, Y., Kilicoglu, H., Roberts, K., & Demner-Fushman, D. (2016). Combining Open-domain and Biomedical Knowledge for Topic Recognition in Consumer Health Questions. *AMIA Annu Symp Proc*, 2016, 914-923.
- Mudassir, M. A., Mor, Y., Munot, R., & Shankarmani, R. (2021). Sentiment Analysis of COVID-19 Vaccine Perception Using NLP. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA).

- Mughli, R. A., Durrant, E., Baia Medeiros, D. T., Shelton, D., Robins, J., Qamar, S. R., O'Keeffe, M. E., & Berger, F. H. (2021). Overnight attending radiologist coverage decreases imaging-related emergency department recalls by at least 90. *Emerg Radiol*, 28(3), 549-555.
- Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.
- Musaoğlu, M., Yuksel, M., Mizikoglu, O., & Arikan, C. (2021). Telemedicine in monitoring pediatric LT patients before and during COVID-19 pandemic. *Pediatr Transplant*, e14138.
- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2019). SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*.
- Naseri, H., Kafi, K., Skamene, S., Tolba, M., Faye, M. D., Ramia, P., Khriouian, J., & Kildea, J. (2021). Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. *J Biomed Inform*, 120, 103864.
- Nasir, A., Shah, M. A., Ashraf, U., Khan, A., & Jeon, G. (2021). An intelligent framework to predict socioeconomic impacts of COVID-19 and public sentiments. *Comput Electr Eng*, 96, 107526.
- Nath, N., Lee, S. H., McDonnell, M. D., & Lee, I. (2021). The quest for better clinical word vectors: Ontology based and lexical vector augmentation versus clinical contextual embeddings. *Comput Biol Med*, 134, 104433.
- Névél, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1), 12.
- Newman-Griffis, D., Divita, G., Desmet, B., Zirikly, A., Rosé, C. P., & Fosler-Lussier, E. (2021). Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *J Am Med Inform Assoc*, 28(3), 516-532.

- Nguyen, A., Lawley, M., Hansen, D., & Colquist, S. (2009). A simple pipeline application for identifying and negating SNOMED clinical terminology in free text.
- Nguyen, A. N., Moore, J., O'Dwyer, J., & Philpot, S. (2015). Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports. *AMIA Annu Symp Proc*, 2015, 953-962.
- Nguyen, A. N., Moore, J., O'Dwyer, J., & Philpot, S. (2016). Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository. *AMIA Annu Symp Proc*, 2016, 964-973.
- Nicolas, J. (2012). To detect and analyze sequence repeats whatever be their origin. *Methods Mol Biol*, 859, 69-90.
- NLPIF. (2021). *2021 Workshop on NLP4IF: Censorship, Disinformation, and Propaganda*. U.S. National Science Foundation. <http://www.netcopia.net/nlp4if/>
- Nothman, J., Qin, H., & Yurchak, R. (2018). Stop word lists in free open-source software packages. Proceedings of Workshop for NLP Open Source Software (NLP-OSS).
- Novo, T. N. F. (2021). MisInfoBot: fight misinformation about COVID on social media.
- Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC Med Inform Decis Mak*, 19(1), 164.
- Oellrich, A., Collier, N., Smedley, D., & Groza, T. (2015). Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. *PLOS ONE*, 10(1), e0116040.
- Omogegbe, N. A. I., Ndaman, I. O., Misra, S., Abayomi-Alli, O. O., & Damaševičius, R. (2020). Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. *Journal of Healthcare Engineering*, 2020, 8839524.
- Oyebode, O., Ndulue, C., Adib, A., Mulchandani, D., Suruliraj, B., Orji, F. A., Chambers, C. T., Meier, S., & Orji, R. (2021). Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media

Comments: Text Mining and Thematic Analysis Approach. *JMIR medical informatics*, 9(4), e22734.

- Patel, P., & Zed, P. J. (2002). Drug-related visits to the emergency department: how big is the problem? *Pharmacotherapy*, 22(7), 915-923.
- Patel, S. J., Chamberlain, D. B., & Chamberlain, J. M. (2018). A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage. *Acad Emerg Med*, 25(12), 1463-1470.
- Patrick, J., & Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc*, 17(5), 524-527.
- Patrick, J. D., Nguyen, D. H., Wang, Y., & Li, M. (2011). A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc*, 18(5), 574-579.
- Pattisapu, N., Anand, V., Patil, S., Palshikar, G., & Varma, V. (2020). Distant supervision for medical concept normalization. *J Biomed Inform*, 109, 103522.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Ping, X. O., Tseng, Y. J., Chung, Y., Wu, Y. L., Hsu, C. W., Yang, P. M., Huang, G. T., Lai, F., & Liang, J. D. (2013). Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. *Telemed J E Health*, 19(9), 704-710.
- Plasek, J. M., Goss, F. R., Lai, K. H., Lau, J. J., Seger, D. L., Blumenthal, K. G., Wickner, P. G., Slight, S. P., Chang, F. Y., Topaz, M., Bates, D. W., & Zhou, L. (2016). Food entries in a large allergy data repository. *J Am Med Inform Assoc*, 23(e1), e79-87.
- R-Core-Team. (2017). *R: A Language and Environment for Statistical Computing*. In <https://www.R-project.org/>
- Rajman, M., & Besançon, R. (1998). Text Mining: Natural Language techniques and Text Mining applications. In S. Spaccapietra & F. Maryanski (Eds.), *Data Mining*

and Reverse Engineering: Searching for semantics. IFIP TC2 WG2.6 IFIP Seventh Conference on Database Semantics (DS-7) 7–10 October 1997, Leysin, Switzerland (pp. 50-64). Springer US.

- Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. *arXiv preprint arXiv:2004.03925*.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1), 86.
- Research, T. M. (2020). *Healthcare Natural Language Processing Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast, 2019 - 2027* (TMRGL18554). <https://www.transparencymarketresearch.com/healthcare-natural-language-processing-market.html>
- Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., & Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc*, 2007, 625-629.
- Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Setzer, A., & Roberts, I. (2008). Semantic annotation of clinical text: The CLEF corpus. Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining.
- Roberts, A., Gaizauskas, R., Hepple, M., & Guo, Y. (2008). Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9 Suppl 11(Suppl 11), S3.
- Roberts, K., Kilicoglu, H., Fiszman, M., & Demner-Fushman, D. (2014). Automatically classifying question types for consumer health questions. *AMIA Annu Symp Proc*, 2014, 1018-1027.
- Robinson, J. R., Valentine, A., Carney, C., Fabbri, D., & Jackson, G. P. (2017). Complexity of medical decision-making in care provided by surgeons through patient portals. *Journal of Surgical Research*, 214, 93-101.
- Robles Hernandez, L. A., Callahan, T. J., & Banda, J. M. (2021). A Biomedically oriented automatically annotated Twitter COVID-19 Dataset. *ArXiv*.

- Roehrick, K. (2021). Valence Aware Dictionary and sEntiment Reasoner (VADER). <https://cran.r-project.org/web/packages/vader/vader.pdf>
- RStudio. (2021). *TensorFlow*TM. In <https://github.com/rstudio/tensorflow>
- Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE*, 16(2), e0245909.
- Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L. W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., & Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Crit Care Med*, 39(5), 952-960.
- Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE*, 16(8), e0254937.
- Sasangohar, F., Dhala, A., Zheng, F., Ahmadi, N., Kash, B., & Masud, F. (2021). Use of telecritical care for family visitation to ICU during the COVID-19 pandemic: an interview study and sentiment analysis. *BMJ Qual Saf*, 30(9), 715-721.
- Savova, G., Pradhan, S., Palmer, M., Styler, W., Chapman, W., & Elhadad, N. (2017). Annotating the Clinical Text – MiPACQ, ShARe, SHARpN and THYME Corpora. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 1357-1378). Springer Netherlands.
- Savoy, O. K. J. (2012). Feature selection in sentiment analysis. *Proc. CORIA*, no. January, 273-284.
- Schmidt, B., Walter, H., Zeitler, N., & Korn, K. (2002). Genotypic drug resistance interpretation systems--the cutting edge of antiretroviral therapy. *AIDS Rev*, 4(3), 148-156.
- Sediqin, M. (2021). *Semantic Analysis of Vaccine and Mask Sentiments in COVID-19 Twitter Data* The University of Wisconsin-Milwaukee.
- SentiStrength. (2014). *Dictionary for Sentiment Analysis*. <https://raw.githubusercontent.com/sifei/Dictionary-for-Sentiment-Analysis/master/slang/acronym.csv>

- SentiStrength. (2018). *Slang Lookup Table*.
<https://raw.githubusercontent.com/felipebravom/StaticTwitterSent/master/extra/SentiStrength/SlangLookupTable.txt>
- Shah, A. D., Bailey, E., Williams, T., Denaxas, S., Dobson, R., & Hemingway, H. (2019). Natural language processing for disease phenotyping in UK primary care records for research: a pilot study in myocardial infarction and death. *J Biomed Semantics*, 10(Suppl 1), 20.
- Shah, A. M., Ali, M., Qayyum, A., Begum, A., Han, H., Ariza-Montes, A., & Araya-Castillo, L. (2021). Exploring the Impact of Linguistic Signals Transmission on Patients' Health Consultation Choice: Web Mining of Online Reviews. *Int J Environ Res Public Health*, 18(19).
- Shamrat, F., Chakraborty, S., Imran, M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 23(1), 463-470.
- Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform*, 7(2), e12239.
- Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of Big Data*, 8(1), 18.
- Shrestha, A., Zikos, D., & Fegaras, L. (2021). An annotated association mining approach for extracting and visualizing interesting clinical events. *Int J Med Inform*, 148, 104366.
- Shrestha, I. (2021). A literature review of Mental Health prediction during different stages of Covid-19.
- Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O., & Chupryna, A. (2020). Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications. 2020 IEEE International Conference on Problems of Infocommunications.
- Smith, P. (2017). *Sentiment analysis of patient feedback* University of Birmingham.

- Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6), 646-651.
- Steinkamp, J. M., Bala, W., Sharma, A., & Kantrowitz, J. J. (2020). Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *Journal of Biomedical Informatics*, 102, 103354.
- Steitz, B. D., Sulieman, L., Warner, J. L., Fabbri, D., Brown, J. T., Davis, A. L., & Unertl, K. M. (2021). Classification and analysis of asynchronous communication content between care team members involved in breast cancer treatment. *JAMIA Open*, 4(3), ooab049.
- Sterling, N. W., Brann, F., Patzer, R. E., Di, M., Koebbe, M., Burke, M., & Schrager, J. D. (2020). Prediction of emergency department resource requirements during triage: An application of current natural language processing techniques. *J Am Coll Emerg Physicians Open*, 1(6), 1676-1683.
- Strauss, J. A., Chao, C. R., Kwan, M. L., Ahmed, S. A., Schottinger, J. E., & Quinn, V. P. (2013). Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Inform Assoc*, 20(2), 349-355.
- Styler, W. F. t., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., & Pustejovsky, J. (2014). Temporal Annotation in the Clinical Domain. *Trans Assoc Comput Linguist*, 2, 143-154.
- Sulieman, L., Gilmore, D., French, C., Cronin, R. M., Jackson, G. P., Russell, M., & Fabbri, D. (2017). Classifying patient portal messages using Convolutional Neural Networks. *J Biomed Inform*, 74, 59-70.
- Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.
- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics.
- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *J Healthc Eng*, 2018, 4302425.

- Sun, Y., Butler, A., Stewart, L. A., Liu, H., Yuan, C., Southard, C. T., Kim, J. H., & Weng, C. (2021). Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials. *J Biomed Inform*, 118, 103790.
- Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., & Zuccon, G. (2013). *Overview of the ShARe/CLEF eHealth Evaluation Lab 2013* Proceedings of the 4th International Conference on Information Access Evaluation. Multilinguality, Multimodality, and Visualization - Volume 8138, Valencia, Spain.
- Surano, F. V., Porfiri, M., & Rizzo, A. (2021). Analysis of lockdown perception in the United States during the COVID-19 pandemic. *Eur Phys J Spec Top*, 1-9.
- Suryawati, E., & Widyanoro, D. H. (2017). Combination of Heuristic, Rule-Based and Machine Learning for Bibliography Extraction. 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME).
- Syed, K., Sleeman, W. t., Hagan, M., Palta, J., Kapoor, R., & Ghosh, P. (2020). Automatic Incident Triage in Radiation Oncology Incident Learning System. *Healthcare (Basel)*, 8(3).
- Talpada, H., Halgamuge, M. N., & Vinh, N. T. Q. (2019). An Analysis on Use of Deep Learning and Lexical-Semantic Based Sentiment Analysis Method on Twitter Data to Understand the Demographic Trend of Telemedicine. 2019 11th International Conference on Knowledge and Systems Engineering (KSE).
- Tang, B., Chen, Q., Wang, X., Wu, Y., Zhang, Y., Jiang, M., Wang, J., & Xu, H. (2015). Recognizing Disjoint Clinical Concepts in Clinical Text Using Machine Learning-based Methods. *AMIA Annu Symp Proc*, 2015, 1184-1193.
- Tange, H. J., Hasman, A., de Vries Robbé, P. F., & Schouten, H. C. (1997). Medical narratives in electronic medical records. *Int J Med Inform*, 46(1), 7-29.
- Tayarani N, M.-H. (2021). Applications of artificial intelligence in battling against covid-19: A literature review. *Chaos, Solitons & Fractals*, 142, 110338.
- Technology Institute. (2017). *Certified Health IT Developers and Editions Reported by Health Care Professionals Participating in the Medicare EHR Incentive Program*

(Health IT Quick-Stat #30, Issue).

<https://www.healthit.gov/data/quickstats/health-care-professional-health-it-devleopers>

- Thorpe Huerta, D., Hawkins, J. B., Brownstein, J. S., & Hswen, Y. (2021). Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis. *SSM Popul Health*, 15, 100851.
- Timimi, F., Ray, S., Jones, E., Aase, L., & Hoffman, K. (2019). Patient-Reported Outcomes in Online Communications on Statins, Memory, and Cognition: Qualitative Analysis Using Online Communities. *J Med Internet Res*, 21(11), e14809.
- Topaz, M., Murga, L., Bar-Bachar, O., Cato, K., & Collins, S. (2019). Extracting Alcohol and Substance Abuse Status from Clinical Notes: The Added Value of Nursing Data. *Stud Health Technol Inform*, 264, 1056-1060.
- Tran, K. V., Phan, H. P., Van Nguyen, K., & Nguyen, N. L.-T. (2020). UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network. *arXiv preprint arXiv:2009.02935*.
- Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008). Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9 Suppl 3(Suppl 3), S2.
- Tsuruoka, Y., McNaught, J., Tsujii, J. i., chi, & Ananiadou, S. (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20), 2768-2774.
- Turchin, A., & Florez Builes, L. F. (2021). Using Natural Language Processing to Measure and Improve Quality of Diabetes Care: A Systematic Review. *J Diabetes Sci Technol*, 15(3), 553-560.
- Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., & Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84, 93-102.
- Tziafas, G., Kogkalidis, K., & Caselli, T. (2021). Fighting the COVID-19 infodemic with a holistic BERT ensemble. *arXiv preprint arXiv:2104.05745*.

- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.
- Uyangodage, L., Ranasinghe, T., & Hettiarachchi, H. (2021). Transformers to Fight the COVID-19 Infodemic. arXiv:2104.12201. Retrieved April 01, 2021, from <https://ui.adsabs.harvard.edu/abs/2021arXiv210412201U>
- Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5), 514-518.
- Uzuner, Ö., Solti, I., Xia, F., & Cadag, E. (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5), 519-523.
- Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5), 552-556.
- Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *J Med Internet Res*, 22(12), e21418.
- Van Vleck, T. T., Chan, L., Coca, S. G., Craven, C. K., Do, R., Ellis, S. B., Kannry, J. L., Loos, R. J. F., Bonis, P. A., Cho, J., & Nadkarni, G. N. (2019). Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression. *Int J Med Inform*, 129, 334-341.
- Vargas, A. N., Maier, A., Vallim, M. B. R., Banda, J. M., & Preciado, V. M. (2021). Negative Perception of the COVID-19 Pandemic Is Dropping: Evidence From Twitter Posts. *Front Psychol*, 12, 737882.
- Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., & Rosé, C. P. (2021). Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *J Biomed Inform*, 121, 103880.
- Vashisth, S., Brethenoux, E., Emmott, S., Davis, M., Norrie, D., & Elliot, B. (2020). *Market Guide for Text Analytics*. G. S. Reports. <https://info.indico.io/2020-gartner-text-analytics>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*.
- Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. *J Biomed Inform*, 88, 11-19.
- Venugopalan, J., Chanani, N., Maher, K., & Wang, M. D. (2017). Combination of static and temporal data analysis to predict mortality and readmission in the intensive care. *Annu Int Conf IEEE Eng Med Biol Soc*, 2017, 2570-2573.
- Verspoor, K. (2021). The Evolution of Clinical Knowledge During COVID-19: Towards a Global Learning Health System. *Yearbook of Medical Informatics*, 30(01), 176-184.
- Wajsbürt, P., Sarfati, A., & Tannier, X. (2021). Medical concept normalization in French using multilingual terminologies.
- Walsh, S. H. (2004). The clinician's perspective on electronic health records and how they can affect patient care. *Bmj*, 328(7449), 1184-1187.
- Wang, H., Li, Y., Hutch, M., Naidech, A., & Luo, Y. (2021). Using Tweets to Understand How COVID-19-Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study. *J Med Internet Res*, 23(2), e26302.
- Wang, J., Zhang, Z., Wang, W., Pan, L., & Chai, X. (2014). Clinical data mining by exploring public MIMIC-II intensive care database. *Zhongguo Yi Liao Qi Xie Za Zhi*, 38(6), 402-406.
- Wang, L. L., & Lo, K. (2020). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, 22(2), 781-799.
- Wang, S. Y., Pershing, S., Tran, E., & Hernandez-Boussard, T. (2020). Automated extraction of ophthalmic surgery outcomes from the electronic health record. *Int J Med Inform*, 133, 104007.
- Wang, X., Liu, W., Chauhan, A., Guan, Y., & Han, J. (2020). Automatic textual evidence mining in covid-19 literature. *arXiv preprint arXiv:2004.12563*.

- Wang, X., & Matthews, M. (2008). Comparing usability of matching techniques for normalising biomedical named entities. *Pac Symp Biocomput*, 628-639.
- Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak*, 19(1), 1.
- Warrer, P., Hansen, E. H., Juhl-Jensen, L., & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br J Clin Pharmacol*, 73(5), 674-684.
- Wei, Q., Ji, Z., Si, Y., Du, J., Wang, J., Tiryaki, F., Wu, S., Tao, C., Roberts, K., & Xu, H. (2019). Relation Extraction from Clinical Narratives Using Pre-trained Language Models. *AMIA Annu Symp Proc*, 2019, 1236-1245.
- Wester, K., Jönsson, A. K., Spigset, O., Druid, H., & Hägg, S. (2008). Incidence of fatal adverse drug reactions: a population based study. *Br J Clin Pharmacol*, 65(4), 573-579.
- Wong, A., Plasek, J. M., Montecalvo, S. P., & Zhou, L. (2018). Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. *Pharmacotherapy*, 38(8), 822-841.
- Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., & Xu, H. (2019). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3), 457-470.
- Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annu Symp Proc*, 2017, 1812-1819.
- Wunnava, S., Qin, X., Kakar, T., Sen, C., Rundensteiner, E. A., & Kong, X. (2019). Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding. *Drug Saf*, 42(1), 113-122.
- Xie, Z., Sato, I., & Sugiyama, M. (2020). Stable weight decay regularization. *arXiv preprint arXiv:2011.11152*.

- Xu, Y., Qiu, X., Zhou, L., & Huang, X. (2020). Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*.
- Xu, Y., Wang, Y., Liu, T., Liu, J., Fan, Y., Qian, Y., Tsujii, J., & Chang, E. I. (2014). Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. *Journal of the American Medical Informatics Association*, 21(e1), e84-e92.
- Xue, Y., Klabjan, D., & Luo, Y. (2019). Predicting ICU readmission using grouped physiological and medication trends. *Artif Intell Med*, 95, 27-37.
- Xue, Z., Yin, D., & Davison, B. D. (2011). Normalizing microtext. Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence,
- Yang, L. W. Y., Ng, W. Y., Foo, L. L., Liu, Y., Yan, M., Lei, X., Zhang, X., & Ting, D. S. W. (2021). Deep learning-based natural language processing in ophthalmology: applications, challenges and future directions. *Curr Opin Ophthalmol*, 32(5), 397-405.
- Yang, X., Bekoulis, G., & Deligiannis, N. (2020). imec-ETRO-VUB at W-NUT 2020 Shared Task-3: A Multilabel BERT-based system for predicting COVID-19 events. Proceedings of the Sixth Workshop on Noisy User-generated Text.
- Yang, X., Bian, J., Hogan, W. R., & Wu, Y. (2020). Clinical concept extraction using transformers. *J Am Med Inform Assoc*, 27(12), 1935-1942.
- Yim, W.-w., Yetisgen, M., Harris, W. P., & Kwan, S. W. (2016). Natural Language Processing in Oncology: A Review. *JAMA Oncology*, 2(6), 797-804.
- Young, I. J. B., Luz, S., & Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform*, 132, 103971.
- Yu, S., Eisenman, D., & Han, Z. (2021). Temporal Dynamics of Public Emotions During the COVID-19 Pandemic at the Epicenter of the Outbreak: Sentiment Analysis of Weibo Posts From Wuhan. *J Med Internet Res*, 23(3), e27078.
- Yu, S., Su, J., & Luo, D. (2019). Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7, 176600-176612.

- Yu, W., Jiang, Z., Chen, F., Hou, Q., & Feng, J. (2021). LV-BERT: Exploiting Layer Variety for BERT. *arXiv preprint arXiv:2106.11740*.
- Zhang, G., Yuan, J., Yu, M., Wu, T., Luo, X., & Chen, F. (2021, Mar). A machine learning method for acute hypotensive episodes prediction using only non-invasive parameters. *Comput Methods Programs Biomed*, 200, 105845.
- Zhang, S., Kang, T., Zhang, X., Wen, D., Elhadad, N., & Lei, J. (2016). Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models. *Journal of Biomedical Informatics*, 60, 334-341.
- Zhang, X., Bellolio, M. F., Medrano-Gracia, P., Werys, K., Yang, S., & Mahajan, P. (2019). Use of natural language processing to improve predictive models for imaging utilization in children presenting to the emergency department. *BMC Med Inform Decis Mak*, 19(1), 287.
- Zhang, X., Kim, J., Patzer, R. E., Pitts, S. R., Chokshi, F. H., & Schrager, J. D. (2019). Advanced diagnostic imaging utilization during emergency department visits in the United States: A predictive modeling study for emergency department triage. *PLOS ONE*, 14(4), e0214905.
- Zhang, X., Zhang, J., Li, J., Gao, Y., Li, R., Jin, X., Wang, X., Huang, Y., & Wang, G. (2021). Relationship between 24-h venous blood glucose variation and mortality among patients with acute respiratory failure. *Sci Rep*, 11(1), 7747.
- Zhao, S., Li, L., Lu, H., Zhou, A., & Qian, S. (2019). Associative attention networks for temporal relation extraction from electronic health records. *J Biomed Inform*, 99, 103309.
- Zhou, L., Lu, Y., Vitale, C. J., Mar, P. L., Chang, F., Dhopeswarkar, N., & Rocha, R. A. (2014). Representation of information about family relatives as structured data in electronic health records. *Appl Clin Inform*, 5(2), 349-367.
- Zhou, L., Plasek, J. M., Mahoney, L. M., Karipineni, N., Chang, F., Yan, X., Chang, F., Dimaggio, D., Goldman, D. S., & Rocha, R. A. (2011). Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. *AMIA Annu Symp Proc*, 2011, 1639-1648.
- Zong, H., Hu, B., Han, Y., Li, Z., & Zhang, X. (2021). Prevalence and Temporal Trends Analysis of Instruments in Posttraumatic Stress Disorder Using Text Mining. *JMIR Ment Health*.

Zou, C., Wang, X., Xie, Z., & Li, D. (2020). Public Reactions towards the COVID-19 Pandemic on Twitter in the United Kingdom and the United States. *medRxiv*.

Zunic, A., Corcoran, P., & Spasic, I. (2020). Sentiment Analysis in Health and Well-Being: Systematic Review. *JMIR Med Inform*, 8(1), e16023.